Bridging the Automated Machine Learning Transparency and Usability Gap for Non-Experts: A User-Centered Design and Evaluation Study

By

Muhammad Alif Danial, BSc (Hons)

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy, August 2024

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Marina Ng, whose patience, motivation, and immense knowledge have been invaluable throughout my Ph.D. journey. Her guidance and unwavering support in both my research and the writing of this thesis have been essential to my development as a researcher. I could not have imagined having a better advisor and mentor for this important chapter of my academic life.

I am also sincerely thankful to my supervisor, Dr. Radu Muschevici, for his critical role in refining my research and writing. His sharp insights and thoughtful feedback have significantly elevated the quality of my work, and his encouragement has kept me focused and motivated throughout this journey.

I am deeply grateful to Professor Tomas Maul, whose guidance has been a steady source of inspiration and direction throughout my Ph.D. His wisdom and dedication to nurturing young researchers have greatly influenced my academic growth.

My heartfelt appreciation goes to my research colleagues, Chia Yi and Shekinah, for their insightful comments, continuous encouragement, and camaraderie. Their perspectives and constructive critiques have been crucial in shaping my research and helping me navigate the challenges of my Ph.D.

Words cannot adequately express my gratitude to my parents, whose unwavering support and belief in me have been the foundation of my strength. Their love and encouragement have been with me every step of the way. A special thanks to my uncle Ray and aunt Kula, whose support during the final stages of my Ph.D. journey was pivotal. Their kindness and encouragement helped me push through the final hurdles. Lastly, I want to express my deepest appreciation to my beloved, Aisyah, who has been by my side throughout this entire Ph.D. journey. Her patience, love, and understanding have been my constant source of comfort and motivation.

Contents

List of Figures

List of Tables

Abstract

The inherent complexity of Artificial Intelligence (AI) and Machine Learning (ML) tools creates significant barriers for non-expert users. Traditional ML workflows require specialized programming

and statistical knowledge, limiting widespread adoption across various domains where these technologies could provide substantial benefits. This research aimed to develop and evaluate VisAutoML, an automated machine learning tool specifically designed to provide non-expert users with a transparent and user-friendly ML development experience for tabular data. The study sought to identify key factors influencing tool acceptance, understand specific challenges faced by non-experts, and create novel design principles to address these challenges. The research employed a five-stage iterative user-centered design methodology. This included a mixed-method study utilizing an extended Technology Acceptance Model (TAM) to identify acceptance factors and user challenges. The design integrated technology-enhanced scaffolding and Explainable Artificial Intelligence (XAI) principles tailored for non-experts, featuring visualizations of activities, demonstration of scaffold functions, contextually relevant support, and progressive disclosure of XAI visualizations. Two versions of VisAutoML were developed and evaluated against both commercial alternatives and established benchmarks. Initial comparison between VisAutoML 1.0 and H2O AutoML showed significantly higher System Usability Scale scores (61.5 vs 38.5) for VisAutoML and a 20.94% increase in correct answers on knowledge assessments. The redesigned VisAutoML 2.0 demonstrated substantial improvements, with 75% of participants completing ML model development tasks in under 5 minutes. User Experience Questionnaire results showed 'good' scores for pragmatic quality (M=1.60, SD=0.912) and 'excellent' scores for hedonic quality (M=1.59, SD=0.899) and overall usability (M=1.60, SD=0.851). Trust measures were moderate (M=26.11, SD=4.67), while perceived explainability ratings were high (M=161.9, SD=36.24). This research contributes to the field by extending the TAM framework for understanding non-expert AutoML requirements, introducing empirically grounded design principles for usable and transparent AutoML systems, and successfully developing VisAutoML 2.0 with demonstrably enhanced usability and transparency. These contributions provide valuable guidance for making ML more accessible to broader audiences, advancing the democratization of AI technologies beyond technical specialists. Future work should explore additional application domains and further refinements of scaffolding and XAI approaches.

1 Introduction

1.1 Background

Machine learning (ML) is a maturing field of artificial intelligence (AI) that is being widely adopted by various real-world domains such as agriculture, banks, and healthcare. Nowadays, many businesses are attempting to implement AI to maintain and grow their competitive advantage. The World Economic Forum estimates that AI could add the equivalent of $2.6 trillion to $4.4 trillion annually across 63 use cases (Yee, 2023). However, not all companies have the resources to develop these applications due to various challenges such as computing cost, development time, feature selection (the process of selecting a subset of relevant variables for model building), model selection (choosing the most appropriate model or algorithm for a task), hyperparameter optimisation (tuning model parameters to improve performance), and model deployment (making a trained model available for use) (Benbya et al., 2021).

Furthermore, highly skilled experts who are proficient in both programming and statistics are required in the development of effective ML solutions (ML models are the output of the machine learning process used for tasks like prediction or classification). This creates a high barrier of entry for non-expert users, defined as those who are neither knowledgeable about ML nor a particular application domain (Bove et al., 2022). To make ML more accessible and usable, there have been developments in Automated ML (AutoML) systems that automatically carry out ML tasks with minimal human effort. AutoML, as defined by (Hutter et al., 2019), is a framework that automates the development of predictive models. AutoML facilitates rapid and scalable development of models with increased performance, and minimal interaction from ML professionals, and results in considerable time and cost savings (Singh & Joshi, 2022).

AutoML has the potential to democratise ML by allowing domain experts (individuals knowledgeable about a specific field but not necessarily ML) and developers to easily develop ML applications in their organisations without relying on already limited and expensive ML professionals. AutoML systems have been proven to perform accurately even under limited time constraints (Giovanelli & Pisano, 2022). AutoML is currently viewed as a tool that assists users in their ML pursuits in an interactive manner (Wang et al., 2019). Current AutoML frameworks such as Google AutoML, H2O AutoML, and Auto-Sklearn focus primarily on algorithmic optimization and automated model selection, neglecting human-centered design principles. These systems prioritize efficiency and performance metrics over interpretability (the degree to which a human can understand a model's predictions) and collaborative features. Studies have shown that "AutoML is still a long way from its major goals, i.e. comprehensive automation of the entire ML workflow" (Khuat et al., 2022).

This shortcoming stems from most AutoML systems operating as black-box tools (systems that generate solutions without providing transparent reasoning or decision processes comprehensible to human users) that generate solutions without providing transparent reasoning or decision processes comprehensible to human users (Xin et al., 2021). Commercial solutions like DataRobot and Azure AutoML, despite their technical sophistication, frequently lack interactive mechanisms that would allow domain experts to incorporate specialized knowledge or contextual understanding into the automation process. The resulting opacity in prevalent AutoML solutions creates a significant barrier to trust, undermining widespread adoption and effective implementation across various domains (Khuat et al., 2022; Wang et al., 2019).

The field of explainable artificial intelligence (XAI), defined as artificial intelligence systems that can explain their reasoning to a human user, define their strengths and limitations, and convey an understanding of how they will behave in the future (Kremers, 2020), seeks to address the black-box

nature of ML models, mainly through visual analytics, defined as "the science of analytical reasoning facilitated by interactive visual interfaces" (Spinner et al., 2020). Interactive visualisations (visual representations that users can interact with to explore and understand information) have been commonly used as a medium for XAI (Langer et al., 2021).

Currently, various interactive visualisation XAI tools explain complicated model components and workflows of AI models based on algorithms such as Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME). Furthermore, the majority of available AI visualisation tools are designed for AI developers and practitioners rather than domain users with limited AI experience (Hohman et al., 2019). However, literature has shown that few users can accurately interpret the visualisations generated by these XAI methods (Kaur et al., 2020). Although there is a rapid development of innovative XAI methods, studies have shown the majority of XAI is ineffective at producing actionable insights and even manipulating user trust (X. Wang & Yin, 2021).

Furthermore, studies have indicated that effective ML-related UIs/AutoML are faced with two main challenges: capability uncertainty and output complexity (Margetis et al., 2021; Yang et al., 2020). Although many design guidelines have been suggested to overcome these challenges, there is a lack of studies that focus on usable visual analytics for AutoML implementation among non-experts (Amershi et al., 2019; Gil et al., 2019). Researchers have long recognised the importance of developing a transparent (ensuring AI systems are transparent and accountable to build trust) and usable AutoML tool (Lee et al., 2019). However, few studies have explored the methods to allow non-expert users to participate in building models via AutoML.

Research has begun to recognise usability, defined as the degree to which a product or system may be utilised by specific users to accomplish concrete objectives with effectiveness, efficiency, and satisfaction in a specific context of usage (Bevan et al., 2016), as an essential requirement for XAI (Sokol & Flach, 2020). Usable XAI requires good algorithms and well-designed user interfaces that connect algorithm capabilities to non-expert needs (Cheng et al., 2019). Further, there is numerous research that provides frameworks and guidelines for developing usable XAI (H. F. Cheng et al., 2019; Miller, 2019). To date, research on usable XAI has mostly focused on the design of conventional user interfaces with little attention paid to interactive visualisations, which are an important component of usable XAI (Wang et al., 2022). This suggests that there is a need to fill the gap by developing VisAutoML, an AutoML platform that provides usable XAI visualisations and supports model development for non-experts.

1.2 Problem Statement

The pervasive advancement and increasing adoption of Artificial Intelligence (AI) and Machine Learning (ML) technologies across diverse real-world domains, including agriculture, finance, and healthcare, underscore their significant economic and societal value. Projections indicate that AI could contribute trillions annually to the global economy (Yee, 2023), driving businesses to implement AI solutions to maintain competitive advantages. However, the realization of this potential is hampered by substantial challenges in the development and deployment of effective ML solutions. These challenges include considerable computing costs, lengthy development times, and the technical complexity associated with crucial steps such as data preprocessing, feature selection, model selection, hyperparameter optimisation, and model deployment (Benbya et al., 2021).

The Challenge for Non-Expert Users

A critical obstacle to the broader application of ML is the necessity for highly skilled experts proficient in both programming and statistics. This requirement creates a significant barrier to entry for non-

expert users, defined in this research as individuals lacking extensive knowledge of either ML or a specific application domain (Bove et al., 2022). This user group, despite their potential to leverage ML for domain-specific insights and tasks, is largely excluded from directly developing and deploying ML models due to the technical demands of traditional workflows. As identified in the literature, non-experts face numerous challenges throughout the ML development lifecycle, including difficulties in data analysis prior to modeling, problem and feature design, ML model evaluation, ML model selection, and ML model performance improvement (Yang et al., 2018; Ramos et al., 2020). These challenges stem from a limited understanding of ML algorithms, reliance on external documentation and scripts, misinterpretation of performance metrics, limited adaptation of learning algorithms, and an overemphasis on data quantity over other crucial factors (Yang et al., 2018). The exclusion of this large potential user base represents a significant missed opportunity for broader AI adoption and innovation.

The Problem of Usability in Existing AutoML Tools

In response to the demand for more accessible ML, Automated Machine Learning (AutoML) systems have emerged, aiming to automate various tasks within the ML pipeline with minimal human intervention (Hutter et al., 2019). AutoML holds considerable promise for democratizing ML, enabling domain experts and developers to build applications without relying solely on scarce and expensive ML professionals. While existing AutoML tools facilitate rapid model development and can achieve high performance (Giovanelli & Pisano, 2022; Santu et al., 2022), a critical analysis reveals significant limitations concerning their usability when employed by non-expert users. Despite aiming for automation, these tools often present complex interfaces and workflows that are overwhelming for users without prior ML experience (Margetis et al., 2021; Yang et al., 2020). Challenges related to "capability uncertainty" and "output complexity" persist, making it difficult for non-experts to effectively interact with the tools and understand the implications of different settings or outputs (Margetis et al., 2021; Yang et al., 2020). This lack of adequate usability is a pertinent problem because it prevents non-experts from effectively utilizing the power of automation. Although design guidelines have been proposed, there remains a notable lack of studies focusing on usable visual analytics specifically for AutoML implementation among non-experts (Amershi et al., 2019; Gil et al., 2019). Researchers have recognized the importance of developing usable AutoML tools (Lee et al., 2019), but few studies have explored methods enabling non-expert users to actively participate in building models via AutoML in a truly user-friendly manner. Usable XAI is increasingly recognized as an essential requirement, necessitating well-designed user interfaces that connect algorithmic capabilities to non-expert needs (Cheng et al., 2019; Sokol & Flach, 2020). However, research on usable XAI has largely focused on conventional user interfaces, with less attention paid to interactive visualisations as a crucial component for non-experts (Wang et al., 2022). The limited usability of existing tools thus constitutes a major barrier for the target non-expert audience.

The Problem of Transparency in Existing AutoML and XAI Tools

Compounding the usability issue is the lack of adequate transparency in existing AutoML tools for non-experts. Many current AutoML frameworks and tools, such as auto-sklearn, H2O AutoML, and TPOT, often function as "black-box" systems (Coors et al., 2021; Rabhi et al., 2021). They automate complex processes and generate models without providing sufficient transparency into the underlying steps, algorithmic choices, and decision-making processes (Khuat et al., 2022; Xin et al., 2021). This fundamental lack of transparency is a significant problem because it hinders widespread adoption and use by non-experts, as they struggle to understand how models arrive at predictions, limiting their ability to trust and effectively utilize the outputs (Kaur et al., 2020). The literature highlights that while the field of Explainable Artificial Intelligence (XAI) seeks to address this black-box nature, often through visual analytics (Spinner et al., 2020), the majority of available AI visualization and XAI tools

are designed for AI developers and practitioners rather than domain users with limited AI experience (Hohman et al., 2019). Consequently, non-experts often struggle to accurately interpret the visualizations generated by existing XAI methods (Kaur et al., 2020), finding many XAI approaches ineffective at producing actionable insights or susceptible to misinterpretation (X. Wang & Yin, 2021). This indicates a significant gap between the technical capabilities of XAI and its effective application for enhancing transparency for non-expert users. The lack of transparency in existing tools thus prevents non-experts from gaining confidence and understanding in the ML models they might build or use.

Therefore, a significant problem exists in the current landscape of ML tools: there is a lack of usable and transparent AutoML tools specifically designed for non-expert users that effectively integrate interactive visual analytics and explainable AI principles to support the entire ML model development process and enhance user understanding, particularly for common and widely applicable tasks involving tabular data. This research aims to address this critical gap by designing, developing, and evaluating such a tool.

1.3 Research Questions and Objectives

This research aims to explore new technological approaches to support non-expert learning and the development of ML models. The project will address the following research questions:

1. How to design a web-based tool that supports ML model development and ML model exploration for non-experts?
2. What XAI visualization design principles support the learning and exploration of ML models for non-experts?
3. How effective will the AutoML tool be in enhancing non-expert ML model development experience and understanding?

These research questions will be addressed through the achievement of the following research objectives:

1. To establish the key requirements for designing an intuitive AutoML prediction platform for non-expert users.
2. To create a prototype of a usable AutoML system capable of conducting regression and classification tasks on tabular data, incorporating the established requirements.
3. To quantify and analyse the user experience, usability, and knowledge gained of the developed AutoML system through empirical evaluation.
4. To synthesize evidence-based design guidelines for developing effective and explainable AutoML tools for non-expert users.

Achieving these objectives will collectively provide the necessary insights and deliverables to answer the posed research questions. Specifically, the investigation of requirements and the creation of the prototype will inform the design aspects (addressing the first research question). The empirical evaluation will provide the data to determine the tool's effectiveness (addressing the third research question). Finally, the synthesis of design guidelines will draw upon findings from the other objectives

to articulate principles for design and explanation (addressing the first and second research questions).

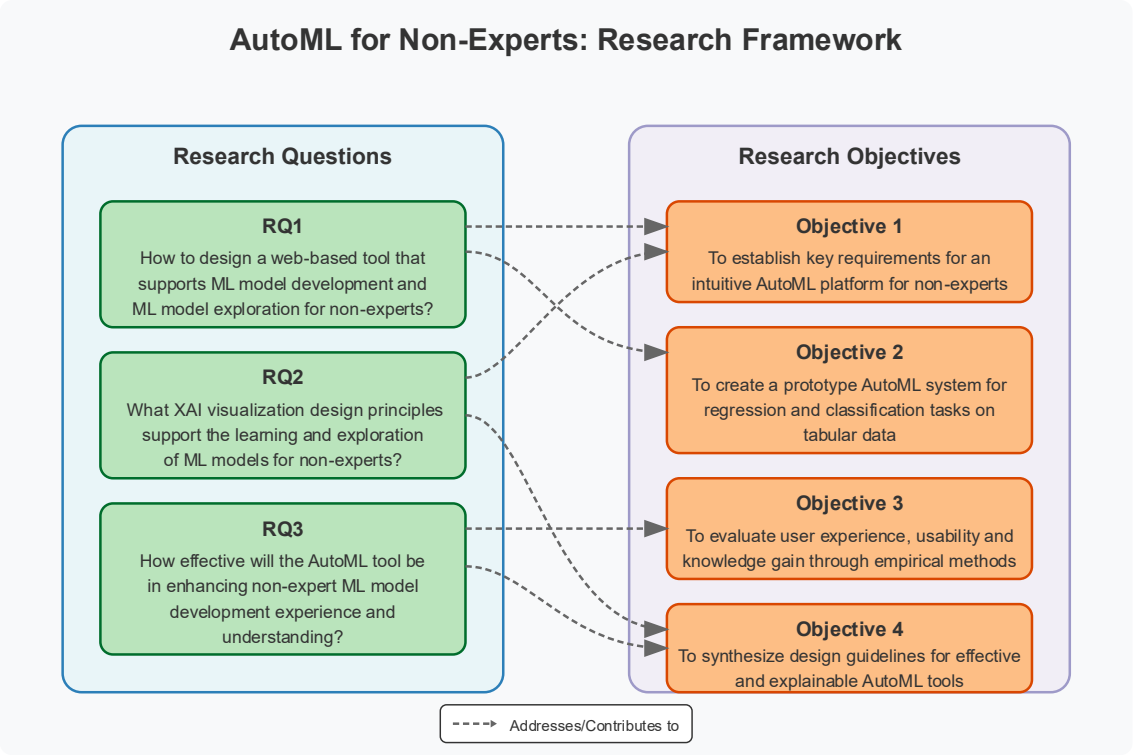


Figure 1 Relationship between each research objective and research question

1.4 Research Scope and Limitations

This research focuses on the design, development, and evaluation of a usable AutoML tool integrated with explainable AI (XAI) visualizations, specifically tailored for non-expert users. The primary aim is to lower the learning curve associated with machine learning (ML) model development and to support users in understanding and evaluating the ML process and its outputs.

To maintain a focused and manageable scope, this study specifically targets tabular data. Tabular datasets are among the most common and widely used formats in real-world ML applications across various sectors such as healthcare, finance, and education.

Moreover, tabular ML development typically requires a sequence of critical and complex tasks such as data cleaning, augmentation, feature scaling, and model selection which pose significant challenges for non-expert users. By focusing the research on tabular data, the project can directly address these challenges, design more effective user assistance mechanisms, and provide concrete, transferable design guidelines.

Additionally, focusing on tabular data allows meaningful comparisons with existing AutoML systems like H2O AutoML and Google AutoML Tables, which predominantly serve tabular use cases. This focus ensures that the research outcomes are both practically relevant and positioned within the current technological landscape. Beyond the choice of data type, the research scope encompasses the following elements:

1. Usability Enhancement in ML Development: The research emphasizes reducing barriers for non-expert users in developing machine learning models. The platform is designed to guide users through the full ML pipeline—including data preprocessing, model selection, training, and evaluation—through a self-guided, supportive interface. Particular attention is given to how users can better understand the steps they are taking and the models they are building, rather than relying on "black box" automation.

2. Development of an Integrated ML and XAI Interface: A central component of the project is the design and development of an interface that combines AutoML functionalities with explainable AI (XAI) visualizations. The system requirements and conceptual framework are derived from a systematic review of peer-reviewed literature, ensuring that the design aligns with validated user needs and best practices in usability and transparency.

3. Evaluation of User Experience and Learning Outcomes: The developed platform is evaluated through a comprehensive set of user studies. These studies apply established instruments such as the User Experience Questionnaire (UEQ) (Schrepp et al., 2017b) to assess user experience across six dimensions, and the System Usability Scale (SUS) (Brooke, 1996) to measure overall usability. Evaluation metrics also include assessments of users' knowledge acquisition and understanding of ML concepts.

4. Mixed-Methods Evaluation Approach: The evaluation combines quantitative and qualitative methods. Quantitative data from questionnaires are complemented by qualitative feedback to gain deeper insights into user behaviour, challenges, and perceptions. This mixed-methods approach ensures a holistic understanding of the platform's effectiveness and user impact.

Limitations:

Given the focused nature of the research, several limitations are acknowledged. First, the scope is confined to tabular data and does not extend to other data modalities such as images, text, or time series. Second, the research primarily targets users with little to no prior ML experience; therefore, findings may not directly generalize to expert user populations. Third, the evaluations were conducted within controlled settings, which, while necessary for consistency, may not capture all complexities of deployment in real-world operational environments.

Overall, this research aims to contribute to the democratization of machine learning by delivering actionable design guidelines for building usable AutoML systems for non-expert users, with a particular emphasis on transparency, guidance, and user empowerment in tabular ML development.

2 Literature Review

2.1 Introduction

This chapter presents a critical review of literature at the intersection of machine learning accessibility, automated systems, and explanatory frameworks to establish the foundation for this thesis's contribution to democratizing AI technologies.

Beginning with an examination of fundamental AI and machine learning concepts, the review systematically analyses Automated Machine Learning (AutoML) and Explainable AI (XAI) approaches through the specific lens of non-expert usability. Rather than treating these as separate technical domains, this chapter synthesizes cross-cutting themes that emerge when considering the practical challenges of making sophisticated ML capabilities accessible to users without specialized training.

The review then presents a structured analysis of existing tools designed for non-expert ML engagement, evaluating their effectiveness against empirically identified user needs. This evaluation reveals significant limitations in current approaches—particularly in balancing automation with meaningful understanding, and in translating complex model explanations into actionable insights for non-technical users. Through this critical assessment, the chapter identifies a persistent disconnect between technical capabilities and practical accessibility that remains inadequately addressed in both research and commercial implementations.

Building on this analysis, the chapter articulates the specific research gaps this thesis addresses: (1) the insufficient integration of explainability principles within user-centered AutoML interfaces; (2) the lack of empirically validated design frameworks specifically calibrated for non-expert ML interaction; and (3) the need for visual analytics approaches that transform model transparency from a technical feature into a meaningful user experience.

The chapter concludes by establishing the theoretical foundation and design principles that guide the development of VisAutoML—an integrated solution that addresses these gaps through a novel synthesis of automated machine learning capabilities with intuitive visual analytics tailored to non-expert cognitive and practical requirements.

2.2 Foundational Concepts

2.2.1 Machine Learning (ML)

Machine learning is a subset of artificial intelligence. It is a vast and expanding field, with applications ranging from image classification, portfolio management, recognition of spoken language, and sentiment analysis. The machine learning process, as illustrated in Figure 2 and supported by recent studies, involves a series of key stages. Initially, data is collected from various sources, including experiments, simulations, and databases. This raw data then undergoes feature processing, which may include normalization and dimensionality reduction, to extract relevant information. As Li et al. (2020) highlight, this processed input is then fed into machine learning models, such as artificial neural networks (ANNs), support vector machines (SVMs), or decision trees, to learn the relationship between the input features and the desired output, which could be labels or predictions. The effectiveness of this learning process is highly dependent on the quality and representation of the input data, as well as the selection and optimization of the machine learning model. Algorithms for machine learning can be divided into supervised, unsupervised, and various other subcategories. Other classification of machine learning methods includes clustering, classification, regression, and decision trees (Alzubi et al., 2018). Machine learning has made tremendous progress with the

advancement of Big Data because the efficiency of machine learning methods is primarily dependent on huge volumes of datasets that are available (Waring et al., 2020). There are 2.5 quintillion bytes of data created every day (Elshawi et al., 2019). As such, Big Data refers to a vast amount of structured and unstructured data that is so massive and complex that it cannot be processed utilising traditional methods (L'heureux et al., 2017).

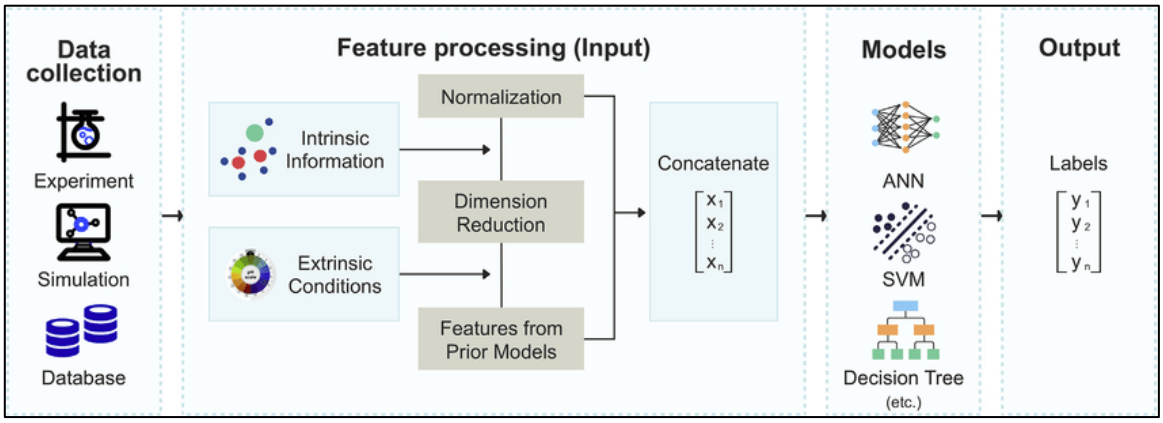


Figure 2 Learning Process in ML (Li et al., 2020)

Hence, it has become extremely difficult to extract usable information from data. On the one hand, machine learning approaches can produce more reliable results when more data is available (Elshawi et al., 2019). On the other hand, this situation raises the possibility of a data science crisis because it is critical to have a growing number of data scientists with in-depth knowledge and skills for them to keep up with harnessing the power of the enormous volumes of data that are produced every day (Roh et al., 2019). The performance of machine learning approaches, however, is very sensitive to a wide range of design variables, which is a substantial hurdle for new users. In a typical machine learning application, for example, professionals must use proper data preprocessing, feature engineering, feature extraction, and feature selection procedures to make the data usable for machine learning. Following these pre-processing phases, data scientists must select an algorithm and optimise hyperparameters to maximise the prediction performance of the final machine learning model (Khalid et al., 2014).

Supervised learning is trained using labelled instances from the labelled dataset, such as an input with known output (Ang et al., 2015). The algorithm is used to learn the mapping function y = f(X) from the input (x) to the output (y). The goal is to accurately approximate the mapping function (f) to anticipate the output (y) given new data as shown in Figure 3. For example, a boolean-function component could include training data points labelled as F (false) or T (true). The goal of supervised learning is to build an estimator that can predict the label of an item given a set of features. The function output can be regression or classification.

Training and testing are two phases of the learning process in a supervised model. An ML algorithm uses samples from the training dataset as input throughout the training process. The learning algorithm detects patterns in the training data and generates an ML model that identifies these patterns. During the testing procedure, the previously produced learning model makes predictions for the data using the execution engine. This method produces labelled data, which provides the final prediction or classified data.

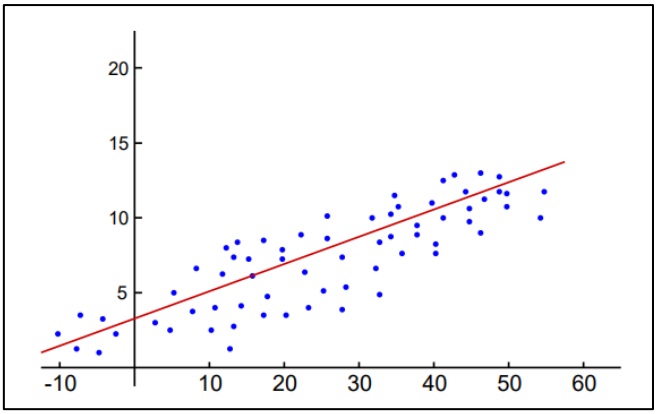


Figure 3 Visual representation of Linear regression (Ang et al., 2015)

The goal of regression is to model the relationship between variables (Kostopoulos et al., 2018). Based on previous data, the algorithm attempts to predict a value for the input. Linear regression is a popular algorithm in this field. These identify the "best-fit line" (Regression line) with the most or closest number of data points. It is the most popular since it is easy to understand and use. Linear regression is a highly adaptable technique that can be used to predict. For instance, house prices in a given area, daytime temperature, and product sales.

Classification techniques are ways of categorizing (dividing) items into groups. The algorithm learns from the provided dataset and categorises each new observation into one of several classes or groups (Saravanan & Sujatha, 2018). In contrast to regression, the outcome variable of classification is a category rather than a value. There are two classification methods as illustrated in Figure 4:

1. Binary Classification: There are only two possible outcomes for the classification.
   -Example: hot OR cold, male OR female, day OR night.
2. Multiclass Classification: There are more than two possible outcomes for the classification.
   -Example: Classification of images, movie categories, shapes.

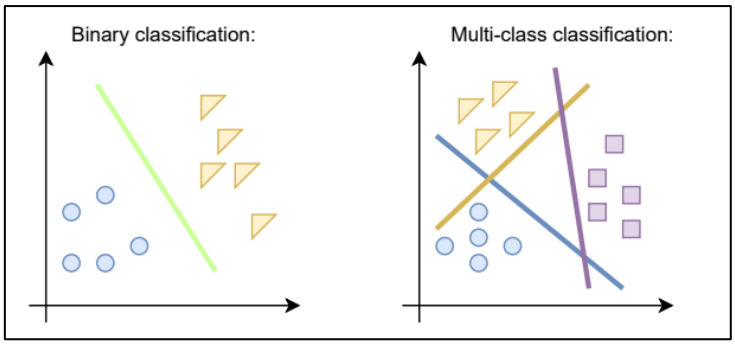


Figure 4 Visual representation of Binary and Multi-class classification

2.2.2 Automated Machine Learning (AutoML)

Open source frameworks and technologies, such as automated machine learning (AutoML), have been developed in response to the desire for machine learning that is easily available, significantly lowering the effort required by users. There are various definitions of AutoML. According to (Hutter et al., 2019) AutoML allows domain experts to automatically develop machine learning applications without the requirement for statistical or machine learning training. AutoML is characterised as a broad field that involves automating any step of the machine learning application process. In general, AutoML seeks to produce data-driven, objective, and automated decisions: the user submits data, and the

framework automatically decides the optimal method for that specific application. As a result, AutoML makes cutting-edge machine-learning approaches available to domain experts who have basic knowledge and are interested in implementing machine learning but lacks the resources to understand the technologies underlying them (Elshawi et al., 2019). There are now a variety of tools and platforms that attempt to automate the machine learning pipeline (M. A. Zöller & Huber, 2021).

Scikit-learn

Scikit-learn, sometimes known as sklearn, is a Python machine-learning library with extensive documentation. Scikit-learn is well-maintained and dependable, providing a diverse set of machine learning methods such as classification, regression, clustering, model selection, and preprocessing. Preliminary work with scikit-learn includes dataset loading and manipulation, imputation of missing values, and preprocessing metrics. Scikit-learn prioritises efficiency, ease of use, documentation, and API consistency (Pedregosa et al., 2011). Many tools and frameworks are based on or expand scikit-learn. The most significant advantage of scikit-learn is its large collection of ML-Algorithms. The majority of them can be utilised with minor code changes.

Tree-Based Pipeline Optimization Tool

TPOT is a Python AutoML tool for automating the construction of machine learning pipelines. It adds its basic regressor, classifier, and cluster algorithms to the scikit learn framework (M. A. Zöller & Huber, 2021). A configuration file can limit TPOT's search space, resulting in a speed advantage (early stopping) (M. A. Zöller & Huber, 2021). Furthermore, this system uses sklearn for its data estimators and manipulators (M. A. Zöller & Huber, 2021). TPOT also allows the specification of a maximum execution time. This framework also allows you to pause and resume execution. TPOT cannot process natural language inputs automatically, nor can it process categorical strings (Olson et al., 2016). Before sending data, the entries must be encoded as an integer. Furthermore, model evaluation is not incorporated into the TPOT process.

Hyperopt-Sklearn

Based on Hyperopt, Hyperopt-Sklearn is an open-source Python framework for fitting classification and regression pipelines. As a result, to depict a big hyperparameter optimization issue, the framework selects an appropriate classifier and preprocessing module (Komer et al., 2019). Additionally, Hyperopt presents a search space that includes modules for categorization, preprocessing, and regression (Komer et al., 2019). The most significant advantage of Hyperopt-Sklearn is that the hyperparameter search can be applied to existing Python programs with minimal changes (Komer et al., 2019). The configuration evaluation cannot be parallelised in Hyperopt-Sklearn. Another limitation is that this framework's pipeline is limited to only one preprocessor and one classification or regression method (Komer et al., 2019).

auto-sklearn

auto-sklearn extends the sklearn framework to develop a machine learning pipeline automatically (M. A. Zöller & Huber, 2021). Numerous supervised machine-learning techniques can be utilised immediately after installing the framework. It also incorporates feature engineering methods such as the imputation of missing data and feature or sample normalization. For common classification and

regression problems, the models employ several sklearn estimator strategies (M. A. Zöller & Huber, 2021). For optimization, Auto-sklearn employs Bayesian search. One of the most significant advantages of this platform is its ease of integration into the existing sklearn toolset, which allows for future expansion. To quickly analyse model performance, Auto-sklearn employs the Bayesian optimization algorithm SMAC3 (M. A. Zöller & Huber, 2021). The inability to interpret input in natural language and the capacity to automatically differentiate between numerical (regression) and categorical (classification) inputs are both disadvantages of this framework (M. A. Zöller & Huber, 2021). As a result, the type must be explicitly specified. Another drawback is that the framework does not support string inputs. As a result, explicit encoding of integer values is required (M. A. Zöller & Huber, 2021).

AutoWeka

Auto-Weka is widely regarded as the first and most innovative machine-learning automation framework (Kotthoff et al., 2019). It was written in Java and built on top of Weka, a prominent machine learning toolkit with a diverse set of machine learning algorithms. For algorithm selection and hyper-parameter optimization, Auto-Weka employs Bayesian optimization using Sequential Model-based Algorithm Configuration (SMAC) and tree-structured parzen estimator (TPE) (Hutter et al., 2011). The Weka optimization algorithm is SMAC by default, although the user can set up the tool to use TPE. SMAC tries to estimate the predictive mean and variance of algorithm performance along the trees of a random forest model to draw a relationship between algorithm performance and a given set of hyper-parameters. The key advantage of applying SMAC is its robustness, which allows it to rapidly dismiss low-performance parameter configurations after evaluating them on a small number of dataset folds. When compared to TPE, SMAC performed better in terms of experimental findings (Hutter et al., 2011).

Google Cloud AutoML

Google's AutoML Tables develop and deploy machine learning models on structured data automatically. It automates model development on a broad variety of data types, including numbers, classes, texts, timestamps, and lists. This tool also detects schema and class distribution, assists in the detection of missing values, and allows for model interpretation. The most significant advantage of this tool is that all of these operations may be completed without any code. Binary classification, multi-class classification, regression, and clustering are all supported by AutoML Table. It also automates input analysis, feature engineering, model selection, model evaluation, and model deployment. Python, Java, and Node.js bindings are available for Google Cloud AutoML Tables (Elshawi et al., 2019).

2.2.3 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (AI) is a recent research area that aims to give end-users transparent AI solutions (Adadi & Berrada, 2018). The goal of XAI is to produce machine learning systems for providing trustworthy and comprehensible explanations for decisions made by black-box models (Kremers, 2020). According to several studies (Kremers, 2020; Schoenborn & Althoff, 2019), explainability is defined mathematically, whereas other studies (Rai, 2020; Ripley, 2007) argue that to increase human understanding, explainability should also include other non-technical components. Whereas the ML and AI groups are focused on explanations to rationalise model decision mechanisms, complicated areas require explanations for how a decision was taken for trust and risk-related issues (Emmert-Streib et al., 2020). As a result, while the ML and AI communities seek mathematical explanations to model behaviour, most domain experts want to understand the

decision-making process behind classifiers. Therefore, different research communities, end-user objectives, and domains contributed to a variety of XAI definitions and ideas (Das & Rad, 2020). Explainable AI can be defined as a collection of techniques and procedures intended to increase the trustworthiness and openness of AI systems. Explanations are characterised as supplementary metadata of an AI model that offers an understanding of the inner workings of a particular AI solution or an AI model as a whole.

XAI is important as it improves transparency and fairness by creating human-understandable justifications and, when used correctly, detects and avoids conflicting examples (Goodfellow et al., 2014). Transparency is imperative in assessing the quality of predictions and reducing errors. Incorrect sample data can convince the wrong image that it is real, preventing the classifier from making the correct categorization. As we increasingly rely on autonomous algorithms to support our daily lives, it is important to ensure the quality of AI algorithms (Kurakin et al., 2016) and transparency in model understanding, and textual or visual reports. The trustability of ML models is a measure of human confidence in the anticipated operation of a particular model in dynamic real-world situations. Thus, understanding 'why a particular option was chosen' is critical for increasing the trust (Rossi, 2018) of end-users, including experts, developers, legislators, and non-experts (Arrieta et al., 2020). Fundamental explanations for classifier prediction are increasingly crucial for stakeholders and governmental organizations as we migrate to a networked AI-driven socioeconomic environment.

Additionally, XAI improves fairness and aids in the mitigation of biases induced into AI decisions by input datasets. Bias in AI algorithms refers to the learned model's disproportionate weight, prejudice, favour, or tendency toward subpopulations of data caused by intrinsic biases in human data collecting and inadequacies from the learning algorithm. Using XAI techniques to learn model behaviour for varied input data distributions, we could learn more about the biases and variations in the data input. This might lead to a powerful AI model (Zou & Schiebinger, 2018). Understanding the input area may allow a better insight into bias mitigation techniques and advocate for more fair models. In XAI, fairness refers to a learned model's capacity to make unbiased and reasonable decisions without favouring a certain subgroup in the dispersion of data input. To better understand fairness in AI, XAI techniques might be used to increase expressive power and provide relevant explanations for feature associations across a variety of data distribution subpopulations. We may understand the subset of characteristics associated with particular class-wise judgments by applying XAI techniques to trace back the prediction biases to the input (Du et al., 2020). Recently, a myriad of XAI approaches was developed to explain the underlying processes of black-box models and their outcomes. XAI techniques are broken down into three levels: explanation level, implementation level, and model dependency level (Alicioglu & Sun, 2022).

An XAI technique's level of explanation indicates whether it is focused on the overall model or a particular instance. The explanation level is divided into two subcategories: local level and global level. The local level explains a model's decisions within a single data instance, while the global level explains an entire model's decision-making process. While some XAI methods, such as the Distillation technique (Tan et al., 2018), Generalized Additive Models (GAM) (Caruana et al., 2015), and Bayesian Rule Lists (BRL) (Letham et al., 2015) offer global-level explanations of an entire model and its underlying processes, others, like Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017), Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017), Deep Learning Important Features (DeepLIFT) (Shrikumar et al., 2017), Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016), offer local explanations.

The implementation level is divided into two categories which are intrinsic and posthoc explanations. Intrinsic explanations, such as Mean Decrease Impurity (MDI) (Breiman, 2002) and Bayesian Rule Lists

(Letham et al., 2015) are produced by the model in terms of how a prediction was developed utilising decision trees, and model parameters. On the other hand, post hoc explanations give insight into black-box models' underlying workings and choices. Many posthoc XAI explainers, such as Saliency Maps (Simonyan et al., 2013), LIME (Ribeiro et al., 2016), Grad-CAM (Selvaraju et al., 2017), Integrated Gradients (Sundararajan et al., 2017), and Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), were developed to convert black-box models into understandable models.

Model dependency is made up of explainers that are both model-specific and model-independent. A single type of algorithm can be explained using model-specific XAI techniques. Intrinsic explanations are model-specific methods, which means they cannot be applied to any model without altering the explanation process (Das & Rad, 2020). In comparison, model-agnostic explanations work with any explanation model and are unaffected by model architecture. Because the majority of explainers that are model-agnostic include post-hoc explanations, these explainers are widely utilised because of their versatility (Adadi & Berrada, 2018). For instance, SHAP (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), ANCHORS (Ribeiro et al., 2018), and LRP (Bach et al., 2015) are the most commonly used explainers that are model-agnostic and post hoc.

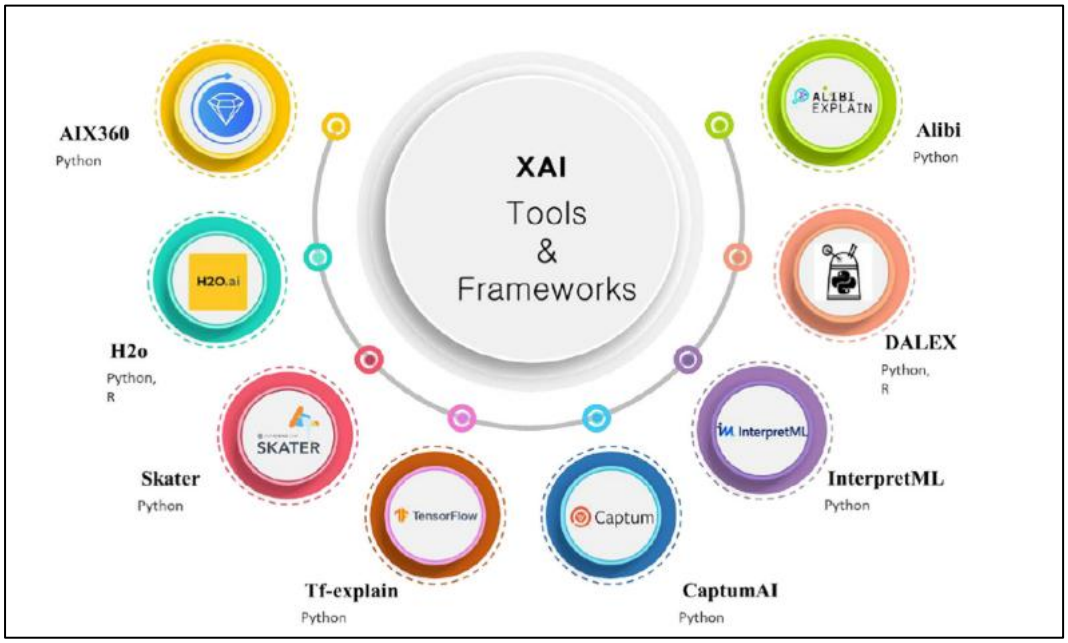


Figure 5 XAI common tools and frameworks (Ding et al., 2022)

As the proliferation of XAI increases, it is becoming increasingly necessary for the academic community to establish a set of tools and frameworks to assist in the implementation and replication of alternate explainability techniques for varying levels of stakeholders. Accordingly, the following are some popular open-source XAI frameworks.

- **AIX360**: This toolkit provides a unified and simple programming layout as well as shared architecture to facilitate various explainability approaches required by a wide range of users. The toolkit provides a few qualities for how "good" an explanation is.

- **H20**: This toolkit provides a detailed insight into how AI algorithms work, both from a local perspective and from a global perspective.

- **Skater**: A Python framework that blends model interpretation for multiple models for machine learning that have practical uses. It allows for both local and global explanations of the learned mechanisms of a black box model in classification or regression tasks.

- **InterpretML**: This is a Python package (Nori et al., 2019) that provides two explanation approaches: glass-box AI models designed for explanation strategies used to understand current AI solutions.

- **Tf-explain**: This framework explains models created with TensorFlow utilising gradient and activation maps.

- **EthicalML-XAI**: This is a machine learning tool for building core explainability in ML systems using an explorative model and data analysis.

- **CaptumAI**: This is a framework for PyTorch models that is open-source and interpretable (Kokhlikyan et al., 2009). It divides explanation approaches into three categories: primary level, layer level, and neuron level. It is simply extensible with additional features and methods and can manage various data types.

- **DALEX**: This is a toolkit developed by (Biecek, 2018) to enable the creation of ethical AI solutions that include a variety of methodologies and bridge the present gap that exists between Black-Box models and XAI. It delivers an interactive model explanation through the use of a model-independent interface and a variety of fairness indicators.

- **Alibi**: This is a Python framework published in (Klaise et al., 2021) that includes excellent explanations of global and local ML algorithms and is built to deal with a variety of data in classification and regression applications.

- **The Explainable AI Toolkit (XAITK)**: This is a toolkit created by the DARPA XAI program, an open-source toolset to help users, academics, and technicians evaluate complicated AI results. It combines a traversable store of independent properties with a unified, widely used software framework.

- **What-If Tool**: A framework for providing an easy-to-understand graphical user interface that makes it possible to comprehend complex AI models for classification or regression. It enables users to make a prediction on a greater number of data samples and view the results in a variety of ways.

Explainable User Interface (XUI)

The XAI process is broken down into two parts, as stated by the DARPA XAI program. It isolates understanding the behaviour of the ML model from showing it to the user by dividing the explainable model and the explanation user interface (Gunning & Aha, 2019). An explanation user interface (XUI) is defined as the total of an XAI system's outputs with which the user could interact directly. To give relevant insights to a specific audience, an XUI may employ the ML model or several explanation-generating systems. The construction of interfaces that enable users to better comprehend ML processes is seen as a major challenge in HCI research (Shneiderman et al., 2016). The XAI process may

be divided into many parts. Murdoch et al. distinguish between an XAI system's descriptive accuracy, prediction accuracy, and relevance. The extent to which the explanation generation approach properly depicts the behaviour of the learned ML model (also known as fidelity) is referred to as descriptive accuracy. Predictive accuracy measures how well the taught ML model identifies the underlying data connections. Both accuracies may be assessed objectively. Contrarily, subjective relevance refers to whether the outputs are displayed in a way that provides insights for a particular audience within a given area of inquiry (Murdoch et al., 2019). Explanatory XUIs attempt to convey a single explanation through a visual or written representation. Exploratory XUIs, on the other hand, allow users to independently investigate the ML model behaviour (Margetis et al., 2021). They work best when inputs can be modified or affected by users. Additionally, Arya et al. distinguish between dynamic and static explanations. Static explanations do not change in response to user input (Arya et al., 2019). Conversely, interactive explanations allow consumers to focus on or request multiple forms of explanations until they are satisfied. Vilone et al. describe interaction as an explanation technique that allows reasoning preceding input to comprehend and answer follow-up queries (Vilone & Longo, 2020). A good explanation infrastructure, according to Moore and Paris, should also be responsive (allowing follow-up queries), flexible (using a variety of explanation techniques), and sensitive (the explanations should take into account the user's knowledge, aim, context, and past interactions) (Moore & Paris, 1991). The following are design principles of interactive XUI based on the work of (Chromik & Butz, 2021).

Responsiveness through Progressive Disclosure

This design principle can be defined as the continuous introduction of subsequent or stepwise functionality based on previous explanations. Literature has suggested that there is a narrow line between providing no explanation and providing excessive explanations (Millecamp et al., 2019). The user's individual preference for cognition affects this threshold. Extraneous information overwhelms those who may be using a simpler mental model of the situation associated ML system.

Sensitivity to the Mind and Context

This design principle can be defined as the adaptation of explanations to the user’s mental processes and situations. The user's explanation demands change as one gain insight and confidence during an engagement process (Liao et al., 2020). Furthermore, users' preexisting assumptions and prejudices impact how they react to various explanation approaches. This necessitates a tailored approach to explaining ML systems.

Complementary Naturalness

This design principle can be defined as the notion of adopting theoretical frameworks in human language to supplement intuitive explanations. Typically, implicit visualisation explanations that can precisely communicate the underlying mechanisms of an AI system are unavailable to non-experts. Thus, post-hoc justifications in natural language are meant to approximate a human explainer's answer in the same scenario (Ehsan et al., 2019). When the condition of a system is ambiguous or unclear, relaying data by textual description may relieve human users' concerns (Robb et al., 2019). The combination of visual cues and written explanations can promote trust and comprehension (Ehsan et al., 2019).

Flexibility through Multiple Ways to Explain

This design principle can be defined as the option of a variety of explanation approaches and modes to allow users to corroborate ideas. Humans develop understanding in a variety of ways. Paez et al. divide them into two categories: subjective understanding acquired via experiences and exemplifications and objective understanding achieved via conceptions and simplified representations (Páez, 2019). In practice, there is frequently no ideal technique for describing anything. For instance, a doctor's clinical diagnosis rarely rests on one level of evidence (Xie et al., 2019). Explanation approaches and modes can thus supplement one another.

XAI Interaction Concepts

Interaction, according to Hornbk and Oulasvirta, characterises the interaction of several constructions. They investigated how the human-computer interaction models defined in HCI research interacted and created seven notions of interaction from this: interaction as information transmission, interaction as dialogue, interaction as control, interaction as experience, interaction as optimum behaviour, interaction as tool usage, and interaction as embodied action (Hornbæk & Oulasvirta, 2017). While according to another study, within a human-agent XAI interaction problem, an explanatory agent provides the motivations behind its own or another agent's decisions (Miller, 2019). As a result, it is focused on how a user interacts with an AI agent through the utilisation of an XUI.

Interaction as (Information) Transmission

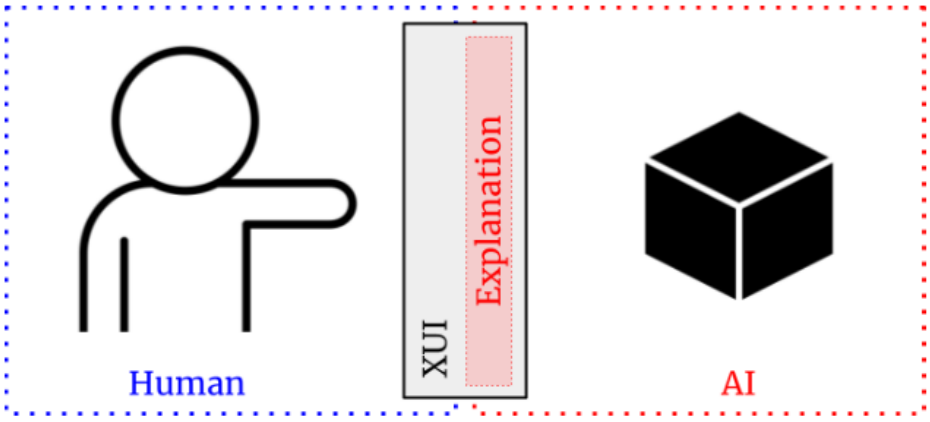


Figure 6 XAI interaction as information transmission

Maximizing information transmission across a noisy medium is the primary idea behind this concept. From a pool of available messages, the interaction selects the best message for dissemination (Hornbæk & Oulasvirta, 2017). This concept adheres to the Shannon-Weaver (Shannon, 1948) communication paradigm, wherein noise is interjected between the transmission of information from the transmitter to the receiver. The purpose of this interaction is to provide consumers with a single comprehensive explanation. Because it can be challenging, if not impossible, to fully explain the complexity of the AI in a way that is understandable to humans, the message is typically cluttered. The XUI serves primarily as a means for communicating this explanation.

Interaction as Dialogue

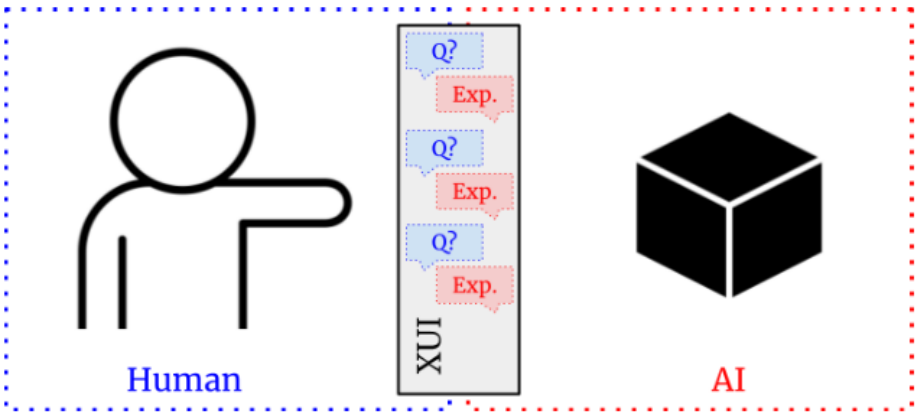


Figure 7 XAI interaction as dialogue

This idea depicts the inputs and outputs data communication loop and human perception/action. Interaction occurs in phases or turns (Hornbæk & Oulasvirta, 2017). It helps to make sure the precise mapping between UI features and user intentions, as well as feedback from the UI, to bridge the execution gap (Anderson, 1988). This concept recognises that one explanation rarely succeeds at the appropriate level of comprehension (Abdul et al., 2018). Rather, it stresses the naturalness and accessibility of explanations (sometimes implicit or simplified). The interaction's purpose is to give users functions that will allow them to progressively construct a conceptual framework of the AI's behaviour.

Interaction as Control

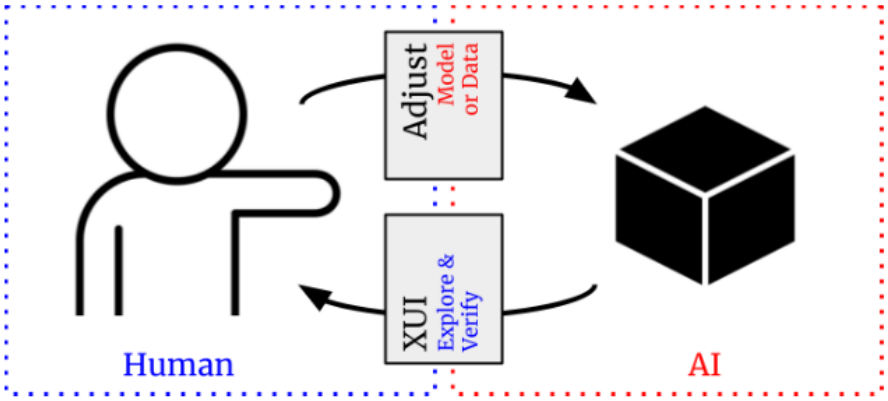


Figure 8 XAI interaction as a control

This approach promotes the human-computer system's efficient and sustainable convergence towards a target state. The interaction, which is based on control theory, aims to adjust a command signal to the desired level and modify its behaviour in response to input. This proposal is similar to the concepts of ML model modification and interactive ML (Dudley & Kristensson, 2018). The ML model sends control signals to the human controller through the XUI. These instruct the controller on how to modify the parameters of the ML model or data such that the model changes its behaviour. The goal of the interaction is to produce the AI behaviour that the controller intends.

Interaction as Experience

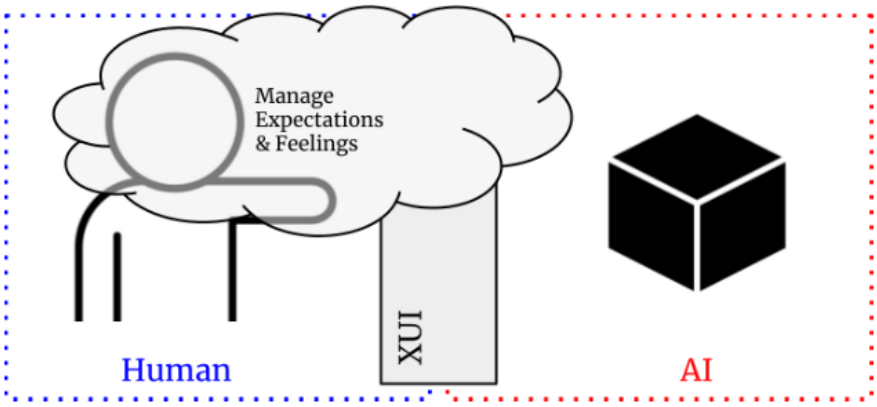


Figure 9 XAI interaction as experience

The emphasis of this concept is on what humans anticipate from a machine. Highly similar to user experience (UX), It comprises an individual's thoughts, feelings, and perceptions before, during, and after contact (Law et al., 2009). When implemented in XAI, this interaction paradigm focuses on controlling humans' preferences and expectations towards AI. Its explanatory aims are around trust (Pilling et al., 2020), satisfaction (Tsai & Brusilovsky, 2019), and persuasiveness (Eiband et al., 2019).

Interaction as Tool Use

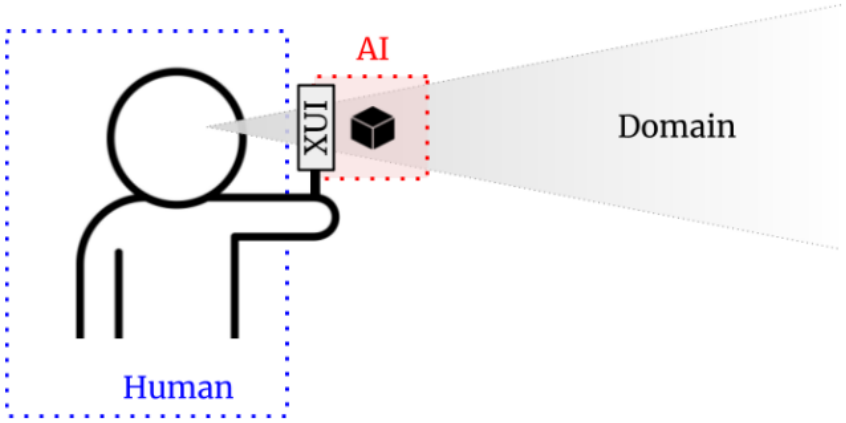


Figure 10 XAI interaction as tool use

This concept is on utilising computers to enhance a user's potential above what the tool itself can offer. Individuals' mental functioning is influenced by the system, according to activity theory. As such, AI may be utilised as a learning tool. Word embeddings, for example, are used as a screening tool in the social sciences to measure societal changes (Garg et al., 2018). This interaction concept, when applied to XAI, assists people in uncovering hidden trends and correlations in domain-specific data. A degree of explanation is essential to promote this learning. The XUI provides insight into an area outside of AI function that would otherwise be challenging to understand. Through this interaction, human cognition can be enhanced.

Interaction as Embodied Action

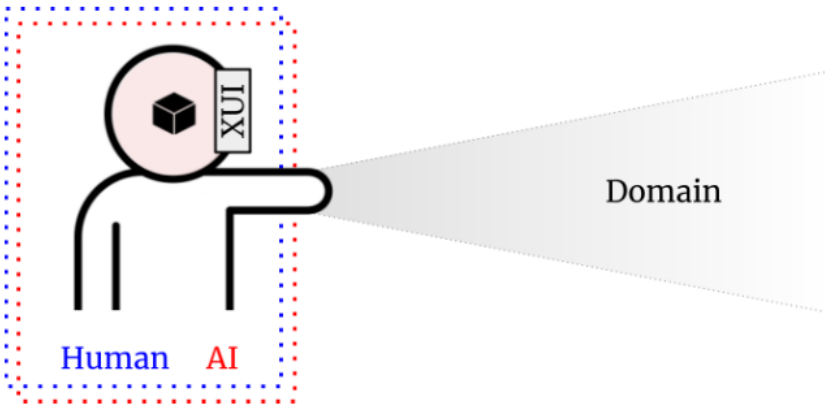


Figure 11 XAI interaction as embodied action

This interaction notion focuses on collaborative combined activity with a machine. Licklider proposed man-computer symbiosis in 1960, stating that "Computers and men are to collaborate in creating judgments along with handling complex circumstances" (Licklider, 1960). Collaboration with AI may allow humans to be magnified. Effective cooperation, on the other hand, extends beyond interaction. This concept also draws on community-based computer-assisted cooperative work (CSCW) notions like mutual objective awareness, proactive goal organization, and shared development monitoring (Oh et al., 2020). Explanations are a critical component of good XAI interaction. Dissatisfaction was attributed to a lack of explanatory transmission (D. Wang, Churchill, et al., 2020). In this approach, XUIs aid in the enhancement of human behaviours. This is especially true for autonomous systems, which have a symbiotic connection. In high-risk settings, autonomous systems have a high degree of autonomy and must communicate what these systems intend to do and the reasoning behind it (Hastie et al., 2018). In such a situation, humans and machines must understand one another's capacities and intended future actions toward a common objective, preferably in real time.

2.2.4 Visual Analytics Overview

Visual analytics is an interdisciplinary field that combines data visualisation, human-computer interaction, and data analysis to facilitate analytical reasoning and decision-making. The field of visual analytics emerged from the need to make sense of increasingly large and complex datasets. Thomas and Cook (2005) defined visual analytics as "the science of analytical reasoning facilitated by interactive visual interfaces." This seminal work laid the groundwork for the field, emphasising the importance of both computational analysis and human insight in the analytical process.

One of the core principles of visual analytics is the integration of automated analysis techniques with interactive visualisations. Keim et al. (2008) proposed a visual analytics process model that outlines key steps including data preprocessing, visualisation, model building, and knowledge generation. This model has been influential in shaping subsequent research in the field. Interactive techniques play a crucial role in visual analytics. Yi et al. (2007) categorized common interaction techniques in information visualisation, such as selection, exploration, and reconfiguration. These techniques enable users to dynamically explore and analyse data, uncovering patterns and insights that might not be apparent in static visualisations.

As datasets grow larger and more complex, scalability has become a significant challenge in visual analytics. Liu et al. (2013) addressed this issue by proposing techniques for visualising high-dimensional data. Evaluation of visual analytics systems remains a complex task due to the involvement of human factors. Lam et al. (2012) provided a comprehensive review of evaluation methods in information visualisation and visual analytics. This work continues to be relevant, offering guidance for researchers on how to assess the effectiveness of their systems. The application of visual analytics spans numerous domains. In the healthcare sector, Rind et al. (2013) demonstrated the use of visual analytics for exploring electronic health records, enabling clinicians to identify patterns in patient data. In the realm of cybersecurity, Shiravi et al. (2012) presented visual analytics techniques for network security monitoring, helping analysts detect and investigate potential threats.

Recent trends in visual analytics include the exploration of novel display technologies and interaction modalities. Isenberg et al. (2013) surveyed the use of large, high-resolution displays in visual analytics. Advancing this line of research, Fonnet and Prié (2021) reviewed the state of immersive analytics, exploring the potential of virtual and augmented reality for data visualisation and analysis. As the field evolves, researchers are also addressing ethical considerations in visual analytics.

The rise of artificial intelligence and machine learning has led to new opportunities and challenges in visual analytics. Endert et al. (2017) discussed the integration of machine learning models into visual analytics workflows, highlighting the potential for human-in-the-loop systems. Building on this work, Spinner et al. (2020) proposed a conceptual framework for explainable artificial intelligence (XAI) in visual analytics, addressing the growing need for interpretability in machine learning models. The intersection of visual analytics and AI has become an increasingly important area of research. Chatzimparmpas et al. (2020) provided a comprehensive survey of visual analytics approaches for explainable artificial intelligence, highlighting the role of visualisation in making complex AI models more interpretable and trustworthy.

Recent work by Hohman et al. (2019) explored the use of visual analytics in deep learning, proposing a framework for visualising and interpreting neural networks. This research demonstrates the potential of visual analytics to demystify complex AI systems and facilitate more informed development and deployment of machine learning models. In the realm of natural language processing, Liu et al. (2018) developed a visual analytics system for analysing and comparing different word embedding models, showcasing how visual analytics can aid in understanding and refining AI techniques for language processing. Looking towards the future, Chen et al. (2020) discussed the challenges and opportunities in visual analytics for artificial intelligence, emphasising the need for new visualisation techniques and interaction paradigms to support the development, debugging, and deployment of AI systems.

The convergence of Artificial Intelligence (AI) and the Internet of Things (IoT), known as AIoT, has created new challenges and opportunities in data analysis and visualisation. This section explores the role of visual analytics in AIoT, highlighting key developments and trends in this emerging field. In the context of AIoT, visual analytics plays a crucial role in making sense of the vast amounts of data generated by interconnected smart devices and AI systems. The ability to visually represent and interact with complex data streams from diverse IoT devices, combined with AI-driven insights, enables decision-makers to gain actionable intelligence from AIoT ecosystems.

One of the primary challenges in AIoT visual analytics is the need to handle real-time, high-velocity data streams. Shi et al. (2016) proposed a visual analytics system for real-time anomaly detection in streaming data, demonstrating the potential of visual analytics to support rapid decision-making in IoT environments. Their approach, which uses compressed graphs to visualise network traffic,

addresses the scalability issues often encountered in large-scale IoT deployments. By enabling analysts to identify patterns and anomalies in real-time data flows, such systems can significantly enhance the security and performance monitoring of AIoT networks.

The integration of machine learning techniques with visual analytics has become increasingly important in AIoT applications. Chen et al. (2020) surveyed visual analytics techniques for machine learning, highlighting approaches that could be particularly relevant for AIoT systems, such as interactive model training and explainable AI visualisations. This integration is crucial in AIoT contexts, where the complexity of AI models and the diversity of IoT data sources can make it challenging for users to understand system behaviours and decision-making processes.

In the realm of smart cities, a key application area for AIoT, Cao et al. (2016) developed a visual analytics system for urban traffic data. Their work showcased how visual analytics can help city planners and decision-makers understand complex patterns in IoT-generated urban data. The system, named Voila, combines spatiotemporal visualisations with anomaly detection algorithms to identify unusual traffic patterns and events. This type of application demonstrates the power of visual analytics in transforming raw IoT data into actionable insights for urban management and planning.

As AIoT systems become more complex, there is a growing need for visual analytics tools that can handle heterogeneous data types. Liu et al. (2017) addressed this challenge by proposing techniques for visualising high-dimensional data, which could be particularly useful for analysing the diverse data generated by AIoT devices. Their survey of advances in high-dimensional data visualisation over the past decade provides valuable insights for researchers and practitioners working on AIoT visual analytics solutions. The ability to effectively visualise high-dimensional data is crucial in AIoT scenarios, where devices may generate diverse data types ranging from simple sensor readings to complex multimedia streams.

The interpretability of AI models in AIoT systems is another area where visual analytics can make significant contributions. Spinner et al. (2020) introduced a visual analytics framework for explainable AI, which could be adapted to help users understand the decision-making processes of AI-powered IoT devices. Their framework, explAIner, combines interactive visualisations with explainable AI techniques to provide users with insights into model behaviour and feature importance. In AIoT contexts, where AI models may be making critical decisions based on IoT data, such explainable visual analytics approaches can enhance trust and facilitate more informed human oversight.

In conclusion, visual analytics plays a crucial role in unlocking the potential of AIoT systems by enabling humans to make sense of complex, high-velocity data streams generated by interconnected smart devices. As AIoT technologies continue to advance, visual analytics will likely become an increasingly important tool for decision-makers across various domains, from smart cities to industrial IoT applications. The ongoing research in this field promises to deliver more sophisticated, scalable, and user-friendly visual analytics solutions that can keep pace with the rapid evolution of AIoT technologies.

2.3 Understanding the Target User: Non-Experts

2.3.1 Defining Non-Expert Users in the Context of AutoML

Research in the field of HCI and ML has revealed that people who develop ML models are either AI practitioners, agents (domain experts), or non-expert users who are neither knowledgeable about ML nor the application domain. Non-expert users, as defined by Bove et al. (2022), are those who are neither knowledgeable about ML nor a particular application domain. This work focuses on the latter non-expert users in developing usable and transparent ML models. We explore the literature on non-expert behaviour towards ML model development and the challenges they faced to ground the requirements for the proposed system.

Non-experts tend to view machine learning algorithms as black-box input-output mechanisms, a mechanism in which data inputs go in, and the prediction and performance metrics are output. This is not an incorrect perception of ML algorithms and will be further implemented in the development process of the proposed system. Within the scope of statistical and topical knowledge, non-experts predominantly do not attempt to use data visualisations or descriptive statistics to understand their data or introspect their models (Yang et al., 2018). This is counter-intuitive to the conventional method of analysing data which is imperative in understanding the relationship between features and the state of the data whether it is biased or not. Moreover, non-experts heavily depended on the documentation of ML tools/APIs, when building a model and relied on publicly available working ML scripts to repurpose them (Yang et al., 2018). This reliance limits their exploration of diverse approaches and innovative model-building techniques.

In the context of AutoML, which aims to automate the development of predictive models (Hutter et al., 2019), the target non-expert user is someone who wants to develop ML applications without requiring statistical or machine learning training. AutoML has the potential to democratise ML by allowing domain experts and developers to easily develop ML applications in their organisations without relying on already limited and expensive ML professionals. For these non-experts to effectively use AutoML tools, the systems need to be designed with their limited prior knowledge and specific challenges in mind, moving beyond being black-box tools that are not understandable from a human's perspective (Xin et al., 2021).

2.3.2 Motivations and Characteristics of Non-Experts

Understanding the motivations and inherent characteristics of non-expert users is a cornerstone for the effective design and implementation of Automated Machine Learning (AutoML) tools. Research delving into the engagement of non-experts with Machine Learning (ML) reveals that their initial forays are often driven by practical, application-oriented goals rather than theoretical curiosity (Yang et al., 2018). These motivations frequently include the desire to extract meaningful insights from datasets they already possess, the need to develop specific, often one-off, ML applications for immediate tasks, or the broader ambition to automate repetitive manual processes within their professional or personal spheres, thereby seeking long-term efficiency gains (Yang et al., 2018). This instrumental perspective underscores that the perceived utility and direct applicability of ML to their specific problems are primary drivers for adoption.

A salient characteristic of non-expert users is their conceptualization of ML algorithms as predominantly input-output mechanisms. This perspective views the ML system as a 'black box' where data is fed in, and predictions or performance metrics are received, without necessarily engaging with or understanding the complex internal computations and decision-making processes (Yang et al., 2018). Correspondingly, non-experts often do not routinely employ standard data analysis techniques, such as data visualisations or descriptive statistics, for the crucial initial exploration of their datasets or for introspecting the trained models (Yang et al., 2018). This contrasts sharply with the typical workflow of experienced data scientists and highlights a significant gap in their data literacy and model understanding practices.

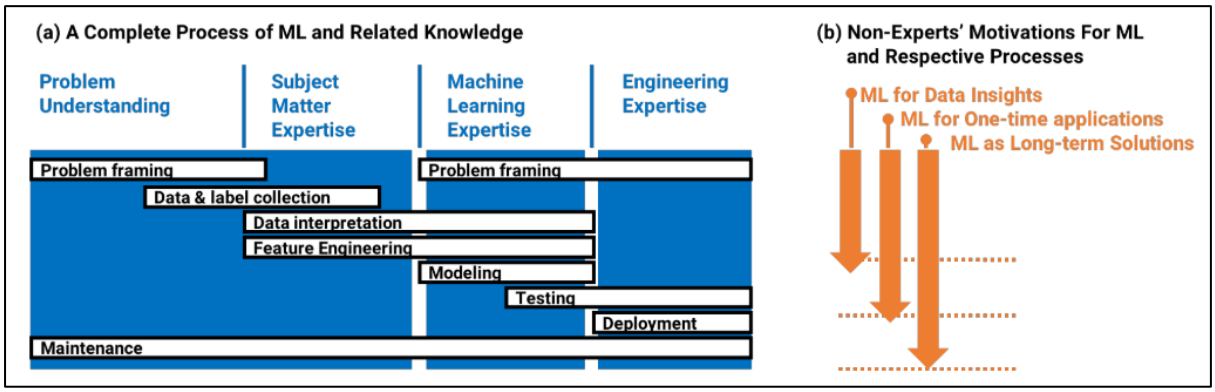


Figure 12 Non-Experts' Motivations For ML (Yang et al., 2018)

Furthermore, the approach non-experts take towards the ML development lifecycle is often marked by a notable dependence on external support structures. They frequently rely heavily on the documentation provided with ML tools and APIs, or they seek out and adapt publicly available working ML scripts to construct their models (Yang et al., 2018). While these resources can provide necessary scaffolding, this reliance can also limit their exposure to alternative methodologies and hinder the development of independent problem-solving skills within the ML domain. It also points to a need for AutoML tools to integrate intuitive guidance and support directly within the user interface, reducing the necessity to constantly refer to external materials.

Beyond the initial model building, non-experts encounter significant challenges in evaluating and improving model performance. They may struggle to fully grasp the inherent limitations of what a particular model can learn from their data and face difficulties in evaluating performance metrics beyond simple accuracy percentages (Ramos et al., 2020; Yang et al., 2018). Metrics such as precision, recall, F1-score, or measures of model generalizability, which are critical for a nuanced understanding of a model's effectiveness, are often overlooked or misinterpreted (Yang et al., 2018). Additionally, converting a real-world problem into a well-defined and feasible ML problem, including appropriate feature design, presents a considerable hurdle, requiring a level of understanding of both algorithm capabilities and data characteristics that non-experts typically lack (Ramos et al., 2020).

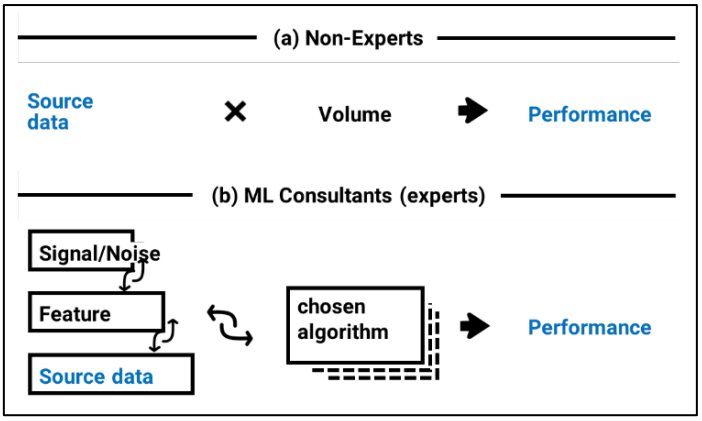


Figure 13 Non-Experts' approaches to improving model performance (Yang et al., 2018)

Collectively, these motivations and characteristics paint a clear picture of the non-expert user as a goal-driven individual seeking practical ML solutions but lacking the foundational knowledge, analytical habits, and technical expertise common among ML practitioners. Their reliance on external guidance and difficulties with model evaluation and problem formulation underscore the critical need for AutoML tools that are not merely automated but are fundamentally designed with a human-centered approach. Such tools must provide intuitive interfaces, integrated educational scaffolding, transparent explanations of automated processes, and accessible interpretations of model outputs and performance to truly democratize ML for this user group.

2.3.3 Key Challenges Faced by Non-Experts in ML Development

The challenges faced by non-experts in machine learning, as identified in the literature, shed light on the nuanced behaviours and misconceptions prevalent in their approach to model development. Non-expert users encounter several significant challenges throughout the Machine Learning (ML) development lifecycle, which hinder their ability to effectively build and utilize ML models. These challenges span various stages, from initial data preparation to model evaluation and improvement, as broadly categorized in Table 3.

Table 1 Major challenges non-experts face in the ML development phase

Major challenges non-experts face	ML development phase
C1: Data analysis prior to ML modelling	Preparation
C2: Problem/feature design	Preparation
C3: ML model evaluation	Evaluation
C4: ML model selection	Modelling
C5: ML model performance and improvement	Modelling

Firstly, non-experts tend to view machine learning algorithms as black-box input-output mechanisms, lacking a comprehensive understanding of their intricate functionalities (Yang et al., 2018). This perception means they may not grasp the underlying processes, which can limit their ability to troubleshoot or adapt models effectively. A primary hurdle, falling under the 'Preparation' phase in Table 3 (C1), is the difficulty in performing adequate data analysis prior to ML modelling. Non-experts

often lack the statistical and analytical skills necessary to understand their data, identify patterns, or assess data quality, which is crucial for successful model development (Yang et al., 2018).

Secondly, there is a notable reliance on documentation and publicly available scripts among non-experts (Yang et al., 2018). This dependence limits their exploration of diverse approaches and innovative model-building techniques. Furthermore, formulating the ML problem and designing appropriate features presents a considerable challenge (Table 3, C2 - 'Preparation' phase). Converting a real-world problem into a solvable ML task requires a deep understanding of both the problem domain and the capabilities and limitations of various ML algorithms. Non-experts may struggle with this conceptual mapping and with the process of feature engineering, which involves transforming raw data into features that improve model accuracy (Khalid et al., 2014; Ramos et al., 2020).

Misinterpretation of model performance metrics represents another significant challenge (Table 3, C3 - 'Evaluation' phase). Non-experts often focus solely on percentage accuracy as the sole metric, overlooking critical metrics like precision, recall, and generalizability (Yang et al., 2018). Model selection (Table 3, C4 - 'Modelling' phase) and evaluation also pose significant difficulties. Non-experts often lack the knowledge to choose the most suitable algorithm for their specific problem and dataset, and they may not fully comprehend the various metrics used to evaluate model performance beyond simple accuracy (Yang et al., 2018). Misinterpreting evaluation results can lead to poor decision-making regarding model deployment or further refinement.

Limited adaptation of learning algorithms is also observed, as non-experts often stick to one approach without exploring alternative algorithms (Yang et al., 2018). Finally, non-experts face challenges related to improving ML model performance (Table 3, C5 - 'Modelling' phase). They may hold misconceptions about how to enhance a model, such as believing that simply increasing the amount of data will suffice, without considering the impact of algorithm choice or feature design (Yang et al., 2018). This lack of understanding regarding the factors influencing model performance limits their ability to iteratively refine their models effectively. Lastly, there is an overemphasis on adding more data as a solution, disregarding the importance of algorithm choice and feature design (Yang et al., 2018).

The literature reveals that non-experts face challenges rooted in limited understanding, reliance on external resources, misinterpretation of metrics, reluctance to explore diverse algorithms, and an overemphasis on data quantity. These challenges, summarized in Table 5 along with proposed user requirements, collectively highlight the need for AutoML tools designed for non-experts to provide robust support, guidance, and transparency throughout the entire ML development process. Addressing these challenges requires specific design considerations, such as those outlined in Table 5, which proposes user requirements based on these identified needs. The aim is to mitigate these challenges by providing educational support, intuitive interfaces, and guidance throughout the model development process, ultimately empowering non-experts to navigate the complexities of machine learning more effectively.

Table 2 Challenges Non-Experts face

Challenge	Description	User Requirement
1. Limited Understanding of ML Algorithms	Non-experts perceive ML algorithms as input-output mechanisms, lacking	Develop an AutoML tool with clear, user-friendly documentation and guided explanations on the functioning of ML algorithms,

	comprehension of their broader functionality.	fostering an intuitive understanding for non-experts.
2. Reliance on Documentation and Public Scripts	Non-experts heavily depend on ML tool documentation and publicly available scripts, limiting their exploration of diverse approaches.	Design an AutoML system with interactive tutorials to encourage diverse model-building approaches and reduce reliance on external documentation.
3. Misinterpretation of Model Performance Metrics	Non-experts focus on percentage accuracy as the sole metric, neglecting crucial metrics like precision, recall, and model generalizability.	Implement an AutoML tool that provides comprehensive performance metrics, accompanied by user-friendly explanations to enhance non-experts' understanding and interpretation of model evaluations.
4. Limited Adaptation of Learning Algorithms	Non-experts tend to stick to one learning algorithm, overlooking the potential benefits of exploring different algorithms.	Develop an AutoML tool with a guided algorithm selection feature, offering insights and recommendations based on the user's dataset characteristics and problem domain.
5. Overemphasis on Adding More Data	Non-experts believe that increasing data quantity alone improves model performance, overlooking the significance of algorithm choice and feature design.	Integrate educational modules within the AutoML system elucidating the importance of algorithm selection and feature engineering, guiding non-experts to make informed decisions beyond data augmentation.

2.4 State of the Art: Existing Tools and Approaches

2.4.1 Existing AutoML Tools and Frameworks

The increasing demand for Machine Learning (ML) applications across various domains, coupled with the scarcity of highly skilled ML professionals, has spurred the development of Automated Machine Learning (AutoML) tools and frameworks. These systems are designed to automate various aspects of the ML pipeline, with the overarching goal of making ML more accessible and efficient. As defined by Hutter et al. (2019), AutoML is a framework that automates the development of predictive models, enabling domain experts to build ML applications without requiring extensive statistical or ML training. This automation encompasses numerous steps within the ML application process,

aiming to produce data-driven, objective decisions with minimal human intervention (Elshawi et al., 2019).

AutoML tools can be categorized based on their target user groups and primary design objectives. Through a comprehensive analysis, several distinct categories emerge, each with specific limitations for non-expert users. These categories include tools oriented towards Domain Experts, Clinician-Focused tools, Research-Oriented Frameworks, and General-Purpose AutoML Platforms. While each category serves valuable purposes within its intended scope, their design often assumes a level of technical or domain-specific knowledge that creates barriers for users with limited ML expertise, a point highlighted by the usability and transparency limitations summarized in Table 6.

The scope of automation within AutoML systems can vary significantly, impacting their accessibility. Santu et al. (2022) propose a taxonomy that classifies AutoML into six levels of automation, ranging from entirely manual processes (Level 0) to fully automated end-to-end workflows (Level 6). Table 4 outlines the key steps involved in the AutoML taxonomy: Task Formulation (TF), Data Visualisation, Cleaning, and Curation (DCC), Prediction Engineering (PE), Feature Engineering (FE), Alternative Models Exploration, Testing, and Validation (ATV), and Result Summary and Recommendation (RSR). Different AutoML tools automate different combinations of these steps, influencing their target user base and the level of expertise required.

Table 3 Steps and descriptions for AutoML taxonomy (Santu et al., 2022)

AutoML Steps	Description
Task Formulation (TF)	Formulating the prediction problem through enquiry with stakeholders.
Data Visualisation, Cleaning, and Curation (DCC)	Analysing data through visualisation, improving the quality of predicting through cleaning data and curating to suit prediction.
Prediction Engineering (PE)	Constructing labels for data points according to the prediction task and creating meaningful training and testing datasets.
Feature Engineering (FE)	Transforming raw data to better represent a prediction task and improve model accuracy
Alternative Models Exploration, Testing, and Validation (ATV)	Explore alternative models, testing and validating the performance against other models to determine the best performing one.
Result Summary and Recommendation (RSR)	Summarise the findings and recommend the most useful tasks to stakeholders.

Figure 14 visually represents this taxonomy, illustrating how various systems align with different levels of automation and their corresponding target users (Access to ML) and efficiency for data scientists. For instance, Level 0 represents manual coding in languages like Python, Java, and C++, requiring full manual control. Level 2 tools like Scikit-learn or Keras automate core ML and ATV steps but require significant manual effort in data preparation and feature engineering. As the automation level increases (e.g., Level 3, 4, and 5), more steps like Feature Engineering (FE) and Prediction Engineering (PE) become automated, potentially increasing efficiency for data scientists but not necessarily improving accessibility for non-experts if the interfaces remain complex or opaque.

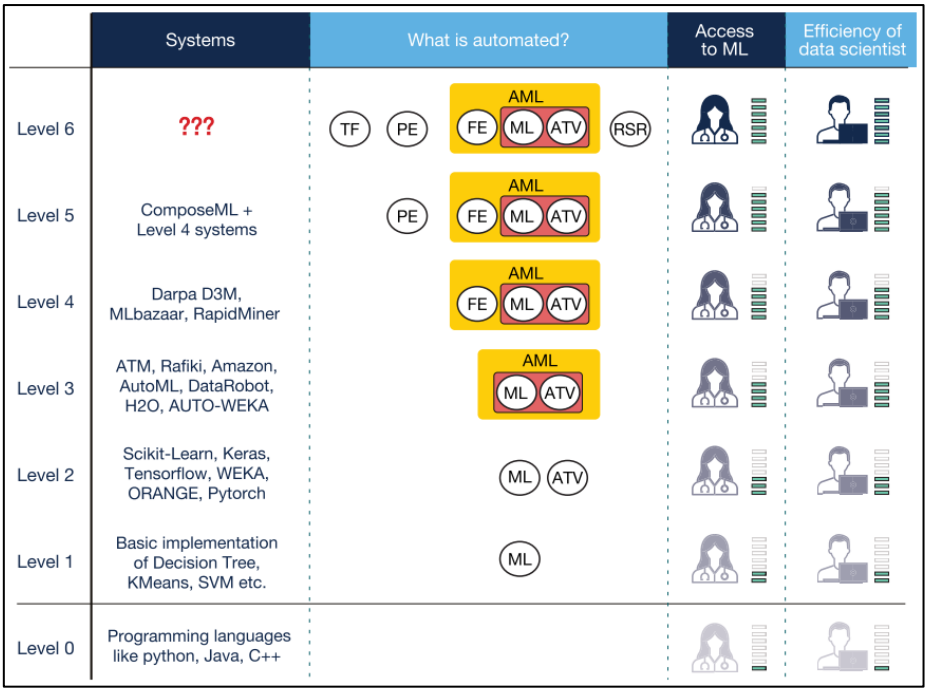


Figure 14 Levels of automation possible for end-to-end machine learning systems (Santu et al., 2022)

Examining specific categories reveals their limitations for non-experts, as detailed in Table 6. Domain Expert-Oriented Tools like Visus (Santos et al., 2019) and XAutoML (M.-A. Zöller et al., 2022) enable users with domain knowledge to build models and explore internal AutoML components. However, as Table 6 indicates, their interfaces can be complex for novices, lacking tailored support, and they often have limited focus on model interpretability and transparent decision processes for users without an ML background. Similarly, Clinician-Focused Tools such as VBrige (F. Cheng et al., 2022) are designed for users with specific domain expertise (e.g., medical data) and some ML interpretation skills, but they are not intuitive for complete beginners and may exclude those without prior exposure to model explanation concepts, as their design is tailored to a specific domain context. Research-Oriented Frameworks like DeepCAVE (Sass et al., 2022), while valuable for analyzing optimization processes, present a high learning curve and complexity that limits accessibility for general users, assuming foundational ML knowledge to interpret visual diagnostics, which is a significant usability limitation noted in Table 6.

General-Purpose AutoML Platforms, while seemingly targeting a broader audience, also present significant barriers for non-experts, as highlighted in Table 6. Tools such as Auto-sklearn (Feurer et al., 2015), H2O AutoML (LeDell et al., 2020), and TPOT (Olson et al., 2016) automate significant portions of the ML pipeline, including model selection and hyperparameter tuning. However, their usability for non-experts is often hindered by complex configuration options and the inherent black-box nature of the generated models. Non-experts may struggle to understand and interpret the results due to insufficient explanation of the algorithmic decisions made during the automation process (Coors et al., 2021). The steep learning curve and model opacity limit their accessibility for users without deep ML understanding, and while some offer visualizations, these are often insufficient to bridge the comprehension gap for true beginners, resulting in outputs that often include black-box models with limited post-hoc explanation features (Table 6).

In summary, while the existing landscape of AutoML tools offers increasing levels of automation across various stages of the ML pipeline, a critical gap remains in providing truly usable and transparent solutions for non-expert users. As evidenced by the limitations detailed in Table 6 across different categories, tools often cater to users with some degree of technical or domain expertise, and the focus on automation can sometimes come at the expense of interpretability and user understanding. Addressing the specific challenges faced by non-experts requires a dedicated approach that integrates user-centered design principles, effective visual analytics, and accessible explainable AI techniques throughout the AutoML workflow.

Table 4 Existing AutoML tools

Tool	Target User	Key Features	Usability Limitations	Transparency Limitations	References
Visus	Domain Experts	Visual analytics for AutoML processes	Interface may be complex for novices; lacks tailored support for non-expert workflows	Limited focus on model interpretability and transparent decision processes	Santos et al., 2019
XAutoML	Domain Experts	Explores internal AutoML components	Not designed for non-expert usability; assumes some ML background	Transparency features not sufficiently adapted to support novice understanding	M.-A. Zöller et al., 2022
VBrige	Clinicians	Bridges ML model explanations to medical data	Designed for users with some ML interpretation skills; not intuitive for complete beginners	May exclude those without prior exposure to model explanation concepts	F. Cheng et al., 2022
DeepCAVE	Researchers	Deep dives into optimization processes and visualizations	High learning curve; complexity of interface and outputs limits accessibility for general users	Assumes foundational ML knowledge to make sense of visual diagnostics	Sass et al., 2022
Auto-sklearn, H20-AutoML, TPOT	Non-experts	Automated ML pipeline generation	General usability for beginners but lacks guidance/explanation for decision-making during pipeline creation	Outputs often include black-box models with limited post-hoc explanation features	Coors et al., 2021

2.4.2 Existing XAI Tools and Visualizations Related Work

The emergence of complex, often opaque, Machine Learning (ML) models has necessitated the development of Explainable Artificial Intelligence (XAI). XAI aims to provide end-users with transparent AI solutions, enabling them to understand the reasoning behind decisions made by these black-box models (Adadi & Berrada, 2018; Kremers, 2020). The goal is to produce systems that offer trustworthy and comprehensible explanations, addressing the critical need for transparency, fairness,

and accountability in AI applications (Arrieta et al., 2020; Du et al., 2020). Explanations are viewed as supplementary metadata that illuminate the inner workings or specific outcomes of an AI model (Das & Rad, 2020).

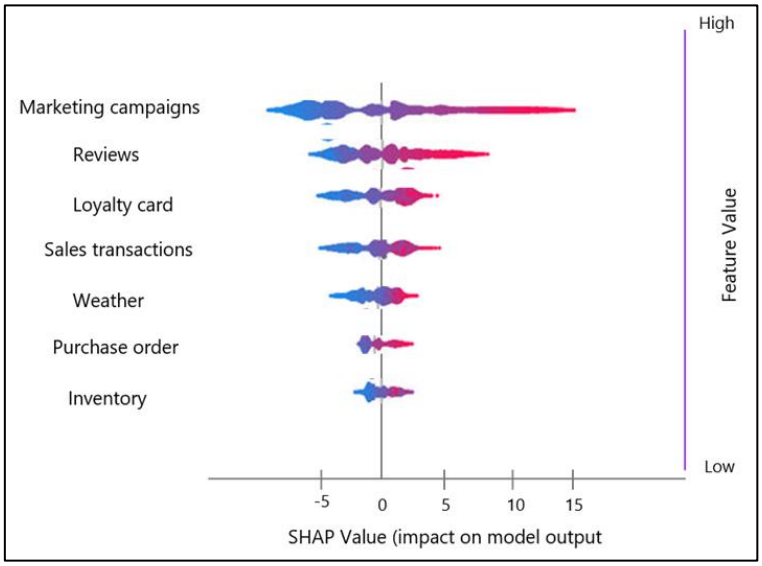


Figure 15 SHAP XAI feature importance chart

Visual analytics has become a commonly used medium for conveying these explanations, leveraging human perceptual and cognitive abilities to facilitate understanding of complex model components and workflows (Langer et al., 2021; Spinner et al., 2019). Interactive visualisations are particularly valuable, allowing users to dynamically explore data and model outputs, uncovering patterns and insights that might not be apparent in static representations (Yi et al., 2007). The intersection of visual analytics and AI, particularly XAI, is an increasingly important area of research, highlighting the role of visualisation in making complex AI models more interpretable and trustworthy (Chatzimparmpas et al., 2020a; Hohman et al., 2018).

Numerous interactive visualisation XAI tools have been developed to explain complicated model components and workflows. These tools often implement various XAI techniques to provide insights into model behaviour, both at a global level (explaining the entire model) and a local level (explaining individual predictions) (Alicioglu & Sun, 2022). Techniques can be intrinsic, where the model itself is interpretable (e.g., decision trees), or post-hoc, applied after training to explain black-box models (Alicioglu & Sun, 2022). Model-agnostic methods, which can be applied to any ML model, are particularly versatile and widely used (Adadi & Berrada, 2018).

Several popular open-source XAI frameworks and tools exist, each offering different capabilities for implementing and visualising explainability techniques. As illustrated in Figure 5, these include AIX360, H2O, Skater, InterpretML, Tf-explain, CaptumAI, DALEX, Alibi, XAITK, and the What-If Tool (Ding et al., 2022). These toolkits provide programming interfaces and shared architectures to facilitate various explainability approaches, often supporting both local and global explanations for classification or regression tasks. They aim to make it easier for developers and researchers to incorporate explainability into their ML workflows.

Many of these tools and frameworks operationalize specific, widely-used XAI algorithms. For instance, Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) and Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017) are prominent model-agnostic techniques that provide local explanations by approximating the black-box model's behaviour. Visualizations are commonly

used to present the outputs of these algorithms, such as feature importance plots showing which features contributed most to a prediction or plots illustrating the relationship between feature values and model output.

Despite the rapid development of innovative XAI methods and tools, a significant challenge remains regarding their accessibility and usability for non-expert users. The majority of available AI visualisation tools and XAI methods are primarily designed for AI developers and practitioners, assuming a certain level of technical literacy and familiarity with ML concepts (Hohman et al., 2019; Alicioglu & Sun, 2021). Literature has shown that few users, particularly non-experts, can accurately interpret the visualisations generated by these XAI methods (Kaur et al., 2020). This suggests that while the technical capabilities for generating explanations exist, the design of user interfaces and visualisations often falls short in effectively communicating these explanations to users without specialized ML expertise.

2.4.3 Human-Centered and Interactive ML Approaches

In contrast to purely automated or black-box Machine Learning (ML) systems, the fields of Human-Centered Computing and Interactive Machine Learning (IML) prioritize the active involvement and understanding of the human user throughout the ML process. These approaches recognize that effective ML solutions often require human expertise, judgment, and interaction, particularly when dealing with complex data, nuanced problems, or when the target users are not ML experts (Shneiderman et al., 2016; Dudley & Kristensson, 2018). The core idea is to design ML systems that are not just accurate or efficient but are also understandable, controllable, and supportive of human decision-making and learning.

Human-Centered ML emphasizes designing systems with the user's needs, capabilities, and context at the forefront. This involves understanding how users interact with data and models, their mental models of AI, and the challenges they face (Liao et al., 2020; Margetis et al., 2021). Principles from Human-Computer Interaction (HCI) are fundamental, focusing on usability, transparency, and effective communication between the human and the AI system (Shneiderman et al., 2016). The goal is to create interfaces and workflows that make ML accessible and empower users to participate meaningfully in the model development and application process.

Interactive Machine Learning (IML) specifically explores how human interaction can be integrated into the ML training and deployment loop. This can involve various forms of interaction, such as providing feedback on model predictions, steering the learning process, or collaboratively refining features (Dudley & Kristensson, 2018). IML systems often employ visual interfaces to make the interaction more intuitive, allowing users to manipulate data, models, and explanations directly (Hohman et al., 2018). This iterative human-in-the-loop approach aims to leverage human intelligence and domain knowledge to improve model performance, address biases, and ensure the model aligns with human values and goals.

Within the realm of making ML accessible to users with limited technical expertise, particularly non-experts, visual tools have played a significant role in human-centered and interactive approaches. These tools capitalize on human cognitive and perceptual capacities, often employing visual programming languages or drag-and-drop interfaces to simplify complex ML workflows (Kahn & Winters, 2018). Examples such as eCraft2learn, an extension to the visual programming language Snap! that provides an easy-to-use interface for manipulating AI cloud services, and Cognimates, a visual programming system utilizing neural networks, illustrate how abstraction and visual

representation can lower the barrier to entry for creating systems with learning capabilities (Kahn & Winters, 2018; Lacerda Queiroz et al., 2021). These tools abstract the complexity of underlying models by showing a visible, composable network of entities reflecting the flow of incoming data and the procedures and manipulations it goes through.

The theoretical underpinnings of these visual and interactive approaches for non-experts often draw from educational theories like Piaget's Constructivist theory and Papert's Constructionist Theory. Piaget's theory suggests that knowledge grows through an individual's interaction with the physical world (Piaget, 1950, as cited in thesis). Building on this, Papert's Constructionist Theory posits that knowledge creation is particularly effective when learners are actively engaged in constructing something tangible (Papert, 1993a, as cited in thesis). Applied to ML, this suggests that providing non-experts with tools to build and experiment with learning machines, and observe their behavior, can foster a deeper understanding of the ML process itself. The computer becomes a tool for invention and, by extension, learning, enabling the creation of projects more complex than might be possible otherwise (Papert, 1993a, as cited in thesis).

These human-centered and interactive approaches, particularly those leveraging visual interfaces and grounded in constructivist learning principles, represent a significant effort to democratize ML. By focusing on usability, transparency, and active human participation, they aim to make ML technologies accessible and understandable to a broader audience, moving beyond the confines of expert practitioners. This is crucial for enabling domain experts and the general public to effectively utilize ML for problem-solving and innovation in their respective fields.

2.5 Critical Analysis of Existing Solutions for Non-Experts

2.5.1 Limitations of Current AutoML Tools for Non-Experts

While Automated Machine Learning (AutoML) tools represent a significant step towards democratizing access to Machine Learning (ML) capabilities, a critical examination of existing platforms reveals considerable limitations that impede their effective use by non-expert users. These tools, often developed with expert users or automation efficiency as primary goals, frequently fail to address the unique needs, knowledge gaps, and interaction preferences of individuals without extensive ML training, creating significant barriers to adoption and successful application.

A pervasive limitation is the implicit assumption of prior technical knowledge embedded within the design of many current AutoML tools. Despite offering simplified interfaces compared to manual coding, platforms often require users to understand fundamental ML concepts, terminology, and evaluation metrics (Alicioglu & Sun, 2021). Tools designed for specific user groups, such as domain experts or clinicians, while tailored to particular data types or problems, still assume a baseline technical literacy that is not universal among all non-experts (Santos et al., 2019; M.-A. Zöller et al., 2022; F. Cheng et al., 2022). This reliance on specialized vocabulary and metrics makes navigating interfaces, configuring options, and interpreting outputs challenging, reinforcing the knowledge barrier rather than lowering it (Liao et al., 2020; Yang et al., 2018).

Furthermore, a significant critique revolves around the transparency, or lack thereof, in many AutoML systems. Driven by the need for speed and performance, these tools often function as "black boxes," automating complex processes without providing clear explanations of the underlying steps or the rationale behind model decisions (Kaur et al., 2020; Coors et al., 2021). This opacity undermines user

trust, as non-experts are left unable to understand why a particular prediction was made or how the model works. Without comprehensible explanations, users cannot critically evaluate the reliability of the results, identify potential biases, or gain actionable insights relevant to their domain, severely limiting the practical utility of the automated process (Ribeiro et al., 2016; Adadi & Berrada, 2018).

The inadequacy of integrated support mechanisms, such as learning scaffolding and contextual guidance, also poses a major limitation for non-experts. Few existing AutoML tools offer structured pathways that progressively build user understanding or provide timely, relevant assistance during complex tasks like feature selection, model configuration, or result interpretation (Yang et al., 2018; Gil et al., 2019). This lack of embedded educational support means non-experts are often left to navigate challenging decisions on their own, hindering their ability to learn effectively, explore alternative approaches, or correctly interpret ambiguous outcomes. The absence of sufficient guidance contributes to user frustration and reduces the likelihood of sustained engagement with the tool.

Finally, the quality and presentation of explanations and feedback within current AutoML tools are frequently insufficient for non-expert comprehension. Visualizations, while present in some platforms, may be designed for technical users, focusing on complex model diagnostics rather than building conceptual understanding for novices (Wang et al., 2021; Hohman et al., 2019). Similarly, system feedback, including error messages, is often technical and uninformative, failing to provide actionable guidance or opportunities for learning from mistakes (Holstein et al., 2019). These shortcomings in accessible explanations prevent non-experts from gaining meaningful insights from the automated process and effectively troubleshooting issues.

In summary, the limitations of current AutoML tools for non-experts stem from design paradigms that prioritize automation and expert-centric features over usability, transparency, and integrated learning support for novice users. The assumption of prior knowledge, the black-box nature of automated processes, the lack of scaffolding and guidance, and the inadequate quality of explanations collectively demonstrate that existing solutions, despite their technical capabilities, fail to adequately empower non-expert users in the ML development lifecycle. Addressing these limitations is crucial for realizing the full potential of AutoML in democratizing access to machine learning.

Table 5 Limitations of Current AutoML Tools and Impact on Non-Experts

Key Limitation for Non-Experts	Description	Supporting Literature	Impact on Non-Expert Users
Assumption of Prior Technical Knowledge	Interfaces and functionalities require understanding of ML concepts, terminology, and metrics.	Alicioglu & Sun, 2021; Santos et al., 2019; M.-A. Zöller et al., 2022; Liao et al., 2020	Steep learning curve; Difficulty navigating the interface; Misunderstanding configuration options; Reinforces knowledge barrier.
Lack of Transparency ("Black Box")	Automated processes and model decisions	Kaur et al., 2020; Coors et al., 2021; Ribeiro et al., 2016; Adadi & Berrada, 2018	Difficulty building trust in the system; Inability to critically evaluate results; Cannot identify biases; Limited actionable insights.

	are not clearly explained.		
Insufficient Integrated Support	Lack of structured learning pathways, contextual help, or guidance during complex tasks.	Yang et al., 2018; Gil et al., 2019; Liao et al., 2020	Hindered learning; Difficulty making informed decisions (e.g., model selection, feature engineering); Struggle with interpreting ambiguous results.
Inadequate Explanation Quality	Visualizations are designed for experts; Error messages are technical and uninformative.	Wang et al., 2021; Hohman et al., 2019; Holstein et al., 2019	Difficulty interpreting model outputs and visualizations; Inability to learn from mistakes; Frustration when encountering errors.

2.5.2 Limitations of Current XAI Tools and Approaches for Non-Experts

While the field of Explainable Artificial Intelligence (XAI) has made significant strides in developing methods and tools to provide transparency for complex Machine Learning (ML) models, a critical evaluation reveals notable limitations regarding their effectiveness and usability for non-expert users. Despite the availability of various XAI techniques and platforms (Figure 5), their design and presentation often fail to align with the cognitive abilities, technical background, and specific needs of individuals without specialized ML expertise.

A primary limitation is that many existing XAI tools and visualizations are predominantly designed for ML practitioners and researchers, rather than for domain users with limited AI experience (Hohman et al., 2019; Alicioglu & Sun, 2021). This expert-centric design results in interfaces and explanations that assume a level of technical understanding, utilizing complex terminology, abstract concepts, and visualizations that are not readily interpretable by non-experts. Consequently, even when explanations are generated, non-expert users may struggle to understand their meaning or significance.

Empirical studies have demonstrated that users, particularly those without an ML background, often face considerable difficulty accurately interpreting the visualizations generated by common XAI methods such as SHAP and LIME (Kaur et al., 2020). Visual representations that are clear and informative to an expert familiar with model internals may appear confusing or meaningless to a novice. This difficulty in interpretation undermines the core purpose of XAI, which is to foster understanding and trust, as unintelligible explanations are functionally equivalent to no explanations for this user group.

Furthermore, a significant limitation for non-domain experts is the lack of actionable insights provided by many XAI approaches (X. Wang & Yin, 2021). While XAI might explain what features influenced a prediction, or how a model behaves, it often does not provide clear guidance on what to do next. For a non-expert user who is also not an expert in the application domain, interpreting technical explanations and translating them into meaningful actions (e.g., improving data quality, adjusting features, or selecting a different model) can be a significant challenge. The explanations may lack the necessary context or practical recommendations relevant to their specific problem or workflow.

Additionally, existing XAI tools are often treated as separate components or are integrated into the ML workflow in ways that do not seamlessly support the non-expert user's journey from data to decision. XAI is most valuable when it is integrated interactively throughout the entire ML development process, allowing users to understand the impact of their choices and the model's behavior at each stage. However, many tools present XAI as a post-hoc analysis step, disconnected from the model building or refinement process, which is less effective for supporting non-expert learning and decision-making within an AutoML context.

In summary, while XAI has advanced the technical capability to explain ML models, the current landscape of XAI tools and approaches is limited in its effectiveness for non-expert users due to expert-centric design, difficulties in interpreting outputs, a lack of actionable insights for non-domain experts, and insufficient integration into the overall ML workflow. Addressing these limitations is crucial for XAI to truly contribute to the democratization of ML and empower non-expert users.

Table 6 Limitations of Current XAI Tools and Impact on Non-Experts

Key Limitation of Current XAI for Non-Experts	Description	Supporting Literature	Impact on Non-Expert Users
Expert-Centric Design	Tools and visualizations are primarily designed for ML practitioners, using complex terminology and abstract concepts.	Hohman et al., 2019; Alicioglu & Sun, 2021	Difficulty navigating interfaces; Explanations are not tailored to their knowledge level; Assumes technical background.
Difficulty Interpreting Outputs	Non-experts struggle to accurately understand the meaning of XAI visualizations and metrics.	Kaur et al., 2020; Yang et al., 2018	Explanations are unintelligible; Cannot verify model reasoning; Reduces trust; Limits critical evaluation.
Lack of Actionable Insights	Explanations do not clearly indicate what steps the user should take based on the insights provided, especially for non-domain experts.	X. Wang & Yin, 2021; Margetis et al., 2021	Cannot translate explanations into practical improvements; Limits ability to refine models or data effectively; Reduces practical utility of XAI.
Insufficient Workflow Integration	XAI is often a separate analysis step rather than seamlessly	Wang et al., 2019; Lee et al., 2019	Hinders understanding of how choices at different stages impact the model; Limits interactive exploration and refinement based on explanations.

	integrated throughout the ML development process.

2.6 Identified Research Gaps and Opportunities

2.6.1 Theoretical Gaps

There is a pressing need for theoretical development that can guide the design of systems that are technically capable and truly accessible and understandable to users without extensive ML expertise. Existing theoretical perspectives, such as traditional models of technology acceptance (e.g., TAM, Davis, 1989), while valuable for understanding general adoption factors like perceived usefulness and ease of use, require extension to specifically account for the nuances introduced by sophisticated automation and the need for understandable explanations in a domain as complex as ML. These models may not fully capture how trust is built or eroded in automated systems that lack transparency, or how the interpretability of intermediate steps and final outputs influences a non-expert's perception and willingness to engage. As noted by Hoffman et al. (2023), trust in AI is multifaceted and influenced by factors like perceived reliability and explainability, highlighting areas where existing theoretical models need further refinement in this specific context.

Theoretical guidance is needed on how to design automated systems that also foster human learning and understanding, especially for users who are new to the domain. Current theories of automation often focus on efficiency gains or the division of labor between human and machine (Lee et al., 2019; Santu et al., 2022), but they rarely provide detailed theoretical principles on how to integrate automation in a way that supports a novice user's cognitive development and mental model formation regarding the automated process itself.

Similarly, theoretical work in XAI has largely focused on the technical generation of explanations or the philosophical underpinnings of interpretability (Miller, 2019; Adadi & Berrada, 2018). There is less theoretical development on how explanations should be structured, presented, and integrated into interactive workflows to effectively support the learning and decision-making processes of non-experts (Liao et al., 2020; Chromik & Butz, 2021). A theoretical framework is needed to articulate how different types of explanations (e.g., local, global, counterfactual) interact with various presentation formats (e.g., visual, textual) to influence non-expert comprehension and trust within an integrated system.

Specifically, a significant theoretical challenge lies in understanding how different levels and types of automation within an AutoML pipeline interact with various forms of XAI to impact a non-expert's cognitive load, mental model formation, and ability to achieve actionable insights. The challenge is not merely automating a process or explaining a result in isolation, but theorizing how the combination of automated steps and integrated explanations can collectively empower a non-expert user to move from raw data to informed decisions and potentially iterative model improvement. Existing theoretical frameworks do not offer a robust basis for predicting how specific design choices in an integrated AutoML-XAI system will influence a non-expert's learning trajectory, trust development, or ability to translate model outputs into practical actions.

This theoretical deficit means that the design of usable and transparent AutoML tools for non-experts is often guided by empirical findings and heuristic principles rather than a strong, predictive theoretical foundation. A theoretical framework that integrates concepts from human-computer interaction, cognitive psychology, educational theory, and ML explainability is needed to provide a principled basis for designing systems that effectively balance automation with human understanding and control for this specific user group. Such a theory could guide the development of design principles and inform the evaluation of system effectiveness beyond traditional performance metrics, focusing on metrics related to user learning, trust, and actionable understanding.

Table 7 Theoretical Gaps

Theoretical Gap Area	Description	Consequence for Non-Expert Users
Integrated Automation & Explanation	Lack of theory on how automating ML steps interacts with integrated XAI to affect non-expert understanding and actionability.	Designs may optimize automation or explanation separately; Difficulty predicting combined impact on user learning and decision-making; Limited theoretical basis for integrated system design.
Trust in Automated Explanations	Insufficient theoretical understanding of how non-experts build and maintain trust in explanations provided by automated ML systems.	Design choices for fostering trust may be ad-hoc; Difficulty ensuring explanations are perceived as reliable and credible by non-experts.
Cognitive Load & Scaffolding	Limited theory on managing cognitive load and providing effective scaffolding for non-experts navigating automated, explained ML workflows.	Interfaces may overwhelm users with information; Scaffolding strategies may not align with non-expert learning processes; Difficulty designing intuitive step-by-step guidance within automation.
Actionable Understanding	Absence of theory on how XAI, integrated into AutoML, facilitates non-experts' ability to derive actionable insights from ML results.	Explanations may be technically correct but not practically useful for non-experts; Difficulty designing systems that guide users towards informed actions based on model outputs.

2.6.2 Methodological Gaps

Traditional evaluation methodologies, often derived from expert-centric contexts or focusing primarily on algorithmic performance, fall short in capturing the nuanced aspects of human-AI interaction and user understanding crucial for this audience. A primary methodological limitation is the predominant emphasis on traditional performance metrics such as accuracy, speed, and computational efficiency when evaluating AutoML tools (Feurer et al., 2015; Elshawi et al., 2019). While these metrics are essential for assessing algorithmic capability, they provide limited insight into how well a non-expert user can actually use the tool, understand the results, or trust the automated process. Evaluating a system solely on how quickly it finds a high-accuracy model does not measure whether a non-expert user comprehends why that model was chosen or what the model's predictions mean in a real-world context.

Furthermore, existing evaluation methodologies for XAI often focus on technical metrics related to explanation fidelity or completeness, or rely on evaluations conducted with expert users (Miller, 2019; Alicioglu & Sun, 2021). Methodologies for assessing how effectively XAI visualizations and explanations support non-expert understanding, trust development, and the ability to derive actionable insights are less developed and standardized. A key theoretical challenge is understanding how XAI facilitates non-experts' ability to move towards informed actions; current evaluation methods often lack the means to robustly measure this translation from explanation to action. Evaluating whether explanations are "technically correct" is different from evaluating whether they are "practically useful" for a non-expert seeking to make decisions based on model outputs.

There is a methodological gap in employing user-centered evaluation approaches that are specifically tailored to the context of non-expert interaction with automated and explained ML systems. While usability metrics like SUS and UEQ (Brooke, 1996; Schrepp et al., 2017a) are valuable for assessing perceived ease of use and user experience, they may not fully capture the unique challenges related to understanding complex, automated processes or interpreting potentially opaque explanations. Methodologies are needed that specifically probe user comprehension of automated steps, mental model formation regarding the ML pipeline, trust in algorithmic decisions, and the ability to effectively utilize explanations to inform actions or refine the process.

Moreover, the iterative nature of designing for non-experts, as suggested by user-centered design principles, requires evaluation methodologies that support iterative refinement (Norman, 2013). This necessitates evaluation methods that can provide actionable insights for design improvements, moving beyond summative assessments to formative evaluations conducted throughout the development lifecycle. Methodologies that combine quantitative measures of usability and understanding with qualitative insights from user interactions, interviews, and observational studies are crucial for gaining a holistic understanding of the non-expert experience.

In conclusion, the methodological landscape for evaluating AutoML and integrated AutoML-XAI systems for non-experts is marked by significant gaps. A reliance on expert-centric or performance-focused metrics, underdeveloped methods for assessing non-expert understanding and actionability of explanations, and a need for more robust user-centered and iterative evaluation approaches collectively highlight the need for new methodologies tailored to this specific context. Developing such methodologies is essential for effectively assessing whether these tools truly empower non-expert users and bridge the gap between complex ML technologies and broader accessibility.

Table 8 Methodical Gaps

Methodological Gap Area	Description	Limitation of Current Approaches
Assessing User Understanding	Lack of standardized methods to measure non-experts' comprehension of automated ML steps and explanations.	Reliance on subjective reports; Difficulty distinguishing perceived understanding from actual understanding; Metrics often focus on explanation fidelity rather than user comprehension.
Evaluating Actionability of XAI	Limited methodologies to assess whether non-experts can translate explanations into practical actions or informed decisions.	Focus on interpreting explanations in isolation; Does not measure the impact of explanations on subsequent user behavior or task performance within a workflow.

Measuring Trust in Automation	Need for robust methods to evaluate how non-experts build and maintain trust in automated and explained ML systems over time.	Trust metrics may be general; Do not specifically probe trust related to the opacity of automated steps or the interpretability of explanations; Longitudinal studies on trust development are scarce.
Integrated System Evaluation	Lack of holistic methodologies to evaluate the combined impact of automation, usability, and explainability on the non-expert experience.	Evaluation components (AutoML performance, XAI fidelity, general usability) are often assessed separately; Fails to capture the synergistic or conflicting effects of integrated elements.

2.6.3 Practical Implementation Gaps

Despite the technical feasibility of automating ML tasks and generating explanations, translating these capabilities into usable and understandable tools for a broad non-expert audience remains a considerable practical challenge.

A key practical implementation gap lies in the design of user interfaces that are truly intuitive and accessible for non-experts. Many existing tools, even those with graphical interfaces, are criticized for their complexity, overwhelming users with technical options and jargon (Yang et al., 2018; Coors et al., 2021). The practical implementation often fails to incorporate fundamental user-centered design principles that prioritize simplicity, clear navigation, and progressive disclosure of information for users without a technical background (Norman, 2013; Nielsen, 2012). This results in tools that, while powerful under the hood, are practically unusable for the intended non-expert audience.

Furthermore, there is a practical gap in implementing effective and usable XAI visualizations and explanations tailored for non-experts. Outputs are often technically accurate but presented in formats that non-experts struggle to interpret (Kaur et al., 2020; Wang et al., 2021). The practical implementation often falls short in translating complex algorithmic concepts into understandable visual or textual explanations that resonate with a non-expert's mental model of the problem domain (Liao et al., 2020; Chromik & Butz, 2021). This practical failure to provide comprehensible explanations limits the non-expert's ability to trust the model and derive actionable insights from its outputs.

Another significant practical implementation gap is the lack of integrated, context-sensitive learning support and guidance within the tools. Non-experts require practical assistance and scaffolding throughout the ML workflow, from data preparation to model evaluation (Yang et al., 2018; Gil et al., 2019). However, existing tools often lack practical implementations of features such as interactive tutorials, context-aware help texts, or automated guidance based on user actions. This leaves non-experts without the necessary practical support to overcome challenges encountered during the ML development process, such as understanding data quality issues or interpreting evaluation metrics (Holstein et al., 2019).

The practical integration of AutoML and XAI within a single, coherent workflow for non-experts also presents challenges. Implementing a system where automated steps are transparently explained, and where XAI is interactively available at relevant points in the process, is a complex practical task. Many implementations treat these as separate components or integrate them in a disjointed

manner, which does not support a non-expert's need for a streamlined and understandable end-to-end experience (Wang et al., 2019; Lee et al., 2019).

These practical implementation gaps collectively demonstrate that despite the theoretical potential and algorithmic advancements, the realization of usable and transparent AutoML tools for non-experts is hindered by shortcomings in user interface design, accessible explanation delivery, integrated learning support, and seamless workflow integration. Addressing these practical challenges is crucial for creating tools that can be effectively adopted and utilized by a broader audience.

Table 9 Practical Implementation Gaps

Practical Implementation Gap Area	Manifestation in Existing Tools	Consequence for Non-Expert Users
Intuitive UI Design	Complex interfaces, technical jargon, overwhelming options, poor navigation.	Difficulty learning and using the tool; Frustration and abandonment; Limited access to powerful features.
Accessible Explanation Delivery	XAI visualizations/text are too technical, not tailored to non-expert understanding; Lack of clear translation of complex concepts.	Inability to interpret model outputs; Lack of trust in results; Cannot gain actionable insights; Explanations are not practically useful.
Integrated Learning Support	Absence of interactive tutorials, context-aware help, or automated guidance within the workflow; Reliance on external documentation.	Struggle to overcome challenges; Limited ability to learn from the tool; Difficulty completing the ML process independently.
Seamless Workflow Integration	Disjointed AutoML and XAI components; Lack of interactive explanations during automated steps; Unclear flow from data to results and insights.	Confusing user journey; Difficulty understanding the overall process; Limited ability to interactively refine models based on explanations.

2.7 Design Considerations and Guiding Principles

2.7.1 Rationale for Design Principles

The proposed design principles for the Automated Machine Learning (AutoML) tool are not arbitrary but are derived directly from the critical analysis of the existing literature, the identified challenges faced by non-expert users in ML development, and the theoretical, methodological, and practical implementation gaps in current AutoML and integrated AutoML-XAI tools. The rationale underpinning these principles is to create a system that effectively addresses these multifaceted issues, thereby making ML more accessible, usable, and transparent for individuals without extensive technical expertise. This requires a deliberate shift towards a human-centered design paradigm that prioritizes the user's experience and understanding throughout the entire ML workflow.

Firstly, the design principles are fundamentally shaped by the need to mitigate the significant challenges non-expert users face throughout the ML development lifecycle, as detailed in Section 2.3.3 and summarized in Tables 3 and 5. Challenges such as limited understanding of ML algorithms, reliance on external documentation, misinterpretation of model performance metrics, limited

adaptation of learning algorithms, and an overemphasis on data quantity necessitate a design approach that provides integrated support and clarifies complex concepts. Principles that promote clear, user-friendly explanations and guided workflows directly address the challenge of limited understanding (Table 5, Challenge 1), fostering an intuitive grasp of ML algorithms. Similarly, principles advocating for interactive tutorials and integrated learning support aim to reduce reliance on external resources (Table 5, Challenge 2) and encourage exploration of diverse approaches (Table 5, Challenge 4). By providing comprehensive and understandable performance metrics with user-friendly explanations, the design principles aim to counter the misinterpretation of evaluation results (Table 5, Challenge 3) and facilitate a better grasp of model performance and improvement (Table 3, C3, C5). Furthermore, integrating educational modules and guidance on algorithm choice and feature design within the tool helps address the misconception about solely relying on data quantity (Table 5, Challenge 5).

Secondly, the design principles are formulated to overcome the specific limitations of current AutoML tools for non-experts, as analyzed in Table 6. Existing tools often suffer from a lack of transparency, assume prior technical knowledge, provide insufficient learning scaffolding and decision support, and offer inadequate explanations (Kaur et al., 2020; Liao et al., 2020; Yang et al., 2018; Gil et al., 2019). Design principles that prioritize transparency aim to move beyond the "black box" nature of many AutoML systems, ensuring users understand the automated process and the resulting models (Coors et al., 2021). Principles focusing on embedded scaffolding and contextual support directly counter the assumption of prior knowledge and insufficient decision support, providing the necessary guidance within the user interface (Akotuko et al., 2021; Puntambekar, 2022). By emphasizing usable and accessible explanation delivery, the principles address the practical implementation gap in providing understandable XAI outputs, ensuring that visualizations and textual explanations are tailored to non-expert comprehension (Kaur et al., 2020; Wang et al., 2021).

Thirdly, the rationale is deeply rooted in addressing the limitations of current XAI tools and approaches when applied to non-experts. XAI outputs are often expert-centric, difficult to interpret, may lack actionable insights, and are not always well-integrated into the ML workflow (Hohman et al., 2019; Kaur et al., 2020; X. Wang & Yin, 2021). Design principles that advocate for accessible explanation delivery and multiple ways to communicate explanations (Chromik & Butz, 2021) aim to make XAI outputs understandable and tailored to non-expert cognitive models. Principles focusing on integrated learning support and seamless workflow integration seek to ensure that XAI is not a disconnected post-hoc step but is interactively available at relevant points in the AutoML process, providing contextualized and potentially actionable insights (Wang et al., 2019; Lee et al., 2019).

Furthermore, the design principles are intended to bridge the critical gaps in the integration of AutoML and XAI for non-experts. The practical implementation gaps, such as disjointed components and unclear flow, lead to a "Confusing user journey; Difficulty understanding the overall process; Limited ability to interactively refine models based on explanations." This highlights the crucial need for a unified design approach. Principles that emphasize visualizing activity sequences and demonstrating scaffold functions (Quintana et al., 2018) aim to create a clear and understandable workflow, guiding the user through the process. Principles promoting engaging user feedback and seamless workflow integration (Mezhoudi, 2013; Yigitbas et al., 2019) are crucial for ensuring that the automated steps are transparently explained and that XAI is interactively available, allowing users to understand the impact of their choices and refine the model based on explanations, thereby alleviating the confusion and difficulty in understanding the overall process and enabling interactive refinement.

Finally, the design principles offer a practical and theoretically informed approach to address the theoretical, methodological, and practical implementation gaps identified in the literature. By focusing

on principles that promote intuitive UI design, accessible explanation delivery, integrated learning support, and seamless workflow integration, the design aims to provide a practical realization of a usable and transparent AutoML tool. These principles are informed by the need for theoretical frameworks that better integrate automation and explanation, guide the development of evaluation methodologies beyond traditional metrics, and address the practical shortcomings in current tool implementations. The rationale is to move beyond heuristic design by grounding the principles in the identified needs and limitations, thereby offering a principled approach to creating effective human-centered AI systems for non-experts (Shneiderman, 2020).

In essence, the rationale for the proposed design principles is to empower non-expert users by transforming complex, opaque, and often disjointed ML processes into an accessible, understandable, and integrated experience. By systematically addressing the challenges faced by this user group and the limitations and gaps in existing tools and theoretical understanding, the design principles aim to democratize access to ML, fostering user understanding, trust, and the ability to leverage AI for their specific needs. The principles serve as a blueprint for designing an AutoML tool that is functional and supportive, transparent, and ultimately empowering for the non-expert user.

2.7.2 Proposed Design Principles

Based on the comprehensive analysis of the literature, the identified challenges faced by non-expert users, and the theoretical, methodological, and practical gaps in existing tools, a set of design principles is proposed to guide the development of a usable and transparent AutoML tool. These principles, summarized in Table 7, are intended to address the specific needs of non-experts and foster an accessible and understandable Machine Learning (ML) development experience. Drawing from established work in technology-enhanced scaffolding and Explainable Artificial Intelligence (XAI) user interfaces, these principles aim to transform the complex and often opaque nature of ML and AutoML into a more intuitive and empowering process.

The design principles are categorized into those focusing on the Overall system and those specific to the XAI component.

Overall System Design Principles

The principles for the overall system focus on providing a supportive and understandable environment for non-expert users throughout the ML workflow.

- DP1. Visualise activity and sequences: This principle emphasizes outlining the procedural and metacognitive processes required to facilitate an ML process. It involves specifying the tasks involved and clarifying the ML development pipeline (Quintana, Zhang, et al., 2018; Wu et al., 2021). The rationale is to provide non-experts with a clear mental model of the steps involved in building an ML model, addressing the challenge of limited understanding of the overall process and the practical gap of unclear workflow integration. Visualizing the sequence of activities helps users navigate the system and understand where they are in the process.

- DP2. Demonstrate scaffold function: This principle focuses on presenting the utility and steps involved throughout the ML process through demonstrations (Puntambekar, 2022; Quintana, Zhang, et al., 2018; Saye & Brush, 2002). This can include example-based development or online demonstrations that show users how to perform specific tasks or utilize features. This directly addresses the challenge of reliance on documentation and public scripts, and the limitation of insufficient learning scaffolding in existing tools, providing practical guidance within the tool itself.

- DP3. Embedded contextually relevant scaffold: This principle advocates for integrating resources based on a conceptual framework directly into the system to facilitate further learner inquiry (Akotuko et al., 2021; Puntambekar, 2022; Saye & Brush, 2002). Examples include hyperlinks specific to an ML task or context-aware help texts. The rationale is to provide just-in-time support that is relevant to the user's current activity, countering the assumption of prior knowledge and the practical gap of limited integrated learning support. This principle supports users in understanding complex concepts as they encounter them.

- DP4. Visible and utilised scaffold: This principle stresses the importance of ensuring that scaffolds are visible and explicitly clarified to learners to promote appropriate usage (Quintana, Reiser, et al., 2018; Sarah, 2022). This involves specifying and explaining the functionality of scaffolds within the system to ensure an effective understanding of how the tool is supporting the user. By making the support mechanisms clear and understandable, this principle aims to improve the usability of the scaffolding itself and ensure that non-experts can effectively leverage the provided assistance.

XAI Component Design Principles

These principles are specifically aimed at making the explanations provided by the XAI component understandable and usable for non-expert users.

- DP5. Progressive explanation disclosure: This principle suggests providing finer granularity of an explanation through subsequent steps following an explanation interaction (Buçinca et al., 2021; Khosravi et al., 2022; Millecamp et al., 2019). Examples include visualizing a specific feature after clicking on it or using tooltips to display the factors of a feature. This addresses the limitation that XAI outputs can be overwhelming by allowing users to explore explanations incrementally, managing cognitive load and tailoring the level of detail to their needs.

- DP6. Natural language rationale: This principle emphasizes complementing visual explanations with textual explanations in natural language to facilitate better understanding (Ehsan et al., 2019, 2021; Wiegreffe & Marasovic, 2021). For instance, providing a natural language explanation of a feature importance chart can help non-experts interpret the visual information. This directly addresses the difficulty in interpreting XAI outputs and the practical gap of accessible explanation delivery by presenting information in a more familiar and understandable format.

- DP7. Multiple ways to communicate an explanation: This principle advocates for providing diverse and related explanations to triangulate insights and understand different angles of explanation (Chou et al., 2022; Páez, 2019; Vilone & Longo, 2021). This could involve visualizing global feature importance alongside local feature importance in relation to each other. Offering multiple perspectives on the model's behavior helps non-experts build a more comprehensive understanding and addresses the limitation that XAI outputs may lack actionable insights by providing richer context. This principle also addresses the practical gap of accessible explanation delivery by catering to different learning preferences and cognitive approaches.

These proposed design principles collectively form a framework for developing an AutoML tool that is functionally capable and designed with the non-expert user firmly in mind, prioritizing usability, transparency, and integrated support throughout the ML development and explanation process.

Table 10 Design Principles for proposed tool

Category	Design principle	Note	Reference
Overall system	**DP1**. Visualise activity and sequences	Outline the procedural and metacognitive processes required to facilitate an ML process. (e.g., Specify the tasks involved, and clarify the ML development pipeline)	(Quintana, Zhang, et al., 2018; Wu et al., 2021)
	DP2. Demonstrate scaffold function	Present the utility and steps involved throughout the ML process. (e.g., Example-based development, online demonstration)	(Puntambekar, 2022; Quintana, Zhang, et al., 2018; Saye & Brush, 2002)
	DP3. Embedded contextually relevant scaffold	Integrate resources based on a conceptual framework into the system to facilitate further learner inquiry. (e.g., Hyperlinks specific to an ML task)	(Akotuko et al., 2021; Puntambekar, 2022; Saye & Brush, 2002)
	DP4. Visible and utilised scaffold	Ensure scaffolds are visible and explicitly clarified to learners to promote appropriate usage. (e.g., Specify and explain the functionality of scaffolds within the system to ensure an effective understanding of the system)	(Quintana, Reiser, et al., 2018; Sarah, 2022)
XAI component	**DP5**. Progressive explanation disclosure	Provide finer granularity of an explanation through subsequent steps following an explanation interaction. (e.g., Visualising a specific feature after clicking on it, tooltips to display the factors of a feature)	(Buçinca et al., 2021; Khosravi et al., 2022; Millecamp et al., 2019)
	DP6. Natural language rationale	Complement visual explanations with textual explanations to facilitate better understanding (e.g., Natural language explanation of a feature importance chart)	(Ehsan et al., 2019, 2021; Wiegreffe & Marasovic, 2021)
	DP7. Multiple ways to communicate an explanation	Provides related explanations to triangulate insights and understand different angles of explanation. (e.g., Visualising global feature importance	(Chou et al., 2022; Páez, 2019; Vilone & Longo, 2021)

		and local feature importance in relation to each other)	

2.8 Summary

This chapter has provided a comprehensive review of the relevant literature and the state of the art pertaining to Automated Machine Learning (AutoML), Explainable Artificial Intelligence (XAI), and Visual Analytics, with a specific focus on their intersection and applicability for non-expert users in Machine Learning (ML) development. The foundational concepts of AI, ML, AutoML, XAI, and Visual Analytics were introduced, establishing the technological context for the research.

A critical aspect explored was the understanding of the target user: non-experts in ML. These users are characterized by limited prior knowledge of ML algorithms and processes, often viewing models as input-output mechanisms and relying heavily on external documentation and scripts (Yang et al., 2018). Key challenges identified include difficulties in data analysis, problem formulation, feature design, model selection, evaluation, and performance improvement. These challenges underscore the significant barriers non-experts face when attempting to engage with traditional or even some existing automated ML workflows.

The state of the art in existing tools and approaches was reviewed, covering AutoML tools and frameworks, XAI tools and visualizations, and human-centered and interactive ML approaches. While tools exist across various levels of automation and XAI methods offer means to generate explanations, a critical analysis revealed significant limitations when applied to non-expert users.

Specifically, current AutoML tools often suffer from a lack of transparency, assume prior technical knowledge, and provide insufficient learning scaffolding and decision support, rendering them largely inaccessible or difficult to use for non-experts (Kaur et al., 2020; Liao et al., 2020; Yang et al., 2018; Gil et al., 2019). Similarly, existing XAI tools, while technically capable, are often expert-centric, produce outputs that are difficult for non-experts to interpret, may lack actionable insights, and are not always seamlessly integrated into the ML workflow (Hohman et al., 2019; Kaur et al., 2020; X. Wang & Yin, 2021).

The analysis further highlighted critical gaps in the integration of AutoML and XAI for non-experts. The current landscape often presents a disjointed experience where automated processes lack transparency, and explanations, if provided, are not well-integrated or usable for non-experts seeking to understand the automated steps or derive actionable insights. This creates a confusing user journey and limits the potential for non-experts to learn from and effectively utilize these technologies.

These limitations and integration gaps collectively point to significant research gaps. Theoretical gaps exist in frameworks that adequately integrate automation, explainability, and usability for non-experts (Kandel et al., 2012; Hoffman et al., 2023). Methodological gaps are apparent in the evaluation approaches used to assess non-expert understanding, trust, and actionability within these systems, often relying on expert-centric or performance-focused metrics (Abdul et al., 2018). Practical implementation gaps stem from the shortcomings in designing intuitive user interfaces, delivering accessible explanations, providing integrated learning support, and ensuring seamless workflow integration for non-experts (Norman, 2013; Wang et al., 2021; Holstein et al., 2019).

Addressing these identified gaps and limitations necessitates a deliberate, user-centered approach to designing AutoML tools for non-experts. The proposed design considerations and guiding principles are a direct response to this need, aiming to provide a framework for developing a system that

prioritizes usability, transparency, and integrated support throughout the ML development and explanation process. By focusing on visualizing workflows, demonstrating scaffolding, embedding contextual help, ensuring visible support, progressively disclosing explanations, providing natural language rationales, and communicating explanations in multiple ways, the goal is to create an AutoML tool that empowers non-experts to effectively leverage ML, fostering understanding and trust in the process and results. This forms the basis for the subsequent development and evaluation of the proposed system.

3 Methodology

3.1 Introduction

This chapter delineates the comprehensive methodology employed to address the research questions and achieve the objectives outlined in Chapter 1. The successful development and rigorous evaluation of an Automated Machine Learning (AutoML) tool, specifically engineered to augment usability and transparency for non-expert users, necessitated a structured and progressive research approach. This methodology was chosen to ensure that the resultant system is not merely technically proficient but also genuinely accommodates the needs and inherent capabilities of its intended user base, a critical consideration given the well-established challenges confronting non-experts in the domain of Machine Learning (ML) development.

The foundational methodological framework underpinning this research is a user-centred design (UCD) paradigm. UCD constitutes an iterative design process wherein the needs and requirements of the end-users occupy a central position throughout the entire design and development lifecycle (Norman, 2013). This stands in contradistinction to alternative approaches that might prioritize technical feasibility or commercial imperatives over user requirements. Considering the specific focus on crafting a tool for non-experts, whose difficulties with ML concepts and processes are extensively documented within the literature (Yang et al., 2018; Ramos et al., 2020), a UCD approach was deemed indispensable to guarantee the tool's accessibility, practical utility, and efficacy in cultivating user comprehension.

The selection of UCD as the guiding methodology is directly justified by the nature of the research problem. Addressing the challenges faced by non-experts, such as their limited understanding of complex ML workflows and reliance on external support (Yang et al., 2018), requires a design process that actively involves the target users. UCD facilitates this involvement, ensuring that the design decisions are informed by empirical data on user behaviour, preferences, and points of difficulty. This is particularly relevant given that non-experts may perceive ML algorithms as 'black boxes' and struggle with tasks like problem formulation and model evaluation (Ramos et al., 2020).

Furthermore, the iterative character of the UCD methodology is particularly well-suited to navigating the complexities associated with designing interactive systems for a diverse user population exhibiting varying levels of technical fluency. This iterative nature facilitates the systematic collection of user requirements at the outset of the design process. By gathering detailed insights into what users need and expect from an AutoML tool, the research can establish a solid foundation for subsequent development efforts.

Following the initial collection of requirements, the iterative nature of UCD supports the subsequent development of prototypes that are directly informed by these requirements. These prototypes serve as tangible representations of the design concepts, allowing for early testing and validation with the target user group. This approach ensures that the design evolves based on concrete feedback, rather than proceeding based on assumptions.

A further crucial aspect facilitated by the iterative UCD methodology is the stringent evaluation of these prototypes, involving representatives drawn from the target users. These evaluations provide essential data on the usability and transparency of the developing system. The insights garnered from each successive evaluation phase serve a vital function: they directly inform and guide subsequent design iterations. This feedback loop ensures that the system undergoes continuous refinement, leading to a product that progressively demonstrates a stronger alignment with user needs and

expressed preferences. This cyclical process of designing, prototyping, and evaluating is fundamental to effectively addressing concerns related to usability and transparency in a user-focused manner.

As visually depicted in Figure 15, the research design is organized around a five-stage UCD framework. The initial stage, designated as Design Research, entailed a thorough and systematic exploration of the target users and their specific requirements. This foundational phase was critical and encompassed a range of activities aimed at comprehending the existing landscape of AutoML and Explainable Artificial Intelligence (XAI) tools, pinpointing the specific challenges encountered by non-experts, and gathering preliminary requirements that would subsequently inform the design trajectory. This initial stage was paramount for anchoring the research firmly in the empirical reality of user needs and for avoiding potentially unfounded assumptions regarding user capabilities or preferences.

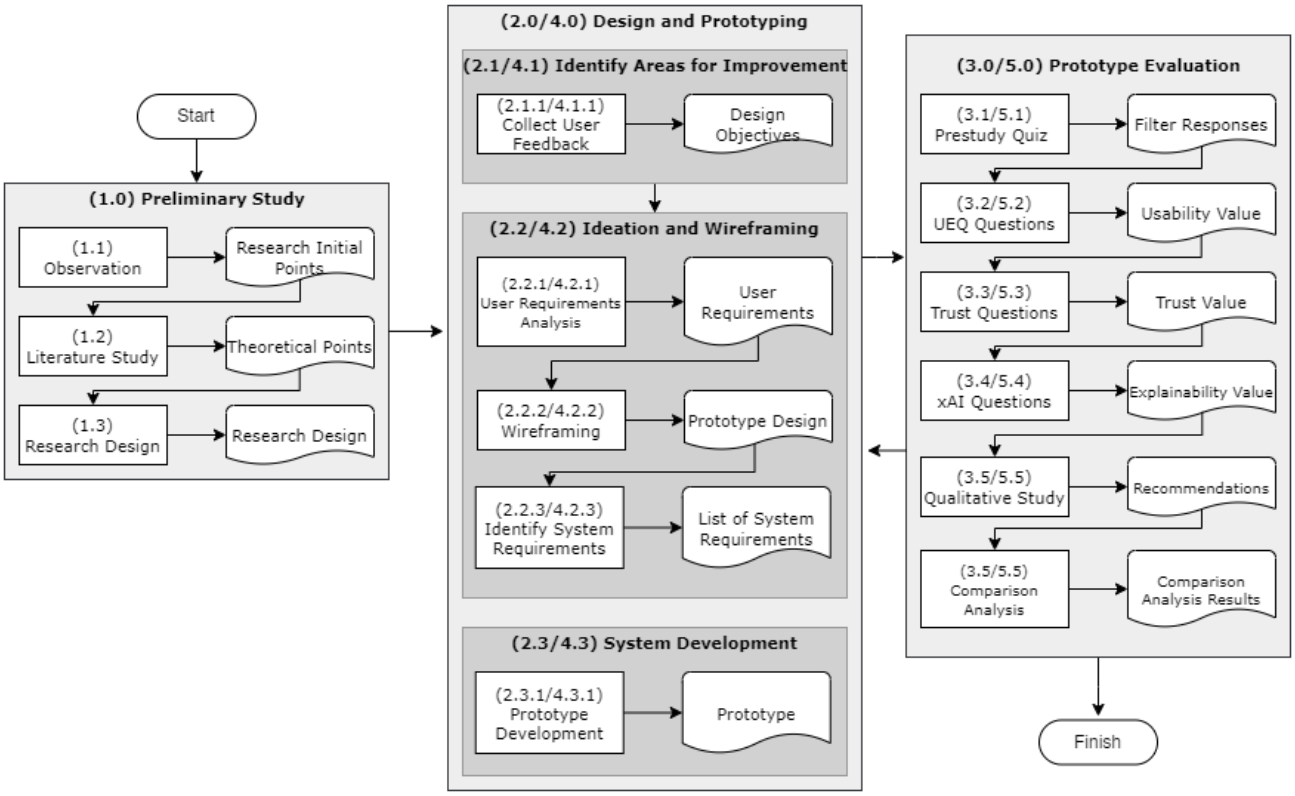


Figure 16 Research design stages in detail

Following the initial phase of design research, the methodology advanced into the Design and Prototyping stages, represented as Stages 2 and 4 within Figure 15. These stages were dedicated to the crucial task of translating the accumulated requirements and insights into tangible design concepts and interactive prototypes. This involved a process of conceptualization, design, and subsequent development of wireframes and prototypes, transitioning progressively from low-fidelity representations to more refined and interactive models. A structured approach to system development, specifically the waterfall model, was adopted within these stages to ensure that the construction of the tool proceeded in an organized manner, guided by the predefined requirements and established design principles. The evaluation stages were conducted in parallel with and following these development efforts.

3.2 Overall Research Design

The research undertaken to develop and evaluate VisAutoML, an AutoML tool for non-experts, was guided by a structured and iterative methodological framework. As articulated in the introduction, the complexities inherent in designing usable and transparent Machine Learning (ML) tools for users with limited technical expertise necessitate an approach centered on the user throughout the development lifecycle. Consequently, the overall research design adopted a user-centred design (UCD) paradigm, which is particularly well-suited for addressing the multifaceted challenges of human-computer interaction in complex domains (Norman, 2013). This approach ensures that user needs, capabilities, and feedback are integral to shaping the final system, moving beyond purely technical considerations.

The rationale for employing a UCD approach is deeply rooted in the research objectives and the characteristics of the target user group. The primary objectives include investigating user requirements, designing and developing a usable system, evaluating the platform, and identifying design guidelines. Achieving these objectives effectively, especially for non-expert users who face challenges such as understanding complex workflows and interpreting results, requires direct engagement with these users throughout the research process. UCD provides the necessary framework for this engagement, ensuring that the design decisions are empirically grounded and responsive to actual user needs, rather than based on assumptions.

As visually represented in Figure 15, the research design is structured as a five-stage iterative process within the UCD framework. This cyclical model emphasizes that the research did not follow a linear path but involved repeated cycles of design, development, and evaluation. The stages are interconnected, with outputs from one stage directly informing the activities of subsequent stages, particularly through feedback loops from evaluation phases back to design and prototyping. This iterative structure is crucial for progressively refining the system based on user insights and addressing usability and transparency issues as they emerge.

The initial stage of the research design, Design Research (Stage 1 in Figure 15), served as the foundational phase for understanding the problem space and the target users. This stage encompassed key activities aimed at investigating and gathering requirements for developing an intuitive AutoML prediction platform for non-expert users. As detailed in Table 1 and Table 2, these activities included conducting a literature review on AutoML tools, usability, and transparency, developing and analyzing an extended Technology Acceptance Model (TAM), performing user surveys and interviews, creating user personas, and analyzing existing AutoML platforms. The insights gleaned from this stage, such as the identification of major challenges non-experts face, directly informed the subsequent design and development efforts.

Following the requirements gathering, the research proceeded to the Design and Prototyping stages (Stages 2 and 4 in Figure 15). These stages were dedicated to translating the user requirements and design principles into a tangible system. The primary objective addressed here was to design and develop a usable AutoML system for non-experts. As outlined in Table 1 and Table 2, activities in these stages involved reviewing best practices in UCD, developing wireframes and prototypes, identifying system requirements, and implementing the system functionality. The adoption of a structured approach like the waterfall model within these development stages (as indicated in the thesis draft's description of system development) provided a systematic way to build the prototype based on the defined specifications derived from the preceding research stage.

Integrated within the iterative UCD framework were the Prototype Evaluation stages (Stages 3 and 5 in Figure 15). These stages were critical for assessing the effectiveness of the developed prototypes in meeting user needs and addressing usability and transparency concerns. The objective here was to evaluate the platform through user studies to determine user experience, usability, and transparency. Table 1 and Table 2 describe activities such as recruiting participants, collecting qualitative and quantitative data through questionnaires and interviews, and analyzing the collected data. The findings from these evaluations, such as usability scores, trust levels, and qualitative feedback, provided essential empirical evidence that fed back into the Design and Prototyping stages (Stage 4), enabling iterative refinement of the system based on user experiences. This iterative evaluation process is key to ensuring the system progressively improves in meeting user requirements.

Table 11 Research activities to be conducted to accomplish each research objective

Research Objectives	Research Activities
To investigate and gather the requirements for developing an intuitive AutoML prediction platform for non-expert users.	1. Conduct a literature review on AutoML tools and their usability and transparency. 2. Developing an extended technology acceptance (TAM) model for AutoML tool use by non-expert users. 3. Applying and analysing proposed TAM model to understand factors affecting user acceptance and behavioural intention in AutoML tool use by non-expert users. 4. Perform user surveys and interviews to understand the needs and expectations of non-expert users in machine learning. 5. Create personas representing the target audience of non-expert users. 6. Identify and analyse existing AutoML platforms. 7. Analyse the gathered data to identify common user requirements and preferences. 8. Develop a user requirements document based on the findings.
To design and develop a usable AutoML system for non-experts to conduct regression and classification tasks on tabular data.	9. Review best practices in user-centred design and usability principles. 10. Develop wireframes and low-fidelity prototypes of the AutoML system based on the identified user requirements. 11. Create a high-fidelity prototype of the AutoML system incorporating user feedback. 12. Develop the backend functionality of the AutoML system to enable regression and classification tasks on tabular data. 13. Integrate machine learning algorithms and automation features into the system.

	14. Design and develop XAI visualisations, such as feature importance plots, SHAP (SHapley Additive exPlanations) values, to aid users in understanding model predictions. 15. Integrate the XAI visualisations seamlessly into the AutoML system's user interface. 16. Document the design and development process of XAI visualisations, including their role in enhancing transparency and interpretability of the machine learning models.
To evaluate the platform through ethnography studies to determine the user experience, usability, and transparency.	17. Recruit a diverse group of non-expert users to participate in ethnography studies. 18. Collect qualitative data through interviews and surveys to capture users' perceptions, experiences, and the knowledge they gain from using the platform. 19. Analyse the collected data to identify patterns, pain points, and areas of improvement in the user experience and usability. 20. Iteratively refine the platform based on the insights obtained from ethnography studies. 21. Summarise the findings in a comprehensive ethnographic report, including recommendations for further enhancements.
To identify design guidelines for an AutoML tool that supports ML model development and explanation to non-expert users.	22. Review existing literature on design principles and guidelines for user-friendly AutoML tools. 23. Analyse the results from mixed-method studies to extract insights into what makes the platform user-friendly and effective for non-expert users. 24. Synthesise the collective knowledge into a set of design guidelines specific to AutoML tools for non-expert users. 25. Validate the design guidelines through usability testing with non-expert users to ensure their practicality and effectiveness.

Table 1 provides a comprehensive overview of the research activities undertaken in the designated timeframes. Each activity was designed to contribute to the overarching objectives of the study. The following table provides a detailed description of the rationale and execution of each activity.

Table 12 Detailed description of purpose and execution for each research activity

Activity	Purpose	Execution
1. Literature Review on AutoML Tools	To understand AutoML tools' usability and transparency.	Conducted a thorough review of relevant literature, considering seminal works and recent advancements in the field.
2. Developing Extended TAM Model	To tailor the Technology Acceptance Model (TAM) for non-expert users.	Extended the TAM model through a rigorous process involving theoretical considerations, and iterative refinement.
3. Analysing Proposed TAM Model	To understand factors influencing user acceptance and behavioural intention in AutoML tool use.	Employed statistical analyses and qualitative assessments to scrutinize the proposed TAM model, drawing insights from user perspectives.
4. User Surveys and Interviews	To understand the needs and expectations of non-expert users in machine learning.	Conducted surveys and interviews, utilising standardized questionnaires and open-ended inquiries to capture a comprehensive spectrum of user insights.
5. Create Personas	To represent the target audience of non-expert users.	Crafted personas based on the amalgamation of user survey data, ensuring a nuanced representation of user demographics and preferences.
6. Identify and Analyse AutoML Platforms	To evaluate existing AutoML platforms.	Systematically reviewed available platforms, considering features, functionalities, and user feedback.
7. Analyse Gathered Data	To identify common user requirements and preferences.	Analysed data from surveys, interviews, and platform analyses, resulting in a comprehensive user requirements document.
8. Develop User Requirements Document	To document findings and insights.	Summarised research outcomes into a comprehensive user requirements document.
9. Review Best Practices in Design and Usability	To inform the design and development process.	Conducted a review of best practices in user-centred design and usability principles.

10. Develop Wireframes and Low-Fidelity Prototypes	To visualise the AutoML system based on user requirements.	Created preliminary visual representations to outline the system's structure and functionality.
11. Create High-Fidelity Prototype	To incorporate user feedback for system refinement.	Developed an advanced prototype, integrating user feedback to optimise the user experience.
12. Develop Backend Functionality	To enable regression and classification tasks on tabular data.	Implemented backend functionality to enhance the AutoML system's capabilities.
13. Integrate ML Algorithms and Automation Features	To enhance system capabilities.	Integrated machine learning algorithms and automation features into the AutoML system.
14. Design and Develop XAI Visualisations	To aid users in understanding model predictions.	Crafted XAI visualisations such as feature importance plots and SHAP values, emphasising transparency and interpretability.
15. Integrate XAI Visualisations	To enhance the AutoML system's user interface.	Seamlessly integrated XAI visualisations into the user interface, ensuring a cohesive user experience.
16. Document Design and Development Process	To provide insights into transparency and interpretability enhancements.	Documented the design and development process of XAI visualisations, emphasising their role in transparency and interpretability.
17. Recruit Non-Expert Users for Mixed-method Studies	To capture user perceptions and experiences.	Recruited a diverse group of non-expert users for participation in mixed-method studies.
18. Collect Qualitative Data through Interviews and Surveys	To understand users' knowledge gain and experiences.	Conducted interviews and surveys to gather qualitative data on users' perceptions and experiences with the platform.
19. Analyse Collected Data	To identify patterns, pain points, and areas of improvement.	Analysed ethnographic data to identify user patterns and areas for platform improvement.
20. Iteratively Refine the Platform	To enhance user experience based on insights.	Implemented iterative refinements to the platform based on ethnography study insights.

21. Summarise Findings in Evaluation Report	To document research outcomes and recommendations.	Compiled a comprehensive evaluation report summarising findings and recommending further enhancements.
22. Review Design Principles for User-Friendly AutoML Tools	To inform user-friendly design guidelines.	Reviewed existing literature on design principles for user-friendly AutoML tools.
23. Analyse Evaluation Results for Design Insights	To extract insights into user-friendly platform elements.	Analysed evaluation study results to extract insights into user-friendly elements of the platform.
24. Synthesise Design Guidelines	To provide actionable design recommendations.	Synthesised collective knowledge into a set of design guidelines specific to AutoML tools for non-expert users.
25. Validate Design Guidelines through Usability Testing	To ensure practicality and effectiveness.	Conducted usability testing with non-expert users to validate the practicality and effectiveness of the design guidelines.

The research activities undertaken throughout this study were planned and executed to align with the iterative nature of the User-Centred Design (UCD) methodology, providing the empirical foundation for the development and evaluation of VisAutoML. As outlined in the overall research design and visually represented in the five-stage framework (Figure 15), these activities were not confined to a single phase but were strategically integrated across the research lifecycle. Table 2 provides a detailed description of the purpose and execution for each of these activities, illustrating how they collectively contributed to achieving the research objectives.

The methodology commenced with the Design Research stage (Stage 1 in Figure 15), a crucial phase dedicated to investigating and gathering the necessary requirements for an intuitive AutoML platform for non-expert users (Table 1, Objective 1). This involved Activity 1, an extensive literature review of AutoML tools, focusing on their usability and transparency, executed through a thorough examination of relevant scholarly work (Table 2). Subsequently, Activities 2 and 3 focused on developing and analysing an extended Technology Acceptance Model (TAM) specifically tailored for non-expert users of AutoML tools, employing statistical and qualitative assessments to understand factors influencing user acceptance and behavioural intention (Table 2). Further deepening the understanding of the target audience, Activities 4 and 5 involved performing user surveys and interviews and creating personas, utilizing standardized questionnaires and open-ended inquiries to capture the diverse needs and expectations of non-expert users in machine learning (Table 2). Concurrently, Activities 6 and 7 entailed identifying and analysing existing AutoML platforms and the gathered data to pinpoint common user requirements and preferences (Table 2). The culmination of this stage was Activity 8, the development of a comprehensive user requirements document based on these findings (Table 2).

Following the foundational requirements gathering, the research transitioned into the Design and Prototyping stages (Stages 2 and 4 in Figure 15), aimed at designing and developing a usable AutoML system (Table 1, Objective 2). This phase was guided by Activity 9, a review of best practices in user-

centred design and usability principles (Table 2). Activities 10 and 11 involved the iterative creation of wireframes and low-fidelity prototypes, followed by a high-fidelity prototype, based on the identified user requirements and incorporating user feedback from evaluation stages (Table 2). Wireframes served as skeletal representations defining the basic structure and essential functionalities, carefully constructed to correspond with user requirements and design principles. Prototypes, interactive models of the user interface, were created based on these wireframes, providing a more concrete assessment of the system's capabilities. The process then transitioned to Activities 12 and 13, developing the backend functionality and integrating machine learning algorithms and automation features into the system (Table 2). Activities 14, 15, and 16 focused on designing, developing, and integrating XAI visualisations, such as feature importance plots and SHAP values, seamlessly into the user interface to aid user understanding of model predictions and documenting this process (Table 2). System requirements, including functional and non-functional aspects, were determined accordingly throughout these stages. The system development adopted the waterfall model within these stages, establishing user requirements, designing wireframes, prototype development, and deployment.

Integrated within the iterative UCD framework were the Prototype Evaluation stages (Stages 3 and 5 in Figure 15), critical for evaluating the platform to determine user experience, usability, and transparency (Table 1, Objective 3). This involved Activity 17, recruiting a diverse group of non-expert users for mixed-method studies (Table 2). Activity 18 focused on collecting qualitative data through interviews and surveys to capture users' perceptions, experiences, and knowledge gained (Table 2). Activity 19 involved analyzing the collected data to identify patterns, pain points, and areas for improvement in user experience and usability (Table 2). This comprehensive review approach was crafted to identify areas that needed enhancement, utilizing well-established transparency and usability testing procedures (e.g., UEQ, Trust Questionnaire, XAI Questionnaire), reinforced by critical user feedback. The adoption of a dual-track methodology for usability and transparency testing, which included both quantitative and qualitative aspects, guaranteed an in-depth understanding of user experiences with the prototype. Following this review, Activity 20 entailed iteratively refining the platform based on the insights obtained from these studies (Table 2). The primary goal was to methodically integrate user perspectives at each stage, aiming to fulfill and exceed user expectations by developing a user-centered experience.

Finally, the research activities contributed to achieving the objective of identifying design guidelines for an AutoML tool that supports ML model development and explanation to non-expert users (Table 1, Objective 4). This involved Activity 22, reviewing existing literature on design principles for user-friendly AutoML tools (Table 2). Activity 23 focused on analysing the results from mixed-method studies to extract insights into what makes the platform user-friendly and effective (Table 2). Activity 24 synthesized this collective knowledge into a set of design guidelines specific to AutoML tools for non-expert users (Table 2). The culmination was Activity 25, validating these design guidelines through usability testing with non-expert users to ensure their practicality and effectiveness (Table 2). Thus, the systematic execution of the activities detailed in Table 2, guided by the iterative UCD framework depicted in Figure 15, provided a holistic and systematic approach to address the challenges of AutoML tool adoption among non-expert users.

3.3 Extended Technology Acceptance Model Study: Methodology

3.3.1 Study Design

In order to investigate non-expert users' perceptions of Machine Learning (ML) model development and to assess the usability and transparency of the proposed VisAutoML system, a comprehensive

study was conducted employing a mixed-methods approach. This design integrated both quantitative and qualitative data collection and analysis techniques, chosen to provide a holistic understanding of user acceptance factors and user experiences with the system. The quantitative component primarily relied on data gathered through a structured online questionnaire, while the qualitative component involved conducting semi-structured interview sessions with representatives of the target user group.

This study was theoretically grounded in an extended version of the Technology Acceptance Model (TAM). TAM is recognized as a highly influential and extensively validated theoretical framework in the field of information systems research, widely applied for describing and predicting an individual's acceptance and usage of new technologies (Rafique et al., 2020; Turner et al., 2010). The core of the traditional TAM posits that an individual's behavioural intention (BI) to use a system is directly influenced by their attitude (AT) towards using that system, which is, in turn, predicted by two fundamental beliefs: perceived usefulness (PU) and perceived ease of use (PEOU) (Davis, 1989).

The application of TAM was deemed particularly appropriate and robust within the specific context of evaluating AutoML tools designed for non-expert users for several key reasons. Firstly, prior research has consistently demonstrated that for non-technical users interacting with complex systems, such as ML platforms, the perceived ease of use is a critical determinant influencing their decision to adopt the technology (Bussone et al., 2015; Torkzadeh & Van Dyke, 2002). Systems perceived as requiring significant cognitive effort or technical skill are less likely to be embraced by users who lack extensive technical expertise. Furthermore, in domains like ML, where the underlying processes can often appear opaque, the perceived transparency and overall usability of a tool become essential prerequisites for fostering user trust and encouraging sustained engagement (Wang et al., 2019).

Secondly, the perceived usefulness of the technology holds equally significant importance in this context. Non-expert users are unlikely to invest time and effort in engaging with AutoML systems unless they perceive clear and tangible benefits. These benefits might manifest in terms of improvements to their work performance, enhanced decision-making capabilities, or positive contributions to their personal learning outcomes (Wang et al., 2019). Previous studies have underscored that user trust in the system, coupled with the perceived value it delivers, act as key mediators influencing technology adoption, particularly among user populations lacking specialized technical knowledge (Binns et al., 2018).

Acknowledging that external factors can influence the core TAM constructs and, consequently, behavioral intention (Davis et al., 1992; Venkatesh & Davis, 2000), this study extended the traditional TAM by incorporating two additional constructs: Perceived Authority (PA) and Perceived Enjoyment (ENJ). AutoML platforms, by their nature, automate complex ML workflows, which can inadvertently reduce user control and visibility over critical model development decisions. This abstraction, while intended to enhance ease of use, can simultaneously introduce significant challenges related to usability and transparency (Wang et al., 2019; Holzinger et al., 2019). Users may experience uncertainty, skepticism, or even distrust towards automated system outputs, particularly when the underlying decision-making processes are perceived as non-transparent.

In this specific context, the inclusion of Perceived Authority (PA) was intended to address aspects of the transparency challenge by capturing the influence of social validation and external endorsements in shaping users' trust. When users perceive the system as endorsed by credible authorities or trusted sources, their willingness to accept automated recommendations and outputs can increase, even in situations where they lack full technical understanding of how those outputs were generated (Rai et al., 2019; Venkatesh & Davis, 2000). Thus, PA was hypothesized to support perceived system credibility, which is essential for promoting trust in semi-opaque AutoML processes and thereby

enhancing perceived transparency. Simultaneously, Perceived Enjoyment (ENJ) was considered critical for addressing usability concerns beyond mere effort reduction. Non-expert users engaging with potentially complex AutoML and XAI interfaces might face cognitive overload. Prior work has indicated that intrinsic motivation, often fostered through enjoyable and engaging interactions, can significantly improve users' willingness to explore, learn, and effectively utilize complex systems (Van der Heijden, 2004; Wang et al., 2019). By promoting positive emotional responses and making the interaction enjoyable, ENJ was hypothesized to encourage sustained engagement with explanatory features and the overall workflow, thus supporting deeper comprehension and better usability outcomes. Consequently, the proposed TAM extension explicitly targeted two primary barriers to the adoption of AutoML and XAI systems by non-experts: enhancing usability through fostering engagement (ENJ), and strengthening transparency through perceived credibility (PA). These extensions aimed to ensure that user acceptance of VisAutoML would be driven not solely by instrumental evaluations of usefulness and ease of use, but also by broader socio-psychological factors critical for effective interaction with complex AI systems.

3.3.2 Proposed TAM model

The Technology Acceptance Model (TAM) stands as a highly influential and widely applied theoretical framework for understanding and predicting individual acceptance and usage of information systems (Rafique et al., 2020). At its core, the conventional TAM posits that a user's behavioural intention (BI) to use a system is primarily determined by their attitude (AT) towards using the system. This attitude, in turn, is predicted by two key beliefs: perceived usefulness (PU) and perceived ease of use (PEOU) (Davis, 1989). Perceived usefulness refers to the degree to which a person believes that using a particular system would enhance their job performance or life quality, while perceived ease of use relates to the degree to which a person believes that using a particular system would be free of effort (Davis, 1989). The standard TAM structure suggests a causal path where PEOU influences PU, and both PU and PEOU influence AT, which then leads to BI.

The proposed research model, depicted in Figure 16, represents a theoretically grounded extension of this conventional TAM structure, adapted specifically to the context of non-expert user acceptance of an AutoML tool like VisAutoML. While the core relationships between PU, PEOU, AT, and BI are retained based on established TAM principles (Davis, 1989; Davis et al., 1992), the model is enhanced by the inclusion of external factors hypothesized to influence these core beliefs and attitudes. TAM acknowledges that external variables can impact BI and actual system use, typically mediated through PU and PEOU (Davis et al., 1992; Venkatesh & Davis, 2000).

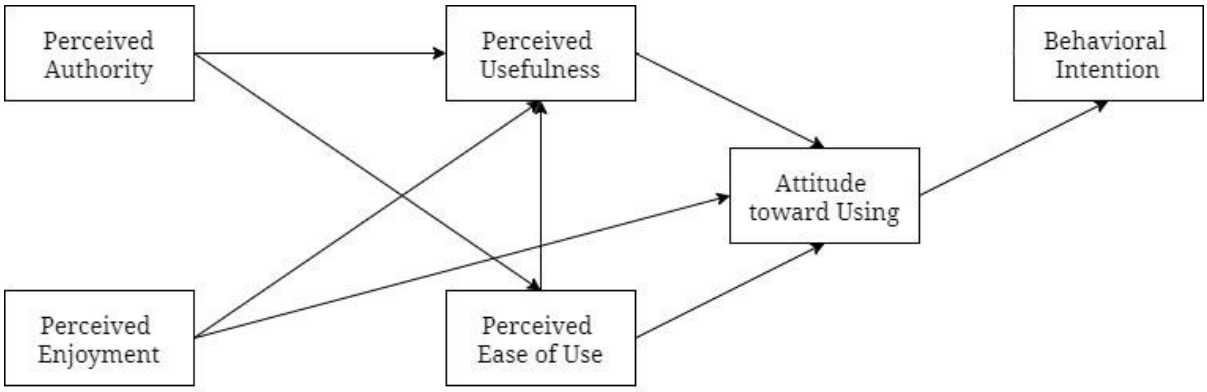


Figure 17 Proposed model

The Technology Acceptance Model (TAM) stands as a highly influential and widely applied theoretical framework for understanding and predicting individual acceptance and usage of information systems (Rafique et al., 2020). At its core, the conventional TAM posits that a user's behavioral intention (BI) to use a system is primarily determined by their attitude (AT) towards using the system. This attitude, in turn, is predicted by two key beliefs: perceived usefulness (PU) and perceived ease of use (PEOU) (Davis, 1989). Perceived usefulness refers to the degree to which a person believes that using a particular system would enhance their job performance or life quality, while perceived ease of use relates to the degree to which a person believes that using a particular system would be free of effort (Davis, 1989). The standard TAM structure suggests a causal path where PEOU influences PU, and both PU and PEOU influence AT, which then leads to BI.

The proposed research model, depicted in Figure 17, represents a theoretically grounded extension of this conventional TAM structure, adapted specifically to the context of non-expert user acceptance of an AutoML tool like VisAutoML. While the core relationships between PU, PEOU, AT, and BI are retained based on established TAM principles (Davis, 1989; Davis et al., 1992), the model is enhanced by the inclusion of external factors hypothesized to influence these core beliefs and attitudes. TAM acknowledges that external variables can impact BI and actual system use, typically mediated through PU and PEOU (Davis et al., 1992; Venkatesh & Davis, 2000).

In this extended model, Perceived Authority (PA) and Perceived Enjoyment (ENJ) are introduced as external constructs, building upon validations by Davis et al. (1992) and Venkatesh & Davis (2000) regarding the influence of external factors. Perceived Enjoyment (ENJ) is defined as the extent to which an activity is perceived to be enjoyable, irrespective of performance consequences (Davis et al., 1992). In the context of an interactive visualization-based AutoML tool, enjoyment derived from the interaction itself is hypothesized to influence both the perceived usefulness (H5) and the attitude towards using the system (H6), suggesting that a more enjoyable experience can make the tool seem more beneficial and foster a positive attitude. Perceived Authority (PA), adapted here to represent a form of social influence, is defined as the extent to which an individual perceives that important others believe they should use the system (Venkatesh & Davis, 2000). This construct is posited to influence perceived usefulness (H7) and perceived ease of use (H8), reflecting the idea that endorsement or perceived social acceptance of the tool can make it seem more useful and easier to adopt.

Figure 16 visually represents these hypothesized relationships. Arrows indicate the direction of influence: PEOU is shown influencing PU (H1) and AT (H2); PU influences AT (H3); AT influences BI (H4). The external factors, ENJ and PA, are shown influencing the core TAM constructs: ENJ influences PU (H5) and AT (H6), while PA influences PU (H7) and PEOU (H8). This structure explicitly models how the unique aspects of an AutoML tool designed for non-experts – potentially offering an enjoyable visual experience and being subject to social influence or perceived credibility – are hypothesized to impact the fundamental drivers of technology acceptance.

The predictive capability of this proposed extended TAM model lies in its ability to forecast non-expert users' behavioural intention to use VisAutoML. By measuring the constructs of Perceived Usefulness, Perceived Ease of Use, Perceived Enjoyment, Perceived Authority, and Attitude toward Using, the model provides a structural basis for understanding the factors that are likely to drive or hinder adoption. If empirical data supports the hypothesized relationships (H1-H8), the model can predict the likelihood of future VisAutoML usage based on users' perceptions across these dimensions. For instance, strong positive perceptions of VisAutoML's ease of use, usefulness, and enjoyability, potentially bolstered by positive social influence, are predicted to lead to a more favourable attitude, which in turn is expected to translate into a higher intention to use the tool in the future. This predictive power is crucial for evaluating the potential success of VisAutoML and identifying which

aspects of the user experience are most critical for promoting adoption among the target non-expert audience. The measurement of these constructs is facilitated by a multi-item questionnaire utilizing a 7-point Likert scale, with specific items designed for each construct (Table 8).

Table 13 Questionnaire structure.

Construct	Items	Measure
Perceived usefulness (PU)	PU1	The system will improve my life/job quality.
	PU2	The system will make my life/job more convenient.
	PU3	The system will make me more effective in my life/job.
	PU4	The system will be useful to me/my job.
	PU5	Tracking and collecting data on myself or the environment will be useful.
	PU6	Using AI and ML to find patterns in my data will be useful.
	PU7	Using visualisation to gain insights into my data will be useful.
Perceived ease of use (PEOU)	PEOU1	It will be easy to track, collect and use data with the system.
	PEOU2	It will be difficult to learn and apply AI and ML models on my own.
	PEOU3	It will be easy to create visualisations using automated visualisations.
	PEOU4	It will be easy for me to become skilful at using the system.
	PEOU5	I have the knowledge necessary to use the system.
	PEOU6	I believe that the system will be easy to use.
Attitude (AT)	AT1	I look forward to using the system.
	AT2	I think that using the system is beneficial to me.
	AT3	I have positive feelings about using the system.
	AT4	I like to learn how AI and ML models work.
Behavioural intention (BI)	BI1	I intend to use the system in the future.
	BI2	I will always try to use the system in my daily life.
	BI3	I plan to use the system frequently.
Perceived authority (PA)	PA1	People whose views I respect support the use of visual analytics.
	PA2	I believe that peer use will increase my perception regarding the reliability of the system.
	PA3	My friends or colleagues will think highly of me if I use the system.
	PA4	I believe the information visualisation should be visible to others.
	PA5	I believe the information visualisation should be fun and interesting to the eye.
Perceived enjoyment (ENJ)	ENJ1	I enjoy interacting with visualisations
	ENJ2	I have fun using visualisations to gain insights into my data
	ENJ3	Using visualisations on data would be interesting

3.3.3 Participants

This section details the methodology employed for recruiting participants and collecting their demographic information for the Extended Technology Acceptance Model (TAM) study. As the study aimed to investigate the perceptions and acceptance of an AutoML tool among non-expert users, the

recruitment strategy focused on identifying and engaging individuals with limited prior experience in Machine Learning (ML) and related technical domains.

For the quantitative component of the study, which involved the administration of an online questionnaire, participants were primarily recruited using a snowball sampling method. This approach was initiated by contacting an initial group of individuals who fit the criteria of being potential non-experts in AI and subsequently requesting them to recommend other individuals from their network with similar characteristics who might be willing to participate. Snowball sampling was deemed suitable for this exploratory phase as it facilitated reaching a population of non-expert users who might not be readily accessible through random sampling methods.

The qualitative component of the study, comprising semi-structured interviews, utilized a purposive sampling strategy. Participants for the interviews were specifically sought out based on predefined criteria aligning with the definition of non-experts in AI and information visualization development. Recruitment efforts for the interviews included the dissemination of emailed advertisements and leveraging word-of-mouth referrals within relevant networks. This purposive approach allowed for the selection of participants whose backgrounds and experiences were most pertinent to the research objectives, enabling in-depth exploration of their anticipated interactions with and perceptions of the proposed system.

Demographic information for both the questionnaire and interview participants was collected to provide context for the study findings and to characterize the sample in relation to the target non-expert population. For the online questionnaire, demographic data, including age, gender, educational background, field of study, and self-reported ML experience, was collected via dedicated self-report items integrated within the questionnaire instrument. For the semi-structured interviews, demographic information was collected directly from participants during the interview sessions. The collection of this data allowed for a description of the participant pool in the results chapter, enabling an assessment of the sample's characteristics in relation to the study's focus on non-expert users. The recruitment strategies and participant selection aimed to ensure that the individuals participating in the study represented the intended non-expert user group, thereby enhancing the relevance of the findings to the research questions concerning the usability and transparency of AutoML for this audience.

3.3.4 Data Collection

Data for the Extended Technology Acceptance Model (TAM) study was collected using a mixed-methods approach, comprising both quantitative and qualitative components. This dual approach was employed to capture a comprehensive understanding of non-expert users' perceptions, acceptance factors, and experiences related to Machine Learning (ML) model development and the proposed VisAutoML system.

The quantitative data was primarily collected through a structured online questionnaire. This instrument was designed to measure participants' perceptions across the constructs of the Extended TAM model: Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Attitude toward Using (AT), Behavioural Intention (BI), Perceived Authority (PA), and Perceived Enjoyment (ENJ). The questionnaire utilized a 7-point Likert scale, ranging from "strongly agree" to "strongly disagree," for items measuring these constructs. As detailed in Table 8, the questionnaire consisted of a specific

number of items for each construct (7 for PU, 6 for PEOU, 4 for AT, 3 for BI, 5 for PA, and 3 for ENJ). In addition to the Likert scale items, the questionnaire also included several multiple-choice, ranking, and subjective open-ended questions to gather further insights into participants' preferences, expectations, and challenges related to interacting with an automated ML system. The online questionnaire was administered to the recruited participants, and responses were recorded for subsequent quantitative analysis.

The qualitative data was collected through semi-structured interview sessions. These interviews were designed to explore participants' anticipated experiences and perspectives on the proposed system in greater depth than the questionnaire could allow. The interviews followed a semi-structured format, guided by questions structured around three main constructs: perceived ease of use of AI (C1), perceived usefulness of visualisations (C2), and platform ease of use (C3). This format provided flexibility, allowing the interviewer to probe deeper into participants' responses and explore emergent themes while ensuring coverage of key areas of interest. The interview sessions were conducted online via Microsoft Teams and typically lasted for approximately one hour each. Participant responses during the interviews were recorded for subsequent qualitative analysis, specifically thematic analysis, to identify common patterns and themes related to user perceptions and requirements.

3.3.5 Data Analysis

The data collected from the online questionnaire and semi-structured interviews were subjected to rigorous analysis using both quantitative and qualitative techniques, aligning with the mixed-methods design of the study. The analytical procedures were selected to address the research objectives by exploring the relationships between the Extended Technology Acceptance Model (TAM) constructs and gaining in-depth insights into non-expert users' perceptions and requirements.

For the quantitative data obtained from the online questionnaire, statistical analysis was conducted using Statistical Package for the Social Sciences (SPSS), version 26. The initial steps involved data standardization to ensure comparability across different items and constructs. Descriptive statistics, including means and standard deviations, were calculated for each questionnaire item and for the aggregated constructs. To assess the internal consistency and reliability of the multi-item scales used to measure the TAM constructs, Cronbach's alpha coefficients were calculated (Yan & Yibing, 2010). Values above 0.7 were considered acceptable, indicating satisfactory reliability. Pearson correlation analysis was then performed to examine the strength and directionality of the linear relationships between the constructs of the proposed Extended TAM model. Finally, linear and multiple regression analyses were conducted to investigate the hypothesized causal relationships between the independent and dependent variables within the model (Field, 2013). This allowed for the determination of the significance and strength of the influence of perceived ease of use, perceived usefulness, perceived enjoyment, and perceived authority on attitude towards using and behavioural intention, thereby testing the study's hypotheses.

The qualitative data, collected through semi-structured interviews and the open-ended questions included in the online questionnaire, were analysed using a thematic analysis approach. This involved reading and re-reading the transcribed interview data and responses to open-ended questions to become thoroughly familiar with the content. The data were then systematically coded, assigning labels to segments of text that represented key concepts or themes related to the research questions

and the constructs of interest (C1, C2, C3 for interviews, and thematic grouping for open-ended responses). Codes were subsequently grouped into broader themes and sub-themes, identifying common patterns, user perspectives, challenges, and suggestions for improvement regarding ML model development and the proposed system. Illustrative quotes were selected from the data to support and exemplify the identified themes in the results presentation.

To ensure the trustworthiness and rigor of the qualitative findings, validity assessment was conducted, focusing on the criteria of transferability, credibility, and confirmability (Petersen & Gencel, 2013; Yin, 2009). Transferability, the extent to which the findings can be applied to other similar contexts, was addressed by providing detailed descriptions of the participant characteristics and the study setting. Credibility, ensuring the factual accuracy and truthfulness of the account, was enhanced through data source triangulation, comparing and cross-validating insights obtained from the interviews and the open-ended questionnaire responses. Confirmability, the degree to which the findings are objective and not influenced by researcher bias, was considered through a process of reflexivity, critically examining the researcher's own assumptions and potential influences on the interpretation of the data.

3.4 Prototype Evaluation Study 1: VisAutoML 1.0 Comparison Study

3.4.1 Study Design

This section details the design of the first prototype evaluation study, a comparison conducted between VisAutoML 1.0 and an existing AutoML tool. The primary objective of this study was to empirically evaluate the usability, transparency, and overall effectiveness of the initial VisAutoML prototype in comparison to an established solution, specifically tailored for non-expert users. This comparative approach aimed to benchmark VisAutoML 1.0 and identify its relative strengths and weaknesses from a user perspective, providing crucial empirical data to inform subsequent design iterations.

The study employed a between-subjects experimental design. This design involves assigning different participants to different conditions, allowing for a direct comparison of outcomes between groups exposed to distinct treatments. In this case, participants were randomly assigned to one of two groups: an experimental group and a control group. The experimental group engaged in developing Machine Learning (ML) models using both the VisAutoML 1.0 prototype and the chosen comparison tool. The control group, conversely, performed the same ML development tasks exclusively using the comparison tool. This between-subjects design was chosen to minimize carryover effects that might occur if participants used both tools sequentially, ensuring that the observed differences in user experience and performance could be more confidently attributed to the specific tool used.

The existing AutoML tool selected for comparison was H2O AutoML. The selection of H2O AutoML was based on specific criteria deemed relevant for a comparison focused on non-expert usability and transparency. These criteria included its graphical user interface (GUI)-based development capabilities, which make it accessible to users without requiring command-line interaction. Furthermore, its inclusion of AutoML functionality, automating key steps of the ML pipeline, and its provision of Explainable Artificial Intelligence (XAI) features were important considerations, aligning with the focus of this research. Its general availability for review and use, along with H2O AutoML's widespread recognition and use within the broader ML domain, provided a relevant and credible benchmark for evaluating a new tool targeting a similar audience.

The evaluation took place within controlled university lecture settings, providing a consistent environment for data collection and minimizing potential confounding variables. The study was conducted over a specific timeframe, spanning March 2023 to April 2023. During the study sessions, participants in both groups were instructed to perform practical ML development tasks designed to simulate realistic use cases. These tasks involved developing ML models for both regression and classification problems, requiring participants to interact with the tools to prepare data, build models, and interpret results. The methodology involved participants applying the respective tools to a common dataset to ensure comparability of the development process and outcomes across both the experimental and control groups. This standardized approach to task execution and dataset usage was critical for obtaining reliable comparative data.

3.4.2 Tool Selection

The empirical evaluation of the VisAutoML 1.0 prototype necessitated a comparative study against an existing AutoML tool. This comparison was crucial for benchmarking the prototype's performance and user experience from the perspective of non-expert users, providing valuable context for its strengths and weaknesses relative to established systems. The selection of an appropriate comparison tool was guided by specific, justified criteria derived from the research objectives and the identified limitations of current AutoML and XAI tools for the target audience. These criteria aimed to ensure the comparison was relevant and provided meaningful insights into the effectiveness of VisAutoML's design approach in addressing documented challenges.

The key selection criteria were:

1. Graphical User Interface (GUI)-based development: A fundamental goal of VisAutoML is to democratize access to Machine Learning (ML) by providing a no-code or low-code visual interface, thereby lowering the barrier to entry for non-experts who may lack traditional programming skills (Hutter et al., 2019; Singh & Joshi, 2022). Consequently, the comparison tool needed to similarly offer a GUI-based approach to ML model development. This criterion ensured that the comparative evaluation could focus on the usability and transparency of the visual workflow and interaction design, rather than being confounded by differences in the technical skill set required to operate the tools (Elshawi et al., 2019; Yang et al., 2018). Evaluating tools with similar interaction paradigms allowed for a more direct assessment of how design choices within a visual environment impact non-expert users.

2. AutoML and Explainable Artificial Intelligence (XAI) functionality: This research posits that effective tools for non-experts must integrate both AutoML capabilities (automating parts of the ML pipeline) and usable XAI features (providing comprehensible model explanations) to enhance transparency and foster trust (Adadi & Berrada, 2018; Ribeiro et al., 2016). Existing AutoML tools often prioritize automation speed over transparency, resulting in "black-box" systems that are difficult for non-experts to understand (Kaur et al., 2020; Coors et al., 2021). While XAI tools exist, they are frequently designed for experts and their outputs may not be interpretable by non-technical users (Hohman et al., 2019; Wang et al., 2021). Therefore, the comparison tool needed to possess functionality in both these areas to enable a direct assessment of how VisAutoML's integrated and user-centred approach to explainability compared to existing methods within an automated ML context. This allowed for evaluation of how the tool's design facilitated understanding of automated decisions and model outputs.

3. Availability for review: Practical constraints of the research necessitated the selection of a tool that was readily accessible and available for empirical evaluation within a controlled study setting. Furthermore, selecting a tool with established use and recognition within the ML domain provided a valuable benchmark, allowing the study's findings to be contextualized within the existing landscape of AutoML tools. The prevalence of the tool in academic literature and commercial use indicated its relevance and provided a basis for comparison against a widely adopted system, lending credibility to the comparative results (Zöller & Huber, 2021).

Following these justified criteria, H2O AutoML was selected as the comparison system for the evaluation study. H2O AutoML aligns well with the specified requirements:

1. It offers a graphical user interface (GUI) that facilitates model development without requiring intricate coding skills, thereby meeting the GUI-based development criterion and making it accessible to users with diverse levels of programming and ML expertise (H2O Official Documentation, n.d.; Zöller & Huber, 2021). This visual interface supports users who are not comfortable with code-based ML development.

2. It provides robust AutoML functionality, automating various steps of the ML pipeline, and incorporates a level of XAI functionality designed to offer explanations for trained models, thus addressing the requirement for both capabilities (H2O Official Documentation, n.d.; Santu et al., 2022). This allows for a comparison of how automation and explanation are presented to users.

3. H2O AutoML is widely used and recognized within the machine learning domain, with significant prevalence in both commercial applications and academic research, fulfilling the availability for review criterion and providing a credible benchmark for comparison (Zöller & Huber, 2021). Its established presence makes it a suitable representative of existing AutoML solutions accessible to a broader user base.

The selection of H2O AutoML based on these criteria allowed for a relevant and informative comparison study, providing valuable insights into the relative strengths and weaknesses of the VisAutoML 1.0 prototype's design and functionality from a non-expert user's perspective. The main goal of this comparison was to empirically inspect and contrast the features, capabilities, and perceived performance of VisAutoML 1.0 against an established tool, thereby contributing to the understanding of effective design strategies for usable and transparent AutoML systems for non-experts.

3.4.3 Participants

This section outlines the methodology for participant recruitment and the collection of demographic and prior experience data for the first prototype evaluation study, the comparison between VisAutoML 1.0 and H2O AutoML. The selection and characteristics of the participants were crucial for ensuring the relevance of the study's findings to the target non-expert user group.

Participants for this comparison study were recruited from university lecture settings. This recruitment approach provided access to a population with diverse academic backgrounds, including Computer Science and Finance, as noted in the results of this study. A total of 82 students participated in this evaluation. Following recruitment, participants were randomly assigned to either the experimental group or the control group. Random assignment was employed to distribute potential confounding variables, such as prior technical experience or inherent learning abilities, evenly between the two conditions, thereby strengthening the internal validity of the comparison.

Demographic information was systematically collected from all participants prior to their engagement with the study tasks. This included data on age, gender, and educational background. The collection method involved the administration of a pre-test survey, which included specific items designed to capture this information through self-report. Gathering these demographic details allowed for a characterization of the study sample, providing context for the interpretation of the results and enabling an assessment of the sample's representativeness in relation to the broader non-expert population.

In addition to basic demographics, data on participants' prior experience with programming and Machine Learning (ML) was also collected using the pre-test survey. Participants were asked to rate their proficiency or experience level in these areas, typically on a Likert scale. This information was essential for confirming that the recruited sample aligned with the study's focus on non-experts and for understanding the baseline technical exposure of the participants in both the experimental and control groups. The collection of both demographic and prior experience data using standardized survey instruments ensured consistency and facilitated the description of the participant cohorts in the results chapter.

3.4.4 Data Collection

This section details the methodology employed for collecting data during the first prototype evaluation study, the comparison between VisAutoML 1.0 and H2O AutoML. A mixed-methods approach was utilized to gather both quantitative and qualitative data, providing a comprehensive assessment of user experiences, usability, transparency, and knowledge gain when interacting with the two AutoML tools.

Quantitative data collection involved the administration of several standardized instruments. Participants completed pre-test and post-test knowledge domain surveys. These surveys consisted of items designed to assess participants' understanding of fundamental concepts related to Machine Learning development, including data preparation steps and algorithms. The questions varied between the pre-test and post-test, but addressed the same core knowledge subject areas, allowing for the measurement of knowledge gain over the course of the study. Usability was quantitatively assessed using the System Usability Scale (SUS) and the User Experience Questionnaire (UEQ). The SUS is a 10-item questionnaire utilizing a 5-point Likert scale, providing a single score reflecting overall perceived usability (Brooke, 1996). The UEQ is a more comprehensive instrument with multiple scales covering pragmatic and hedonic quality dimensions, typically using a 7-point Likert scale (Schrepp et al., 2017a). These questionnaires were administered after participants had completed the study tasks with the respective tools.

Qualitative data collection methods were employed to capture richer insights into participants' experiences, perceptions, and suggestions for improvement. Open-ended questions were included in

the questionnaires to allow participants to freely express what they liked and disliked about the tools and to provide suggestions for enhancement. Additionally, semi-structured interviews were conducted with a subset of participants. These interviews followed a flexible format, guided by key themes related to usability and transparency, allowing researchers to probe deeper into participants' interactions with the tools and gather detailed feedback on their experiences. The qualitative data collected through these methods provided valuable context for interpreting the quantitative results and identifying specific areas for refinement in the VisAutoML prototype.

3.4.5 Data Analysis

This section describes the analytical procedures applied to the data collected during the first prototype evaluation study, the comparison between VisAutoML 1.0 and H2O AutoML. Consistent with the mixed-methods approach employed for data collection, both quantitative and qualitative analysis techniques were utilized to interpret the findings related to usability, transparency, and knowledge gain.

Quantitative data analysis was conducted using Statistical Package for the Social Sciences (SPSS), version 26. The data obtained from the pre-test and post-test knowledge domain surveys, the System Usability Scale (SUS), and the User Experience Questionnaire (UEQ) were subjected to statistical tests to enable comparison between the experimental and control groups. To assess differences in usability metrics (UEQ scales and SUS scores) and knowledge gain scores between the two groups, independent samples t-tests were performed. Additionally, one-way between-groups ANOVA was conducted to further examine the significance of disparities in knowledge gain following the intervention. Prior to conducting t-tests or ANOVA, Levene's test for equality of variances was employed to check the assumption of homogeneity of variances, ensuring the appropriate statistical test was interpreted. Descriptive statistics, including means and standard deviations, were calculated for all quantitative measures to summarize the data within each group.

Qualitative data analysis was performed on the responses gathered from the open-ended questions included in the questionnaires and the transcripts from the semi-structured interviews. A thematic analysis approach was applied to this qualitative data. This involved systematically reading through the responses and transcripts to identify recurring patterns, ideas, and sentiments expressed by participants regarding their experiences with VisAutoML 1.0 and H2O AutoML. The process included initial coding of the data, grouping codes into broader themes, and interpreting these themes in relation to the study objectives concerning usability and transparency. This qualitative analysis provided rich, contextual insights that complemented the quantitative findings, helping to explain observed differences between the groups and identify specific aspects of the tools that influenced user experience.

3.5 Prototype Evaluation Study 2: VisAutoML 1.0 Usability and Transparency Evaluation

3.5.1 Study Design

This section describes the design of the second prototype evaluation study, which focused on an in-depth assessment of the VisAutoML 1.0 prototype's usability and transparency. Situated within the iterative User-Centred Design (UCD) framework (Figure 15, Stage 3), the primary objective of this

study was to gain a detailed understanding of non-expert users' experiences when interacting with the initial version of the tool. This evaluation aimed to identify specific strengths, weaknesses, and areas for improvement, providing crucial empirical data to inform the subsequent redesign phase (Figure 15, Stage 4).

The study employed a single-group evaluation design. Unlike the preceding comparison study which involved multiple tools, this evaluation concentrated solely on participants' interactions with VisAutoML 1.0. This design was chosen to allow for a focused and detailed examination of user experience, usability, and transparency metrics specifically within the context of the VisAutoML prototype, without the confounding factor of comparing it directly against another system during the evaluation tasks.

Participants in this study were instructed to perform practical tasks designed to simulate a typical Machine Learning (ML) development workflow using the VisAutoML 1.0 tool. These tasks involved engaging with a pre-loaded dataset (specifically, a built-in Titanic dataset, as mentioned in the original draft) and utilizing the tool's functionalities to build and interact with an ML model. The tasks were structured to guide participants through key stages of the ML process supported by VisAutoML 1.0, such as data loading, potentially data review, model development (e.g., selecting features for a classification task), and interacting with the resulting Explainable Artificial Intelligence (XAI) visualizations to understand model predictions and feature importance. The specific nature of the tasks, focusing on classification and interaction with XAI, was intended to elicit user feedback on the tool's core features designed to enhance usability and transparency for non-experts.

An important aspect of the methodology for this main evaluation study was the preceding conduct of iterative pilot studies. As noted in the original draft, these pilot studies involved a smaller number of participants and were instrumental in refining the study design, data collection instruments, and procedures prior to the main evaluation. Insights gained from the pilot phases, such as optimizing study duration and improving instructions, directly informed the final methodology implemented for this in-depth evaluation of VisAutoML 1.0, aiming to enhance data quality and the overall evaluation process.

3.5.2 Participants

This section details the methodology employed for recruiting participants and collecting their demographic and prior experience data for the in-depth evaluation study of the VisAutoML 1.0 prototype. As the study aimed to assess the usability and transparency of the tool among non-expert users, the recruitment strategy focused on engaging individuals with limited prior experience in Machine Learning (ML) and related technical domains, consistent with the target audience definition.

Participants for this evaluation study were recruited through the Amazon Mechanical Turk platform. This online crowdsourcing marketplace was utilized to access a diverse pool of potential participants. An initial number of participants were recruited for the study. Following data cleaning and screening procedures, a subset of these participants were retained for the final evaluation analysis.

Demographic information was systematically collected from all participants who completed the study. This included data on age, gender, and educational attainment. The collection method involved the administration of a questionnaire prior to participants engaging with the main study tasks. This

questionnaire included specific self-report items designed to capture these demographic characteristics. Gathering these details allowed for a description of the study sample in the results chapter, providing essential context for interpreting the findings and assessing the sample's characteristics.

In addition to basic demographics, data on participants' prior experience with Machine Learning (ML) was also collected using the same pre-task questionnaire. Participants were asked to rate their proficiency or experience level in ML, typically through a self-assessment on a Likert scale. This information was crucial for confirming that the recruited sample aligned with the study's focus on non-experts, defined as individuals with limited prior ML exposure.

Furthermore, a screening procedure was implemented to enhance data quality and ensure participants met certain criteria. This involved administering a short quiz based on the instructions provided for the study tasks. Participants were required to answer all questions correctly to be included in the final study sample. This methodology aimed to identify and exclude participants who may not have fully understood the task requirements or were not engaging seriously with the study, thereby contributing to the reliability of the collected data.

3.5.3 Data Collection

This section describes the methodology employed for collecting data during the second prototype evaluation study, the in-depth assessment of VisAutoML 1.0. A mixed-methods approach was utilized to gather both quantitative and qualitative data, providing a comprehensive evaluation of user experiences, usability, and transparency when interacting with the VisAutoML 1.0 prototype.

Quantitative data collection involved the recording of task completion time and the administration of several standardized questionnaires. Participants' time taken to complete the designated Machine Learning (ML) development task using VisAutoML 1.0 was recorded. This metric provided a quantitative measure of the tool's efficiency from a user perspective. Usability was quantitatively assessed using the User Experience Questionnaire-Short Form (UEQ-S). The UEQ-S is a condensed version of the UEQ, consisting of 8 items typically measured on a 7-point Likert scale, designed to efficiently capture key dimensions of user experience, including pragmatic and hedonic quality (Schrepp et al., 2017b). Transparency was quantitatively evaluated using two instruments: the Trust Questionnaire and the Explainable Artificial Intelligence (XAI) Questionnaire. The Trust Questionnaire, comprising 7 items on a 5-point Likert scale, aimed to measure participants' confidence in and reliance on the VisAutoML 1.0 system (Hoffman et al., 2023). The XAI Questionnaire, consisting of 30 items on a 7-point Likert scale, was used to assess users' perceptions of the explainability of the XAI features provided by the tool (Silva et al., 2023). These questionnaires were administered after participants had completed the study tasks.

Qualitative data collection methods were employed to capture richer insights into participants' experiences, perceptions, and suggestions for improvement. While the primary quantitative data provided metrics on usability and transparency, qualitative feedback offered valuable context and depth. This included responses to open-ended questions that may have been part of the questionnaires, allowing participants to articulate their thoughts and experiences in their own words. The qualitative data collected through these methods provided valuable insights into specific aspects of the VisAutoML 1.0 prototype that influenced user experience and helped identify areas for

refinement.

3.5.4 Data Analysis

This section describes the analytical procedures applied to the data collected during the second prototype evaluation study, the in-depth assessment of VisAutoML 1.0. Consistent with the mixed-methods approach employed for data collection, both quantitative and qualitative analysis techniques were utilized to interpret the findings related to usability, transparency, and user experience.

Quantitative data analysis was performed on the task completion time and the responses from the standardized questionnaires (UEQ-S, Trust Questionnaire, and XAI Questionnaire). Descriptive statistics, including means, standard deviations, and ranges, were calculated to summarize the task completion time data, providing an overview of the tool's efficiency. For the questionnaire data, descriptive statistics were computed for each item and for the aggregated scales (pragmatic quality, hedonic quality, overall usability, trust, and explainability). To assess the internal consistency and reliability of the multi-item scales, Cronbach's alpha coefficients were calculated for the UEQ-S, Trust Questionnaire, and XAI Questionnaire (Yan & Yibing, 2010). Reliability coefficients were examined to ensure the scales provided consistent measurements.

To explore the relationships between the key constructs of usability, trust, and explainability, Pearson correlation analysis was conducted. This statistical technique was used to determine the strength and direction of the linear association between the scores obtained from the UEQ-S, Trust Questionnaire, and XAI Questionnaire. Correlation coefficients and their significance levels were examined to understand how perceived usability related to perceived trust and explainability within the VisAutoML 1.0 context.

Qualitative data analysis was conducted on any textual responses obtained from open-ended questions included in the questionnaires. A thematic analysis approach was applied to these qualitative data. This involved systematically reviewing the responses to identify recurring themes, patterns, and insights related to participants' experiences, perceptions of usability and transparency, and suggestions for improvement. The process included coding the data and grouping codes into broader themes to provide a rich, contextual understanding that complemented the quantitative findings.

3.6 Prototype Evaluation Study 3: VisAutoML 2.0 Usability and Transparency Evaluation

3.6.1 Study Design

This section details the design of the third prototype evaluation study, focusing on an in-depth assessment of the refined VisAutoML 2.0 prototype's usability and transparency. Situated within the iterative User-Centred Design (UCD) framework (Figure 15, Stage 5), the primary objective of this study was to evaluate the effectiveness of the improvements implemented based on the findings from the preceding evaluation stages. This evaluation aimed to determine the user experience, usability, and transparency of the enhanced tool, providing crucial empirical data on its performance and user acceptance.

The study employed a single-group evaluation design. Similar to the in-depth evaluation of VisAutoML 1.0, this study concentrated solely on participants' interactions with the VisAutoML 2.0 prototype. This design was selected to facilitate a focused and detailed examination of user experience, usability, and transparency metrics specifically within the context of the refined VisAutoML system. The goal

was to assess the impact of the design iterations on user perceptions and interactions without the complexities of a comparative setup during the evaluation tasks.

Participants in this study were instructed to perform practical tasks designed to simulate a typical Machine Learning (ML) development workflow using the VisAutoML 2.0 tool. Specifically, participants were given the following tasks:

1. Access VisAutoML using a provided link and launch the application.

2. On the home page, start a quick tour to familiarize themselves with the interface.

3. By following the guides, create a classification ML model using the built-in Titanic dataset, specifying 'Name' as the ID column and removing irrelevant columns.

4. On the model evaluation / predict & explain page, explore all five tabs and determine: the most important feature; the relationship (negative/positive correlation) between 'survived' and 'passenger class'; and the relationship (negative/positive correlation) between 'survived' and 'sex'.

These tasks involved engaging with a pre-loaded dataset (specifically, a built-in Titanic dataset) and utilizing the tool's functionalities to build and interact with an ML model. The tasks were structured to guide participants through key stages of the ML process supported by VisAutoML 2.0, such as data loading, model development (e.g., building a classification model), and interacting with the resulting Explainable Artificial Intelligence (XAI) visualizations to understand model predictions and feature importance. The specific nature of the tasks, focusing on classification and interaction with XAI, was intended to elicit user feedback on the tool's core features designed to enhance usability and transparency for non-experts, particularly those features that were refined in the transition from VisAutoML 1.0 to 2.0.

3.6.2 Participants

This section details the methodology employed for recruiting participants and collecting their demographic and prior experience data for the in-depth evaluation study of the refined VisAutoML 2.0 prototype. Consistent with the preceding evaluations, the recruitment strategy focused on engaging individuals with limited prior experience in Machine Learning (ML) and related technical domains, aligning with the definition of the target non-expert user group.

Participants for this evaluation study were recruited through the Amazon Mechanical Turk platform. This online crowdsourcing marketplace was utilized to access a broad pool of potential participants. An initial number of participants were recruited for the study, from which a final sample was retained for analysis following data screening and quality control procedures.

Demographic information was systematically collected from all participants who completed the study. This included data on age, gender, educational attainment, and field of study. The collection method involved the administration of a questionnaire prior to participants engaging with the main study tasks. This questionnaire included specific self-report items designed to capture these demographic characteristics. Gathering these details allowed for a comprehensive description of the study sample in the results chapter, providing essential context for interpreting the findings and assessing the sample's characteristics.

In addition to basic demographics, data on participants' prior experience with Machine Learning (ML) was also collected using the same pre-task questionnaire. Participants were asked to rate their proficiency or experience level in ML, typically through a self-assessment on a Likert scale. This information was crucial for confirming that the recruited sample aligned with the study's focus on non-experts, defined as individuals with limited prior ML exposure.

Furthermore, a screening procedure was implemented to enhance data quality and ensure participants met certain criteria, including attentive engagement with the study instructions. This involved administering a short quiz based on the instructions provided for the study tasks. Participants were required to correctly answer all questions on this quiz to be included in the final study sample. This methodology aimed to identify and exclude participants who may not have fully understood the task requirements or were not engaging seriously with the study, thereby contributing to the reliability and validity of the collected data.

3.6.3 Data Collection

This section describes the methodology employed for collecting data during the third prototype evaluation study, the in-depth assessment of the refined VisAutoML 2.0. Consistent with the preceding evaluation phases and the mixed-methods approach of the overall research design, data collection integrated both quantitative and qualitative techniques. This comprehensive approach aimed to capture a detailed understanding of user experiences, usability, and transparency when interacting with the enhanced VisAutoML 2.0 prototype.

Quantitative data collection involved the recording of task completion time and the administration of several standardized questionnaires. Participants' time taken to complete the designated Machine Learning (ML) development task using VisAutoML 2.0 was systematically recorded. This metric provided a direct, objective measure of the tool's efficiency from a user perspective. Usability was quantitatively assessed using the User Experience Questionnaire-Short Form (UEQ-S). The UEQ-S is a validated instrument designed to efficiently capture key dimensions of user experience, including pragmatic (e.g., efficiency, perspicuity) and hedonic (e.g., stimulation, novelty) quality, typically measured on a 7-point Likert scale (Schrepp et al., 2017b).

Transparency was quantitatively evaluated using two specific instruments: the Trust Questionnaire and the Explainable Artificial Intelligence (XAI) Questionnaire. The Trust Questionnaire, comprising multiple items typically rated on a Likert scale (e.g., 5-point), aimed to measure participants' confidence in and reliance on the VisAutoML 2.0 system (Hoffman et al., 2023). The XAI Questionnaire, consisting of a larger set of items rated on a Likert scale (e.g., 7-point), was specifically designed to assess users' perceptions of the explainability of the XAI features provided by the tool (Silva et al., 2023). These standardized questionnaires were administered to participants after they had completed the study tasks with the VisAutoML 2.0 prototype.

Qualitative data collection methods were employed to capture richer, more nuanced insights into participants' experiences, perceptions, and suggestions for improvement that quantitative measures alone might not fully reveal. This included the collection of responses to open-ended questions. These questions were designed to allow participants to articulate, in their own words, what they liked and disliked about the VisAutoML 2.0 tool and to provide specific suggestions for enhancement. The qualitative data collected through these methods provided valuable context for interpreting the quantitative results and identifying specific aspects of the refined prototype that influenced user experience, transparency perceptions, and overall satisfaction.

3.6.3 Data Analysis

This section describes the analytical procedures applied to the data collected during the third prototype evaluation study, the in-depth assessment of the refined VisAutoML 2.0. Consistent with the mixed-methods approach employed for data collection, both quantitative and qualitative analysis techniques were utilized to interpret the findings related to usability, transparency, and user experience.

Quantitative data analysis was performed on the recorded task completion time and the responses from the standardized questionnaires (UEQ-S, Trust Questionnaire, and XAI Questionnaire). Descriptive statistics, including means, standard deviations, and frequency distributions (e.g., for task completion time categories), were calculated to summarize the data for each quantitative measure. This provided an overview of the tool's efficiency and participants' perceptions across the usability and transparency constructs. To assess the internal consistency and reliability of the multi-item scales, Cronbach's alpha coefficients were calculated for the UEQ-S, Trust Questionnaire, and XAI Questionnaire (Yan & Yibing, 2010). Reliability coefficients were examined to ensure the scales provided consistent measurements of the intended constructs.

To explore the relationships between the key constructs of usability, trust, and explainability within the context of VisAutoML 2.0, Pearson correlation analysis was conducted. This statistical technique was used to determine the strength and direction of the linear association between the scores obtained from the UEQ-S (representing usability dimensions), Trust Questionnaire, and XAI Questionnaire. Correlation coefficients and their significance levels were examined to understand how perceived usability related to perceived trust and explainability for the refined prototype.

Qualitative data analysis was conducted on the textual responses obtained from the open-ended questions included in the questionnaires. A thematic analysis approach was applied to these qualitative data. This involved systematically reviewing the responses to identify recurring themes, patterns, and insights related to participants' experiences, perceptions of the refined VisAutoML 2.0's usability and transparency, and suggestions for further improvement. The process included initial coding of the data, grouping codes into broader themes, and interpreting these themes to provide a rich, contextual understanding that complemented the quantitative findings. This qualitative analysis was crucial for gaining deeper insights into specific aspects of the refined prototype that influenced user experience and for identifying actionable areas for future development.

3.7 System Development Methodology

This section details the methodology employed for the development of the VisAutoML prototypes, encompassing both VisAutoML 1.0 and its subsequent iteration, VisAutoML 2.0. Situated within the iterative User-Centred Design (UCD) framework (Figure 15), the system development activities primarily occurred during the Design and Prototyping stages (Stages 2 and 4). The methodology aimed to translate the user requirements and design principles identified in the preceding research phases into a functional and interactive software tool for evaluating usability and transparency among non-expert users.

The overall approach to system development within these stages adopted aspects of the waterfall model. This provided a structured sequence of development phases, beginning with clearly defined user and system requirements established during the Design Research phase (Stage 1). Following

requirement specification, the process moved through design (wireframing and prototyping), implementation (coding), and implicitly, testing phases, although the primary evaluation was conducted in dedicated Prototype Evaluation stages (Stages 3 and 5). While the UCD framework is iterative, the waterfall model provided a systematic approach for building each prototype version based on a defined set of specifications derived from the user-centred insights.

The implementation of the VisAutoML prototypes leveraged a combination of established web development frameworks and Machine Learning (ML) libraries. The system was primarily developed as a web application, utilizing React JS for the frontend user interface. React JS, a popular JavaScript library, was chosen for its component-based architecture, which facilitated the creation of reusable UI elements and enhanced the interactivity of the application. The Material-UI (MUI) library was also employed to provide pre-designed components, contributing to a modern and user-friendly interface aesthetic.

The backend functionality was predominantly built using the Django framework, a high-level Python web framework. Django's Model-View-Template (MVT) design pattern provided a structured approach for handling data storage, business logic, and serving content. Data storage and management were facilitated by Django's Object-Relational Mapping (ORM), connecting to a database. Django exposed a set of RESTful APIs that the React JS frontend and other components utilized to retrieve and manipulate data.

For the implementation of ML algorithms and data processing, Python libraries such as Pandas and Scikit-learn (Sklearn) were integral. Pandas was used for efficient data manipulation and analysis, including tasks like data cleaning, handling missing values, scaling, encoding, and splitting datasets. Sklearn provided a wide range of ML algorithms for implementing classification and regression tasks, as well as tools for model training and evaluation. The system incorporated various algorithms for automated model selection, evaluating them against the dataset to identify the best-performing one.

A separate component, utilizing the Flask framework, was responsible for generating the interactive Explainable Artificial Intelligence (XAI) visualizations. Flask, a lightweight Python web framework, was used in conjunction with visualization libraries such as Plotly and Shap. Plotly Dash was specifically employed to create interactive web-based visualizations that allowed users to explore SHAP values and model explanations in detail through interactive components like dropdown menus, sliders, and graphs. This separation of visualization logic in a Flask component, communicating with the Django backend via APIs, provided flexibility and optimized performance for rendering complex visualizations.

The iterative nature of the UCD methodology was reflected in the development process through the creation of VisAutoML 1.0 and its subsequent redesign into VisAutoML 2.0. The development of VisAutoML 2.0 was directly informed by the evaluation findings and identified areas for improvement from VisAutoML 1.0. This iterative development cycle ensured that the refined prototype incorporated enhancements aimed at improving usability, transparency, and user experience based on empirical evidence from the target user group. The design principles guided the implementation throughout these development stages, ensuring that the technical realization aligned with the user-centered goals.

3.8 Summary

This chapter has presented the comprehensive methodology employed to address the research questions and achieve the objectives of developing and evaluating VisAutoML, an Automated Machine

Learning (AutoML) tool designed for non-expert users. The research was fundamentally guided by an iterative user-centred design (UCD) framework, structured into distinct stages of design research, design and prototyping, and prototype evaluation (Figure 15). This UCD approach was deemed essential to ensure that the resulting system is genuinely usable, transparent, and aligned with the specific needs and capabilities of the target non-expert audience (Norman, 2013).

The methodology commenced with a thorough Design Research phase (Stage 1), focused on investigating and gathering user requirements. This involved a mixed-methods approach, including an Extended Technology Acceptance Model (TAM) study theoretically grounded in established TAM principles (Davis, 1989; Venkatesh & Davis, 2000). The TAM study utilized both quantitative data from an online questionnaire and qualitative insights from semi-structured interviews to understand factors influencing technology acceptance and to explore non-expert perceptions and requirements. Complementary activities included a literature review of existing tools and user challenges, user surveys, and the creation of user personas, collectively providing a rich foundation of user-centered requirements.

The research then proceeded through iterative Design and Prototyping stages (Stages 2 and 4), where the VisAutoML prototypes (VisAutoML 1.0 and 2.0) were developed. The system development methodology adopted aspects of the waterfall model within these iterative cycles, moving from defined requirements and design principles to implementation. The implementation leveraged established web development frameworks such as React JS and Django, alongside key Machine Learning and visualization libraries including Pandas, Scikit-learn, Flask, Plotly, and Shap, to build a functional and interactive web application capable of performing AutoML tasks and generating Explainable Artificial Intelligence (XAI) visualizations.

Integral to the UCD framework were the Prototype Evaluation stages (Stages 3 and 5), which involved empirical studies to assess the usability and transparency of the developed prototypes. These included a comparison study between VisAutoML 1.0 and an existing AutoML tool (H2O AutoML) to benchmark the initial prototype. Subsequent in-depth evaluations of both VisAutoML 1.0 and the refined VisAutoML 2.0 were conducted. These evaluations employed mixed-methods approaches, collecting quantitative data through standardized questionnaires (e.g., UEQ, SUS, Trust Questionnaire, XAI Questionnaire) and objective measures like task completion time, alongside qualitative data from open-ended questions and interviews (Brooke, 1996; Schrepp et al., 2017a; Hoffman et al., 2023; Silva et al., 2023).

The data collected across all evaluation studies were subjected to rigorous quantitative and qualitative analysis techniques. Quantitative analysis included descriptive statistics, reliability analysis (e.g., Cronbach's alpha), correlation analysis (e.g., Pearson), and comparative statistical tests (e.g., t-tests, ANOVA) where applicable (Yan & Yibing, 2010; Field, 2013). Qualitative data from interviews and open-ended questions were analyzed using thematic analysis to identify key user perceptions, challenges, and suggestions. Validity assessments were conducted to ensure the trustworthiness of the qualitative findings (Petersen & Gencel, 2013).

In summary, the methodologies employed in this chapter, guided by the iterative UCD framework, provided a systematic and user-centered approach to developing and evaluating VisAutoML. The combination of requirements gathering, iterative development, and rigorous mixed-methods evaluation, with specific studies focusing on TAM-based acceptance factors, comparative performance, and in-depth usability and transparency assessments, has generated the necessary empirical data. This comprehensive methodological foundation provides the basis for the results

presented in Chapter 5, which will detail the findings derived from these studies and their implications for addressing the challenges of making AutoML usable and transparent for non-expert users.

4 System Development

4.1 VisAutoML 1.0

This section details the initial phase of system development, focusing on the design and implementation of the VisAutoML 1.0 prototype. Following the comprehensive user requirements analysis conducted in the preceding stage, which illuminated the significant challenges non-expert users face with existing Machine Learning (ML) tools, VisAutoML 1.0 was conceptualized and built as a foundational iteration of the proposed Automated Machine Learning (AutoML) tool. Its development was a direct response to the identified need for a system that enhances both usability and transparency for this specific user group. The primary purpose of developing VisAutoML 1.0 extended beyond simply creating a functional tool; it critically served as a tangible artifact for empirical evaluation, specifically designed to validate the user requirements and design principles derived from the foundational research in a practical setting.

A key objective for developing VisAutoML 1.0 was its intended role as a comparison tool. As rigorously outlined in the research methodology, a significant component of the evaluation process involved a comparative study between the VisAutoML 1.0 prototype and an existing, recognized AutoML tool deemed suitable for non-experts, namely H2O AutoML. The rationale for this comparison stemmed from the critical analysis of existing AutoML tools, which revealed persistent limitations in terms of usability, transparency, and integrated support mechanisms for non-expert users (Table 6). Tools like H2O AutoML, while powerful in automation, often present complex interfaces and generate outputs that can be difficult for non-experts to fully understand or trust (Coors et al., 2021; Rabhi et al., 2021). Therefore, the development of VisAutoML 1.0 was specifically geared towards creating a system that implemented the core functionalities of regression and classification on tabular data, comparable to existing tools, but crucially incorporated the novel user-centred design principles and integrated XAI features identified as essential for improving the non-expert experience. This deliberate design allowed for a direct, empirical benchmarking to provide quantitative and qualitative data on VisAutoML 1.0's relative performance, usability, transparency, and effectiveness in enhancing user understanding when juxtaposed with an established solution. The comparison study was thus critical for empirically validating whether the design choices made based on the foundational research translated into a perceptibly better and more understandable user experience for the target audience compared to the status quo represented by existing tools.

Furthermore, VisAutoML 1.0 served as a crucial prototype for model validation within the iterative User-Centred Design (UCD) framework. As the first tangible realization of the theoretical requirements and synthesized design principles, its evaluation was paramount for testing their practical effectiveness. The evaluation methodology for VisAutoML 1.0 was specifically designed to gather detailed feedback from target users, utilizing instruments such as the User Experience Questionnaire (UEQ-S) to assess usability dimensions like pragmatic and hedonic quality, the Trust Questionnaire to gauge user confidence, and the XAI Questionnaire to evaluate perceived explainability This rigorous evaluation process was essential to validate whether the implemented features and interface elements effectively addressed the identified challenges of usability and transparency for non-experts in practice. The findings from this validation process, provided critical empirical evidence on the prototype's strengths and weaknesses from the user's perspective, such as task completion efficiency, perceived interface quality, trust levels in automated outputs, and the comprehensibility of XAI visualisations. This critical evaluation served as the empirical basis for the subsequent redesign phase, directly informing the objectives and implementation of VisAutoML 2.0 by highlighting specific areas requiring improvement based on user feedback and performance data. Thus, the development of VisAutoML 1.0 was a deliberate and essential step in the UCD cycle, designed explicitly to be evaluated

both as a comparison tool against existing solutions and validated as a prototype to gather the necessary empirical data for iterative refinement and the development of a more effective tool in subsequent versions.

4.1.1 Foundational Research for VisAutoML 1.0 Design

The design and development of VisAutoML 1.0 were fundamentally guided by a comprehensive and multi-faceted user requirements analysis. Recognizing the specific challenges faced by non-expert users in navigating the complexities of machine learning (ML) workflows (Bove et al., 2022; Yang et al., 2018), this foundational research phase aimed to identify the core needs, expectations, and pain points of the target audience. This systematic inquiry was critical to ensuring that the resulting prototype would not merely automate ML tasks but would do so in a manner that enhanced usability and transparency, thereby lowering the barrier to entry for individuals without extensive technical expertise. The insights gleaned from this initial stage were instrumental in defining the functional and non-functional requirements of the system and in shaping the underlying design philosophy that prioritised a user-centred approach throughout the development lifecycle (Huang & Chiu, 2016; Williams, 1986).

The requirements analysis process encompassed a variety of interconnected methods. An extensive review of existing literature provided a theoretical grounding, highlighting established design principles for user-friendly interfaces and identifying the limitations of current AutoML tools, particularly concerning transparency and interpretability for non-experts (Alicioglu & Sun, 2021; Hohman et al., 2019; Wang et al., 2019). Concurrently, direct engagement with potential users through surveys and semi-structured interviews offered invaluable qualitative and quantitative data on their perceptions, challenges, and preferences when interacting with ML concepts and tools (Yang et al., 2018). This empirical data was further enriched by an analysis of existing AutoML platforms, which served as benchmarks to understand current capabilities and identify opportunities for innovation (Elshawi et al., 2019; Zöller & Huber, 2021). The application of an extended Technology Acceptance Model (TAM) provided a structured framework for investigating the factors influencing user acceptance and behavioural intention towards adopting such a tool, revealing the significant interplay between perceived usefulness, ease of use, enjoyment, and authority (Davis, 1989; Venkatesh & Davis, 2000).

Collectively, the findings from these research strands directly informed the synthesis of a set of design principles tailored for non-expert AutoML systems, emphasizing aspects such as visual guidance, scaffolding, and accessible explanations (Chromik & Butz, 2021; Sharma & Hannafin, 2007). These principles were then translated into the concrete design of the VisAutoML 1.0 prototype. The insights into user challenges, such as difficulty with terminology and reliance on external documentation (Yang et al., 2018), motivated the inclusion of features like guided workflows, embedded contextual help, and simplified visualisations. The understanding of user motivations and preferred interaction styles, gained through interviews and surveys, influenced the development of an intuitive interface and the initial implementation of Explainable AI (XAI) components aimed at demystifying model outputs (Ribeiro et al., 2016). Furthermore, the creation of user personas served as a constant reference point, ensuring that design decisions remained aligned with the diverse backgrounds and needs of the intended user base. The subsequent subsections elaborate on each of these foundational research components and illustrate their specific impact on the design and features of the VisAutoML 1.0 prototype.

4.1.1.1 Influence of Literature Review on Prototype Features

The initial phase of the user requirements analysis for VisAutoML 1.0 was significantly shaped by a comprehensive review of existing literature, as detailed in Chapter 2. This review provided a crucial theoretical foundation by exploring the landscapes of Automated Machine Learning (AutoML), visual analytics, Explainable Artificial Intelligence (XAI), and the documented challenges faced by non-expert users within these domains (Alicioglu & Sun, 2021; Hohman et al., 2019; Wang et al., 2019; Yang et al., 2018). A key insight derived from this review was the prevalent "black-box" nature of many existing AutoML systems, which, despite their automation capabilities, often lack the transparency necessary for users to understand the underlying processes and build trust in the model outputs (Kaur et al., 2020; Khuat et al., 2022; Xin et al., 2021). This identified gap, particularly the lack of transparency in existing AutoML solutions, directly motivated the integration of XAI functionalities into the VisAutoML 1.0 prototype. The aim was to move beyond simple automation by including components designed to demystify the ML process and its outputs for non-experts (Spinner et al., 2019).

Furthermore, the literature review underscored the critical role of effective visual analytics in making complex data and model information comprehensible to a broader audience (Keim et al., 2008; Yi et al., 2007). Research indicated that while visualisations are employed in some ML tools, they are often not specifically tailored for non-experts and may assume a level of technical literacy that the target user group lacks (Hohman et al., 2019; Wang et al., 2021). This understanding was pivotal in the design of the visual interface and XAI visualisations for VisAutoML 1.0, leading to an emphasis on clarity, simplicity, and the strategic use of progressive disclosure (DP5 in Table 23) to prevent user cognitive overload. The review also highlighted the specific challenges non-experts encounter throughout the ML development pipeline, from initial data preparation to model evaluation (Yang et al., 2018), which are comprehensively summarized in Table 5. These challenges, such as difficulties with data analysis prior to modelling (C1) and the misinterpretation of model performance metrics (C3), directly informed the design of guided workflows (DP1 in Table 23) and the inclusion of simplified explanations for model evaluation metrics within the prototype interface.

The analysis of existing AutoML tools revealed that while some platforms offer graphical user interfaces, they often require a baseline technical knowledge or lack sufficient scaffolding mechanisms tailored for true beginners (Auto-sklearn, H2O AutoML, TPOT). This comparative understanding, highlighting the usability limitations of current tools for non-experts (Table 6), reinforced the necessity for VisAutoML 1.0 to incorporate robust scaffolding mechanisms (Sharma & Hannafin, 2007). This included designing features that provide step-by-step guidance (DP2 in Table 23), embedded contextual help (DP3 in Table 23), and visible cues (DP4 in Table 23) integrated throughout the user journey, particularly in critical stages such as data import, review, and model development. Consequently, the literature review served as a foundational blueprint, systematically identifying critical user needs and existing technological limitations that VisAutoML 1.0 was designed to address through its user-centred design and integrated explainability features.

4.1.1.2 Prototype Design Informed by Understanding Non-Experts

A deep understanding of the target user group, non-experts in Machine Learning (ML), was paramount in shaping the design of the VisAutoML 1.0 prototype. As elaborated in Section 2.3, non-experts are characterized by limited prior knowledge of ML algorithms and processes, often perceiving them as "black-box" input-output mechanisms (Yang et al., 2018). This fundamental characteristic directly influenced the design decision to prioritize an intuitive and simplified interface, minimizing the need

for users to engage with complex technical details or code. The prototype's drag-and-drop interface for model development, for instance, was a direct response to the need to abstract away the complexities of traditional ML programming, making the process more approachable for beginners.

Moreover, the research highlighted that non-experts often do not habitually perform crucial data analysis steps, such as using visualisations or descriptive statistics, prior to model building or for model introspection (Yang et al., 2018). This challenge (C1 in Table 5) informed the design of the Dataset Review page, which automatically calculates and displays key dataset details and provides a dropdown for visualising data distribution. This feature was included to encourage users to explore their data in a simplified, visually-driven manner, addressing a key gap in their typical workflow. The prototype also aimed to counter the reliance on external documentation (Yang et al., 2018) by embedding visual cues, hover texts, and pop-ups with relevant information throughout the system providing just-in-time support directly within the interface.

The challenges non-experts face in interpreting model performance metrics beyond simple accuracy (C3 in Table 5) and understanding how to improve model performance (C5 in Table 3) were also critical considerations. The Model Review page of the VisAutoML 1.0 prototype was designed to present model evaluation metrics and XAI visualisations in a more accessible format, although the initial version's explainability was later identified as an area for improvement. The design incorporated different tabs for impact, impact relationship, and model metrics, aiming to break down the information logically. The inclusion of feature importance visualisations and components like the "what-if" input were intended to provide users with insights into model decisions in a more understandable way, directly influenced by the need to demystify the "black-box" nature of ML for this audience. Thus, the comprehensive understanding of non-expert characteristics, motivations, and challenges, as detailed in the foundational research, profoundly shaped the user-centred design and feature set of the VisAutoML 1.0 prototype.

4.1.1.3 Findings from Existing AutoML Tools Applied to Prototype

The analysis of existing Automated Machine Learning (AutoML) tools and summarized in Table 6, provided crucial empirical context and informed the design strategy for the VisAutoML 1.0 prototype. This examination highlighted that while current AutoML platforms offer increasing levels of automation across the ML pipeline (Santu et al., 2022, Figure 14, Table 4), they often fall short in providing truly usable and transparent experiences for non-expert users. Specific limitations identified included the assumption of prior technical knowledge, the generation of non-interpretable "black-box" models, and insufficient integrated support mechanisms.

The design of VisAutoML 1.0 directly aimed to address these shortcomings. Recognizing that many existing tools require users to understand fundamental ML concepts and terminology (Alicioglu & Sun, 2021; Liao et al., 2020), the prototype was designed with an emphasis on simplifying the user interface and workflow. The sequential navigation menu, for instance, was a deliberate design choice to guide users through the ML process step-by-step, reducing the cognitive load associated with complex, non-linear interfaces found in some existing tools. Furthermore, the analysis revealed that existing general-purpose AutoML platforms often generate models that lack transparency and readily understandable explanations (Coors et al., 2021; Rabhi et al., 2021). This limitation was a primary driver for integrating

XAI visualisations directly into the VisAutoML 1.0 workflow, aiming to provide users with insights into model decisions that were often absent or difficult to access in other tools.

The lack of sufficient learning scaffolding and contextual guidance in existing tools (Yang et al., 2018; Gil et al., 2019) was another key learning applied to the prototype design. While existing tools might automate steps, they often do not explain why certain automated choices were made or provide guidance on how to interpret intermediate results. VisAutoML 1.0 incorporated elements like hover texts and pop-ups to provide contextual information, aiming to offer a more supportive learning environment than typically found in other platforms. The comparison study with H2O AutoML in the evaluation phase was specifically designed to empirically validate whether VisAutoML 1.0's design, informed by these identified limitations of existing tools, resulted in a more usable and understandable experience for non-experts. Thus, the critical analysis of existing AutoML tools served as a vital source of empirical evidence, highlighting the specific gaps that the VisAutoML 1.0 prototype was explicitly designed to bridge.

4.1.1.4 Synthesis of Design Principles

The comprehensive user requirements analysis, encompassing the literature review, understanding of non-expert characteristics, and analysis of existing tools, culminated in the synthesis of a set of design principles specifically tailored for developing usable and transparent AutoML systems for non-experts. These principles, served as the guiding framework for the subsequent wireframing and prototype design of VisAutoML 1.0. Each principle was deliberately translated into specific design elements and functionalities within the prototype's interface.

For instance, the principle of Visualising activity and sequences (DP1 in Table 23) directly influenced the sequential navigation menu present on every page of the VisAutoML 1.0 prototype. This visual pipeline aimed to outline the procedural steps involved in the AutoML process, addressing the non-expert challenge of understanding the overall workflow. The principle of Demonstrate scaffold function (DP2 in Table 23) was reflected in the inclusion of tutorial videos accessible from the home page and the intended functionality of pop-ups to explain specific features. These elements were designed to present the utility and steps involved in using the tool's features through demonstration.

The principle of Embedded contextually relevant scaffold (DP3 in Table 23) guided the integration of informational icons and intended pop-ups throughout the interface, particularly on pages like Dataset Review and Model Development. These elements were designed to provide users with contextually relevant information and guidance directly within the workflow. Furthermore, the principle of Visible and utilised scaffold (DP4 in Table 23) emphasized ensuring that these support mechanisms were clearly visible and their functionality explained, reinforcing their intended use by non-expert users.

For the XAI component, the principle of Progressive explanation disclosure (DP5 in Table 23) influenced the design of interactive visualisations on the Model Review page. Features like the "what-if" input component and the dynamic updating of contributions tables and plots based on user input were designed to provide finer granularity of explanations through interaction. The principle of Natural language rationale (DP6 in Table 23) was intended to be implemented through supplementary textual explanations accompanying visualisations, such as the description above the column impact visualisation. Finally, the principle of Multiple ways to communicate an explanation (DP7 in Table 23)

informed the inclusion of various XAI visualisations and components within different tabs on the Model Review page, offering diverse perspectives on model behavior and predictions. Thus, the synthesized design principles served as a critical bridge between the user requirements analysis and the concrete design and implementation of the VisAutoML 1.0 prototype.

4.1.1.5 Impact of Extended TAM Model Study

The Extended Technology Acceptance Model (TAM) study, provided a structured framework for understanding the psychological and social factors influencing non-expert users' potential acceptance of an AutoML tool. The findings from this mixed-methods study, encompassing quantitative analysis of questionnaire data and qualitative insights from interviews and open-ended questions, had a direct impact on shaping the features and design priorities of the VisAutoML 1.0 prototype. The study revealed that Perceived Usefulness (PU), Perceived Ease of Use (PEOU), and Perceived Enjoyment (ENJ) were significant predictors of users' attitude towards and intention to use the system.

Specifically, the strong influence of Perceived Usefulness (PU) on attitude and intention underscored the critical need for VisAutoML 1.0 to demonstrate tangible benefits to users. This finding, consistent with prior TAM research (Bakhit Jaafreh, 2018; Lee et al., 2003; Patil, 2017), motivated the inclusion of features that highlight the utility of the tool for tasks relevant to non-experts, such as predicting trends or segmenting data. The emphasis on presenting clear model evaluation metrics and XAI visualisations was partly driven by the need to demonstrate the model's performance and the insights it could provide, thereby enhancing the perceived usefulness.

The significant influence of Perceived Ease of Use (PEOU) reinforced the importance of designing an intuitive and easy-to-navigate interface. The drag-and-drop functionality for model development and the sequential workflow were direct responses to this requirement for a system that is perceived as easy to learn and use. Furthermore, the study highlighted that PEOU was influenced by Perceived Authority (PA), suggesting that perceived social influence or credibility could impact how easy users found the system to use. This indirectly supported the need for a polished and professional interface design in VisAutoML 1.0 to enhance perceived credibility.

The inclusion of Perceived Enjoyment (ENJ) as a significant factor influencing both PU and attitude highlighted the importance of making the ML development process engaging and enjoyable for non-experts. This finding, consistent with work by Yang et al. (2018), motivated the incorporation of interactive elements, particularly within the XAI visualisations, to make the exploration of model results a more positive experience. The suggested features based on the correlated constructs from the TAM study, outlined in Table 15, such as built-in data exploration and visualisation tools and customisation options, directly informed the feature set included or planned for VisAutoML 1.0, demonstrating the tangible impact of the Extended TAM model study on the prototype's design.

4.1.2.6 Survey and Interview Findings

The qualitative and quantitative data collected through user surveys and semi-structured interviews provided rich, direct insights into the needs, preferences, and challenges of the target non-expert users, profoundly influencing the design of the VisAutoML 1.0 prototype. The survey results, for instance, strongly emphasized that a "Good user interface" was the top priority for a user-friendly system, followed by "No programming required". This directly supported the design focus on creating

a visually appealing and intuitive graphical interface with minimal need for coding, as manifested in the drag-and-drop features and simplified workflow of the prototype.

Interview findings further elaborated on what constituted a user-friendly system, with participants highlighting the importance of "clear design and intuitive user experience". This qualitative feedback reinforced the need for the streamlined layout and navigation implemented in VisAutoML 1.0. The surveys also revealed user preferences for learning how to use the system, with a strong inclination towards "step-by-step tutorial in the system" and "video tutorials". This directly informed the inclusion of video tutorials on the home page and the planned embedding of visual cues and hover texts throughout the prototype to provide integrated guidance.

Insights into the types of data non-experts would track (primarily financial-related data and the current limitations they experienced with data recording tools (often related to poor system design, provided valuable context for the design of the data handling and visualization features. While not fully realized in the initial 1.0 prototype, these findings suggested the importance of robust data import and review capabilities and the need for clear visualisations of relevant data types. Interviewees' perspectives on the perceived usefulness of visualisations, including the desire for aesthetically pleasing and semi-automated visualisations accessible on both desktop and mobile platforms, also influenced the design of the XAI visualisations in the Model Review page, emphasizing interactivity and clarity. The suggested features based on survey findings (Table 21) and interview findings (Table 22), such as intuitive UI design, interactive tutorials, visual explanations, and built-in data exploration tools, were directly considered and, where feasible, incorporated into the VisAutoML 1.0 prototype, demonstrating the tangible impact of user feedback on the design.

4.1.2.7 User Persona-Driven Findings

The creation and analysis of detailed user personas, played a pivotal role in grounding the design of the VisAutoML 1.0 prototype in the specific needs and characteristics of the target non-expert audience. Personas, such as Emily Johnson (marketing professional), Jonathan (business student), and Kimberly Tanaka (sociology doctoral student), served as tangible representations of the diverse backgrounds, goals, challenges, and preferences identified through the preceding requirements gathering activities (Hudson, 2013). By focusing on these archetypes, the design process could move beyond abstract requirements to consider how real users with specific motivations and limitations would interact with the tool.

The challenges faced by these personas, such as a lack of technical knowledge in AI, difficulties with complex terminology and syllabi, and grappling with the intricacies of data analysis and model interpretation, directly informed the design emphasis on simplicity and accessibility in the VisAutoML 1.0 prototype. The streamlined workflow, sequential navigation menu, and drag-and-drop interface for model building were designed with these challenges in mind, aiming to minimize cognitive load and make the ML process less intimidating for users like Emily, Jonathan, and Kimberly.

Furthermore, the diverse goals and motivations of the personas, ranging from gaining insights into consumer behaviour and automating manual processes to conducting academic research, highlighted the need for a tool that could support various applications. This influenced the core functionality of VisAutoML 1.0 to handle common ML tasks like regression and classification on tabular data, relevant to the potential use cases of the personas. The design of the data import and review pages was informed by the personas' need to work with their own datasets and understand their structure and

quality, while the model review pages and XAI visualisations were designed to help them interpret results and gain the insights necessary to achieve their goals.

The preferences expressed by the personas, such as a desire for user-friendly interfaces, clear learning paths, visually engaging explanations, customisable templates, and integration with familiar systems, also directly shaped the prototype's features. The inclusion of tutorial videos, pre-set datasets, and interactive XAI visualisations in VisAutoML 1.0 can be seen as responses to these stated preferences for guided learning and accessible, engaging interactions. While the initial prototype may not have fully realized all aspects of these preferences (leading to subsequent iterations), the personas provided a crucial user-centred compass, ensuring that the design decisions for VisAutoML 1.0 were consistently aligned with the needs and expectations of its intended non-expert users.

4.1.2 Design Principles and Application

As highlighted in section 2.13, technology-enhanced scaffolding emerges as a crucial facet, providing support for the learning of ML concepts within the system. Suwastini et al.'s taxonomy of technology-enhanced scaffolding—procedural, conceptual, metacognitive, and strategic—serves as a guide in integrating scaffolding mechanisms into the system. The application of these scaffolding types addresses varying aspects of the learning environment, from procedural orientation to metacognitive processes, ensuring a comprehensive support system for users.

Furthermore, insights from Chromik et al. (2021), Mohseni et al. (2021), Gomez et al. (2021), and Wang et al. (2019) have significantly influenced the XAI interaction concepts and design principles incorporated into the system. The work spans from proposing evaluation frameworks and stakeholder-centric design considerations to establishing links between human reasoning processes and XAI approaches. These contributions serve as a robust foundation for creating an interface that addresses the challenges faced by non-experts and facilitates accessible ML model development.

Design principles and a wireframe prototype are presented to illustrate the intended system. For this project, the design principles are extracted from five design studies (Chromik & Butz, 2021; Gil et al., 2019; Mohseni et al., 2021; Sharma & Hannafin, 2007; D. Wang, Yang, et al., 2019). The relevant and similar design principles were merged and those that were not related were removed. Table 23 summarises the design principles for the proposed system.

Table 14 Design Principles for proposed tool

Category	Design principle	Note	Reference
Overall system	**DP1**. Visualise activity and sequences	Outline the procedural and metacognitive processes required to facilitate an ML process. (e.g., Specify the tasks involved, and clarify the ML development pipeline)	(Quintana, Zhang, et al., 2018; Wu et al., 2021)
	DP2. Demonstrate scaffold function	Present the utility and steps involved throughout the ML process. (e.g., Example-based development, online demonstration)	(Puntambekar, 2022; Quintana, Zhang, et al.,

			2018; Saye & Brush, 2002)
	DP3. Embedded contextually relevant scaffold	Integrate resources based on a conceptual framework into the system to facilitate further learner inquiry. (e.g., Hyperlinks specific to an ML task)	(Akotuko et al., 2021; Puntambekar, 2022; Saye & Brush, 2002)
	DP4. Visible and utilised scaffold	Ensure scaffolds are visible and explicitly clarified to learners to promote appropriate usage. (e.g., Specify and explain the functionality of scaffolds within the system to ensure an effective understanding of the system)	(Quintana, Reiser, et al., 2018; Sarah, 2022)
XAI component	**DP5**. Progressive explanation disclosure	Provide finer granularity of an explanation through subsequent steps following an explanation interaction. (e.g., Visualising a specific feature after clicking on it, tooltips to display the factors of a feature)	(Buçinca et al., 2021; Khosravi et al., 2022; Millecamp et al., 2019)
	DP6. Natural language rationale	Complement visual explanations with textual explanations to facilitate better understanding (e.g., Natural language explanation of a feature importance chart)	(Ehsan et al., 2019, 2021; Wiegreffe & Marasovic, 2021)
	DP7. Multiple ways to communicate an explanation	Provides related explanations to triangulate insights and understand different angles of explanation. (e.g., Visualising global feature importance and local feature importance in relation to each other)	(Chou et al., 2022; Páez, 2019; Vilone & Longo, 2021)

4.1.3 System Requirements

The system requirements, which include software, functional, and non-functional components, were directly developed from the thorough analysis of user requirements. This methodical procedure

guarantees that the resulting system requirements are closely linked with the recognised needs and expectations of non-expert users in the field of machine learning. A comprehensive list of the minimum specifications was established within the domain of software requirements. This specification ensures that the system is capable of fulfilling the basic software requirements necessary for optimal performance.

The functional requirements were established to clearly express the precise functions and capabilities of the proposed system. These requirements act as a detailed plan for how the system will function, specifying the necessary features that directly address the user needs found in the user requirements analysis.

Concurrently, non-functional requirements were specified to define the properties and traits of the proposed system. These non-functional characteristics encompass factors such as performance, dependability, and usability, which stem from a comprehensive examination of user expectations and prioritise user-centric goals. The system requirements operate as an intermediary between the user requirements and the practical implementation of *the VisAutoML* tool.

The following are the software requirements of the system:

1. Internet connection
2. JavaScript-enabled web browser

The following are the development tools used:

1. Visual Studio Code
2. Django
3. React JS
4. Flask
5. Plotly
6. NPM
7. Sklearn

4.1.3.1 Functional Requirements Specifications

The functional requirements are requirements that define the function or behaviour of the proposed system. Table 24 displays the functional requirements of the proposed system.

Table 15 Functional requirements

No.	Description
FR1	The system will display a home page
FR2	The system will display a list of projects with details such as the learning task, the model used, and the overall score
FR3	The system will allow users to create a new project based on a classification or regression task
FR4	The system allows users to name, open, and delete existing projects
FR5	The system has two video tutorials for regression and classification use cases
FR6	The system will embed visual cues like hover texts throughout the system
FR7	The system has a navigation menu consisting of five pages which are home, import dataset, dataset review, model development, and model review pages
FR8	The system allows users to import a dataset (.csv file)
FR9	The system allows users to choose a pre-set dataset (.csv file)
FR10	The system allows users to return to a previous page using a back button on every page except for the home page
FR11	The system allows users to navigate to the next page using the next button on every page except for the model review

FR12	The system calculates and displays the details of the dataset like total rows, total columns, empty rows, empty columns, data type, percent empty and if a column is fit for use
FR13	The system allows users to specify a column to visualise the data distribution
FR14	The system displays a histogram distribution visualisation of a chosen column within a dataset
FR15	The system allows users to develop an ML model using a drag-and-drop interface
FR16	The system allows users to choose between an automated or a manual model selection
FR17	The system accepts feature input based on four boards which are the prediction column, identifier column, columns not to use, and columns to use boards
FR18	The system allows only one column to be specified in the prediction column board
FR19	The system will not allow a prediction to be run without designating the prediction column
FR20	The system specifies the identifier, columns to use, and columns to not use boards to be optional boards to be filled
FR21	The system automatically updates the dropdown with either regression or classification models based on the user's choice
FR22	The system accepts unit input for regression learning task
FR23	The system accepts label input for classification learning task
FR24	The system automatically pre-processes the data in the background for model development
FR25	The system automatically presents XAI visualisations in three different tabs which are impact, impact relationship, and model metrics
FR26	The system displays feature importance, what if, contributions table, and contributions plot XAI visualisation on the impact tab
FR27	The system displays feature importance and impact relationship XAI visualisation on the impact relationship tab
FR28	The system displays model metrics, predicted vs actual, and plot vs feature XAI visualisation on the model metrics tab for regression
FR29	The system displays model metrics and confusion matrix XAI visualisation on the model metrics tab for classification
FR30	The system allows users to return to the home page at the end of the model review
FR31	The system allows users to rerun a model using updated changes

4.1.3.2 Non-Functional Requirements Specifications

The non-functional requirements are requirements that define the attributes of the proposed system such as the usability and reliability of the system. Table 25 displays the non-functional requirements of the IoT prototyping toolkit.

Table 16 Non-Functional requirements

No.	Description of Non-Functional Requirements
NFR1	**Product Requirement** • The system deploys onto a domain and requires an internet connection
NFR2	**Reliability** • The system must be able to read and write data from the database • The system shall run a Django and Flask server simultaneously
NFR3	**Usability** • The interface should be intuitive and usable for non-experts with the help of visual cues, tutorials, and instructions
NFR4	**Performance** • The system must be able to interactively respond to inputs • The system should support running machine learning algorithms
NFR5	**Extensibility** • The system shall support future extension with new functionalities
NFR6	**Legal**

• The system should not have any legal issues involving intellectual property infringement, user privacy violation, and any use of restricted technology

4.1.5 Use Case Diagram

The detailed relationship between the user and the system is clearly displayed through thoughtfully developed use case diagrams. The project clearly specifies the parties involved as the server and the user. The intricate relationship between different system functions and their corresponding users can be observed in the subsequent use case diagram. Table 26 provides a complete explanation of the specific links between actors and their corresponding use cases.

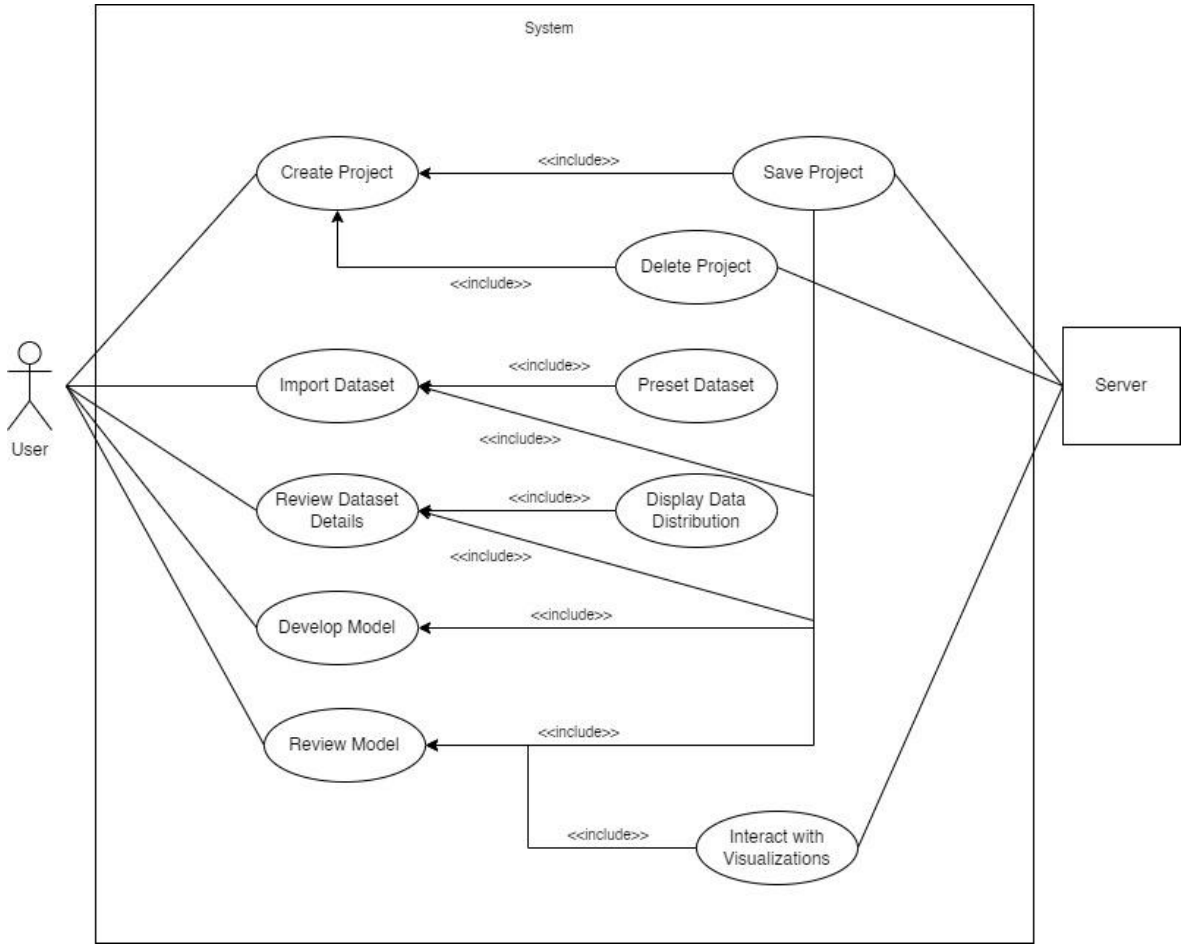


Figure 18 Use Case Diagram

Table 17 Use Case Description

No.	Use Case	Actor	Relationship	Description
UC1	Create Project	User	Association	The use case creates a new project
UC2	Save Project	User & Server	Association & Include UC1, UC4, UC6, UC8, UC9, UC10	Project is saved automatically throughout the system

UC3	Delete Project	User & Server	Association & Include UC1	Project is deleted on the main page
UC4	Import Dataset	User	Association	Users can import dataset
UC5	Pre-set Dataset	User	Association & Include UC4	Users can choose a pre-set dataset
UC6	Review Dataset	User	Association	Users can review the dataset chosen
UC7	Display Distribution	User	Association & Include UC6	Users can visualise the data distribution of the dataset
UC8	Develop Model	User	Association	Users can develop the model
UC9	Review Model	User	Association	Users can review the model through XAI
UC10	Interact with Visualisations	User & Server	Association & Include UC9	User can interact with the visualisations which are updated by the server

4.1.6 System Architecture

This section will elaborate on the architecture of the system. The system is a web application built on the Django framework which is based on a model-view-template (MVT) design pattern and uses a declarative approach to enable the development of complex web applications with minimal coding. The Django framework is responsible for handling data storage and management. The Django framework uses object-relational mapping (ORM) to connect to the SQLite database and exposes a set of RESTful APIs that the React JS frontend and the Flask framework use to retrieve and manipulate the data.

The React JS front is responsible for rendering the user interface of the web application and for handling the user interactions. The front-end APIs are provided by the Django framework to retrieve the data and uses React components and libraries, such as Redux and Redux-Saga, to manage the application state and handle asynchronous operations. The front end uses React Router to enable navigation between different pages of the web application and uses the CSS pre-processor, Sass, to define the styles of the UI elements.

The Flask framework is responsible for running the interactive visualisation of the data. The Flask framework is connected to the Django framework through the APIs and uses the Plotly visualisation library to generate the visualisations. The Flask framework also provides a web server that can serve the static assets of the visualisation, such as the JavaScript and CSS files, and that can handle the user interactions with the visualisation, such as hover events and data filtering. The system architecture is shown in Figure 23 below.

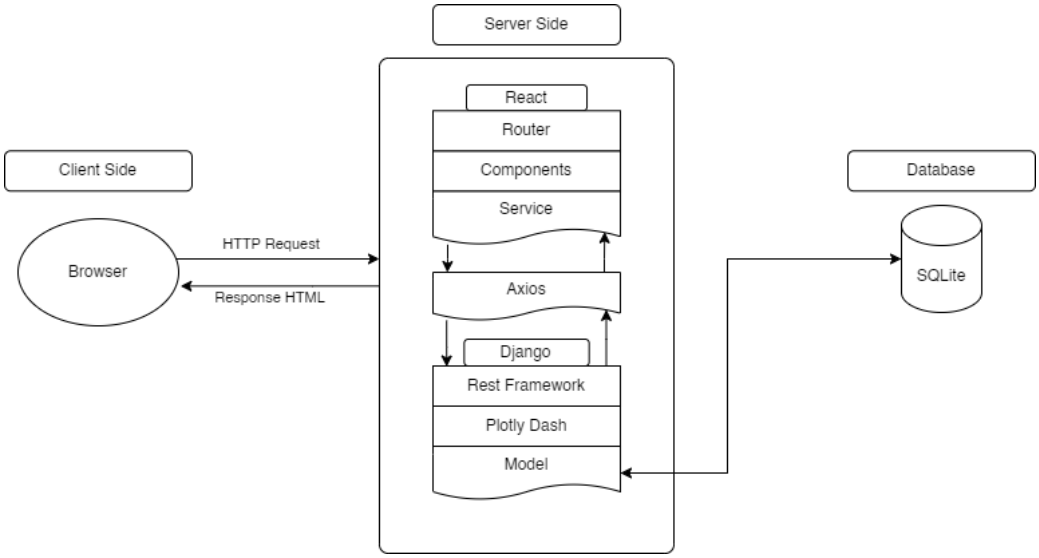


Figure 19 Proposed tool system architecture

4.1.7 Wireframing Process

The wireframing process in crafting *VisAutoML* involved a strategic approach guided by the core objectives of creating an intuitive, visually coherent interface aligned with the needs of non-expert users. Adobe XD served as the primary design tool for this crucial stage.

1. Defining the wireframing objectives was the initial step, grounded in user requirements, design principles, and insights from the user requirements analysis. This provided a clear focus on creating wireframes that reflected the envisioned system's architecture while prioritizing intuitiveness and accessibility.
2. Component identification and placement followed, involving the thorough mapping of essential UI elements, navigation menus, interactive buttons, and instructional prompts. The objective was to ensure an intuitive and visually cohesive layout guiding users through the ML development pipeline.
3. Iterative design refinements were implemented throughout the process, incorporating feedback loops and usability considerations. Collaborative features in Adobe XD facilitated seamless iterations, refining the wireframes based on stakeholder input and insights gathered from surveys and interviews.
4. Incorporating design principles was a crucial aspect, with wireframes reflecting principles such as visualising activity sequences, demonstrating scaffold functions, and embedding contextually relevant scaffolds. This integration ensured that the wireframes represented structure and encapsulated guiding principles enhancing user understanding.
5. Responsive design considerations were integrated to address the diversity in user devices and screen sizes. The wireframes were crafted to adapt seamlessly across desktops, tablets, and mobile devices, ensuring a consistent and user-friendly experience.

The wireframing process using Adobe XD was a dynamic journey, driven by user-centric design principles, a clear understanding of non-expert user requirements, and the overarching goals of creating an accessible and intuitive *VisAutoML* tool. The resulting wireframes set the visual foundation for the subsequent prototyping phase, encapsulating the essence of a system that aligns seamlessly with the needs and expectations of its intended audience. The diagram below displays the designed wireframes using Adobe XD.

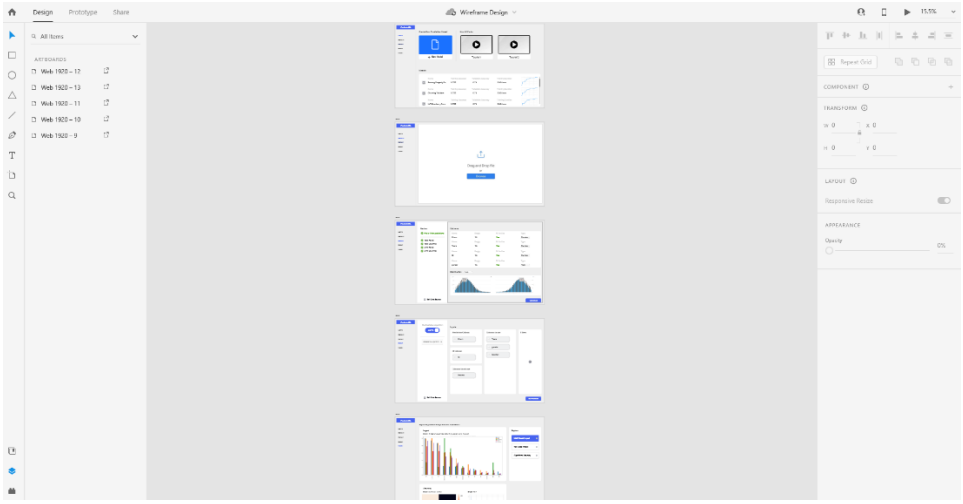


Figure 20 Wireframes designed using Adobe XD

4.1.8 Wireframe and Prototype Design

The user interface design is intended to be intuitive and easy to use for users to interact and develop a model with the system. Based on the design principles in 3.5, the first iteration of the system was developed to include five sections which are separated by page. The functionality of each page will be elaborated with the interfaces. The wireframes for this initial design were methodically crafted using Adobe XD. These wireframes served as the foundation for the development of the web-based prototype, ensuring that the envisioned design principles and functional elements were translated into an interactive and user-friendly interface. The first iteration of the system was developed to include five sections which are separated by page. The functionality of each page will be elaborated with the interfaces. The following are the titles of each page:

1) Home page
2) Dataset import page
3) Dataset review page
4) Model development page
5) Model explain page

The home page allows users to watch tutorial videos, create a new project, delete an existing project, and open an existing project. The navigation menu on the left side is present on every page to illustrate the steps involved in the AutoML pipeline (**DP1**). If the tutorial video buttons are clicked, a popup should appear with the video automatically played (**DP2**). The models section will display a table specifying the details of the model including the type of learning task, the name of the project, and the accuracy score.

Users can open and delete existing models by clicking on the corresponding buttons within the table. Clicking on the new model button will display a popup to input the name of the project and allow the user to choose between a regression or classification project. Once the create button is clicked on the popup, the user will be navigated to the dataset import page.

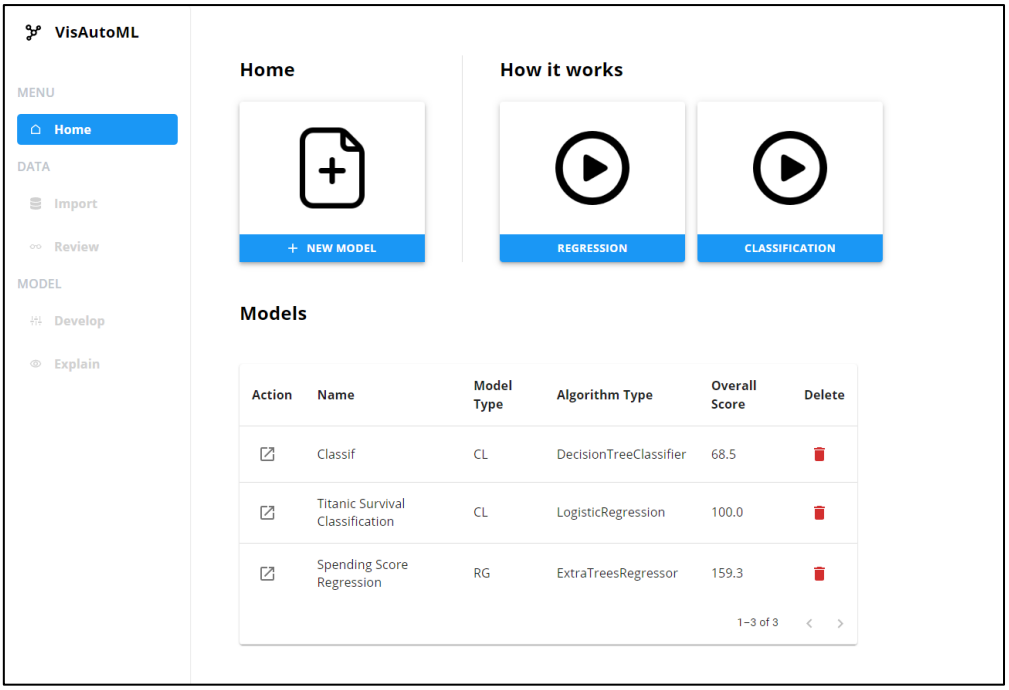


Figure 21 Home page prototype interface

The dataset import page allows users to specify the dataset to be used for the ML model development. A user can either import a .csv file or choose a sample dataset from the pre-set datasets by clicking the dropdown. If a user did not import or specify a pre-set dataset the next button to progress to the next step will not be clickable. The user can also return to the previous page by clicking on the back button and the model will not be saved in the database. Once a dataset has been chosen, the user will be navigated to the dataset review page.

Figure 22 Dataset import prototype interface

The dataset review page will display details of the dataset such as the number of rows, columns, unfit rows, and unfit columns. When a user is navigated to this page, a popup will be displayed to provide relevant information regarding data preparation in ML and explain the buttons and icons to explain a function or button (**DP3, DP4**). The dataset review page automatically checks whether

the dataset is fit for predictions based on certain criteria such as the total number of rows, r > (number of columns*30). Furthermore, it should present the data type of each column within a table and allow the user to add a description for each column. Next, a dropdown should be available for the user to visualise the distribution of the data within a specific column. A histogram of the data should illustrate the distribution of the data for all data types like string, integer, and float. The user also has the option to return to the previous page by clicking on the back button. Once the user is satisfied with the results and clicks on next, the user will be navigated to the dataset review page.

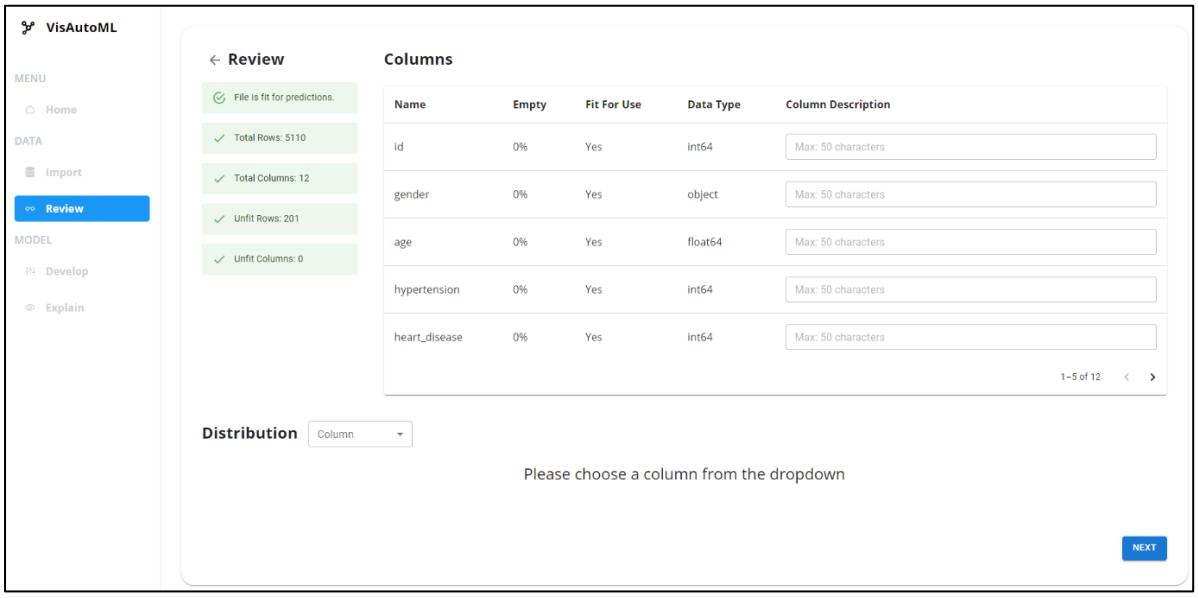


Figure 23 Dataset review prototype interface

The model development page will allow users to program the model intuitively. In this iteration, the interaction for programming the model utilises a drag-and-drop interface to cater to non-experts. When a user is navigated to this page, a popup will be displayed to provide relevant information regarding model development in ML and explain the prepared buttons and icons to explain a function or button within the system (**DP3)** Furthermore, the auto switch allows the automation of model selection and if the toggled manual, the users can explore and use other models within the dropdown. Next, the users can program the model using the cards which represent a column.

Each card can be dragged into a board like the prediction column, ID column, columns not to use, and columns to use boards. On the far right, based on the learning task specified in the beginning, the user must input the unit of the prediction column for a regression task or the labels of data within the prediction column for a classification task. On each board and input field there are information icons that allow users to hover over and read a specific component of the interface to guide and scaffold their development experience (**DP4**). The user also has the option to return to the previous page by clicking on the back button. Once the user fills the prediction column with a card and inputs the appropriate field, the next button is clickable and will navigate the user to the model explanation page.

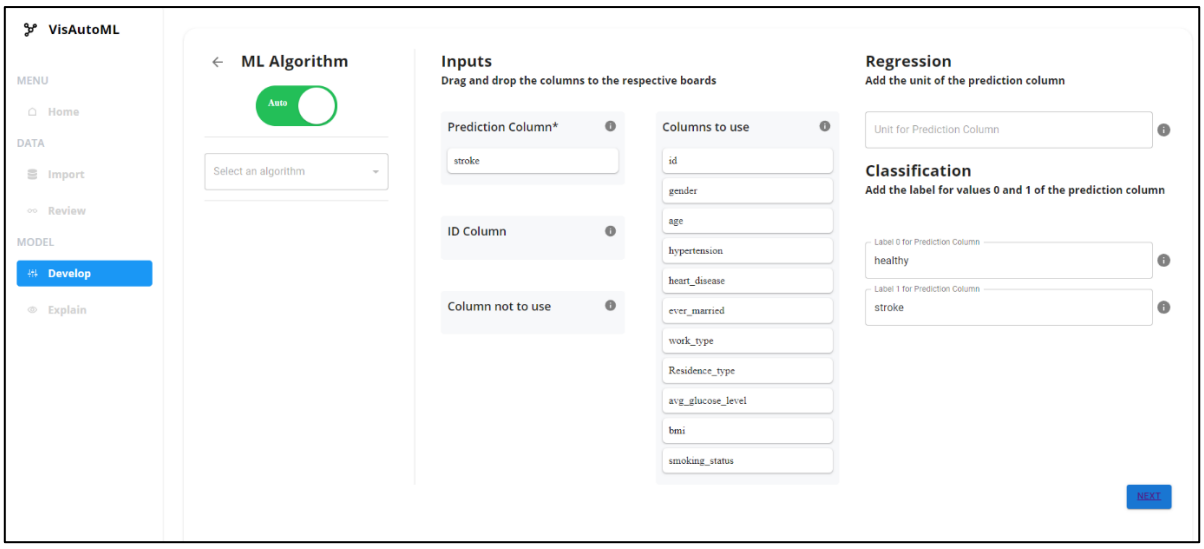


Figure 24 Model development prototype interface

The model review page will display interactive XAI visualisations and components of the model which are separated by three tabs dependent on the type of learning task (i.e., regression or classification). Each model review tab is accompanied with instructions at the top of each visualisation to assist users (**DP3, DP4**). If the learning task is regression, the tabs are: 1) impact; 2) impact relationship; 3) regression metrics. On the other hand, if the learning task is classification, the tabs are: 1) impact; 2) impact relationship; 3) classification metrics.

The XAI visualisations and components within the tab's impact and impact relationship for the regression and classification learning task are similar. The XAI visualisations and components for the impact tab are: 1) column impact; 2) what if input component; 3) contributions table; 4) contributions plot. The XAI visualisations and components for the impact relationship tab, on the other hand, are: 1) column impact; 2) impact relationship. However, the metrics tabs for each learning task provide different XAI visualisations and components. For instance, the regression metrics tab provides a model summary table with different sets of metrics specific to a regression learning task and two XAI visualisations which are: 1) predicted vs actual; 2) plot vs feature. The classification metrics tab, on the other hand, provides a model performance metrics table with metrics specific to a classification learning task and a single XAI visualisation which is a confusion matrix. Below is an overview of the impact tab.

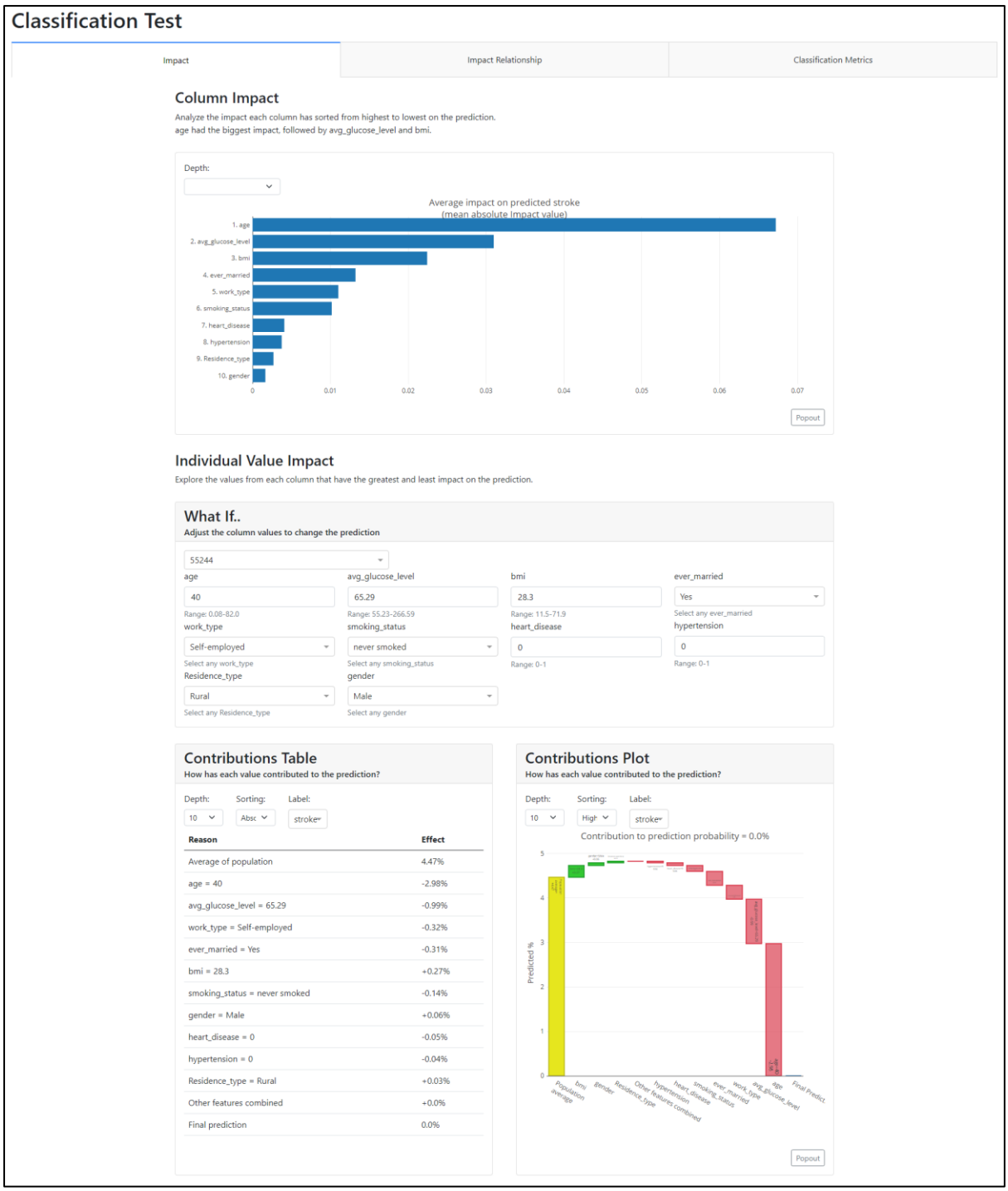


Figure 25 Model review prototype interface

The impact tab is divided into two sections which are the column impact and the individual value impact. The column impact visualisation is essentially a feature importance visualisation based on calculated Shap values. A user can interact with the visualisations as it is dynamically interactive. The column impact visualisation displays a horizontal bar chart of the columns based on their impact on the prediction. Above the visualisation, users are supplemented with textual explanations describing the visualisation explanation (**DP6**). A user can choose the depth of the columns by clicking on the dropdown and choosing a preferred depth. The user could also enlarge the chart using the "popout" button at the bottom of the visualisation. As the visualisations are built on Plotly, users can also zoom, select, pan, and download the visualisations as a .png image file.

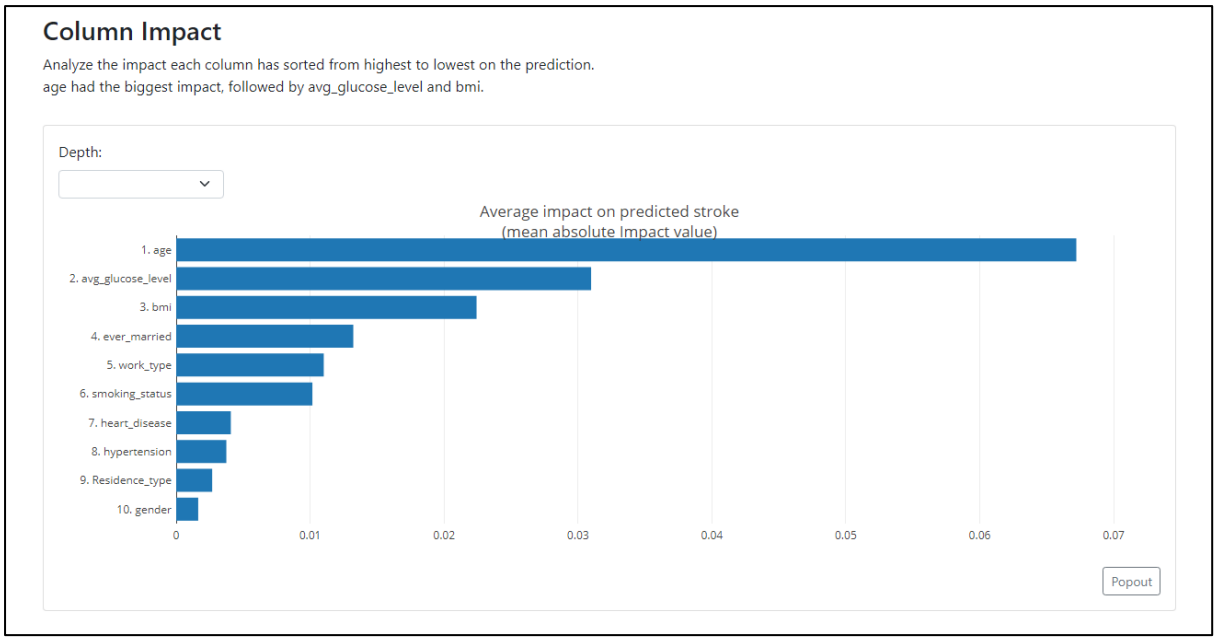


Figure 26 Model review prototype interface

The individual value impact section is a group of visualisation components that dynamically update based on input. The visualisation components are the what-if input component, contributions table, and the contributions plot. The section is displayed below in Figure 31. The what-if input component allows the user to input different indexes of the dataset which equates to a row of the dataset. Once an index is chosen, the input fields in the component dynamically update to the corresponding index's value for each column (**DP5**). Furthermore, the contributions table and contributions plot dynamically update to the chosen index's value. The input fields of the what if component also display the range of values for each column below each field and at the top of each input field is the name of the column. Users can manipulate the values of each column by typing in a value or clicking on the arrow buttons to increase or decrease the row's values. Any modification of the input field is dynamically updated and reflected on the contributions table and contributions plot (**DP7**).

For the contributions table, users are shown the effect percentage of each column within the chosen row. Additionally, users can choose the depth of the columns which means the number of columns to show (**DP7**). Users can also choose the sorting sequence by absolute, high to low, low to high, and importance. If the learning task is classification, users can choose to visualise the effect percentage of the positive or negative label (**DP7**).

The contributions plot is a vertical bar chart visualisation coloured according to the chosen label. The effect percentage for each column is visualised with the name of the column at the bottom of the chart. Similar to the contributions table, users can choose the depth of the columns which means the number of columns to show. Users can also choose the sorting sequence by absolute, high to low, low to high, and importance (**DP7**). If the learning task is classification, users can choose to visualise the effect percentage of the positive or negative label. The user could also enlarge the chart using the "popout" button at the bottom of the visualisation. As the visualisations are built on Plotly, users can also zoom, select, pan, and download the visualisations as a .png image file.

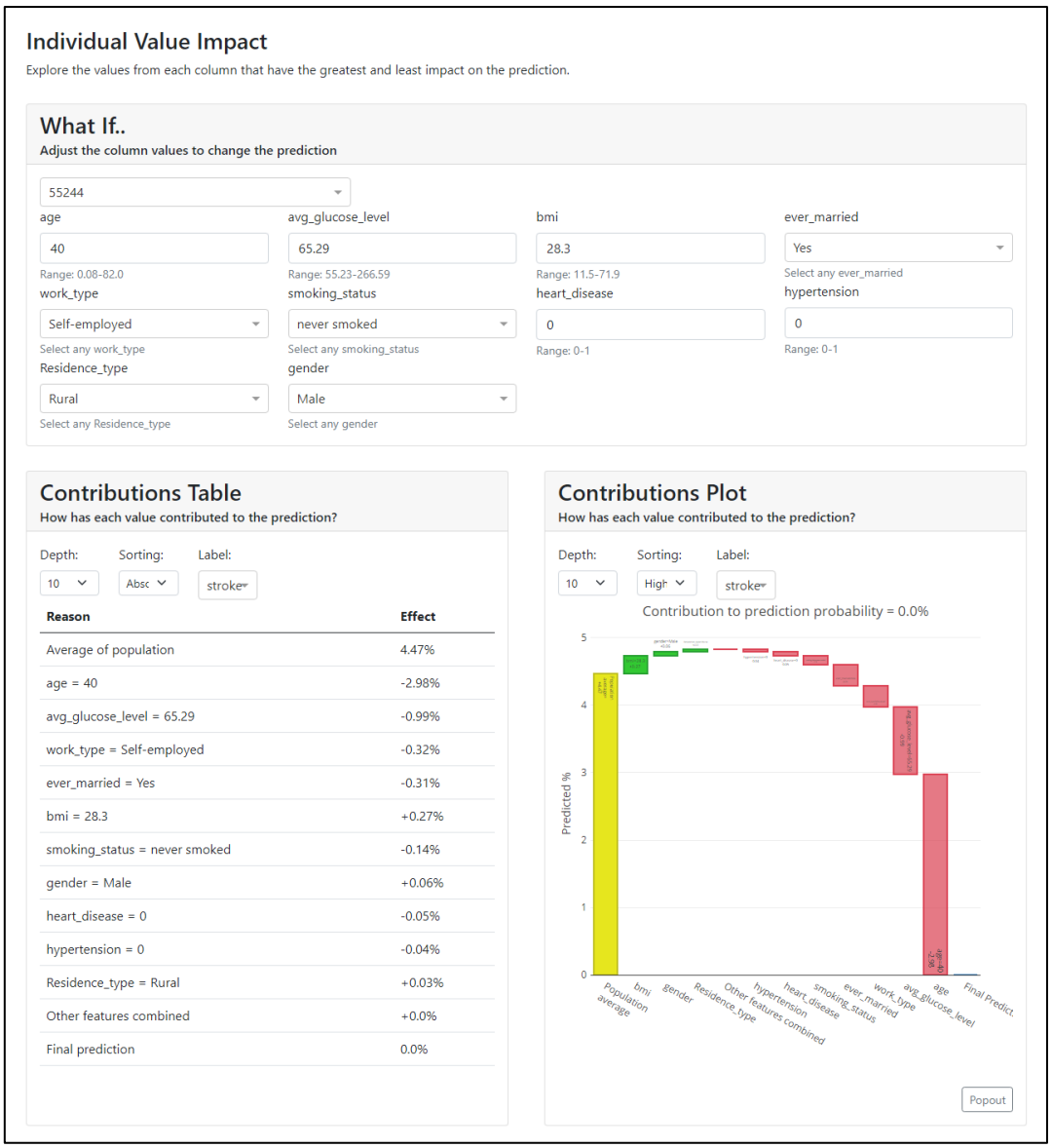


Figure 27 Model review prototype interface

The impact relationship tab consists of two XAI visualisation components which are the column impact and impact relationship visualisations. The column impact visualisation is similar to the one in the impact tab. However, on this tab, the visualisation is dynamically connected to the impact relationship visualisation. When a user clicks on a column in the visualisation, that specific column is visualised in the impact relationship column as shown in the figure below (**DP5,DP7**).

The impact relationship visualisation is a dynamic visualisation that automatically shifts between a scatter plot and a violin plot if the chosen column is numerical or categorical respectively. It features a label dropdown similar to the one in the contributions table and plot. The colour dropdown allows the user to specify a column to include in the visualisation for exploring the correlation. The scatter plot and violin plot values show tooltips when hovered on indicating the ID value and the Shap value for each data point. Users can also explore a single ID or row by clicking on the ID dropdown and choosing a row. A black circle will identify the position of that specific data instance for the chosen ID row. Users can also remove outliers by clicking on the button below. The arrangement of the visualisation can also be sorted according to frequency, Shap impact, and alphabetically when the chosen column is category-based. The user could also enlarge the chart using the “popout” button at

the bottom of the visualisation. As the visualisations are built on Plotly, users can also zoom, select, pan, and download the visualisations as a .png image file.

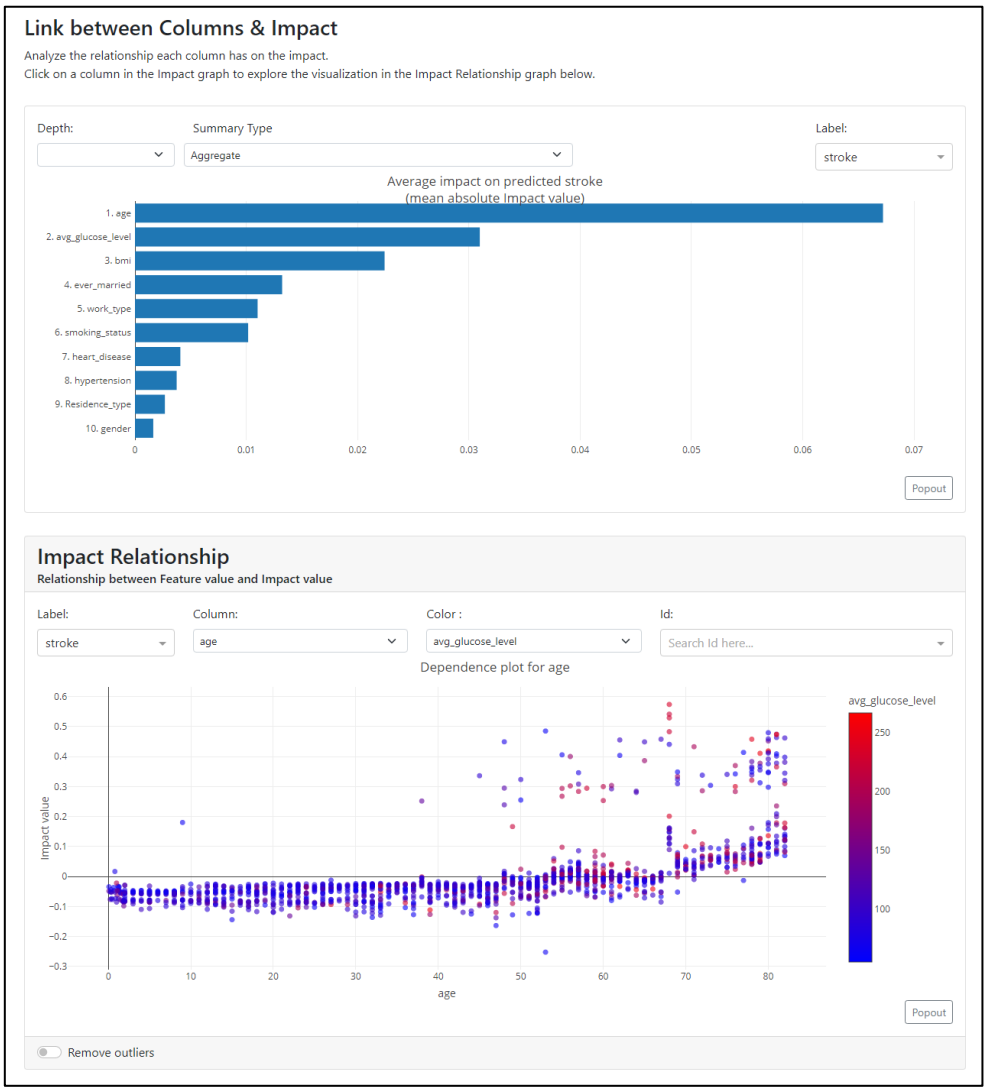


Figure 28 Model review prototype interface

As the model metrics tabs are different for regression and classification learning tasks, the regression model metrics tab is discussed below. The regression model metrics tab consists of three components which are the model summary, predicted vs actual visualisation, and plot vs feature visualisation. The model summary component displays the regression model scores such as mean squared error, root mean squared error, mean absolute error, mean absolute percentage error, and r squared within a table. Users can hover over each score to get a description and definition of each metric (**DP3**).

The predicted vs actual visualisation presents a dynamic scatter plot visualisation of the predicted values against the actual values. The visualisation allows users to toggle a logarithmic axis for both the y and x-axis. Users can also hover over each data point to identify the ID and Shap value (**DP5**). Similar to other visualisations, users could enlarge the chart using the "popout" button at the bottom of the visualisation. As the visualisations are built on Plotly, users can also zoom, select, pan, and download the visualisations as a .png image file.

The plot vs feature visualisation presents another scatter plot visualisation of the residuals and features. This visualisation allows users to choose which feature to visualise and analyse against the

residuals to find correlations (**DP7**). Users can also hover over each data point to identify the ID, feature value, and Shap value. Furthermore, users can choose what data to visualise on the y-axis such as residual difference, residuals log ratio, and observed data. The visualisation also includes a Winsor input value to remove outliers from the plot. Similar to other visualisations, users could enlarge the chart using the "popout" button at the bottom of the visualisation. As the visualisations are built on Plotly, users can also zoom, select, pan, and download the visualisations as a .png image file.

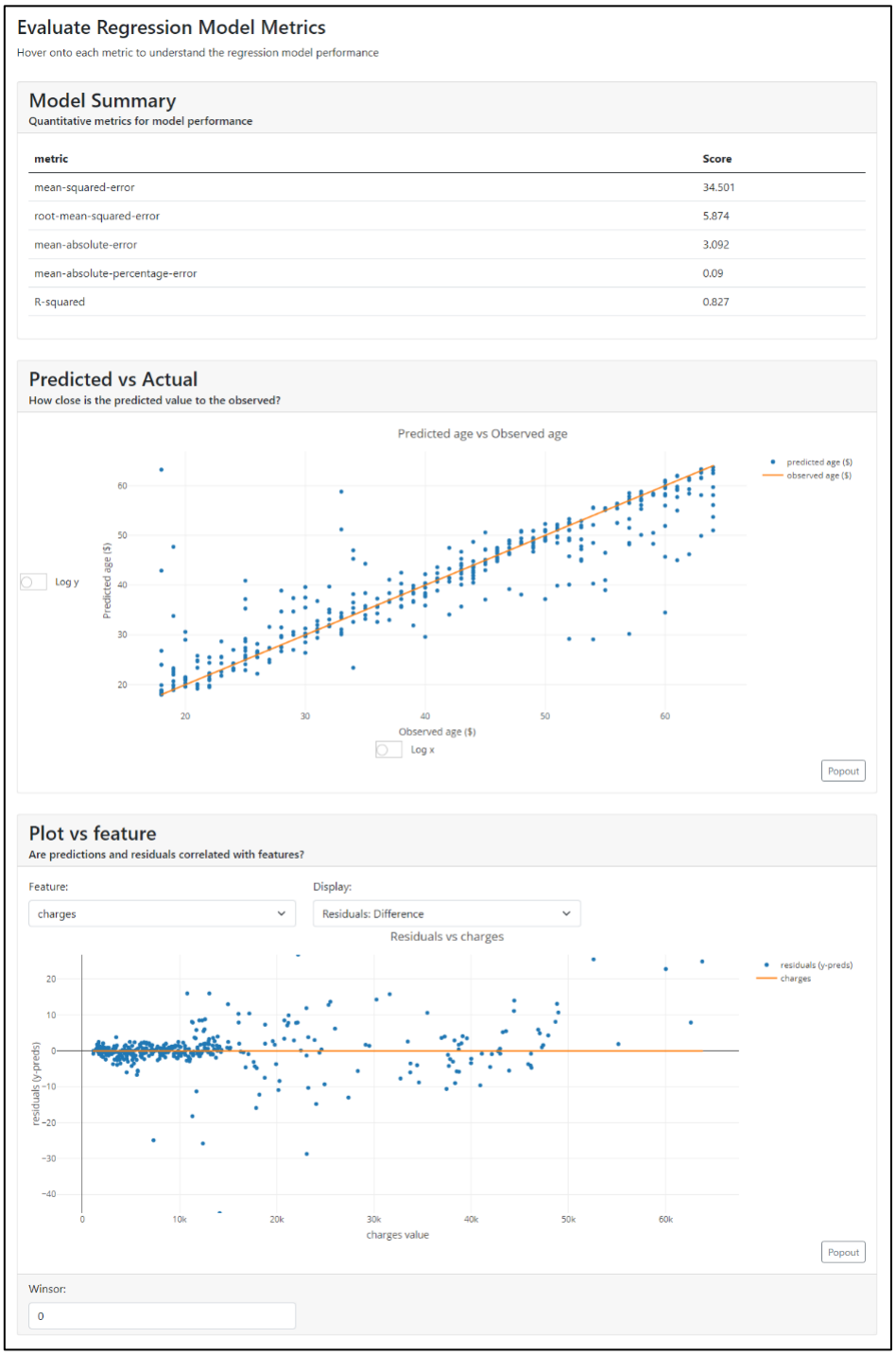


Figure 29 Model review prototype interface

The classification model metrics tab consists of two components which are the model performance metrics, and the plot confusion matrix. The model summary component displays the classification model scores such as accuracy, precision, recall, f1, roc auc score, and log loss within a table. Users can hover over each metric to obtain a description and definition for each score (**DP3**).

The confusion matrix visualisation presents a dynamic heat map visualisation of the predicted values against the actual values based on the positive and negative labels. There are four quadrants within the visualisation. Each quadrant presents a percentage and value of each instance. Users can toggle between highlighting either the percentage or the value frequency with the highlight percentage button at the bottom. Users can also hover over the visualisation component title to obtain a definition and explanation of confusion matrices (**DP3**). Similar to other visualisations, users could enlarge the chart using the "popout" button at the bottom of the visualisation. As the visualisations are built on Plotly, users can also zoom, select, pan, and download the visualisations as a .png image file.

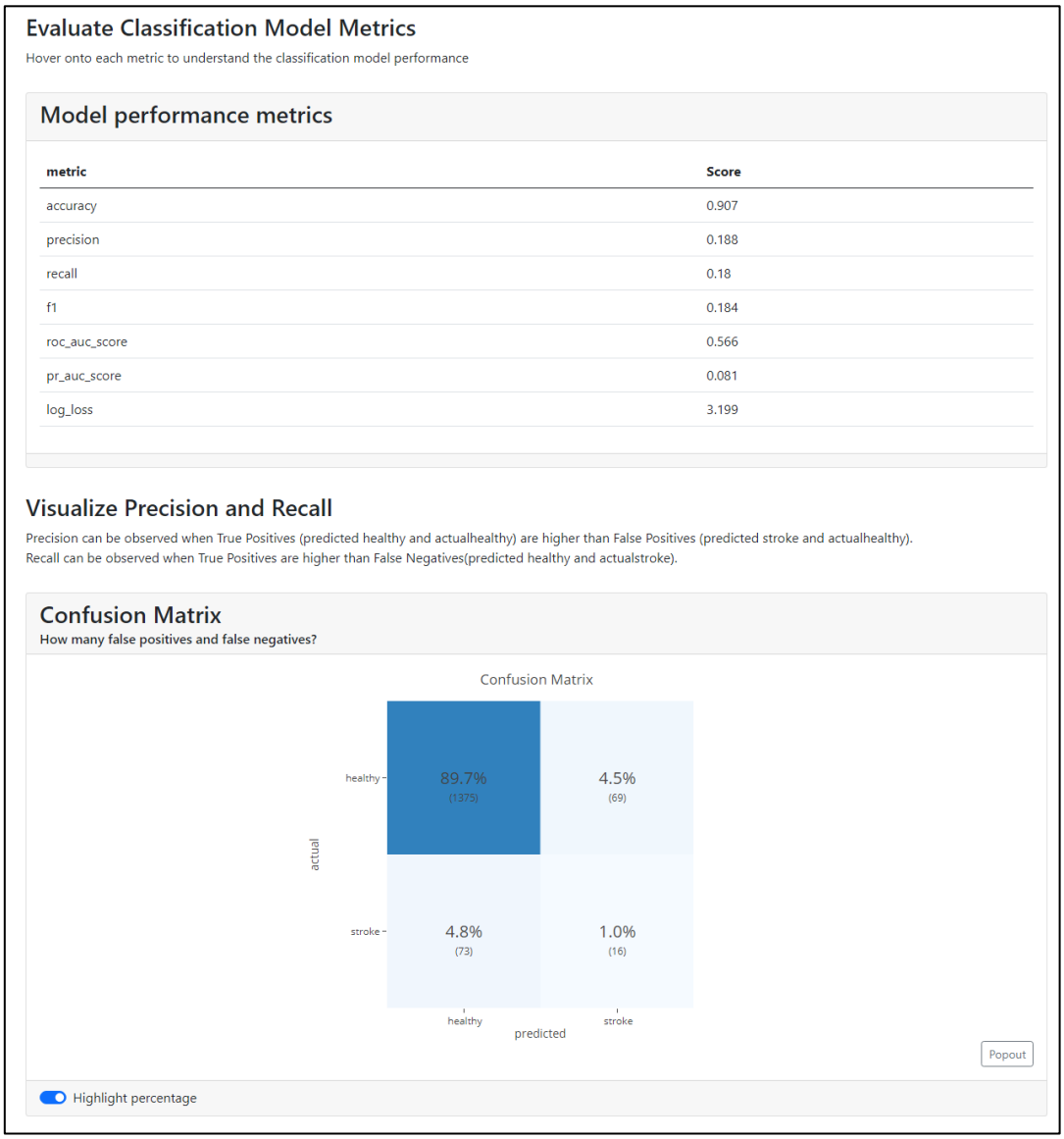


Figure 30 Model review prototype interface

4.1.9 Prototype Development

The VisAutoML system was mainly developed using React JS and Django. To develop the front end of the system, ReactJS was used with the MUI library. The front end was built on the following react components in Figure 35. Each component represents a UI element from React such as web pages, popups, list items, charts, and tables. For the backend of the system, the data to be stored in the database was defined in Django's "models.py" file as shown below. The model types define the model types to be saved with the corresponding short form used. The model class defines the columns in the model database such as model name, model type algorithm name, etc. The "views.py" file (as shown in Figure 36) in Django was used to handle all users' requests and return the appropriate responses. The ModelViewSet class handles the display, creation of new models, and deletion of models requests in the web app.

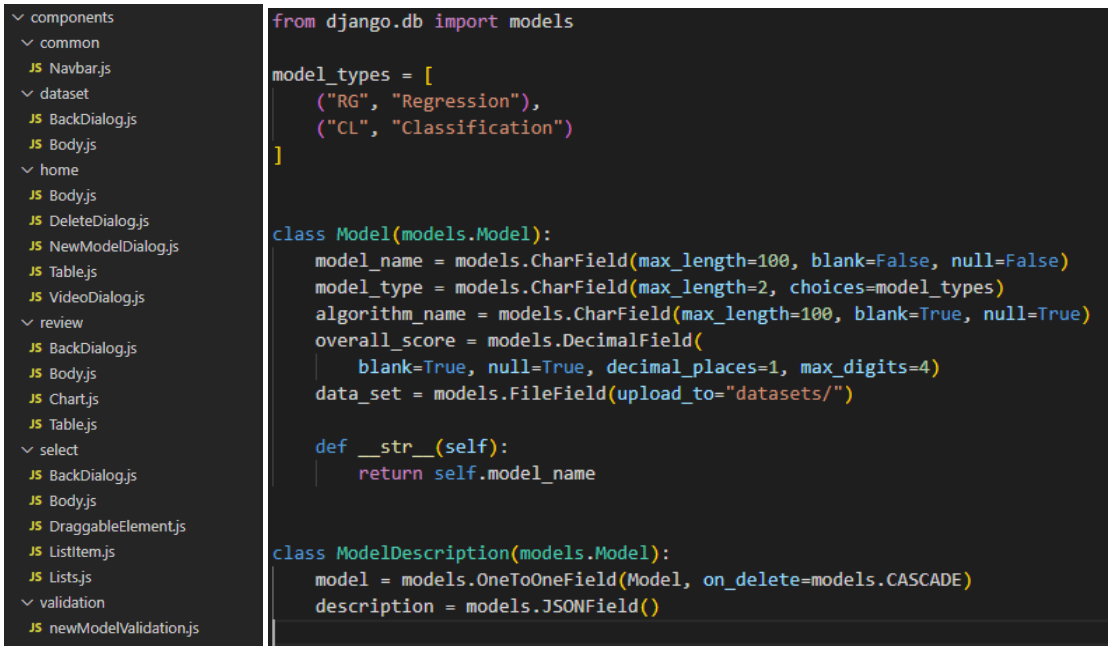


Figure 31 React components (left) Django models.py (right)

```
class ModelViewSet(viewsets.ViewSet):

    def list(self, request):
        models = Model.objects.all()
        serializer = ModelSerializer(models, many=True)
        return Response(serializer.data)

    def create(self, request):
        try:
            serializer = ModelSerializer(data=request.data)
            if serializer.is_valid():
                model = serializer.save()
                result = get_review(model.data_set.path)
                description = ModelDescription.objects.create(
                    model=model, description={})
                description_serializer = ModelDescriptionSerializer(description)
                return Response(
                    {"response": result, "model": serializer.data, "description": description_serializer.data})
            return Response(serializer.errors, status=status.HTTP_400_BAD_REQUEST)
        except Exception as e:
            traceback.print_exc()

    def destroy(self, request, pk):
        Model.objects.get(id=pk).delete()
        models = Model.objects.all()
        serializer = ModelSerializer(models, many=True)
        return Response(serializer.data)
```

Figure 32 Snapshot of Views.py file

In addition, Pandas was used for data manipulation and analysis of imported data. Pandas facilitated the easy cleaning and transformation of the data, handling missing values, scaling numerical values, encoding categorical columns, and splitting the dataset into input features and target labels. Below is a snapshot (Figure 37) of the data preparation steps done with pandas. Scikit-learn (Sklearn) was used to implement the machine learning pipeline, from data preparation to model training and evaluation. for the machine learning algorithms and tools. Sklearn provided a wide range of algorithms and tools for training and evaluating the models. The algorithms used in this iteration are: 1) logistic regression; 2) random forest classifier; 3) gradient boosting classifier; 4) decision tree classifier; 5) LGBM classifier; 5) XGBC classifier; 6) random forest regression; 7) gradient boosting regression; 8) bagging regressor; 9) extra trees regressor. For an automatic model selection, each algorithm is evaluated against the dataset and the one with the highest overall score is used for the XAI visualisation.

```
def prepare_model(drop, IDColumn):
    df = pd.read_csv(train_csv)
    has_header = csv.Sniffer().has_header(open(train_csv).read(2048))

    # id column set
    if (IDColumn != ""):
        IDColumn = IDColumn.replace(' ', '_')
        df.set_index(IDColumn, drop=True, inplace=True)
        df.index.name = IDColumn

    # predict to columns
    result = predict.replace(' ', '_')

    # convert list drop
    if drop != []:
        converter = lambda x: x.replace(' ', '_')
        drop = list(map(converter, drop))
        drop

    # space to underscore for all headers
    if has_header == False:
        df.columns = ['co_' + str(i + 1) for i in range(len(df.iloc[0].values))]
    df.columns = df.columns.str.replace(' ', '_')

    # drop unused columns
    if drop != []:
        df.drop(columns=drop, axis=1, inplace=True)
```

Figure 33 Snapshot of data preparation steps

```
b = []
a = [LogisticRegression, RandomForestClassifier, GradientBoostingClassifier, DecisionTreeClassifier,
     LGBMClassifier, XGBClassifier]

for i in a:
    model = i().fit(x_train, y_train.values.ravel())

    # testing training accuracy
    from sklearn import metrics
    from sklearn.metrics import balanced_accuracy_score
    from sklearn.metrics import average_precision_score
    from sklearn.metrics import roc_auc_score
    from sklearn.metrics import brier_score_loss

    y_pred = model.predict(x_test)
    x = metrics.accuracy_score(y_test, y_pred)
    y = balanced_accuracy_score(y_test, y_pred)
    z = average_precision_score(y_test, y_pred)
    z1 = roc_auc_score(y_test, y_pred)
    z2 = brier_score_loss(y_test, y_pred)
    b.append((x + y + z1) / 3 / 0.01)
    print(x, y, z, z1, z2)

print(b)
best_score = max(b)
print("best_score ", str(best_score))
index = b.index(best_score)
```

Figure 34 Snapshot of sklearn ML code

The XAI dynamic visualisations are based on Sklearn, Plotly, Shap, and Dash libraries. The models will be developed and the Shap values would be obtained using the shap.explain() function. Plotly Dash was used to create a web application that allows users to interact with the SHAP values and explore the model's explanations in more detail. This can be done using Dash components such as dropdown menus, sliders, and graphs. The Plotly Dash's "app.layout" and "app.callback" functions were used to define the layout and behaviour of the web application and use the SHAP values and the model's predictions to update the visualisations in real time. Below is a snapshot of a function implementing the confusion matrix plot using Plotly Dash.

```
fig = go.Figure(data, layout)
annotations = []
for x in range(cm.shape[0]):
    for y in range(cm.shape[1]):
        top_text = f"{cm_normalized[x, y]}%" if percentage else f"{cm[x, y]}"
        bottom_text = f"{cm_normalized[x, y]}%" if not percentage else f"{cm[x, y]}"
        annotations.extend([
            go.layout.Annotation(
                x=fig.data[0].x[y],
                y=fig.data[0].y[x],
                text=top_text,
                showarrow=False,
                font=dict(size=20)
            ),
            go.layout.Annotation(
                x=fig.data[0].x[y],
                y=fig.data[0].y[x],
                text=f" <br> <br> <br>({bottom_text})",
                showarrow=False,
                font=dict(size=12)
            )]
        )
longest_label = max([len(label) for label in labels])
fig.update_layout(annotations=annotations)
fig.update_layout(margin=dict(t=40, b=40, l=longest_label*7, r=40))
return fig
```

Figure 35 Snapshot of Plotly Dash code

4.2 VisAutoML 2.0

Continuing within the framework of the UCD development process, VisAutoML 2.0 will focus on refining and optimizing the user experience based on the insights collected from the initial version. This iterative approach is central to the UCD methodology, ensuring that user feedback and evaluations actively inform the enhancement of the proposed tool. This chapter will explore the iterative design process, considering feedback from users, usability evaluations, and transparency assessments to guide the evolution of VisAutoML. The subsequent sections include a summary of previous stage, areas for improvement, redesign objectives, redesign methodology, wireframing process, prototyping process and expected outcomes. This UCD-centric approach ensures VisAutoML 2.0 is a tool that evolves in tandem with user needs, preferences, and expectations.

4.2.1 Areas for Improvement

The comprehensive evaluation of VisAutoML 1.0 has provided insights into the tool's strengths and areas for improvement. Identified across various areas such as usability, transparency, and user experience, these areas for enhancement offer valuable guidance for future refinements. The following sections detail these areas for improvement.

Usability has been identified as a crucial area for refinement in VisAutoML. Despite the noted efficiency of the tool, users reported challenges during the loading phase, finding it confusing. A significant step towards enhancing the tool's usability would be to address this issue by providing more informative error messages. This would make the loading phase more intuitive and user-friendly, alleviating user confusion. To further increase usability, the tool could also incorporate a progress indicator during the loading phase to keep users informed of the process.

Another aspect of usability that could benefit from further refinement is the tool's interface. While users generally appreciated its intuitiveness, they suggested that it could be made more beginner friendly. Implementing a dual-mode interface catering to both beginners and experts could enhance this aspect. The beginner mode could offer a simplified, guided experience, while the expert mode could provide more control and customisation options. This would cater to the varying needs of users, making the tool more accessible.

The onboarding process, according to the users, could be clearer. Simplifying the onboarding process by providing a step-by-step guide or interactive tutorial could help reduce the initial learning curve associated with the tool. This would likely make the tool more accessible to novice users and ensure they can navigate the tools with clarity and confidence from the outset.

The tool's pragmatic quality, referring to its practicality and efficiency, was rated 'Below Average' compared to the benchmark. To enhance this aspect, certain features could be made more prominent or easier to use. This could involve refining the workflow to make it smoother and more logical, or even adding shortcuts for frequently used actions. These improvements could make the tool more practical and efficient to use, thereby enhancing its pragmatic quality.

Hedonic quality, relating to the tool's ability to provide a satisfying and enjoyable user experience, was also identified as an area for improvement. Adding features that increase user engagement or enjoyment could help in this regard. This could include interactive elements or gamification features that make the tool more engaging. Additionally, refining existing features to make them more satisfying to use could also enhance the tool's hedonic quality.

Overall usability, as measured by the User Experience Questionnaire (UEQ), was rated 'Below Average' compared to the benchmark. Addressing the identified challenges and enhancing the tool's pragmatic and hedonic quality could significantly improve overall usability. This could involve usability testing with a diverse group of users to gain further insights into potential areas of improvement.

Transparency emerged as a key theme from the user feedback. While VisAutoML’s explainability was rated relatively high, users suggested that the tool's outputs could be clearer and more comprehensible. A potential refinement could be to enhance the existing Explainable AI (XAI) visualisations to provide more detailed, intuitive, and easy-to-understand explanations of the tool's outputs. This could involve using simpler language, clearer graphics, or interactive elements that allow users to explore the outputs in-depth.

Trust also emerged as an important factor in user satisfaction and adoption. Users expressed a moderate level of trust in VisAutoML. Enhancing user trust could involve improving transparency and ensuring the tool's outputs are reliable and accurate. This could be supported by providing clear explanations of how the tool arrives at its outputs, using XAI visualisations, and by demonstrating the tool's accuracy through case studies or benchmark tests.

Interactivity was another aspect that users appreciated, particularly the tool's drag-and-drop feature. Extending this interactivity to other areas of the tool could further enhance the user experience. For instance, dynamic elements could be incorporated into the XAI visualisations, allowing users to adjust parameters and immediately see the effects on the model results. This could foster a deeper understanding of the machine learning process and the tool's functionalities.

The feedback highlighted the importance of clear guidance and user support. Building on this, tooltips and help text could be incorporated into the tool's interface features and XAI visualisations to provide context-specific assistance. This could involve explaining complex terms or concepts, or providing guidance on interpreting the visualisations. This would make the tool's outputs more accessible and understandable, reducing potential confusion.

The introduction of an Expert/Beginner mode aligns with the feedback on the tool's usability and ease of use. The Beginner mode could offer a simplified development experience, focusing on key insights and using layman's terms. The Expert mode could provide a more advanced development experience, allowing experienced users to delve deeper into the data.

Users highlighted the importance of clear and comprehensive performance metrics. Building on this, the XAI visualisations could be enhanced to provide a more comprehensive overview of performance metrics. This could involve visualising key metrics in a clear and intuitive manner, and enabling side-by-side comparisons of different XAI visualisations.

The documentation of VisAutoML was identified as an area of concern. To address this, the documentation could be expanded to include more detailed explanations of the XAI visualisations, including how to interpret them and how they contribute to the machine learning process. This would provide valuable guidance and support to users, increasing their confidence in using the tool.

Lastly, the interconnectedness of trust, explainability, and usability suggests that improvements in one area could impact the others. Therefore, a holistic approach that considers all these aspects could be most effective in enhancing the overall user experience of VisAutoML.

Table 18 Identified areas for improvement

Area for Improvement	Suggested Improvements
Loading Phase	More informative error messages, progress indicator
Onboarding Process	Simplified process with a step-by-step guide or interactive tutorial
Pragmatic Quality	Make certain features more prominent or easier to use, refine workflow, add shortcuts
Hedonic Quality	Add interactive elements or gamification features, refine existing features
Overall Usability	Usability testing with diverse users, address identified challenges
Transparency	Enhance XAI visualisations for clearer, more comprehensible outputs
Trust	Improve transparency and reliability of outputs
Interactivity	Extend interactivity to XAI visualisations
Guidance and Support	Incorporate tooltips and help text into interface features and XAI visualisations
Expert/Beginner Mode	Offer simplified development experience for beginners and detailed ones for experts

Performance Metrics	Enhance XAI visualisations to provide a comprehensive overview of model output
Documentation	Expand documentation to include detailed explanations of XAI visualisations

In conclusion, the feedback provided by users offers invaluable insights into the areas of VisAutoML that require further refinement. Addressing these areas, specifically enhancing the XAI visualisations and improving the tool's usability, trustworthiness, and user-friendliness, could lead to a marked enhancement in the overall user experience. Such improvements are expected to increase user satisfaction and encourage wider adoption of the tool.

4.2.2 Redesign Objectives

Based on the comprehensive feedback and insights gathered from various stakeholders, a set of redesign objectives for VisAutoML has been identified. These objectives are strategically aimed at enhancing the overall usability, transparency, and user experience of the tool. Each redesign objective focuses on a specific area for improvement, ensuring that the tool meets the evolving needs of non-experts more effectively.

The first objective targets the loading phase of VisAutoML. Users reported confusion during this phase, and the redesign aims to provide more informative error messages and a progress indicator. Using progress bars or spinners can improve user patience by giving them a sense of time and progress. Nielsen (1993) discusses how progress indicators manage user expectations by visually representing the time remaining, which can reduce anxiety and improve overall user satisfaction during waiting periods. Clear and informative error messages can reduce user frustration and help them understand what went wrong. Shneiderman and Plaisant (2004) emphasise the importance of informative error messages in guiding users through problems, which helps reduce frustration and improves the overall user experience by providing clear guidance on how to resolve issues.

The onboarding process is the next area of focus. The redesign will simplify this process by implementing a step-by-step guide or interactive tutorial. Interactive tutorials or walkthroughs can significantly enhance the onboarding experience by guiding users through the initial setup. Anderson et al. (2009) reviewed how step-by-step guides and interactive tutorials help new users understand and engage with systems more effectively, reducing the learning curve and increasing initial satisfaction. Engaging users with interactive content during onboarding can also improve their understanding and retention. Mayer (2009) outlines how interactive tutorials enhance learning by providing immediate feedback and engaging multiple senses, which improves user understanding and retention of information.

Improving the tool's pragmatic quality is another objective. Making frequently used features more prominent can streamline workflows and enhance usability. Norman (2013) discusses the importance of feature prominence in design, emphasising that making key features easily accessible can significantly improve usability and streamline user workflows. Card, Moran, and Newell (1983) explore the psychology of human-computer interaction and highlight how optimized workflows can reduce cognitive load, making tasks easier and faster to complete. Implementing keyboard shortcuts or quick access menus can improve efficiency. Nielsen (1993) emphasises the role of shortcuts in usability,

showing that providing quick access to frequently used functions can greatly enhance efficiency and user satisfaction.

Enhancing the tool's hedonic quality is also a key objective. This will involve adding interactive elements or gamification features that increase user engagement and enjoyment and refining existing features to make them more satisfying to use. Incorporating interactive elements can increase user engagement and satisfaction. Csikszentmihalyi (1990) discusses the concept of "flow" and how interactive elements can create an engaging user experience that promotes satisfaction and immersion. The addition of game design elements also makes tasks more engaging and enjoyable. Deterding et al. (2011) define gamification and demonstrate how incorporating game-like elements can increase user motivation, engagement, and enjoyment in non-game contexts.

Transparency is another area targeted for enhancement. The tool's XAI visualisations will be enhanced to provide clearer and more comprehensible outputs. This will make it easier for users to understand and interpret the tool's outputs, increasing the tool's overall transparency. Tufte (1983) highlights the importance of clear and detailed visualisations in conveying complex information effectively, which enhances understanding and decision-making. Also creating interactive visualisations allow users to gain a deeper understanding of the model outputs. Heer and Shneiderman (2012) show how interactive visualisations can help users explore data in a more meaningful way, leading to deeper insights and better comprehension.

Boosting user trust in VisAutoML is a key objective of the redesign. Trust can be improved by making the model's decision-making process more visible and understandable to users. Ribeiro et al. (2016) discuss methods for explaining classifier predictions, which can help users understand and trust the model's outputs. Also, providing detailed explanations of the model’s decision-making process can boost user trust. Doshi-Velez and Kortz (2017) emphasise the importance of explanation in AI for accountability, which supports the need for clear explanations in VisAutoML.

Interactivity is another focus of the redesign. Feedback highlighted users' appreciation for the tool's drag-and-drop feature, indicating the value they place on interactivity. The incorporation of more dynamic elements into the XAI visualisations, would allow users to adjust parameters and immediately see the effects on the model results. Shneiderman (1997) and Heer & Shneiderman (2012) provide insights into the benefits of direct manipulation and interactive dynamics in user interfaces. Amershi et al. (2014) discuss the role of humans in interactive machine learning, emphasising the importance of user control and interaction, which supports the need for dynamic elements in XAI visualisations.

Providing clearer guidance and user support is also a key objective. To address this, tooltips and help text will be incorporated into the interface features and XAI visualisations. This will provide users with context-specific assistance, reducing potential confusion and enhancing the user experience. Clear guidance and support can significantly enhance usability. Nielsen (1993) and Cooper et al. (2007) emphasise the importance of usability engineering and interaction design in creating user-friendly interfaces. Norman (2002) also discusses the principles of user-centred design, which supports the incorporation of context-specific assistance to improve user experience.

The introduction of an Expert/Beginner mode aligns with the feedback on the tool's usability and ease of use. The redesign will introduce a beginner mode offering a simplified development experience, while an expert mode will provide more detailed options. Nielsen (1993) and Schneiderman & Plaisant (2004) highlight the importance of usability and designing for different user skill levels, supporting the need for both beginner and expert modes. Norman (2002) also emphasises the need for

accommodating different user expertise levels, supporting the introduction of an expert mode with advanced features.

Providing a more comprehensive overview of performance metrics is another objective. To achieve this, the XAI visualisations will be enhanced to provide a comprehensive overview of the model output, visualising key metrics in a clear and intuitive manner. This will assist users in understanding their models' performance. Doshi-Velez & Kim (2017) and Lipton (2018) discuss the importance of interpretability and clear presentation of model metrics, supporting the need for comprehensive and intuitive visualisations of performance metrics.

Finally, the documentation of VisAutoML will be expanded. This will involve providing more detailed explanations of the XAI visualisations. Gregor and Jones emphasise that thorough documentation is crucial for users to understand and effectively apply complex systems. Their framework highlights that detailed explanations help users grasp AI/ML concepts, thereby enhancing the usability of VisAutoML. This is also shown in Amershi et al. who provided guidelines for designing human-AI interactions, which stress the importance of clear documentation in making AI systems understandable and usable.

Table 19 Redesign objectives based on area of improvement

Area for Improvement	**Redesign Objectives**	**Detailed Actions**
Loading Phase	Enhance user understanding during the loading phase	Develop more informative and user-friendly error messages; Implement a progress indicator to keep users informed about loading status
Onboarding Process	Simplify the onboarding process for new users	Create a step-by-step guide or interactive tutorial to guide users through the initial setup and usage of the tool
Pragmatic Quality	Improve the tool's practicality and efficiency	Make certain features more prominent or easier to use; Refine the workflow to optimize user journey; Add shortcuts for frequently used actions
Hedonic Quality	Enhance user engagement and enjoyment	Incorporate interactive elements or gamification features; Refine existing features to make them more satisfying and appealing to use
Transparency	Improve the clarity and comprehensibility of tool's outputs	Enhance the XAI visualisations to provide clearer, more detailed, and interactive explanations of the tool's outputs
Trust	Boost user trust in model outputs	Improve transparency and reliability of outputs; Demonstrate how VisAutoML arrives at its outputs; Provide evidence of the tool's accuracy through case studies or benchmark tests
Interactivity	Extend the tool's interactivity	Incorporate more dynamic elements into the system and XAI visualisations, allowing users to adjust

		parameters and immediately see the effects on the model results
Guidance and Support	Provide clearer guidance and user support	Incorporate tooltips and help text into the interface features and XAI visualisations to provide context-specific assistance
Expert/Beginner Mode	Cater to both beginners and experts	Introduce a beginner mode offering a simplified development experience; Provide an expert mode with more detailed controls and settings
Performance Metrics	Provide a comprehensive overview of performance metrics	Enhance the XAI visualisations to provide a comprehensive overview of the model output, visualising key metrics in a clear and intuitive manner
Documentation	Expand the documentation of VisAutoML	Provide more detailed explanations of the XAI visualisations in the documentation, offering valuable guidance and support to users

The outlined redesign objectives serve as a comprehensive roadmap towards enhancing the user experience with VisAutoML. Each objective, informed by user feedback, targets specific areas of improvement ranging from improving the loading phase to expanding the tool's documentation. The objectives also aim to enhance transparency, trust, interactivity, and support, all crucial elements for a satisfying and engaging user experience. By implementing more informative error messages, simplifying the onboarding process, refining the tool's workflow, enhancing XAI visualisations, and expanding the tool's documentation, among other initiatives, the redesign aims to make VisAutoML more efficient, user-friendly, and engaging. These objectives also underscore the value of a user-centred design approach, which prioritizes the needs and preferences of users. The proposed changes are expected to improve the tool's usability and functionality and boost user satisfaction and adoption rates. Moving forward, these objectives will guide the development process, ensuring that VisAutoML continues to evolve in line with user needs and preferences.

4.2.3 Revised Design Principles

Based on the feedback and insights gathered from users in stage 5, the design principles guiding the development of VisAutoML have been revisited and refined. This process is crucial to validate the findings from user feedback and to identify specific areas where improvements are necessary. The primary goal is to enhance the usability, transparency, and overall user experience of VisAutoML, ensuring it meets the evolving needs of non-experts more effectively.

As part of this revision, a new design principle, (DP5) engaging and informative user feedback, has been introduced. This principle is grounded in the understanding that effective user feedback mechanisms are essential for maintaining user engagement and satisfaction. Research has shown that providing clear, real-time updates and interactive elements can significantly enhance user experience and reduce uncertainty during interactions with digital tools (Sutcliffe, 2022; Yigitbas et al., 2019).

Engaging and informative feedback helps in keeping users informed about the status of ongoing processes, such as loading and error handling, and contributes to a more enjoyable and engaging user experience. This principle aligns with findings from user experience studies, which highlight the

importance of feedback in improving user satisfaction and trust in digital systems (Amershi, Weld, Vorvoreanu, Fourney, Nushi, Collisson, Suh, Iqbal, Bennett, Inkpen, et al., 2019; Chatzimparmpas et al., 2020).

Table 20 Revised design principles

Category	Design principle	Note	Reference
Overall system	**DP1**. Visualise activity and sequences	Outline the procedural and metacognitive processes required to facilitate an ML process. (e.g., Specify the tasks involved, and clarify the ML development pipeline)	(Quintana, Zhang, et al., 2018; Wu et al., 2021)
	DP2. Demonstrate scaffold function	Present the utility and steps involved throughout the ML process. (e.g., Example-based development, online demonstration)	(Puntambekar, 2022; Quintana, Zhang, et al., 2018; Saye & Brush, 2002)
	DP3. Embedded contextually relevant scaffold	Integrate resources based on a conceptual framework into the system to facilitate further learner inquiry. (e.g., Hyperlinks specific to an ML task)	(Akotuko et al., 2021; Puntambekar, 2022; Saye & Brush, 2002)
	DP4. Visible and utilised scaffold	Ensure scaffolds are visible and explicitly clarified to learners to promote appropriate usage. (e.g., Specify and explain the functionality of scaffolds within the system to ensure an effective understanding of the system)	(Quintana, Reiser, et al., 2018; Sarah, 2022)
	DP5. Engaging and Informative User Feedback	Provide timely, clear, and relevant information to users about their actions and the system's state. This design principle aims to enhance user engagement, ensure smooth interaction with the system, and reduce uncertainty.	(Mezhoudi, 2013; Sutcliffe, 2022; Yigitbas et al., 2019)
XAI component	**DP6**. Progressive explanation disclosure	Provide finer granularity of an explanation through subsequent steps following an explanation interaction. (e.g., Visualising a specific feature after clicking on it, tooltips to display the factors of a feature)	(Buçinca et al., 2021; Khosravi et al., 2022; Millecamp et al., 2019)
	DP7. Natural language rationale	Complement visual explanations with textual explanations to facilitate better understanding (e.g., Natural language explanation of a feature importance chart)	(Ehsan et al., 2019, 2021; Wiegreffe & Marasovic, 2021)
	DP8. Multiple ways to communicate an explanation	Provides related explanations to triangulate insights and understand different angles of explanation. (e.g., Visualising global feature importance and local feature importance in relation to each other)	(Chou et al., 2022; Páez, 2019; Vilone & Longo, 2021)

The revised design principles now comprehensively address key areas for enhancement, including user onboarding, tool practicality, user engagement, clarity of outputs, trust in model outputs, interactivity, user support, and documentation. By applying these principles, the redesign efforts aim to create a more intuitive, transparent, and enjoyable AutoML tool that better supports non-expert users in their

machine learning tasks. The table below shows the corresponding design principle for each redesign objective.

Table 21 Redesign objectives with corresponding design principle

Redesign Objectives	Detailed Actions	Corresponding Design Principle
Enhance user understanding during the loading phase	Implement flow diagrams displaying the visual pipeline of ML development stages	DP1. Visualise Activity and Sequences
	Create interactive process maps for exploring different stages of the ML pipeline	DP1. Visualise Activity and Sequences
	Develop video walkthroughs guiding users through each phase of the ML process	DP2. Demonstrate Scaffold Function
Simplify the onboarding process for new users	Develop interactive tutorials for step-by-step guidance	DP2. Demonstrate scaffold function
	Provide example-based learning with sample datasets and pre-built models	DP2. Demonstrate scaffold function
	Implement contextual help overlays for brief descriptions and usage tips	DP3. Embedded Contextually Relevant Scaffold
Improve the tool's practicality and efficiency	Implement contextual tooltips with explanations and additional resources	DP3. Embedded Contextually Relevant Scaffold
	Integrate in-system guides triggered by specific actions	DP3. Embedded Contextually Relevant Scaffold
	Refine the workflow to optimize user journey	DP4. Visible and utilised scaffold
Enhance user engagement and enjoyment	Provide real-time feedback on user actions	DP5. Engaging and Informative User Feedback
	Use progress indicators to inform users about the current stage of their ML project	DP5. Engaging and Informative User Feedback
	Incorporate interactive elements	DP5. Engaging and Informative User Feedback
Improve the clarity and comprehensibility of tool's outputs	Implement layered information allowing users to click for more detailed explanations	DP6. Progressive explanation disclosure
	Include expandable sections for deeper dives into explanations	DP6. Progressive explanation disclosure
	Enhance XAI visualisations to provide clearer, more detailed, and interactive explanations	DP8. Multiple Ways to Communicate an Explanation
Boost user trust in model outputs	Complement visual data with clear, concise natural language explanations	DP7. Natural Language Rationale
	Provide scenario-based explanations for specific predictions	DP7. Natural Language Rationale
Extend the tool's interactivity	Present both global and local feature importance visually	DP8. Multiple ways to communicate an explanation
	Combine visual aids, textual descriptions, and audio/video content for explanations	DP8. Multiple ways to communicate an explanation
	Allow users to adjust parameters and immediately see the effects on the model results	DP5. Engaging and Informative User Feedback
Provide clearer guidance and user support	Place help icons next to complex features	DP4. Visible and utilised scaffold
	Provide quick start guides for initial familiarization with the tool	DP4. Visible and utilised scaffold
	Incorporate tooltips and help text into the interface features and XAI visualisations	DP3. Embedded contextually relevant scaffold

Cater to both beginners and experts	Implement video walkthroughs guiding users through each phase of the ML process	DP1. Visualise activity and sequences
	Create cross-linked explanations for navigating between related concepts	DP8. Multiple ways to communicate an explanation
	Introduce a beginner mode offering a simplified development experience	DP2. Demonstrate scaffold function
Provide a comprehensive overview of performance metrics	Use comparative visuals to show how features impact the model overall versus specific predictions	DP8. Multiple ways to communicate an explanation
	Enhance the XAI visualisations to provide a comprehensive overview of the model output	DP8. Multiple ways to communicate an explanation
	Visualise key metrics in a clear and intuitive manner	DP6. Progressive explanation disclosure
Expand the documentation of VisAutoML	Integrate textual descriptions and scenario-based explanations in the documentation	DP7. Natural language rationale
	Provide more detailed explanations of the XAI visualisations	DP7. Natural language rationale
	Offer valuable guidance and support to users through comprehensive documentation	DP3. Embedded Contextually Relevant Scaffold

4.2.4 Redesign Methodology

The redesign methodology for VisAutoML is structured into distinct phases. Each phase plays a crucial role in shaping the redesign, with every subsequent phase building upon the outputs of the previous one.

The first phase is the defining of areas for improvement. This phase involves an exhaustive review of the tool, with a keen focus on user feedback and data patterns. This analysis serves to highlight the aspects of VisAutoML that users may find challenging or unsatisfactory. By identifying these areas of improvement, the redesign process is given a clear direction, and a foundation is laid for the subsequent phases.

Once the areas of improvement have been duly identified, the process transitions into the second phase, which involves setting the redesign objectives. This phase takes the areas of improvement identified in the first phase and translates them into clear, actionable objectives. These objectives serve as the roadmap for the redesign process, guiding every decision and action in the subsequent phases.

Following ideation and concept development, the process moves into the third phase, wireframing. Here, interactive, low-fidelity wireframes of the proposed changes are created. These wireframes serve as visual representations of how the proposed changes will work in practice, allowing for an exploration and refinement of the concepts.

The final phase of the redesign methodology is implementation. This phase involves accurately and effectively applying the proposed changes to the actual tool based on the wireframes. It is in this phase that the redesign comes to fruition, marking the culmination of the process.

In conclusion, the redesign methodology for VisAutoML is a comprehensive, sequential process that involves defining areas for improvement, setting redesign objectives, ideation and concept development, wireframing, and implementation. Each phase is instrumental in ensuring that the

redesign enhances the user experience and meets the needs and preferences of the tool's user base. The table below summarises each phase and its procedure.

Table 22 Redesign methodology

Phase	Procedure
Defining Areas for Improvement	An exhaustive review of the tool is conducted, focusing on user feedback and data patterns. This phase identifies aspects of VisAutoML that users find challenging or unsatisfactory, thereby setting the direction for the redesign process.
Setting Redesign Objectives	This phase involves developing objectives based on the areas of improvement identified in the first phase. The objectives serve as a roadmap for the redesign process, guiding every decision and action in the subsequent phases.
Wireframing	Interactive, low-fidelity wireframes of the proposed changes are created. The wireframes serve as visual representations of how the proposed changes will work in practice, allowing for exploration and refinement of the concepts.
Implementation	The proposed changes are accurately and effectively applied to the actual tool based on the wireframes. This phase marks the culmination of the redesign process.

In conclusion, the redesign methodology for VisAutoML is a comprehensive and sequential process that is structured into distinct phases: defining areas for improvement, setting redesign objectives, ideation and concept development, wireframing, and implementation. Each phase serves a unique purpose and collectively, they ensure a systematic and efficient approach towards the redesign. This methodical approach ensures that each step of the redesign is well-defined and transitions smoothly into the next, maintaining a clear focus on the ultimate goal of enhancing the user experience. It beautifully encapsulates the principles of User-Centered Design (UCD), emphasising the importance of understanding and addressing user needs throughout the redesign process. By adhering to this methodology, the redesign of VisAutoML is expected to effectively address identified areas of improvement and align the tool more closely with user needs and preferences.

4.2.5 Wireframing Redesign

The high-fidelity wireframing process represents a pivotal component in the strategic redesign of the tool. This process, which is conducted using Adobe XD, is grounded in the concepts, redesign objectives, and design principles that were developed and established during the Design and Prototyping Stage 1. This process aims to create a highly detailed and accurate representation of the final design, a key step in translating abstract ideas into a tangible and functional product. High-fidelity wireframes are a cornerstone in creating a user-centric design. They offer a comprehensive view of the tool, incorporating exact dimensions, colours, typography, and images, thereby presenting a realistic preview of the tool prior to its development.

Home Page

The home page, being the first point of interaction for users, is redesigned with the objective of creating an intuitive and welcoming environment. The redesign is guided by the objectives developed earlier, which are aimed at enhancing the user experience and the tool's functionality.

A significant addition to the home page is an easy/expert mode switch in the navigation menu. This feature caters to both beginners and experts, aligning with the redesign objective of accommodating varying levels of user expertise. Beginners can navigate the tool with simplified controls and settings, while experts can access more detailed functionalities as required.

To enhance the onboarding process, tooltips are introduced throughout the home page. These tooltips align with the redesign objective of providing clearer guidance and user support. They offer brief explanations and instructions to guide users through the tool's features and functionalities.

The home page also includes a dashboard, providing a comprehensive overview of the models developed by the user. This feature aligns with the redesign objective of enhancing the tool's pragmatic quality by improving its practicality and efficiency. The dashboard displays key information about the models, such as performance metrics, model types, and average scores, providing users with a quick and easy overview of their work.

Documentation buttons are prominently displayed on the home page, providing easy access to comprehensive resources about the tool and its functionalities. This feature aligns with the redesign objective of expanding the documentation of the tool, providing users with more detailed explanations and guidance.

Buttons directing users to more information and a video tutorial are also incorporated into the home page. These features align with the redesign objective of enhancing user engagement and enjoyment (hedonic quality) by providing diverse learning resources and interactive elements. The high-fidelity prototype redesign of the home page is shown in the diagram below.

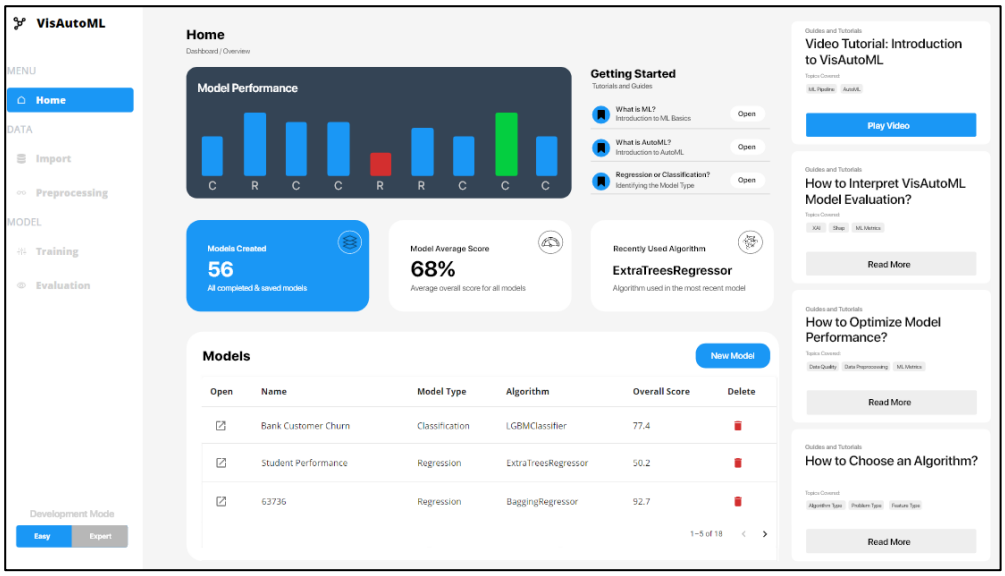


Figure 36 Wireframe of Home Page for VisAutoML 2.0

Data Import Page

The Data Import Page functions as a pivotal component of the tool, offering a platform for users to upload their data for analysis. The redesign of this page has been guided by an explicit focus on enriching user experience and augmenting the tool's functionality, primarily through the introduction of features tailored to non-expert users. This approach enhances the tool's accessibility and user-friendliness, thus broadening its appeal to a diverse user base.

A noteworthy feature introduced is the provision of built-in datasets, which can be loaded effortlessly with a click. This addition is specifically designed to cater to non-expert users or beginners, providing them with an opportunity to explore and learn using these ready-to-use datasets. This feature effectively lowers the entry barrier for beginners, allowing them to familiarize themselves with the tool's functionalities without the necessity of providing their datasets initially. This strategic move aligns with the redesign objective of making the tool accessible to a wide range of users, from novices to experts. While beginners can leverage these built-in datasets, experts retain the flexibility to upload their custom datasets.

To further enrich user guidance and support, tooltips and easy-to-follow instructions have been incorporated into each section of the Data Import Page. These elements are designed to assist non-expert users who may require additional guidance while navigating the tool. The tooltips, particularly beneficial for users who opt for the easy mode, offer succinct explanations and step-by-step instructions, facilitating a smooth user journey through the data import process. This approach aligns with the redesign objective of enhancing user guidance and support, ensuring users feel assisted and confident when utilising the tool.

Another significant addition to the Data Import Page is the data preview section, located at the bottom of the page. This innovative feature provides a concise overview of the uploaded data, offering crucial information about the data structure, variables, and initial observations. This transparency is particularly advantageous for non-expert users who may not possess a comprehensive understanding of data analysis. It aligns with the redesign objective of bolstering the tool's transparency by offering users a clear and comprehensible view of their data. The high-fidelity prototype redesign of the data import page is shown in the diagram below.

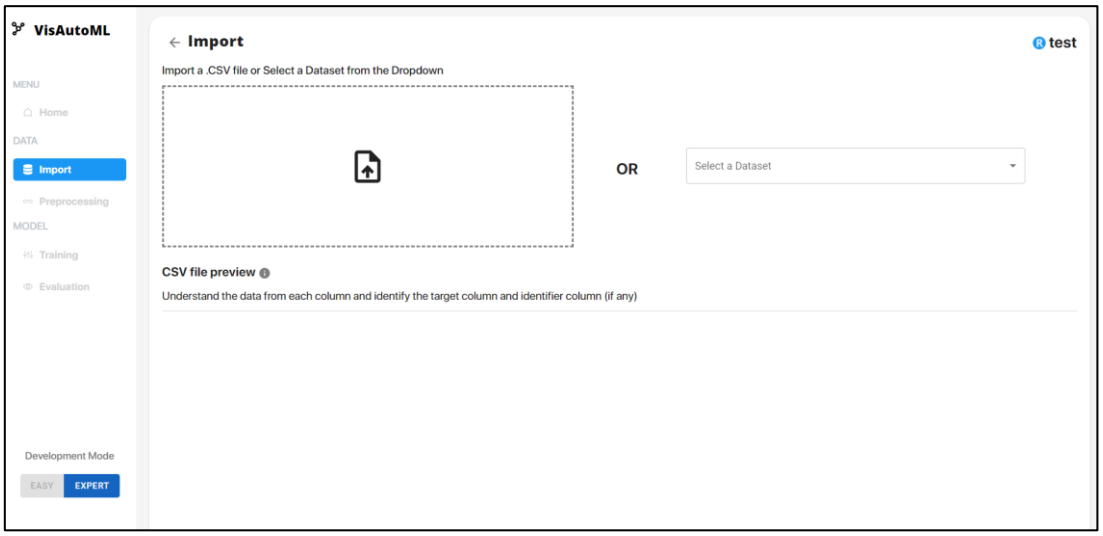


Figure 37 Wireframe of Data Import Page for VisAutoML 2.0

Data Preprocessing Page

The Data Preprocessing Page is a crucial aspect of the tool where users can prepare their data for subsequent analysis. Two new sections, the Data Quality Section and the Data Editor Section, have been added to this page. These additions aim to enhance the tool's transparency and build trust with users by providing them with a clear and comprehensive view of their data.

The Data Quality Section is a novel addition designed to provide users with interactive visualisations of their data. In line with the redesign objective of improving the tool's transparency, this section allows users to inspect their data quality comprehensively. Users can interact with automated histograms and distribution visualisations to gain a deep understanding of their data. Any main issues with the data are flagged with alerts, ensuring the users are well-informed about the quality of their data. This transparency can enhance trust between the users and the tool, as users are given full visibility of their data's quality. This is crucial for non-expert users who may not have the technical skills to evaluate their data quality independently.

Another significant addition to the Data Preprocessing Page is the Data Editor Section. This section, aligning with the redesign objective of making the tool more user-friendly, allows users to edit their data directly within the tool. Non-expert users can benefit significantly from this feature, as it eliminates the need for additional data editing tools and simplifies the data preparation process. Users can perform a range of editing tasks, from simple tasks such as renaming variables to more complex tasks like handling missing data. This user-friendly design approach can make the tool more accessible to non-expert users, thereby broadening its user base.

To further support non-expert users, tooltips have been added to both the Data Quality Section and the Data Editor Section. These tooltips, available in easy and expert modes, provide users with step-by-step guidance on how to use the tool effectively. The easy mode tooltips provide straightforward instructions, while the expert mode tooltips offer more detailed guidance. This caters to the needs of both non-expert and expert users, ensuring the tool remains versatile and adaptable to different user needs. The high-fidelity prototype redesign of the data preprocessing page is shown in the diagram below.

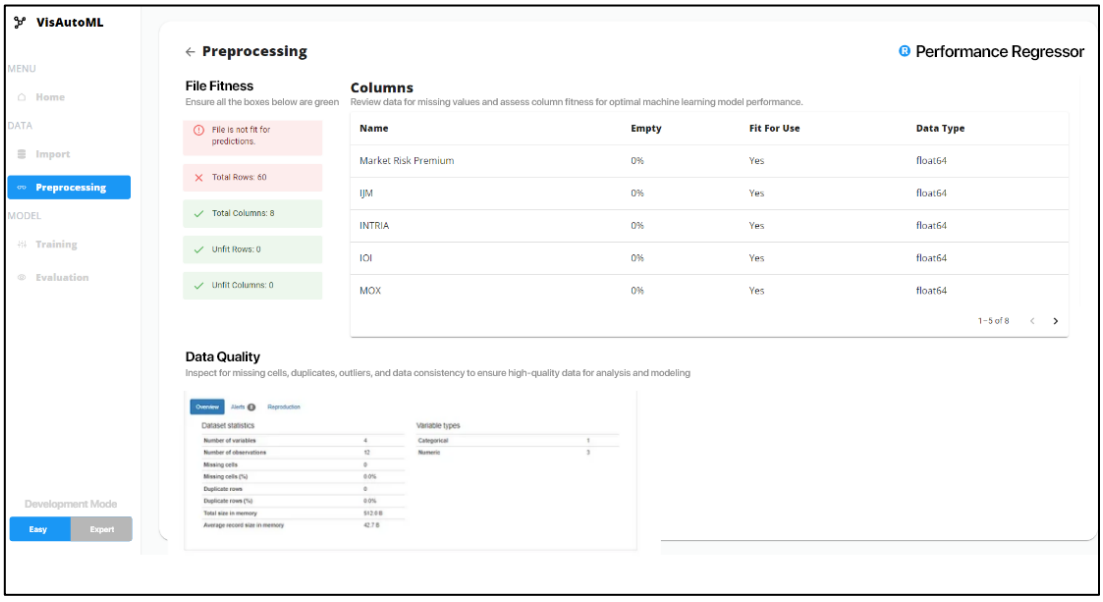


Figure 38 Wireframe of Data Preprocessing Page for VisAutoML 2.0

Model Training Page

The Model Training Page, a critical component of the tool, facilitating users in training their ML models effectively. The guided onboarding steps have been incorporated to streamline the onboarding process for the users. The essence of these steps lies in their alignment with the research objective of simplifying the tool's initial interaction for new users. By facilitating a step-by-step guide, this feature significantly reduces the learning curve associated with the tool, instilling confidence even in non-expert users. It enables them to navigate through the tool with relative ease and understanding, thereby fostering a sense of self-efficacy and autonomy.

The addition of the data split percentage dragger button manifests the research objective of pragmatic quality. This feature, by virtue of its interactivity, offers users the ability to adjust the data split ratio with precision- a task that is often perceived as complex, particularly by non-expert users. The simplification of this process enhances the tool's practicality and efficiency and heightens its accessibility and user-friendliness.

The Data Viewer section has been introduced, keeping in mind the research objectives of transparency and trust. This feature provides users with the capacity to preview their data while simultaneously mapping columns for prediction, identifier columns, and columns to ignore. For non-expert users, this feature proves to be particularly beneficial, as it enhances their understanding of their data structure. Moreover, by making the tool's operations more transparent, it fosters an increased level of trust in the tool.

Interactive tooltips have been integrated across the page, aiming to enhance the guidance and support provided to the users. These tooltips, which offer concise explanations and instructions for each section, align with the research objective of providing clearer guidance and user support. They prove to be a valuable source of immediate assistance to users, especially non-expert ones who might require additional guidance when navigating through the tool. The diagram below shows the redesigned model training page.

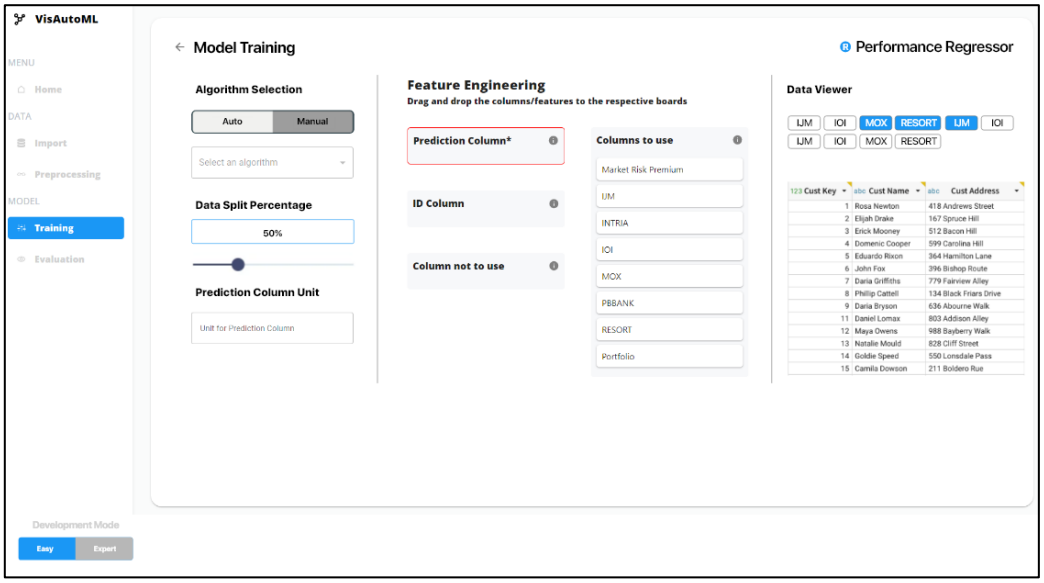


Figure 39 Wireframe of Model Training Page for VisAutoML 2.0

A loading screen equipped with XAI visualisation animations for the next page has been added to improve the hedonic quality of the tool, and to enhance the transparency and trust among users. These XAI visualisation animations augment user engagement and provide clearer, more detailed, and

interactive explanations of the tool's outputs. For non-expert users, this feature provides a deeper understanding of each functionality and the characteristic of each visualisation, thereby fostering trust and facilitating transparency.

Lastly, the progress bar and the Cancel and Evaluate Model buttons have been added to improve the pragmatic quality and enhance the user experience during the loading phase. The progress bar provides real-time updates on the model training process, imparting a sense of control and understanding to the users. The buttons provide users the flexibility to cancel the process or evaluate the model at any point, offering more control to users, especially non-experts. The diagram below show the added loading screen.

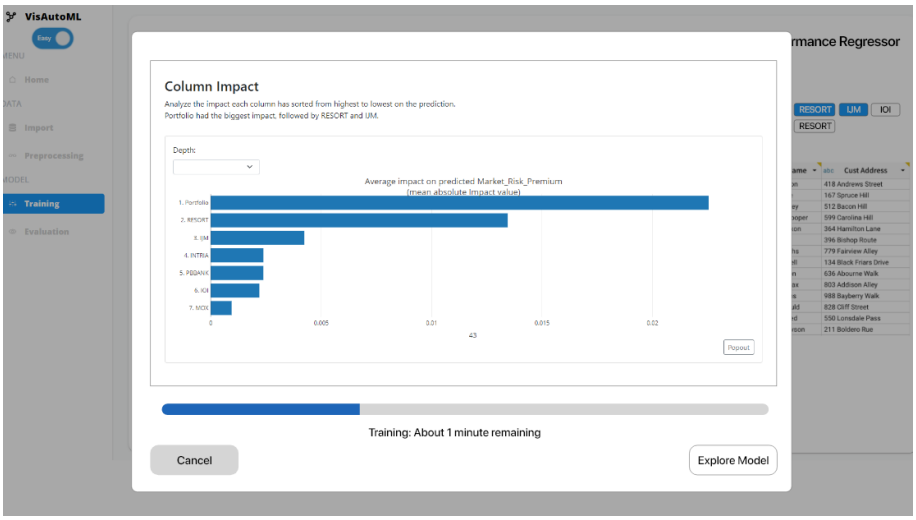


Figure 40 Wireframe of Loading Screen for VisAutoML 2.0

Model Evaluation

The redesigned Model Evaluation page incorporates five tabs, each redesigned to enhance the user's understanding and interaction with the model's predictions, thereby fostering greater trust, transparency, and interactivity. This is achieved through three significant improvements:

1. Interactive Visualisations: Every visualisation has been reimagined to be interactive. Users can now engage with the data more directly by selecting, highlighting, and zooming in on specific points of interest. This interactivity makes data exploration more engaging and provides users with a deeper understanding of the data, enhancing both the pragmatic and hedonic qualities of the tool.

2. Tooltips: To assist users in navigating the visualisations, tooltips have been incorporated. Whenever a user hovers over a data point, a tooltip appears to provide additional context or information. This feature enhances the tool's transparency and trustworthiness, offering an extra layer of information to help users understand the visualisations better, especially those who may be less familiar with data interpretation.

3. Onboarding Process: To ensure a smooth user experience, especially for first-time users, an onboarding path has been introduced. This path guides users through each feature and output of the visualisations with a series of prompts, tips, and explanations. This step-by-step guidance aims to reduce potential confusion and increase understanding, resulting in a more user-friendly and intuitive tool.

These improvements are a part of each section of the Model Evaluation page, including the Feature Importance Tab, Classification/Regression Stats Tab, Individual Predictions Tab, What If Tab, and Feature Dependence Tab. Each section offers its unique insights and interactions, all designed with the aim of improving trust, transparency, interactivity, and user experience. Below are the applications of the redesign for each tab and justifications for each in improving the user experience and transparency of the tool.

The first tab, the Feature Importance Tab exhibits an interactive XAI visualisation, the Feature Importance Chart, which provides a clear depiction of the relative importance of each feature in the model. The interactivity of the chart allows users to delve deeper into each feature's role, promoting transparency and trust in the model's predictions. It also enhances the hedonic quality of the tool by making the exploration of complex data more engaging. The Feature Importance tab is displayed in the diagram below.

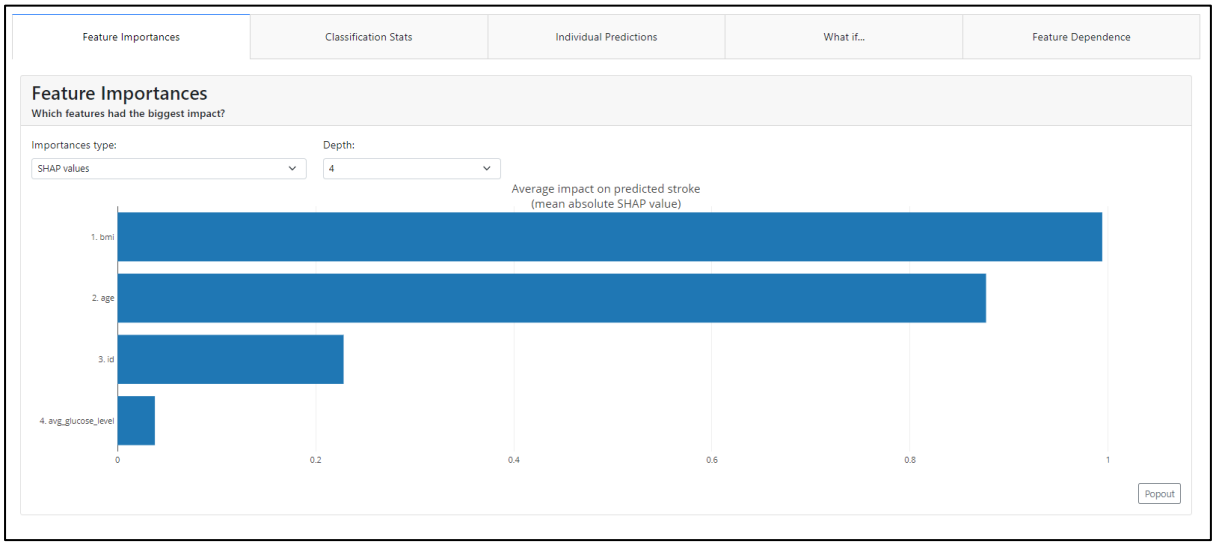


Figure 41 Wireframe of Feature Importance tab for VisAutoML 2.0

Following this, the Classification/Regression Stats Tab offers an extensive evaluation of the model's performance. This section presents crucial performance metrics and includes visualisations that elucidate the model's accuracy. By providing both quantitative and qualitative insights into the model's performance, it enhances the pragmatic quality of the tool. It also fosters trust by offering a comprehensive understanding of the model's performance. The Classification/Regression Stats tab is displayed in the diagram below.

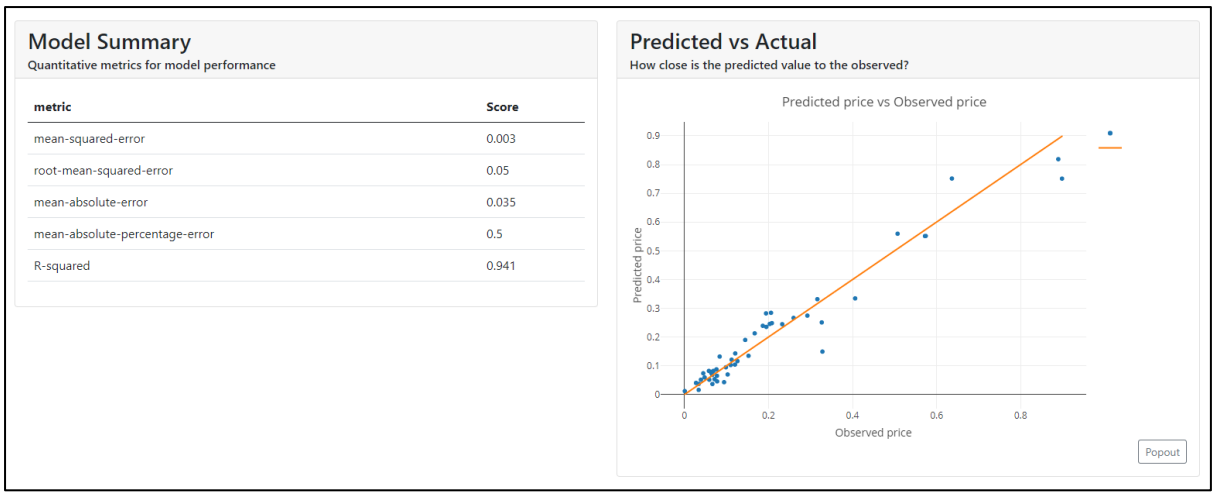


Figure 42 Wireframe of Regression Stats tab for VisAutoML 2.0

Further enhancing the tool's interactivity is the Individual Predictions Tab. This section allows users to interactively explore the predictions for individual rows of data using the Contributions Plot. This XAI visualisation is designed to present the justification behind each prediction, thereby fostering transparency and trust in the model's predictions. It also aligns with the redesign objective of improving hedonic quality by making the tool more engaging. The Individual Predictions tab is displayed in the diagram below.

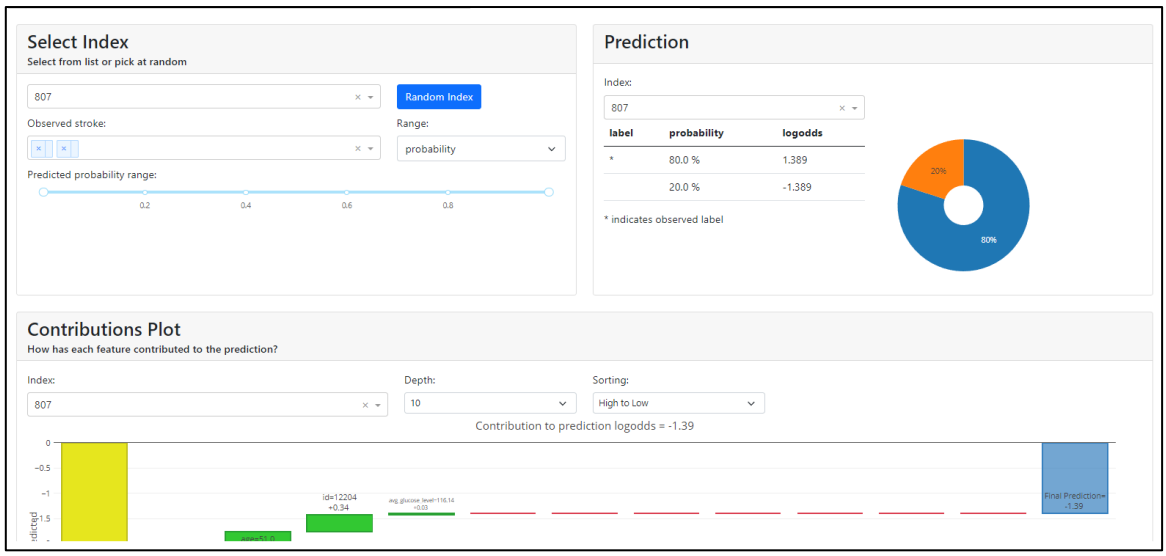


Figure 43 Wireframe of Individual Predictions tab for VisAutoML 2.0

The What If Tab offers a unique exploration experience to the users. It allows them to manipulate the values for a column and observe in real-time how these manipulations affect the model's prediction. This interactive exploration makes the tool more engaging, thereby improving its hedonic quality, and fosters transparency and trust by enabling users to understand how changes in input data influence predictions. The What If tab is displayed in the diagram below.

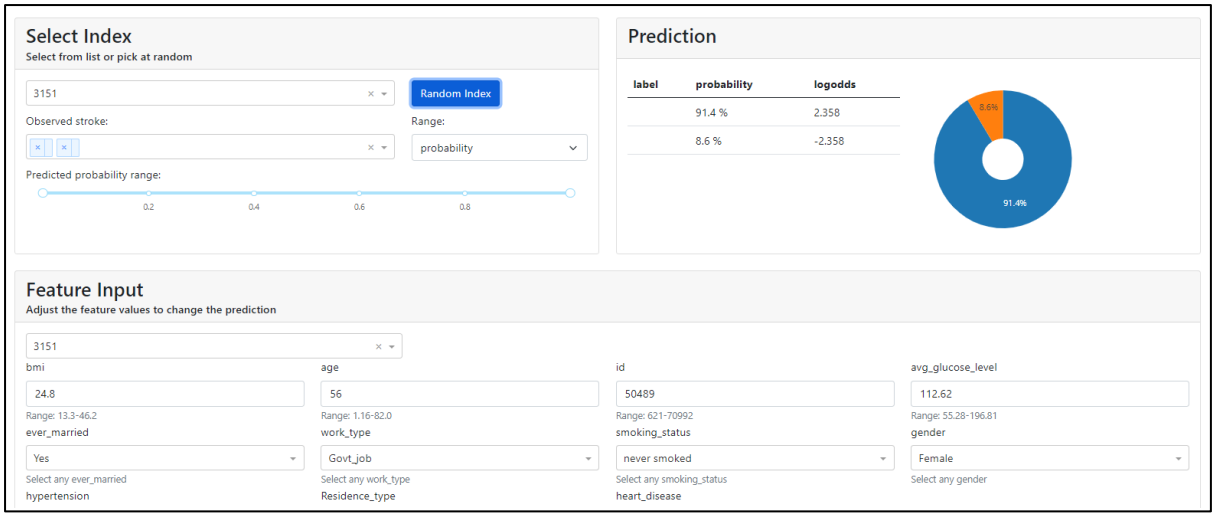


Figure 44 Wireframe of What If.. tab for VisAutoML 2.0

Finally, the Feature Dependence Tab presents a Feature Importance Visualisation alongside a SHAP Dependence Visualisation. These visualisations enable users to explore the group of values that significantly affect the model's predictions. This section caters to the redesign objective of enhancing pragmatic quality by offering detailed insights into the factors influencing the model's predictions. Simultaneously, it improves the tool's hedonic quality by providing an interactive and engaging way to understand these factors. The Feature Dependence tab is displayed in the diagram below.

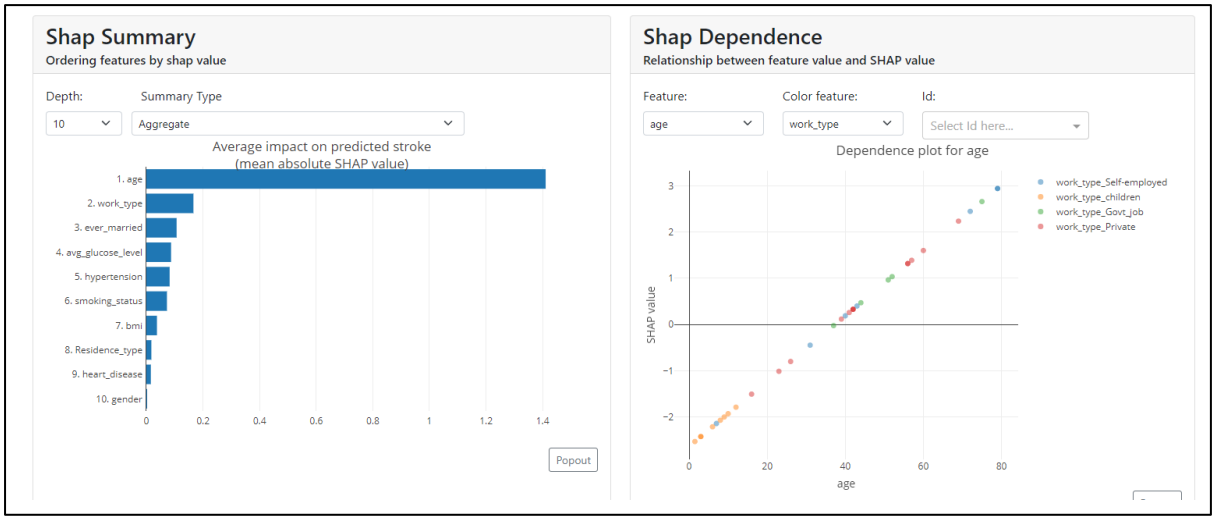


Figure 45 Wireframe of Feature Dependence tab for VisAutoML 2.0

In conclusion, the wireframing process guided the comprehensive redesign of five distinct pages: Home page, the Data Import page, the Data Preprocessing page, Model Training page, and Model Evaluation page. These redesigns, rooted in the redesign objectives of improved user interface, interaction, and trustworthiness, are anticipated to significantly elevate the tool's transparency and usability. This sets the stage for the forthcoming sections on the system requirements, which will delve into the technicalities of bringing these redesigns to life.

4.2.6 System Requirements

The system requirements for VisAutoML 2.0, encompassing software, functional, and non-functional components, are the culmination of a rigorous, iterative process. This process began with a thorough analysis of user requirements and feedback from VisAutoML 1.0, followed by the establishment of clear redesign objectives to address identified needs and enhance the user experience.

Building upon these redesign objectives, an extensive ideation phase was undertaken to generate a wide range of potential solutions and improvements. These ideas were then refined and developed into concrete concepts, each evaluated against the redesign objectives to ensure alignment with user needs and expectations.

These concepts served as the foundation for the wireframe redesign phase, where they were visualised and further refined. This iterative process of ideation, concept development, and wireframe redesign ensured that all enhancements and new features were carefully thought out and tailored to the needs of non-expert users in the machine learning domain.

The functional requirements serve as a comprehensive guide to the system's capabilities, articulating the enhancements and new features that have been integrated based on the ideation and concept development phases. Concurrently, the non-functional requirements outline the characteristics and properties of the proposed system. The system requirements for VisAutoML 2.0, serve as a bridge between the redesign objectives and the practical implementation of *the VisAutoML* tool, effectively encapsulating the enhancements and improvements that have been integrated to elevate the user experience and the tool's effectiveness.

The following are the software requirements of the system:

1. Internet connection
2. JavaScript-enabled web browser

The following are the development tools used:

1. Visual Studio Code
2. Django
3. React JS
4. Flask
5. Plotly
6. NPM
7. Sklearn

4.2.6.1 Functional Requirements

The functional requirements are requirements that define the function or behaviour of the proposed system. Table 41 displays the functional requirements of the proposed system.

Table 23 Functional requirements

No.	Description
FR1	The system will display a home page
FR2	The system will display a list of projects with details such as the learning task, the model used, and the overall score
FR3	The system will allow users to create a new project based on a classification or regression task
FR4	The system allows users to name, open, and delete existing projects
FR5	The system has one video tutorial for the AutoML tool
FR6	The system will embed visual cues like hover texts throughout the system
FR7	The system has a navigation menu consisting of five pages which are home, import dataset, dataset review, model development, and model review pages
FR8	The system allows users to import a dataset (.csv file)
FR9	The system allows users to choose a pre-set dataset (.csv file)
FR10	The system allows users to return to a previous page using a back button on every page except for the home page
FR11	The system allows users to navigate to the next page using the next button on every page except for the model review
FR12	The system calculates and displays the details of the dataset like total rows, total columns, empty rows, empty columns, data type, percent empty and if a column is fit for use
FR13	The system displays interactive visualisations for data quality and alerts for bad data quality
FR14	The system displays a data editor section to allow users to edit the imported dataset
FR15	The system allows users to develop an ML model using a drag-and-drop interface
FR16	The system allows users to choose between an automated or a manual model selection
FR17	The system accepts feature input based on four boards which are the prediction column, identifier column, columns not to use, and columns to use boards
FR18	The system allows only one column to be specified in the prediction column board
FR19	The system will not allow a prediction to be run without designating the prediction column
FR20	The system specifies the identifier, columns to use, and columns to not use boards to be optional boards to be filled
FR21	The system automatically updates the dropdown with either regression or classification models based on the user's choice
FR22	The system accepts unit input for regression learning task
FR23	The system accepts label input for classification learning task
FR24	The system automatically pre-processes the data in the background for model development
FR25	The system automatically presents XAI visualisations in five different tabs which are Feature Importance, Classification/Regression Stats, Individual Predictions, What If, Feature Dependence.
FR26	The system displays feature importance XAI visualisation on the impact tab
FR27	The system displays model summary and confusion matrix/predicted vs actual visualisation on the Classification/Regression Stats tab
FR28	The system displays Prediction and Contributions plot on Individual Predictions tab
FR29	The system displays Prediction and Feature Input components on the What If tab.
FR30	The system displays a Shap Summary and Shap Dependence plot on the Feature Dependence tab.
FR31	The system allows users to return to the home page at the end of the model review
FR32	The system allows users to rerun a model using updated changes
FR33	The system displays a loading screen during model training before the model evaluation section
FR34	The system consists of an Easy/Expert development mode toggle for all five pages which triggers tooltip guides throughout the tool
FR35	The system includes a Data Viewer section to peek into imported dataset on the model training page
FR36	The system includes a data split percentage draggable button on the model training page
FR37	The system includes a data viewer section on the data import page to display the imported data
FR38	The system displays a dashboard on the home page which shows the model performance, model type distribution, model average score, latest used algorithm and models created
FR39	The system includes documentation button links which are accessible on the home page

FR40	The system displays a welcome popup to prompt a user to choose easy/expert development mode

4.2.6.2 Non-Functional Requirements Specifications

The non-functional requirements are requirements that define the attributes of the proposed system such as the usability and reliability of the system. Table 42 displays the non-functional requirements of the IoT prototyping toolkit.

Table 24 Non-functional requirements

No.	Description of Non-Functional Requirements
NFR1	**Product Requirement** • The system deploys onto a domain and requires an internet connection
NFR2	**Reliability** • The system must be able to read and write data from the database • The system shall run a Django and Flask server simultaneously
NFR3	**Usability** • The interface should be intuitive and usable for non-experts with the help of visual cues, tutorials, and instructions
NFR4	**Performance** • The system must be able to interactively respond to inputs • The system should support running machine learning algorithms
NFR5	**Extensibility** • The system shall support future extension with new functionalities
NFR6	**Legal** • The system should not have any legal issues involving intellectual property infringement, user privacy violation, and any use of restricted technology

4.2.7 Prototype Development

The VisAutoML tool's redesign implementation was developed using a combination of React JS and Django, two powerful frameworks for frontend and backend development, respectively. This combination enabled the creation of a highly interactive and efficient system. The frontend interface was developed using ReactJS, a popular JavaScript library known for its efficiency and flexibility. React's component-based architecture allowed the creation of reusable UI elements, which enhanced the consistency and maintainability of the code. The Material-UI (MUI) library was used in conjunction with ReactJS to provide a set of pre-designed components. These components, such as web pages, pop-ups, list items, charts, and tables, were visually represented in the wireframe redesign phase in 6.7. These components made the development process smoother and faster, and also ensured a modern and user-friendly interface.

On the other hand, the backend, responsible for data management and logic, was built using Django, a high-level Python framework that encourages rapid development and clean, pragmatic design. Django's "models.py" file was instrumental in defining the structure of the database. It outlined the model types to be stored, typically represented by their corresponding shortforms, and the model class, which defined the columns in the model database. Columns included information like model name, model type, algorithm name, etc.

The "views.py" file in Django, depicted in Figure 55, played a crucial role in handling user requests and returning suitable responses. Specifically, the ModelViewSet class was responsible for managing the display of existing models, the creation of new models, and the deletion of models in the web application. This class was key in facilitating user interactions with the tool's backend.

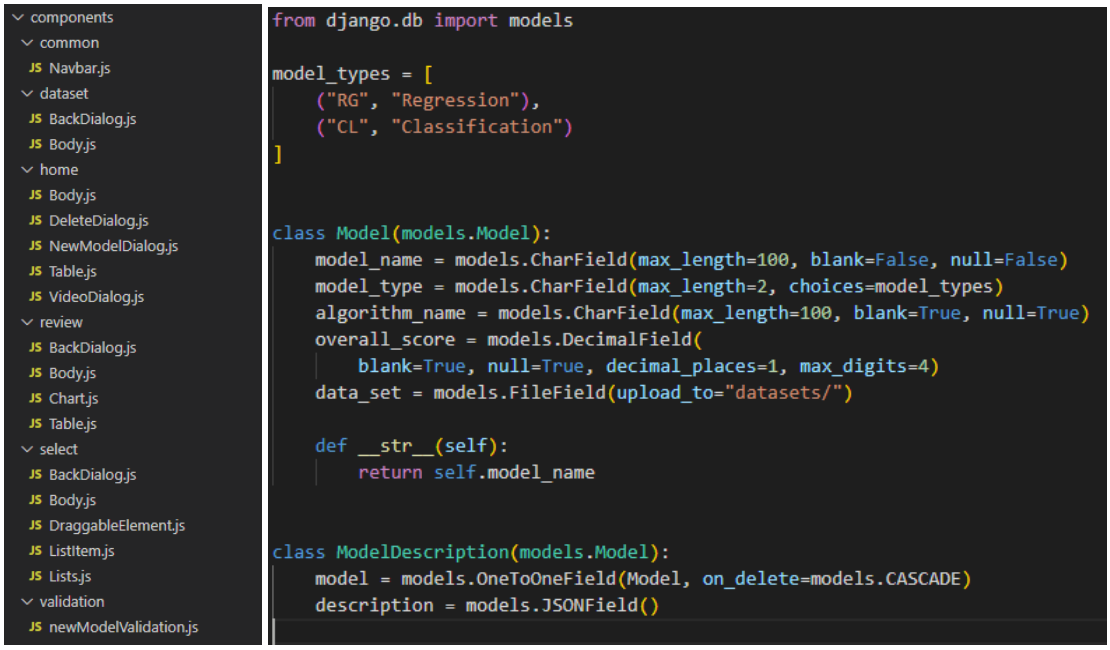


Figure 46 React components (left) Django models.py (right)

```python
class ModelViewSet(viewsets.ViewSet):

    def list(self, request):
        models = Model.objects.all()
        serializer = ModelSerializer(models, many=True)
        return Response(serializer.data)

    def create(self, request):
        try:
            serializer = ModelSerializer(data=request.data)
            if serializer.is_valid():
                model = serializer.save()
                result = get_review(model.data_set.path)
                description = ModelDescription.objects.create(
                    model=model, description={})
                description_serializer = ModelDescriptionSerializer(description)
                return Response(
                    {"response": result, "model": serializer.data, "description": description_serializer.data})
            return Response(serializer.errors, status=status.HTTP_400_BAD_REQUEST)
        except Exception as e:
            traceback.print_exc()

    def destroy(self, request, pk):
        Model.objects.get(id=pk).delete()
        models = Model.objects.all()
        serializer = ModelSerializer(models, many=True)
        return Response(serializer.data)
```

Figure 47 Snapshot of Views.py file

Alongside React JS and Django, other critical components in the development of the VisAutoML were Pandas and Scikit-learn, both of which are powerful Python libraries used extensively in data science.

Pandas is renowned for its robust data manipulation and analysis capabilities. It was employed in the initial stages of the machine learning pipeline, particularly for data preprocessing. This involved cleaning the data, handling missing values, scaling numerical values, encoding categorical columns, and splitting the dataset into input features and target labels. These steps, all crucial in shaping the raw data into a form suitable for machine learning, are illustrated in Figure 57.

Post data preparation, Scikit-learn (Sklearn) was brought into play. This library is a go-to solution for many data scientists due to its comprehensive collection of machine learning algorithms and tools. It was used throughout the remainder of the machine learning pipeline, encompassing model training, evaluation, and selection.

The tool was designed to support a wide range of algorithms, which were incorporated into the Sklearn pipeline. The algorithms used in this iteration included logistic regression, random forest classifier, gradient boosting classifier, decision tree classifier, LGBM classifier, XGBC classifier, random forest regression, gradient boosting regression, bagging regressor, and extra trees regressor.

To ensure the selection of the most effective model, an automatic model selection feature was incorporated. This feature evaluated each algorithm against the dataset, and the one yielding the highest overall score was chosen for the XAI visualisation. This automated process eliminated the need for manual selection, thereby increasing efficiency and reducing the likelihood of human error.

```
def prepare_model(drop, IDColumn):
    df = pd.read_csv(train_csv)
    has_header = csv.Sniffer().has_header(open(train_csv).read(2048))

    # id column set
    if (IDColumn != ""):
        IDColumn = IDColumn.replace(' ', '_')
        df.set_index(IDColumn, drop=True, inplace=True)
        df.index.name = IDColumn

    # predict to columns
    result = predict.replace(' ', '_')

    # convert list drop
    if drop != []:
        converter = lambda x: x.replace(' ', '_')
        drop = list(map(converter, drop))
        drop

    # space to underscore for all headers
    if has_header == False:
        df.columns = ['co_' + str(i + 1) for i in range(len(df.iloc[0].values))]
    df.columns = df.columns.str.replace(' ', '_')

    # drop unused columns
    if drop != []:
        df.drop(columns=drop, axis=1, inplace=True)
```

Figure 48 Snapshot of data preparation steps

```
b = []
a = [LogisticRegression, RandomForestClassifier, GradientBoostingClassifier, DecisionTreeClassifier,
     LGBMClassifier, XGBClassifier]

for i in a:
    model = i().fit(x_train, y_train.values.ravel())

    # testing training accuracy
    from sklearn import metrics
    from sklearn.metrics import balanced_accuracy_score
    from sklearn.metrics import average_precision_score
    from sklearn.metrics import roc_auc_score
    from sklearn.metrics import brier_score_loss

    y_pred = model.predict(x_test)
    x = metrics.accuracy_score(y_test, y_pred)
    y = balanced_accuracy_score(y_test, y_pred)
    z = average_precision_score(y_test, y_pred)
    z1 = roc_auc_score(y_test, y_pred)
    z2 = brier_score_loss(y_test, y_pred)
    b.append((x + y + z1) / 3 / 0.01)
    print(x, y, z, z1, z2)

print(b)
best_score = max(b)
print("best_score ", str(best_score))
index = b.index(best_score)
```

Figure 49 Snapshot of sklearn ML code

The XAI dynamic visualisations are based on Sklearn, Plotly, Shap, and Dash libraries. The models will be developed and the Shap values would be obtained using the shap.explain() function. Plotly Dash was used to create a web application that allows users to interact with the SHAP values and explore the model's explanations in more detail. This can be done using Dash components such as dropdown menus, sliders, and graphs. The Plotly Dash's "app.layout" and "app.callback" functions were used to define the layout and behaviour of the web application and use the SHAP values and the model's predictions to update the visualisations in real time. Below is a snapshot of a function implementing the confusion matrix plot using Plotly Dash.

```
fig = go.Figure(data, layout)
annotations = []
for x in range(cm.shape[0]):
    for y in range(cm.shape[1]):
        top_text = f"{cm_normalized[x, y]}%" if percentage else f"{cm[x, y]}"
        bottom_text = f"{cm_normalized[x, y]}%" if not percentage else f"{cm[x, y]}"
        annotations.extend([
            go.layout.Annotation(
                x=fig.data[0].x[y],
                y=fig.data[0].y[x],
                text=top_text,
                showarrow=False,
                font=dict(size=20)
            ),
            go.layout.Annotation(
                x=fig.data[0].x[y],
                y=fig.data[0].y[x],
                text=f" <br> <br> <br>({bottom_text})",
                showarrow=False,
                font=dict(size=12)
            )]
        )
longest_label = max([len(label) for label in labels])
fig.update_layout(annotations=annotations)
fig.update_layout(margin=dict(t=40, b=40, l=longest_label*7, r=40))
return fig
```

Figure 50 Snapshot of Plotly Dash code

The dynamic visualisations for Explainable Artificial Intelligence (XAI) in the tool are underpinned by the integration of Sklearn, Plotly, Shap, and Dash libraries. These libraries collectively facilitate the creation of a comprehensive and interactive web application, empowering users to delve deeper into the model's explanations. Model development and the computation of Shapley Additive Explanations (SHAP) values are executed using Sklearn and Shap libraries, respectively. The shap.explain() function is specifically employed to derive SHAP values, which provide a measure of the contribution of each feature to the prediction for each instance.

The interactive web application is constructed using Plotly Dash, a high-level Python framework built specifically for data visualisation. Dash components such as dropdown menus, sliders, and graphs enable users to interact with the SHAP values and explore the model's explanations in more detail.

The layout and behaviour of the web application are defined using Plotly Dash's "app.layout" and "app.callback" functions. These functions also facilitate the real-time updating of visualisations based on the SHAP values and the model's predictions.

An example of how these libraries are used in practice is illustrated in the snapshot above, which depicts a function implementing a confusion matrix plot using Plotly Dash. This plot provides an effective visualisation of the performance of the model, offering insights into the number of correct and incorrect predictions.

5 Findings

5.1 Introduction

This chapter presents the findings derived from the research methodologies detailed in Chapter 3. The studies conducted aimed to address the research questions and objectives by investigating non-expert users' perceptions of Machine Learning (ML) development, gathering requirements for the VisAutoML tool, and evaluating its usability and transparency through iterative prototype assessments. The results presented herein provide empirical evidence regarding the effectiveness of the proposed design principles and the potential of VisAutoML to enhance the ML development experience for non-expert users.

The methodologies employed, guided by the iterative user-centred design (UCD) framework (Figure 15), included an Extended Technology Acceptance Model (TAM) study, a comparison study between VisAutoML 1.0 and an existing AutoML tool, and in-depth usability and transparency evaluations of both the initial VisAutoML 1.0 prototype and the refined VisAutoML 2.0. The data collected through mixed-methods approaches, including questionnaires and interviews, were subjected to rigorous quantitative and qualitative analysis techniques, as described in Chapter 3.

This chapter is structured to present the results from each of these studies in dedicated sections. Section 5.2 will present the findings from the Extended Technology Acceptance Model (TAM) study, detailing insights into factors influencing non-expert users' acceptance of an AutoML tool. Section 5.3 will present the results of the comparison study between VisAutoML 1.0 and H2O AutoML, providing a benchmark assessment of the initial prototype's performance in terms of usability, transparency, and knowledge gain. Section 5.4 will present the findings from the in-depth usability and transparency evaluation of VisAutoML 1.0, offering a detailed analysis of user experiences with the initial prototype. Subsequently, Section 5.5 will present the results of the evaluation of the refined VisAutoML 2.0 prototype, assessing the impact of the design iterations on usability and transparency. Finally, Section 5.6 will provide a summary of the key findings across all studies presented in this chapter.

5.2 Extended Technology Acceptance Model (TAM) Study

5.2.1 Participant Demographics

A total of 73 participants participated in the questionnaire. The snowball sampling method was used to recruit the participants. After selecting an initial group of individuals among non-experts in the field of AI, the participants then indicated other potential members with similar characteristics to participate in the study. The demographic information is shown in Table 9. There were 36 males and 37 females. There were 57 participants under 30 years old, 2 between 31 and 40, 11 between 41 and 50, and 3 above 51 years of age. The mean participant age was 27 years old (SD = 10.14). The education levels were secondary school and above. Of these, 18 held a diploma, 38 held a bachelor’s degree, 10 held a master’s degree, and 1 held a doctorate. Academically, the majority of participants (89%) were from a Computer Science background, while the remaining participants were from Engineering, Business, Psychology, and Finance fields (Figure 17). In terms of machine learning experience, 65.8% of participants rated themselves at the beginner level (rating 1), and 34.2% rated themselves slightly more experienced (rating 2), with none reporting higher levels of expertise (Figure 18).

Table 25 Demographic information of the participants.

Gender		Age		Education	
Item	**No.**	**Item**	**No.**	**Item**	**No.**
Female	37	Under 30	57	Secondary School	6
Male	36	Between 31 and 40	2	Diploma	18
		Between 41 and 50	11	Bachelor's Degree	38
		Above 51	3	Master's Degree	10
				Doctorate Degree	1
Total	73		73		73

The sample size of 73 participants is considered adequate for a qualitative Technology Acceptance Model (TAM) study, particularly at the exploratory stage. Prior TAM research typically employs sample sizes ranging from around 50 to 200 participants to reliably detect trends in user perceptions (Turner et al., 2010; Rafique et al., 2020). Given the study's focus on non-expert users' perceptions of machine learning model development, the demographic profile aligns well with the intended population: participants had generally low levels of machine learning expertise but varied academic qualifications and fields of study. While the sample was heavily skewed towards participants with a Computer Science background, the presence of individuals from Business, Psychology, and Finance introduces some interdisciplinary diversity. However, the concentration of technically-oriented participants could slightly bias perceptions of system usability and transparency, potentially making the results less generalizable to completely non-technical populations. Furthermore, the snowball sampling method, while practical for reaching non-expert users, could introduce a degree of selection bias, as participants may have recommended peers with similar backgrounds. Despite these considerations, the sample was reasonably diverse and suitably representative for the primary goal of evaluating VisAutoML's perceived usefulness, ease of use, and adoption likelihood among non-expert users with varied but low-to-moderate technical backgrounds.

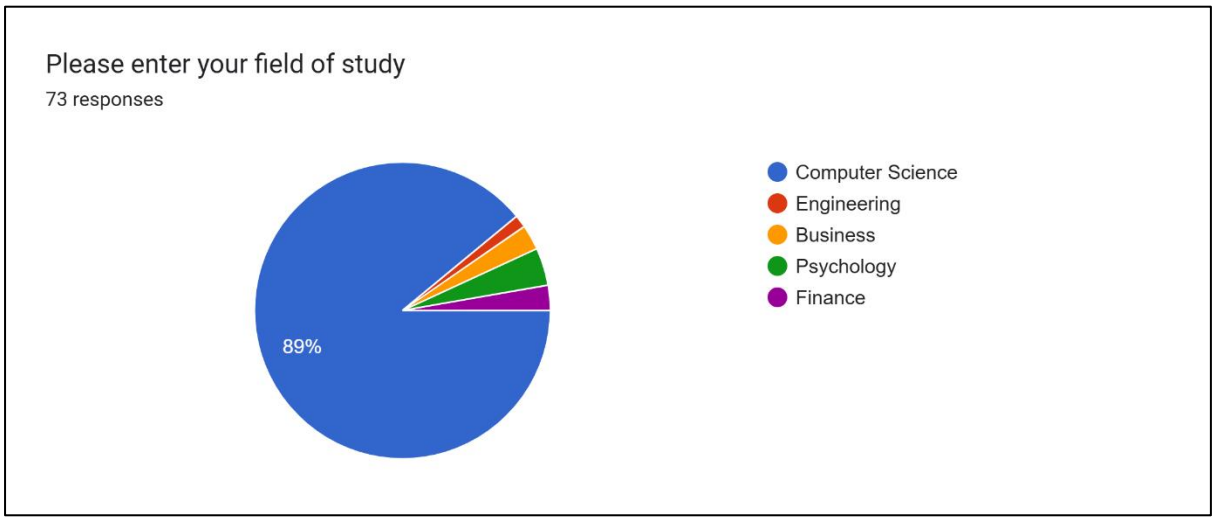


Figure 51 Participant's field of study

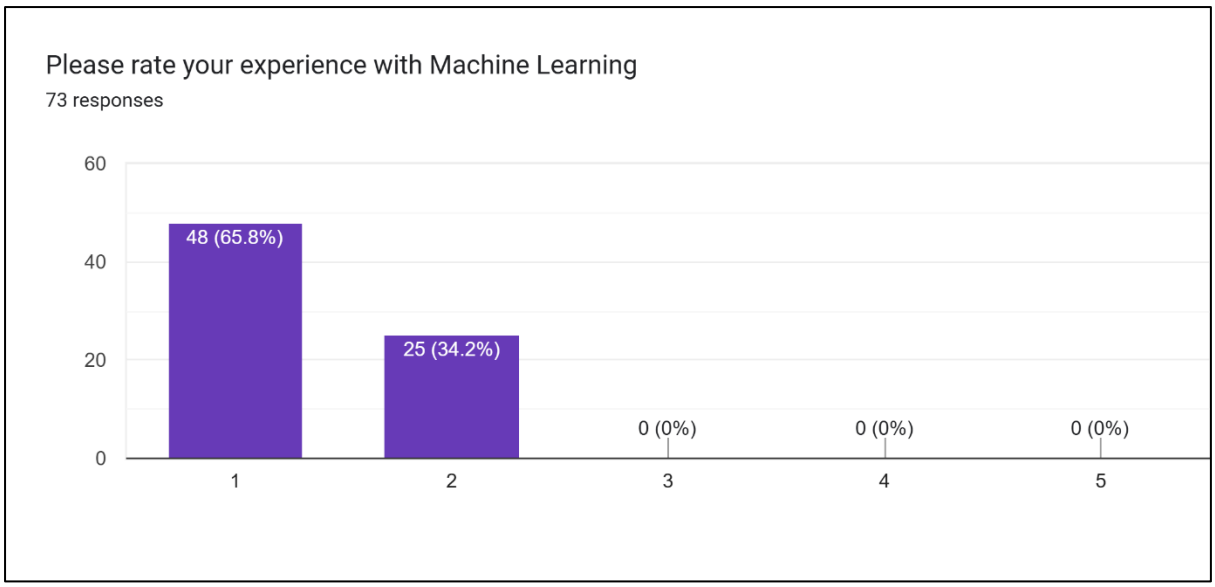


Figure 52 Participant's experience with Machine Learning

5.2.2 Quantitative Results

5.2.2.3 Standardization and Reliability Analysis

To analyse the effectiveness of the questionnaire, data standardization was conducted on the results. The descriptive statistics for each question and construct were calculated and documented. The Cronbach's alpha coefficient is used as an estimate of the reliability of the questionnaire (Yan & Yibing, 2010). The Cronbach's alpha coefficient for each construct was calculated to show the convergent validity and internal reliability of each construct. The results are shown in Table 10. From the table, it is observed that the average of all factors is greater than 5, suggesting the assumptive constructs were typical. As for the convergent validity and internal reliability, the Cronbach's alpha for each factor is greater than 0.7 which is acceptable (>0.7 (Yan & Yibing, 2010)). From this initial analysis, it is concluded that the data are reliable measures for their constructs.

Table 26 Questionnaire standardization and reliability analysis.

Construct	Question	AVG	SD	Construct Avg	Cronbach's Alpha
PU	PU1	6.04	1.01	6.14	0.93
	PU2	6.14	0.90		
	PU3	6.05	0.96		
	PU4	6.10	0.96		
	PU5	6.10	1.00		
	PU6	6.23	0.99		
	PU7	6.32	0.83		
PEOU	PEOU1	6.18	1.08	5.44	0.75
	PEOU2	5.53	1.42		
	PEOU3	5.68	1.31		
	PEOU4	5.30	1.23		
	PEOU5	4.63	1.33		

	PEOU6	5.32	1.31		
AT	AT1	6.03	1.07	5.93	0.87
	AT2	5.95	1.00		
	AT3	5.92	0.95		
	AT4	5.84	1.13		
BI	BI1	5.84	1.08	5.51	0.87
	BI2	5.33	1.19		
	BI3	5.37	1.24		
PA	PA1	5.58	1.20	5.74	0.86
	PA2	5.85	1.09		
	PA3	5.47	1.27		
	PA4	5.60	1.30		
	PA5	6.19	0.95		
ENJ	ENJ1	6.01	0.98	6.09	0.88
	ENJ2	6.04	0.92		
	ENJ3	6.21	0.99		

5.2.2.4 Correlation Analysis

To discuss the strength and directionality of the correlation between two variables, correlation analysis was applied to the data. Pearson correlation analysis was applied to identify the correlations among the constructs of the theoretical model. The averages for each construct category were used in the Pearson correlation analysis and are shown in Table 11.

Table 27 Correlation between constructs. **p < 0.01.

	PU	PEOU	AT	BI	PA	ENJ
PU	1.000					
PEOU	0.655**	1.000				
AT	0.762**	0.671**	1.000			
BI	0.545**	0.683**	0.746**	1.000		
PA	0.692**	0.603**	0.720**	0.583**	1.000	
ENJ	0.719**	0.661**	0.721**	0.616**	0.667**	1.000

1. As shown in Table 11, there was a strong positive correlation between the perceived usefulness of using the system and respectively the perceived ease of use (r=0.655, p<0.01), attitude towards using the system (r=0.762, p<0.01), the social influence (r=0.692, p<0.01) and perceived enjoyment (r=0.719, p<0.01). Additionally, there was a moderate positive correlation between the perceived usefulness and the behavioural intention of using the system (r=0.545, p<0.01). In descending order, this indicated that the higher level of utility demonstrated by the system, the better the attitudes, the perceived enjoyment, the social influence, the perceived ease of use, and the willingness of using the system.

2. As for the perceived ease of use of using the system, there was a strong positive correlation with respectively the perceived usefulness (r=0.655, p<0.01), the attitude towards using the system (r=0.671, p<0.01), the willingness of using the system (r=0.683, p<0.01), the social

influence of using the system (r=0.603, p<0.01), and the perceived enjoyment (r=0.661, p<0.01). In descending order, this indicated that the more intuitive the system is designed, the higher the willingness, attitudes, the perceived enjoyment, level of usefulness, and the social influence is perceived from using the system.

3. As indicated in Table 11, the attitude towards using the system was strongly correlated with respectively the perceived usefulness (r=0.762, p<0.01), the perceived ease of use (r=0.671, p<0.01), the willingness of using the system (r=0.746, p<0.01), the social influence of using the system (r=0.720, p<0.01), and the perceived enjoyment (r=0.721, p<0.01). In descending order, this indicated that the attitude towards using the system is strongly influenced by the level of usefulness perceived, willingness of using the system, the perceived enjoyment, the social influence, and the perceived ease of use.

4. As for the willingness of using the system, there was a strong positive correlation with the attitude towards using the system (r=0.746, p<0.01), the perceived ease of use (r=0.683, p<0.01), and the perceived level of enjoyment (r=0.616, p<0.01). Additionally, there was a moderate positive correlation between the willingness of using the system and respectively, the perceived authority (r=0.583, p<0.01), and the perceived level of usefulness (r=0.545, p<0.01). In descending order, this indicated that the willingness of using the system is strongly influenced by the attitudes, ease of use, level of enjoyment, social influence, and the perceived level of usefulness of using the system.

5.2.2.5 Regression Analysis

Regression analysis is used to explore the relationship between independent and dependent variables. In this study, linear regression was used to determine the strength of the relationship between the dependent and independent constructs to infer the causal relationship. When the effect of more than one independent variable is determined by the dependent variable the regression then becomes multiple regression. Table 11 shows the results of a multiple regression analysis on the intention to use / the BI construct. With an R^2 value of 0.636, therefore, the hypothetical model has an explanation capability of 63%, and the p-value of 0.001 is less than 0.05, which reaches the significance level (Weng et al., 2018). $R^2 = 0.636$, F (23.408), $p < .05$, Adj. $R^2 = .609$.

Table 28 Model Summary

	R	R Square	Adjusted R Square	Std. Error of the Estimate	F	Sig.
BI	.797[a]	.636	.609	.655	23.408	.001[b]

a. Dependent Variable: BI
b. Predictors: (Constant), PU, PEOU, AT, PA, ENJ

Multiple regression analysis was conducted on the remaining constructs to validate the hypothesis and conduct path verification of the hypothetical model. The veracity and reliability of the model were examined by testing the hypothesised relationships between the constructs as shown in Figure 17 and Table 13.

Table 29 Analysis of the significance of path coefficient.

Hypothesis	Beta β	Sig.	R	R^2	Supported?

(H1) PEOU → PU	0.229	0.029	0.790	0.625	Y
(H2) PEOU → AT	0.208	0.039	0.815	0.664	Y
(H3) PU → AT	0.428	<0.001	0.815	0.664	Y
(H4) AT → BI	0.746	<0.001	0.746	0.556	Y
(H5) ENJ → PU	0.357	0.002	0.790	0.625	Y
(H6) ENJ → AT	0.275	0.013	0.815	0.664	Y
(H7) PA → PU	0.315	0.003	0.790	0.625	Y
(H8) PA → PEOU	0.603	<0.001	0.603	0.364	Y

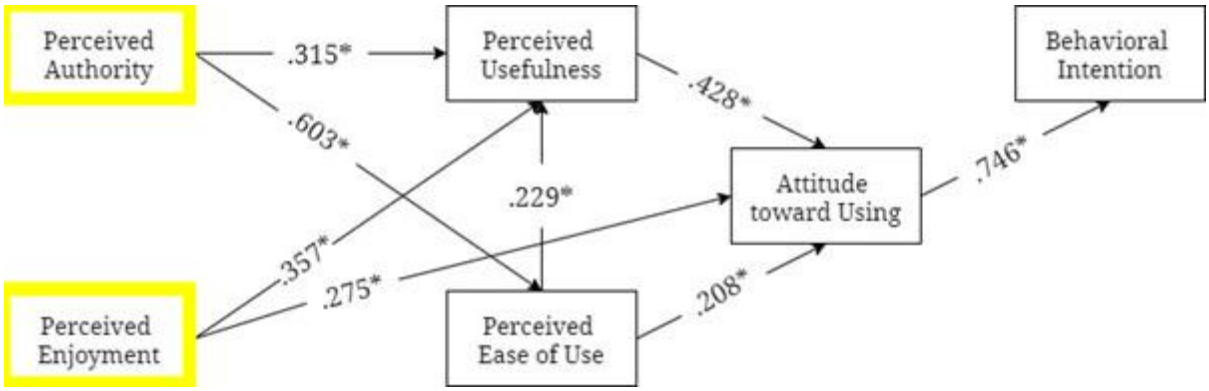


Figure 53 Path verification, *p <0.05.

As depicted in Table 13, PEOU has a positive influence on PU and AT (c=0. 208*, 0.229**). PU has a positive influence on AT (c=0.428*). AT has a positive influence on BI (c=0.746*). ENJ has a positive influence on PU and AT (c=0.357*, 0.275*). PA has a positive influence on PU and PEOU (c=0.315*, 0.603*). Furthermore, the significance for each construct is less than 0.05 which is significant. This confirms the findings from the correlation analysis. Therefore, all hypotheses were supported.

Summarizing the findings, the intention of using the proposed system was significantly affected by the attitude towards using the system, which is consistent with past studies (Bakhit Jaafreh, 2018; Lee et al., 2003; Patil, 2017). The attitude towards using the system was directly affected by perceived ease of use, perceived usefulness, and perceived enjoyment.

This suggests that (1) participants would prefer an intuitive software that is easy to use and (2) participants would only choose to use and continue using the system if they are convinced on the utility and gained benefit it adds in their daily life and (3) participants preferred an enjoyable development. The findings that users would continue using the proposed system if there was an inherent benefit are consistent with (Yang et al., 2018). The TAM model findings indicate the need to focus on improving these 3 constructs in the build phase to encourage positive user attitude towards the system. Of the three constructs, perceived usefulness shows a much stronger effect than perceived ease of use and perceived enjoyment. This implies that the utility the system provides to the user will encourage user's attitude when interacting with the system.

Moreover, the findings revealed that the perceived usefulness is significantly affected by the perceived enjoyment, perceived ease of use and perceived authority. This implies that (1) the greater

the enjoyability of the system the greater the perceived usefulness of the system and (2) the higher the intuitiveness of the system the higher the perceived usefulness of the system and (3) the greater the social influence towards using the toolkit, the higher the perceived usefulness of the system. Of the three, perceived enjoyment shows the strongest effect than perceived ease of use and perceived authority. This suggests that the level of enjoyment experienced by the user directly affects the perceived level of usefulness the system provides.

Lastly, the findings showed that the perceived ease of use shares a strong correlation and linear relationship with perceived authority. This suggests that the higher the social influence of a system onto a user the higher the perceived ease of use suggested of the system. It can be concluded, that from the extended TAM model analysis, greater focus must be allocated to developing the system such that:

I. The system includes beneficial use cases such as preinstalled datasets, useful knowledge and meaningful insights.

II. The system offers exploratory analysis through a variety of visualisations.

III. The system offers high levels of expressiveness, allowing users to customise the projects according to their needs.

IV. The system is built with a robust UI/UX design such that the system is highly usable, intuitive, and enjoyable.

5.2.2.6 Questionnaire Responses

The questionnaire also included several multiple-choice, ranking, and subjective questions to understand the participant's preferences for when they potentially interact with the system. Based on the qualitative data collected, answers for each question were analysed and grouped in respective thematic themes.

When participants were asked "What constitutes a user-friendly system to you?" and had to rank their answers among "Good user interface", "No programming required" and "No hardware wiring" from highest to lowest priority. The option with the highest frequency as first choice was "Good user interface" (49 responses). Next was "No programming required" (35 responses) as second choice with highest frequency and last was "No hardware wiring" as third choice with highest frequency (31 responses).

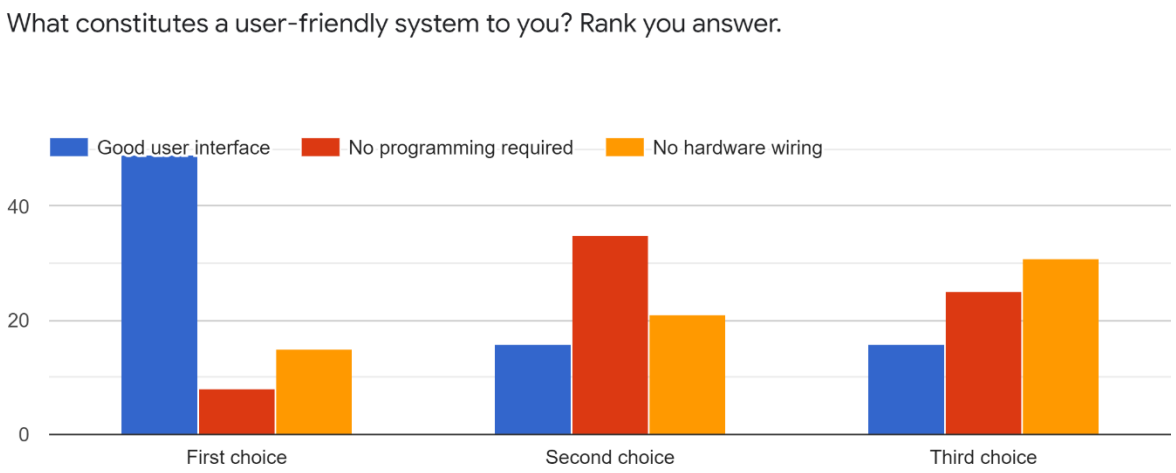


Figure 54 User friendly characteristic.

Furthermore, the questionnaire asked participants their preference between types of automated visualisation implementations or provide their own ideas on the matter. The majority of participants opted for a fully automated process (36 responses). The second most opted answer was the semi-automated process (34 responses). Three opinions were voiced, two, were the incorporation of customised solutions, and one was a combination of the two types of implementations.

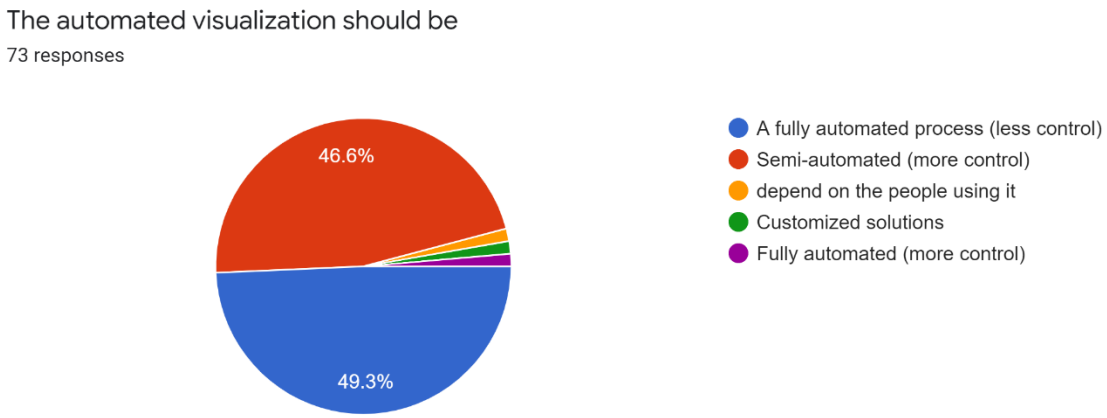


Figure 55 Learning AI.

Next, the questionnaire asked participants their preferred method in learning how to interact with the system or provide their own opinions on the subject. The majority of participants opted for a step-by-step tutorial in the system (45 responses). The second most opted answer was through the use of video tutorials (20 responses). The least opted option, opting the guide of an instruction manual (7 responses). One opinion suggested a system that did not require any tutorials or external guidance when interacting, suggesting a highly intuitive and easily understandable system.

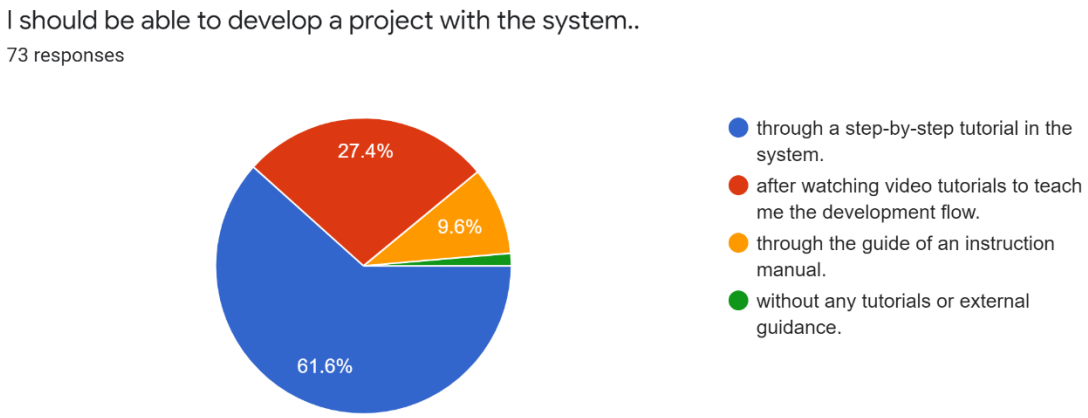


Figure 56 Learning how to use the system.

Further, when participants were asked "What main aspect of your life or job do you/would you track?", the majority answered financial-related data (63%, 46 responses). 16% answered health-related data (12 responses), 10% answered personal emotional data (8 responses), 6% answered personal goal tracking (5 responses) and 2% answered social media-related data (2 responses). The responses are recorded in Table 16.

Table 30 Questionnaire responses analysis.

Q. What main aspect of your life or job do you/would you track?		
Answer category	Responses	Percentage
Financial data	46	63%
Health related	12	16%
Emotions	8	10%
Goal tracking	5	6%
Social media related	2	2%
Total	73	100%

Next, the questionnaire asked participants about their preferred tool/device to track their data. The majority of participants opted for mobile applications (55 responses). The second most opted answer was tied between web applications and wearable devices (8 responses for each option). Two opinions were provided which are the use of Trello and the utilisation of cloud services.

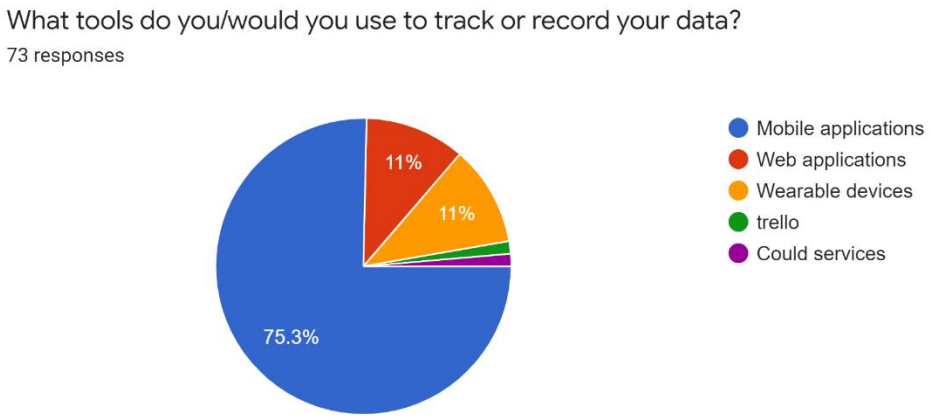


Figure 57 Learning how to use the system.

Additionally, when participants were asked "What are current limitations associated with the tools you would use to record data?", the majority answered bad system design or poor user interface design (41%, 30 responses). 19% answered limited functionality and options (14 responses), 12% answered high cost/ expensive (9 responses), 4% answered privacy and safety concerns (3 responses), 4% answered lack of domain knowledge (3 responses) and 16% answered none or provided no answer (12 responses). The responses are recorded in table 17.

Table 31 Questionnaire responses analysis.

Q. What are the current limitations associated with the tools you would use to record data?		
Answer category	Responses	Percentage
Not user friendly	30	41%
Limited functionality	14	19%
Expensive	9	12%
Privacy & safety concerns	3	4%
Lack domain knowledge	3	4%
Prone to errors	2	2%
None	12	16%
Total	73	100%

Next, when participants were asked "What would you like to learn/ discover from the collected data about yourself or your environment?", the majority answered health-related insights such as calorie intake, sleeping hours, weight management, anxiety levels, exercise data and mental health (32%, 24 responses). 28% answered financial-related insights such as spending habits and investment performance (21 responses). 27% answered productivity related insights such as study hours, screen time, time optimization, decision making, and life quality (20 responses). 5% answered business related insights such as customer interest, cash flow, and cost optimization (4 responses). 1% answered traffic-related insights (1 response) and another 1% answered safety-related insights, specifically natural disaster warning related (1 response). 2% failed to provide an answer. The responses are recorded in Table 18.

Table 32 Questionnaire responses analysis.

Q. What would you like to learn/ discover from the collected data about yourself or your environment?		
Answer category	Responses	Percentage
Health insights	24	32%
Financial insights	21	28%
Productivity insights	20	27%
Business insights	4	5%
Traffic insights	1	1%
Safety insights	1	1%
None	2	2%
Total	73	100%

Lastly, when participants were asked "How and where would you like to have your information visualised and displayed?", the majority answered on mobile-based applications or smartphone-related platforms (72%, 53 responses). 24% answered both mobile and desktop-based platforms (18 responses) and 2% answered a tangible type of visualisation such as living room and room projections (2 responses). The responses are recorded in Table 19.

Table 33 Questionnaire responses analysis.

Q. What would you like to learn/ discover from the collected data about yourself or your environment?		
Answer category	Responses	Percentage
Mobile based	53	72%
Mobile and desktop based	18	24%
Projection	2	2%
Total	73	100%

5.2.3 Qualitative Results

5.2.3.1 Online Interview Sessions

The interview questions were structured around four constructs, where each construct is intended to explore a specific requirement category of the proposed system. The constructs are perceived ease of use of AI (C1), perceived usefulness of visualisations (C2), and platform ease of use (C3).

The rationale for each construct serves a certain aspect of the proposed system. Perceived ease of use of AI sets to understand participant's notions and opinions on learning and applying AI functionality to the proposed system. Perceived usefulness of visualisations aims to explore the participants perception of the utility of visualisations in the proposed system. Finally, the platform ease of use construct aims to explore the participant's main factors in determining the usability of a system.

Participants were invited to take part in the semi-structured interviews about their perceptions and views on the proposed system with questions structured around the three constructs, for instance, "

Do you think it is easy to learn AI on your own?" (C1), "How would you use these visualisations to draw better insights?" (C2) and "What would discourage you from using a software?" (C3).

Participants were recruited through purposive sampling with emailed advertisements and through word-of-mouth between June 2021 and July 2021. Each interview session lasted for about an hour on Microsoft Teams. The interviews were all conducted in English.

5.2.3.1.1 Participant Demographic

A total of 13 participants took part in the semi-structured interview. The demographic information is shown in Table 20. There were 3 males and 10 females. There were 11 participants between 20 and 25 years old, and 2 participants above 25 years of age. The mean participant age was 20 years old (SD = 7.06). The education levels were secondary school and above. Of these, 1 held a diploma, 9 held a bachelor's degree, and 1 held a master's degree. Academically, the majority of participants (76.9%) were from a Computer Science background, with smaller representations from Engineering, Business, Psychology, and Finance fields (Figure 24). Regarding machine learning experience, 61.5% of participants rated themselves at the beginner level (rating 1), and 38.5% rated themselves as slightly more experienced (rating 2), with none reporting higher levels of expertise (Figure 25). Participants were classified as non-experts in the contexts of AI and information visualisation development.

Table 34 Questionnaire responses analysis.

Gender		Age		Education	
Item	**No.**	**Item**	**No.**	**Item**	**No.**
Female	10	Between 20 and 25	11	Secondary School	2
Male	3	Above 25	2	Diploma	1
				Bachelor's Degree	9
				Master's Degree	1
Total	13		13		13

The sample size of 13 participants is suitable for a semi-structured interview study within a qualitative Technology Acceptance Model (TAM) investigation, where the goal is to explore perceptions in-depth rather than achieve statistical generalization. Given the focus on non-expert users' perceptions of machine learning model development, the demographic profile of the interviewees aligns well with the intended population. While a majority had a Computer Science background, the presence of participants from non-technical disciplines contributes to some diversity of perspectives. The concentration of younger participants (mostly aged 20–25) may influence the generalizability of the findings to older user groups. Nevertheless, the relatively homogenous machine learning experience levels ensure that the study captures the intended novice-user viewpoint for evaluating the usability, perceived usefulness, and transparency of the VisAutoML system. An exploratory approach was applied in the analysis of the interview findings, with the data read and re-read, then coded in accordance with the three main constructs (C1–C3). The findings are summarised according to the respective constructs.

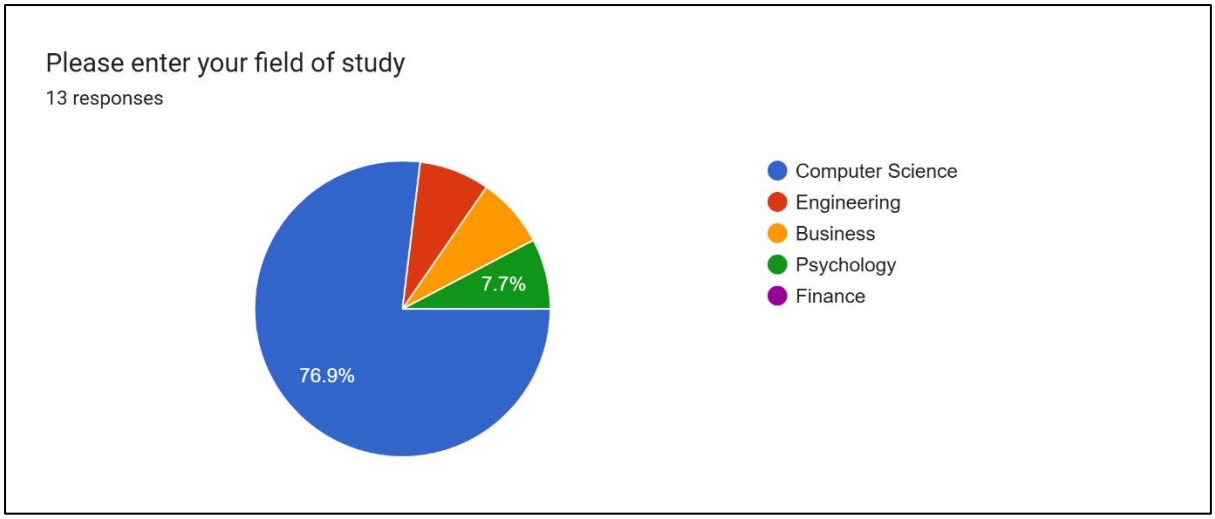


Figure 58 Participant's Field of Study

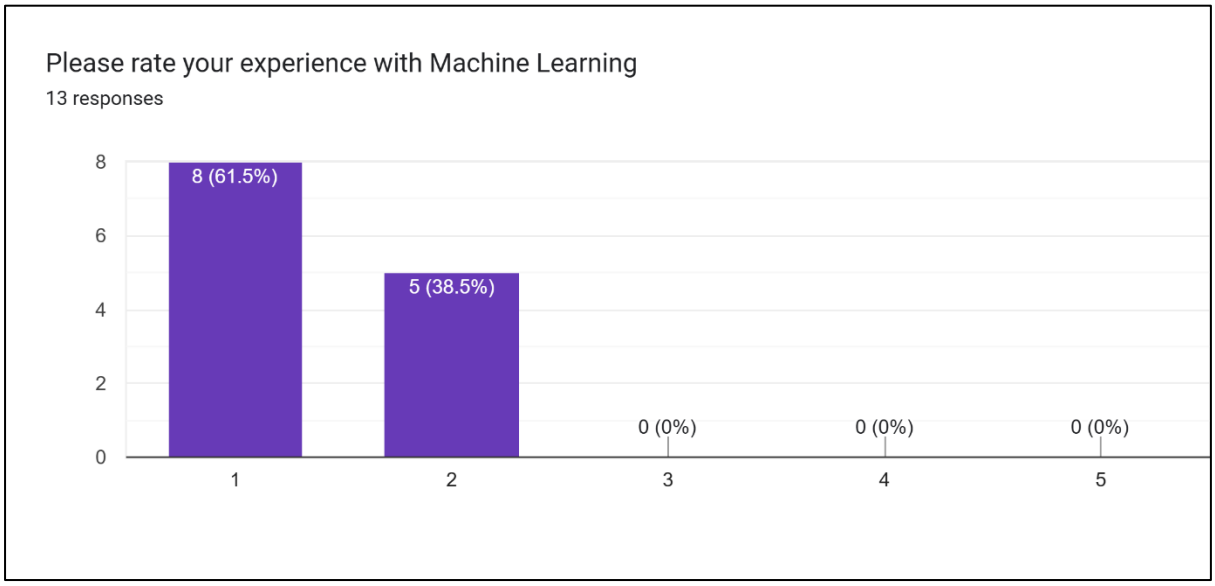


Figure 59 Participant's Experience with Machine Learning

5.2.3.1.2 Results

C1 - Perceived Ease of Use of AI. The perceived ease of use of AI was found to be average as participants lacked the technical application knowledge, which is expected from non-experts e.g., *"What do I understand about AI? I'm not sure, honestly" (P2) and "I don't really have much knowledge on AI maybe just surface knowledge, I know it's about machine learning" (P5). H*owever, participants predominantly expressed their willingness to learn AI e.g., *"Although I obviously don't know much about AI […] but I would love to learn it. Why not? Right" (P11) and "I don't have so much understanding of it […] I think learning it can give me many advantages for my future" (P12)*. Participants also expressed the biggest hurdles when trying to learn AI being terminology, difficult syllabus, different field of study, and no foundational knowledge of AI. Furthermore, participants predominantly agreed that the basic knowledge of AI would be easy to learn by oneself either through video tutorials or books. However, several exceptions mentioned a preference for learning AI in a classroom with educators e.g., *"I think to self-learn it would be difficult, I think I would prefer consulting an expert to teach and explain to me" (P9)*. Lastly, participants outlined their personal expectations towards an AI development system e.g., *"I think creating an AI application by referring to template or*

examples would help" (P1) and "I like how Microsoft integrates all its software together, would expect to see that in an AI system somehow" (P9). Generally, most recommendations were based on existing software systems the participants were already familiar with such as Microsoft, Google, and technical systems.

C2 - Perceived Usefulness of Visualisations. Most participants found visualisations to be beneficial and useful for the interpretation of collected data and that visualisations should be aesthetically pleasing to attract the interest of the viewer. Furthermore, participants provided useful applications of visualisations to their proposed projects e.g., *"I would make a visualisation that uses bar charts and emoticons and based on the set parameters it would show a happy or sad emoticon." (P3) and "visualise blood sugar spike using bar charts and to compare spikes on a daily basis" (P6).* Additionally, participants predominantly preferred to be able to access their visualisations on both desktop and mobile-based applications. Finally, participants expressed a majority preference for semi-automated visualisations, giving the participants customisability and smart prediction of suitable visualisations.

C3 – Platform Ease of Use. Participants generally acknowledge that a user-friendly system consists of good user interface, clear design and intuitive user experience e.g., *"For me, I think a user-friendly system should be not too confusing, not too tedious and the interface has to be simple" (P7) and "the software should have intuitive user interface and minimalist design"* (P11). Most participants mentioned, the usefulness/utility and level of user-friendliness of the system encourages them to continue using a system e.g., "if the system functions well and is beneficial, and makes life easier I would continue using it" (P9) and *"it has to be easy to use, I can get what I want quickly and it is intuitive" (P2).* Lastly, participants mentioned several factors that discourage them from continue using a system, such as advertisements, intimidation by difficult functions, bad user experience, and not leading to any results.

The results of the interview study suggest that participants generally had positive attitudes towards the proposed system, were aware of its importance, and had good theoretical knowledge of the technologies involved based on mainstream AI applications. Based on responses for C1, it is found that participants acquire a basic level of understanding of AI. However, lack the technical and foundational knowledge required for development. Participants also outlined the biggest hurdles when trying to understand AI e.g., (difficult syllabus, confusing terminology, not interesting) and recommendations to make it more intuitive e.g., (curation by example, clear flow of system, good user interface, and useful integration with third-party services). Additionally, for C2, it is evident that participants found visualisations to be beneficial and should be interesting to the viewer. Furthermore, participants proposed the use of the most common chart types in their visualisations e.g., (bar chart, pie chart, line chart, and line graphs). Next, participants expressed a majority preference towards a customisable visualisation generator/ semi-automated visualisation. Besides that, participants also expressed a predominant preference towards visualisations accessible from both desktop and mobile platforms. Finally, in C3, participants outlined several factors associated with a user-friendly system e.g., (good user interface, clear design, online documentation, and intuitive user experience) and factors that hinder usability e.g., (messy interface, errors/bugs, bad user experience, information overload, intimidated by complex functions and non-beneficial functionality).

5.2.4 Derived Requirements

The analysis of the Extended Technology Acceptance Model (TAM) reveals valuable insights into the factors influencing users' intention to use a proposed system. The findings, summarised in Table 14, indicate that Perceived Ease of Use (PEOU) has a positive influence on both Perceived Usefulness (PU) and Attitude Towards using the system (AT), as evidenced by significant coefficients (c=0.208*, 0.229**). PU, in turn, positively influences AT (c=0.428*), and AT has a strong positive influence on Behavioural Intention (BI) to use the system (c=0.746*). Additionally, Enjoyment (ENJ) positively influences PU and AT (c=0.357*, 0.275*), while Perceived Authority (PA) positively influences PU and PEOU (c=0.315*, 0.603*). Table 14 below summarises the findings from the extended TAM model.

Table 35 Summary of findings from extended TAM model

Construct	Positive Influence	Coefficient	Significance	Key Insights
PEOU	PU, AT	0.208*, 0.229**	< 0.05	Positive influence on PU and AT, suggesting users prefer an intuitive system.
PU	AT	0.428*	< 0.05	PU positively influences AT, emphasising the importance of perceived usefulness.
AT	BI	0.746*	< 0.05	Strong positive influence on BI, indicating a significant impact on users' behavioural intention to use the system.
ENJ	PU, AT	0.357*, 0.275*	< 0.05	Enjoyment positively influences PU and AT, highlighting the role of user enjoyment in system acceptance.
PA	PU, PEOU	0.315*, 0.603*	< 0.05	Positive influence on PU and PEOU, emphasising the importance of perceived authority.

The significance level for each construct being less than 0.05 reinforces the robustness of the findings, aligning with the results obtained from correlation analysis. Consequently, all hypotheses formulated are supported, suggesting a comprehensive understanding of the interplay among these constructs.

Summarizing the findings, users' intention to use the proposed system is significantly affected by their attitude towards using the system, consistent with prior studies (Bakhit Jaafreh, 2018; Lee et al., 2003; Patil, 2017). This attitude is directly influenced by perceived ease of use, perceived usefulness, and perceived enjoyment, implying that participants prefer an intuitive software that is easy to use and provides utility and enjoyment in their daily lives.

Furthermore, users' continued usage is contingent upon perceiving inherent benefits in the system, emphasising the importance of utility. Perceived Usefulness demonstrates a more substantial effect than Perceived Ease of Use and Perceived Enjoyment, underlining the system's utility as a critical factor in fostering positive user attitudes. The analysis also highlights the interconnectedness of constructs, with Perceived Usefulness significantly affected by Perceived Enjoyment, Perceived Ease

of Use, and Perceived Authority. Among these, Perceived Enjoyment exerts the strongest influence, emphasising the crucial role of user experience and enjoyment in enhancing perceived usefulness.

Additionally, the analysis reveals a strong correlation between Perceived Ease of Use and Perceived Authority, suggesting that a system's social influence directly affects its perceived ease of use. This indicates the importance of incorporating social influence elements to enhance the perceived usability of the system. The Extended TAM model analysis underscores the necessity to focus on enhancing Perceived Usefulness, Perceived Enjoyment, and Perceived Ease of Use during the system development phase. Specific recommendations include incorporating beneficial use cases, offering exploratory analysis through visualisations, ensuring high expressiveness for customisation, and designing a robust UI/UX for usability, intuitiveness, and enjoyment. These considerations align with the broader literature on technology acceptance and user experience, providing valuable insights for system developers and designers. Table 15 below outlines the suggested features for the proposed tool based on each correlated construct with rationale.

Table 36 Suggested features based on correlated constructs

Feature	Correlated Construct	Rationale
Intuitive User Interface (UI) Design	PEOU, PA	Enhances perceived ease of use and incorporates social influence, making the tool accessible for non-expert users.
Interactive Tutorials and Help Guides	PEOU	Provides interactive tutorials and help guides to enhance users' understanding and navigation of the tool.
Visual Explanations of Model Output	PU, AT	Enhances the perceived usefulness of the tool and contributes to a positive attitude towards using the system.
Built-in Data Exploration and Visualisation Tools	ENJ, PU	Adds an enjoyable aspect to the user experience and enhances the perceived usefulness by facilitating a better understanding of the data.
Preconfigured Use Cases and Datasets	PU, AT	Simplifies the user's initial interaction, positively impacting both perceived usefulness and the attitude towards using the system.
Customisation Options	PU, ENJ	Allowing users to customise their projects enhances both perceived usefulness and enjoyment, contributing to a positive user experience.
In-Tool Support for Exploratory Analysis	PU, ENJ	Provides in-tool support for exploratory analysis, aligning with users' desire for an enjoyable development experience and adding to the perceived usefulness.

5.2.4.1 Survey

The findings from the survey conducted with non-expert users provided valuable insights into their preferences, expectations, and challenges related to interacting with an automated machine learning (AutoML) system. The questionnaire encompassed multiple-choice, ranking, and subjective questions aimed at comprehending participants' preferences and expectations when using such a system. The qualitative data collected were analysed, and responses were grouped into respective themes for a comprehensive understanding.

One of the key aspects explored in the survey was participants' opinions on what constitutes a user-friendly system. Participants were asked to rank the importance of different characteristics, and the majority (49 responses) emphasised the significance of a "Good user interface." This finding aligns with the constructs of Perceived Ease of Use (PEOU) and Perceived Authority (PA), reinforcing the importance of a visually appealing and intuitively navigable system for non-expert users.

In terms of learning preferences, participants overwhelmingly preferred a step-by-step tutorial in the system (45 responses), emphasising the importance of guidance in the learning process. This aligns with the construct of Perceived Ease of Use (PEOU) and suggests that a systematic tutorial approach contributes positively to users' perceived ease of use and overall system satisfaction.

The participants' responses regarding their main aspects of life or job tracking revealed that financial-related data was the most common choice (63%, 46 responses). This preference aligns with the perceived usefulness (PU) construct, emphasising that users find value in tracking financial data as it contributes to their daily lives.

Furthermore, participants indicated their preferred tools/devices for tracking data, with the majority opting for mobile applications (55 responses). This preference for mobile applications aligns with the trend of mobile-first usage and suggests that incorporating mobile-friendly features into the VisAutoML enhance user acceptance.

Regarding current limitations associated with tools for recording data, the majority of participants (41%, 30 responses) identified bad system design or poor user interface design as a significant limitation. This underscores the importance of focusing on a user-friendly design to address the perceived ease of use (PEOU) construct.

Participants also expressed their preferences for the visualisation of information, with a majority (72%, 53 responses) indicating a preference for mobile-based applications or smartphone-related platforms. This aligns with the perceived ease of use (PEOU) construct, as participants find mobile-based platforms more accessible and user-friendly. In conclusion, the survey findings highlight the importance of prioritizing good user interface design, providing step-by-step tutorials, and incorporating mobile-friendly features in VisAutoML for non-expert users. These insights can guide the development of a system that aligns with users' preferences, enhances usability, and addresses the specific needs of non-expert users in the machine learning domain. The suggested features based on findings from the survey are listed in the table below.

Table 37 Suggested features based on survey findings

Feature	Rationale	Correlated Construct
Intuitive User Interface (UI) Design	Participants ranked "Good user interface" as the top priority, emphasising the significance of a visually appealing and intuitively navigable system for non-expert users. Model-friendliness would prompt the tool to be responsive, considering mobile applications might not be suitable for conducting full-fledged ML development.	Perceived Ease of Use (PEOU), Perceived Authority (PA)
Step-by-Step Tutorial	A majority of participants preferred learning through a step-by-step tutorial, highlighting the importance of guidance in the learning process for improved perceived ease of use.	Perceived Ease of Use (PEOU)

Mobile-Friendly Features	Participants overwhelmingly preferred mobile applications for tracking data, indicating the importance of incorporating mobile-friendly features to enhance accessibility and user-friendliness. Model-friendliness would prompt the tool to be responsive, considering mobile applications might not be suitable for conducting full-fledged ML development.	Perceived Ease of Use (PEOU)
Automated Visualisation Process	The majority preferred a fully automated process for visualisation, aligning with user preferences for simplicity and ease of use in handling complex data visualisation tasks.	Perceived Ease of Use (PEOU)
Customisation Options	Users expressed interest in customisation options, indicating the need for flexibility in adapting the tool to their specific needs, contributing to a positive user experience.	Perceived Usefulness (PU)
Financial Data Tracking	Considering the high percentage of participants interested in tracking financial-related data, the tool should provide capabilities for effective financial data tracking.	Perceived Usefulness (PU)
User-Friendly System Design	Identified as a limitation in existing tools, addressing bad system design or poor user interface design is crucial to enhance perceived ease of use and overall user satisfaction. Model-friendliness would prompt the tool to be responsive, considering mobile applications might not be suitable for conducting full-fledged ML development.	Perceived Ease of Use (PEOU)
Incorporation of Mobile Applications	Given the preference for mobile-based platforms for information visualisation, the inclusion of mobile applications is essential for aligning with user preferences. Model-friendliness would prompt the tool to be responsive, considering mobile applications might not be suitable for conducting full-fledged ML development.	Perceived Ease of Use (PEOU)

5.2.4.2 Interview

The interview findings reveal valuable insights into the perceptions and expectations of non-expert users regarding the development and use of AI systems. Three main themes emerged from the interviews, each correlated with specific constructs from the Technology Acceptance Model (TAM), shedding light on participants' attitudes towards the proposed system and suggesting features that align with their preferences and challenges.

Theme C1 - Perceived Ease of Use of AI:

Participants generally exhibited an average level of perceived ease of use regarding AI, highlighting their lack of technical application knowledge. While acknowledging their limited understanding, participants expressed a strong willingness to learn AI, indicating a positive attitude towards incorporating AI into their workflow. The major hurdles identified included challenging terminology, difficult syllabus, different field of study, and no foundational knowledge of AI. Recommendations for

enhancing ease of use included self-learning through video tutorials or books, with some participants expressing a preference for classroom learning with educators. The participants outlined personal expectations towards an AI development system, emphasising the importance of templates, examples, and integration with familiar software systems such as Microsoft and Google.

Theme C2 - Perceived Usefulness of Visualisations:

Participants recognized the usefulness of visualisations for interpreting collected data, emphasising the need for aesthetically pleasing visuals to attract viewer interest. They provided practical applications of visualisations to their proposed projects, showcasing a clear understanding of the potential benefits. The preference for accessing visualisations on both desktop and mobile applications was predominant. Furthermore, participants expressed a majority preference for semi-automated visualisations, emphasising the importance of customisability and smart prediction of suitable visualisations.

Theme C3 - Platform Ease of Use:

Participants associated a user-friendly system with good user interface, clear design, and an intuitive user experience. The usefulness and user-friendliness of a system were highlighted as key factors encouraging continued usage. Conversely, factors such as advertisements, difficult functions, bad user experience, and lack of tangible results were identified as deterrents. Recommendations for an ideal system included simplicity, clear flow, good user interface, and integration with third-party services. Participants stressed the importance of an online documentation system and expressed the need for a system that makes life easier and is beneficial.

Based on these themes, the following table outlines suggested features as requirements for an AI development system, correlating each with the relevant TAM construct:

Table 38 Suggested features based on interview findings

Suggested Feature	Rationale	Correlated Construct
Clear Learning Paths with Tutorials	Participants expressed a preference for step-by-step tutorials to address the hurdle of difficult syllabus and terminology in AI learning.	Perceived Ease of Use (PEOU)
Customisable Templates for AI Development	Participants emphasised the usefulness of templates, aligning with their expectations for an AI development system.	Perceived Ease of Use (PEOU)
Integration with Familiar Software Systems	Participants recommended integration with familiar systems like Microsoft and Google, enhancing perceived usefulness.	Perceived Ease of Use (PEOU), Perceived Authority (PA)
Aesthetically Pleasing Visualisations	Visualisations should be aesthetically pleasing to attract viewer interest, contributing to perceived usefulness.	Perceived Usefulness (PU)
Semi-Automated Visualisation Generator	Participants expressed a majority preference for semi-automated visualisations, providing customisability and smart predictions.	Perceived Usefulness (PU)
Cross-Platform Accessibility for Visualisations	The ability to access visualisations on both desktop and mobile platforms aligns with participants' preferences.	Perceived Usefulness (PU)
User-Friendly Interface and Design	A clear, intuitive user interface and design were consistently highlighted as crucial for a user-friendly system.	Perceived Ease of Use (PEOU)

Online Documentation System	Participants stressed the importance of an online documentation system for easy reference and understanding.	Perceived Ease of Use (PEOU)
Integration with Third-Party Services	Integration with third-party services aligns with participants' expectations and preferences for familiarity in the system.	Perceived Ease of Use (PEOU), Perceived Authority (PA)

5.2.4.3 User Persona

In designing an AutoML system for non-experts with a focus on user-centred design (UCD), the creation of a representative user persona plays a pivotal role in ensuring the system's alignment with the needs, preferences, and challenges of its intended users (Hudson, 2013; Williams, 1986). The user personas presented below is created based on the synthesis of findings derived from user interviews, survey responses, and literature reviews, providing a comprehensive understanding of the expectations and requirements of non-expert users in the domain of AI development.

User Persona 1: Emily Johnson

Background:

Emily Johnson, a 30-year-old female marketing professional, emerges as a distinct user persona reflecting the characteristics and aspirations of non-expert individuals seeking to integrate AI into their professional endeavours. With a bachelor's degree in Marketing, Emily embodies the archetype of a user lacking profound technical expertise in AI development but harbouring a keen interest in harnessing AI capabilities for marketing insights and data-driven decision-making.

Goals and Motivations:

Emily's overarching goals encompass the utilisation of AI to gain profound insights into consumer behaviour, enhance marketing campaigns, and facilitate informed decision-making within the marketing domain. Her motivation lies in the recognition of AI's potential to elevate the efficacy of marketing strategies and contribute to organizational success.

Challenges:

Despite her professional acumen in marketing, Emily grapples with the challenges associated with a lack of technical knowledge in AI and encounters difficulties in navigating the complex syllabus and terminologies intrinsic to AI development.

Attitudes and Behaviours:

Emily exhibits a positive disposition towards learning AI, expressing a willingness to invest time in self-guided learning endeavours. Additionally, she conveys a preference for visually engaging and aesthetically pleasing visualisations for effective communication within the marketing context.

Preferences:

In the realm of system preferences, Emily leans towards a user-friendly AI development system that offers clear learning paths through step-by-step tutorials. She values customisable templates to simplify AI development processes and emphasises the importance of integration with familiar software systems such as Microsoft and Google for a seamless workflow.

Expectations from the AI Development System:

Emily's expectations from the AI development system revolve around the provision of clear learning paths with beginner-friendly tutorials, the availability of aesthetic and customisable visualisations for effective communication of marketing insights, and seamless integration with familiar tools and platforms. She underscores the significance of a user-friendly interface complemented by online documentation for ease of reference.

User Persona 2: Jonathan

Background:

Jonathan, a 22-year-old male undergraduate student pursuing a degree in Business Studies, emerges as a distinct user persona reflecting the characteristics and aspirations of a non-expert individual seeking to integrate AI into their academic and future professional endeavours. With a keen interest in data analysis and decision-making, Jonathan represents the archetype of a user who lacks profound technical expertise in AI development but is eager to harness its capabilities for business insights and strategic planning.

Goals and Motivations:

Jonathan’s overarching goals encompass the utilisation of AI to gain a deeper understanding of business trends, enhance data-driven decision-making, and develop innovative solutions to complex business problems. His motivation lies in the recognition of AI's potential to transform the business landscape and contribute to his success as a future business professional.

Challenges:

Despite his academic background in Business Studies, Jonathan grapples with the challenges associated with a lack of technical knowledge in AI and encounters difficulties in navigating the complex syllabus and terminologies intrinsic to AI development. He seeks to bridge this gap and develop the necessary skills to effectively leverage AI in his academic and professional pursuits.

Attitudes and Behaviours:

Jonathan exhibits a positive and proactive attitude towards learning AI, expressing a strong willingness to invest time and effort in self-guided learning endeavours. He conveys a preference for visually engaging and intuitive visualisations that can effectively communicate business insights and support decision-making processes.

Preferences:

In the realm of system preferences, Jonathan leans towards a user-friendly AI development system that offers clear learning paths through step-by-step tutorials and interactive exercises. He values customisable templates and pre-built models to simplify the AI development process and emphasises the importance of integration with familiar business software and data analysis tools.

Expectations from the AI Development System:

Jonathan's expectations from the AI development system revolve around the provision of clear and structured learning pathways, with a focus on practical applications of AI in the business context. He

seeks a system that offers a seamless integration with the tools and platforms he already uses, enabling a streamlined workflow and the ability to leverage AI-powered insights directly within his business studies and decision-making processes. Additionally, Jonathan values the availability of comprehensive online documentation and community support to address any challenges he may encounter during his AI learning journey.

User Persona 3: Kimberly Tanaka

Background:

Kimberly Tanaka, a 28-year-old female doctoral student in the field of Sociology, emerges as a distinct user persona reflecting the characteristics and aspirations of a non-expert individual seeking to integrate AI into their academic research endeavours. With a strong background in qualitative research methods, Kimberly represents the archetype of a user who lacks profound technical expertise in AI development but is eager to harness its capabilities to enhance her research processes and uncover novel insights.

Goals and Motivations:

Kimberly's overarching goals encompass the utilisation of AI to streamline her research workflows, automate time-consuming tasks, and derive deeper insights from her qualitative data. Her motivation lies in the recognition of AI's potential to revolutionize the field of social sciences, allowing her to uncover hidden patterns, identify emerging trends, and inform her academic publications and policy recommendations.

Challenges:

Despite her experience in qualitative research, Kimberly grapples with the challenges associated with a lack of technical knowledge in AI and encounters difficulties in navigating the complex terminologies and methodologies intrinsic to AI development. She seeks to bridge this gap and develop the necessary skills to effectively leverage AI in her academic pursuits.

Attitudes and Behaviours:

Kimberly exhibits a curious and open-minded attitude towards learning AI, expressing a willingness to invest time and effort in self-guided learning endeavours. She conveys a preference for intuitive and visually engaging interfaces that can seamlessly integrate with her existing research workflows and data analysis tools.

Preferences:

In the realm of system preferences, Kimberly leans towards a user-friendly AI development system that offers clear and accessible learning paths, with a focus on practical applications in the social sciences. She values the availability of pre-trained models and customisable templates that can be tailored to her specific research needs, enabling her to leverage AI capabilities without the burden of extensive technical programming.

Expectations from the AI Development System:

Kimberly's expectations from the AI development system revolve around the provision of comprehensive tutorials, interactive examples, and hands-on exercises that cater to her non-technical background. She seeks a system that offers a seamless integration with the qualitative data analysis

tools she currently uses, allowing her to incorporate AI-powered insights directly into her research processes. Additionally, Kimberly values the availability of online documentation, community forums, and expert support to address any challenges she may encounter during her AI learning journey.

In crafting these user personas, the synthesis of data gathered from interviews, surveys, and literature reviews provided a nuanced understanding of each persona's background, goals, challenges, and preferences. These personas encapsulate the diverse facets of a non-expert user's aspirations and needs, serving as a valuable reference for the development of a user-centred AI system tailored to the requirements of non-expert users.

5.2.5 Validity assessment

In addressing the validity of this research, particularly given its mixed-methods and user-centered nature, several facets warrant explicit discussion. Transferability, which concerns the degree to which the findings can be applied to other similar contexts or populations, was a key consideration. While acknowledging the specific scope of this study (focused on tabular data, specific tasks, and a particular participant demographic as detailed subsequently), efforts were made to enhance potential transferability. The Methodology chapter provides detailed descriptions of the participant recruitment processes, including the use of platforms like Amazon Mechanical Turk and snowball sampling, alongside comprehensive demographic profiles of participants across all evaluation studies. Furthermore, the study describes the controlled laboratory setting in which the prototype evaluations were primarily conducted and the specific methodologies employed for data collection and analysis. This descriptive richness allows future researchers and practitioners to assess the congruence between the context of this study and their own when considering applying the findings and the empirically validated design principles. While direct statistical generalization to all non-expert populations or different ML contexts is not claimed, the detailed contextual information provided facilitates informed judgments about the potential transferability of the insights to similar scenarios involving non-expert users interacting with GUI-based AutoML tools for tabular data.

The credibility of the research, concerned with the accuracy and truthfulness of the findings and ensuring they are free from bias or misrepresentation, was a central focus throughout the research design. Credibility is about establishing the truthfulness and authenticity of the data and ensuring that the findings are not influenced by any biases, preconceptions, or misrepresentations. The adoption of a mixed-methods approach, integrating both quantitative (questionnaires) and qualitative (interviews, open-ended questions) data collection and analysis techniques, significantly enhanced credibility through triangulation of data sources. As highlighted in the Findings chapter, consistency was observed between quantitative results, such as the improvements in usability metrics (UEQ, SUS) and perceived transparency (Trust, XAI questionnaires) across prototype versions, and the qualitative feedback gathered from user interviews and open-ended questionnaires. For instance, qualitative comments praising the intuitive interface and guided workflow aligned with high quantitative usability scores for VisAutoML 2.0, and qualitative feedback on the clarity of visualizations corroborated improved perceived explainability. This convergence of evidence from different methods provides a more robust and credible account of user experiences and perceptions, mitigating the potential biases inherent in relying solely on a single data source or methodology [cf. Creswell & Plano Clark, 2017]. The rigorous data analysis methods employed, including statistical analysis for quantitative data and thematic analysis for qualitative data, further contributed to the credibility of the findings.

Furthermore, confirmability, which addresses the objectivity of the findings and their independence from researcher bias, was addressed through conscious efforts towards reflexivity. Confirmability is

the degree to which the findings of a study are objective and independent of the researcher's influence or bias. The researcher actively engaged in examining their own assumptions, perspectives, and potential influences throughout the research process, from the initial conceptualization and design of the VisAutoML tool based on user requirements to the analysis and interpretation of the empirical data. This process of reflexivity, while not eliminating researcher influence entirely, allows for an increased awareness of potential biases and their impact on the research outcomes [cf. Braun & Clarke, 2006]. Additionally, the qualitative findings and their interpretations are intended for publication and will be subjected to peer review by domain experts. This external validation process serves as a critical check on the confirmability of the research, as independent reviewers assess the extent to which the findings are supported by the data and the interpretations are objective and well-justified within the academic community.

5.3 VisAutoML 1.0

The prototype evaluation stage of the proposed VisAutoML tool signifies a critical stage in the iterative development process, aimed at a thorough assessment of the system's functionality, usability, and transparency. This stage adopts a comprehensive evaluation strategy, integrating a comparison study, usability testing, transparency testing, and qualitative studies involving open-ended questions and interviews. The incorporation of these methodologies seeks to offer a holistic understanding of the tool's performance, user experience, and transparency aspects.

Within the comparison study, a detailed analysis will be conducted, drawing parallels between the proposed *VisAutoML* tool and an existing tool tailored for non-experts. This study aims to uncover nuanced insights into user preferences, system performance, and potential differentiators contributing to a more effective and user-friendly ML tool.

The usability testing component involves a rigorous examination of the proposed *VisAutoML* tool's user interface, navigation, and overall user experience. Leveraging the standardized User Experience Questionnaire (UEQ) questionnaire to quantitatively measure user experience and usability. Additionally, qualitative insights will be gathered through detailed user responses and interviews, providing a deeper understanding of the user experience.

Transparency testing is another crucial aspect, assessing the system's clarity and understandability using tools such as the Trust Questionnaire and the Explainable Artificial Intelligence (xAI) Questionnaire. This evaluation aims to provide insights into users' confidence in the decision-making processes of the VisAutoML.

Qualitative studies, incorporating open-ended questions and interviews, will be employed to delve deeper into users' perspectives, preferences, and potential areas of improvement. These qualitative insights complement quantitative metrics, offering rich information for refining the tool.

5.1.1 Comparison Study

5.1.1.1 Participant Demographic

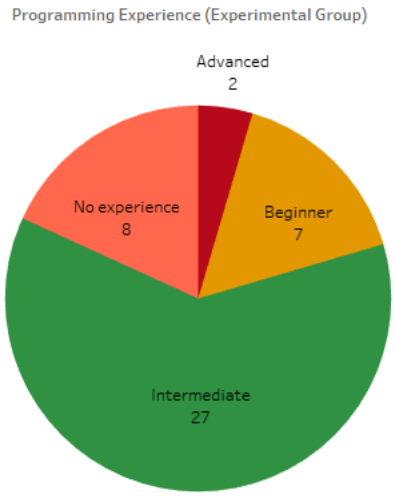
The user trial comprised four distinct groups, totalling 82 participants, with group sizes ranging from 10 to 20 individuals. The participants were exclusively students from the University of Nottingham Malaysia, with two groups representing the School of Computer Science and two groups from the School of Finance. The demographic composition of the participants was carefully considered, acknowledging factors such as age, gender, and education background.

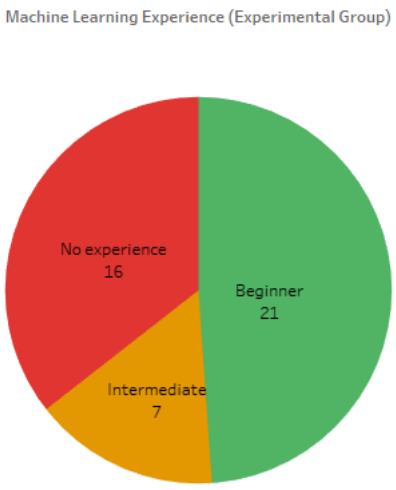
Individual participant ages ranged between 19 and 27 years old, with a mean (M) age of 21.39 years and a standard deviation (SD) of 1.49 years. In terms of gender distribution, 39 participants identified as female, and 43 participants identified as male. The age distribution revealed that the majority of participants, 81 in total, fell between the ages of 19 and 25, with only one participant above the age of 25.

Education backgrounds of the participants varied, with 22 individuals having completed secondary school, 11 possessing a diploma, 46 holding a bachelor's degree, and 3 having attained a master's degree. This diverse representation in education levels ensured a varied and comprehensive participant pool, enhancing the robustness and applicability of the study findings across different academic backgrounds.

Table 39 Participant Demographic

Gender		Age		Education	
Item	**No.**	**Item**	**No.**	**Item**	**No.**
Female	39	Between 19 and 25	81	Secondary School	22
Male	43	Above 25	1	Diploma	11
				Bachelor's Degree	46
				Master's Degree	3
Total	82		82		82





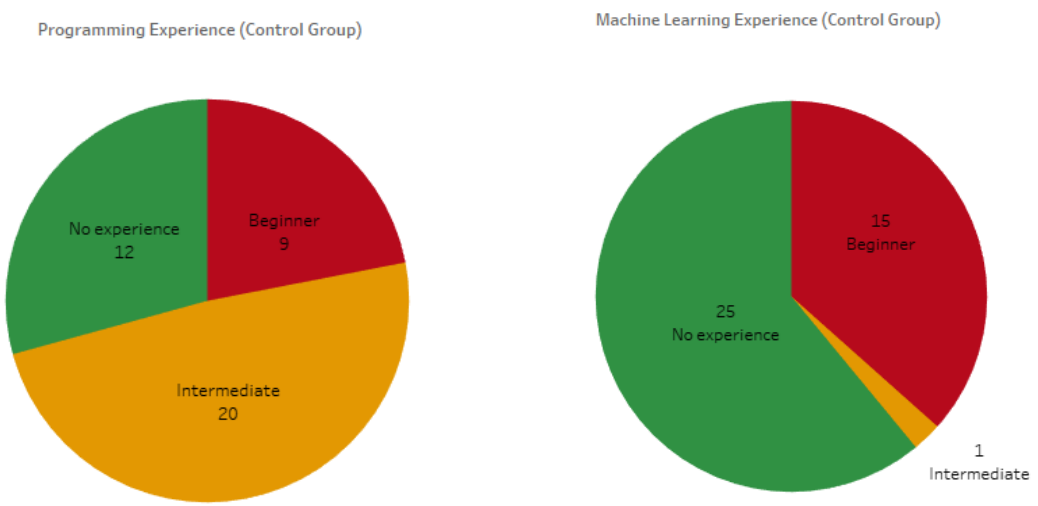


Figure 60 Programming and Machine Learning Experience for the Experimental and Control Group

The experimental group comprised participants demonstrating diverse levels of machine learning and programming proficiency, all falling within the non-expert category. Concerning machine learning experience, the distribution was as follows: 16 participants possessed no prior experience, 21 were beginners, and 7 were at an intermediate level. In parallel, participants' programming experience varied, with 27 individuals at an intermediate level, 7 at a beginner's level, and 2 with advanced experience, while 8 participants had no programming background.

Similarly, the control group featured participants with a comparable composition, albeit with slight distinctions. Notably, none of the participants in the control group had advanced programming experience. Regarding machine learning experience, the control group exhibited a nuanced distribution, with only 1 participant possessing intermediate experience, while a larger cohort (25 participants) had no prior exposure to machine learning. Importantly, none of the participants in either group were classified as experts or advanced practitioners in machine learning, aligning with the defined parameters of non-experts within this study.

5.1.1.2 Usability

The usability evaluation utilised the User Experience Questionnaire (UEQ) and the System Usability Scale (SUS) to gauge participants' experiences comprehensively. In the experimental group, the mean score for pragmatic quality was 0.83, indicative of a moderately positive perception. However, a standard deviation of 1.082 suggested considerable variability among participants. The p-value of 0.331, while not statistically significant, warrants attention for potential improvements. For hedonic quality, the mean score was 0.80, signifying a generally positive response, with a standard deviation of 1.030. The overall experience received a mean score of 0.82, indicating a favourable perception, with a moderate standard deviation of 0.93. These findings suggest that participants generally viewed the VisAutoML positively, with moderate variability. The SUS score for the experimental group was 61.5, reflecting a moderate level of perceived usability. This indicates that participants found the VisAutoML reasonably usable, though with room for potential enhancements.

Conversely, the control group exhibited a less positive response across UEQ metrics. For pragmatic quality, the mean score was -0.5, indicating a less positive perception, with a standard deviation of 1.419. Hedonic quality scored -0.122, reflecting a somewhat less positive response, with a standard deviation of 1.303. The overall experience received a mean score of -0.311, suggesting a less positive perception, with a standard deviation of 1.3. These results indicate a comparatively less favourable

view of H2O AutoML in the control group. The SUS score for the control group was 38.5, signifying a lower perceived usability compared to the experimental group. This suggests that participants in the control group found H2O AutoML less usable than VisAutoML in the experimental group.

Table 40 Usability score differences between experimental and control groups

Group	UEQ - Pragmatic Quality	UEQ - Hedonic Quality	UEQ - Overall Experience	SUS Score
Experimental	0.83 (SD: 1.082, p=0.331)	0.80 (SD: 1.030, p=0.315)	0.82 (SD: 0.93, p=0.286)	61.5
Control	-0.5 (SD: 1.419, p=0.434)	-0.122 (SD: 1.303, p=0.399)	-0.311 (SD: 1.3, p=0.398)	38.5

A comparison between the experimental and control groups reveals that the experimental group generally exhibited more positive responses across UEQ metrics. Additionally, the SUS scores further reinforced this trend, with the experimental group demonstrating a higher perceived usability level. A summarised comparison between the two groups is provided in the table above.

Independent t-tests were conducted on the eight distinct UEQ metrics using Statistical Package for the Social Sciences (SPSS), version 26, developed by IBM in Chicago, Illinois, USA. The mean UEQ scores served as a measure of usability, providing a quantitative basis to compare the two ML platforms. Across all UEQ metrics, VisAutoML demonstrated statistically superior performance. This superiority was evidenced by consistently higher mean scores than H2O AutoML, with the differences reaching statistical significance in all instances. These findings suggest that users may be likely to have a more positive user experience with VisAutoML in comparison to H2O AutoML.

The robustness of these results is further strengthened by the Levene's test for equality of variances. In most cases, the assumption of homogeneity of variances was met, as indicated by non-significant Levene's test results. This was, however, not the case for the UEQ6 metric, where the assumption was violated. This discrepancy warrants further investigation but does not necessarily undermine the overall findings of the study. The detailed t-tests are shown in the table below.

Table 41 Detailed t-tests between groups

Item	F	Sig.	t	Sig. (2-tailed)	Mean Difference
UEQ1	3.463	.066	4.324	.000	1.34146
UEQ2	.094	.760	4.306	.000	1.51220
UEQ3	3.101	.082	3.543	.000	1.09756
UEQ4	1.911	.171	3.977	.000	1.39024
UEQ5	1.496	.225	3.309	.000	.92683
UEQ6	4.195	.044	3.948	.000	1.29268
UEQ7	1.594	.210	2.169	.000	.68293
UEQ8	.085	.771	2.678	.000	.80488

In conclusion, the findings present strong evidence in favour of VisAutoML in terms of user experience quality. The current study provides a solid foundation for further research and can guide decision-making when considering the use of these platforms.

5.1.1.3 Knowledge Gain

Assessing knowledge gain serves as a crucial metric to evaluate the transparency of VisAutoML. By contrasting post-test scores with pretest scores, the extent of knowledge gain within the two groups was measured. The questions devised for the study were thoughtfully designed to inquire on elements of the machine learning development process, encompassing data preparation steps and algorithms, mainly, knowledge that was accessible across both platforms.

The acquired data was carefully organized and input into the SPSS, version 26. a pretest analysis was conducted using an independent t-test. The results provided a quantitative measure of the initial knowledge state for the two groups, setting a baseline for assessing the platforms' performance. The pretest scores revealed a statistically significant difference between the two groups ($t(80)=3.793$, $p<.001$). Specifically, the mean score for the group using VisAutoML was significantly higher than that of the group using H2O AutoML, with a mean difference of 2.51220 (95% CI [1.19396, 3.83043]). The Levene's test was not significant ($F=.001$, $p=.978$), indicating that the assumption of homogeneity of variances was met. This suggests that the variability in scores within each group was comparable, lending further credibility to the findings.

The pretest results indicate that the VisAutoML group had a significantly higher initial knowledge state compared to the H2O AutoML group. This establishes a foundational understanding of the groups' starting points, which is crucial for the subsequent analysis of knowledge gain associated with the use of the two platforms. Subsequently, another one-way between-groups ANOVA was conducted on post-test data to examine the significance of knowledge gain disparities between the experimental and control groups following the intervention.

Following the pretest, a post-test was conducted to measure the knowledge gained from using the VisAutoML and H2O AutoML platforms. Independent t-tests were utilised to compare the mean scores of the two groups after their respective experiences with the platforms. The post-test results displayed a statistically significant difference ($t(80)=3.793$, $p<.001$), similar to the pretest findings. The VisAutoML group outperformed the H2O AutoML group with a mean difference of 2.51220 (95% CI [1.19396, 3.83043]). The Levene's test was not significant ($F=.001$, $p=.978$), confirming that the assumption of equal variances was met. This indicates that the dispersion or spread in the post-test scores for each group was equivalent.

The experimental group displayed an increase in correct answers percentage from 56.3% to 77.24%, whereas the control group's percentage rose from 42.48% to 59.15%. In conclusion, interaction with VisAutoML yielded significantly enhanced learning outcomes, as evidenced by a greater knowledge gain in the experimental group compared to the control group, further highlighting the efficacy of VisAutoML over H20AutoML.

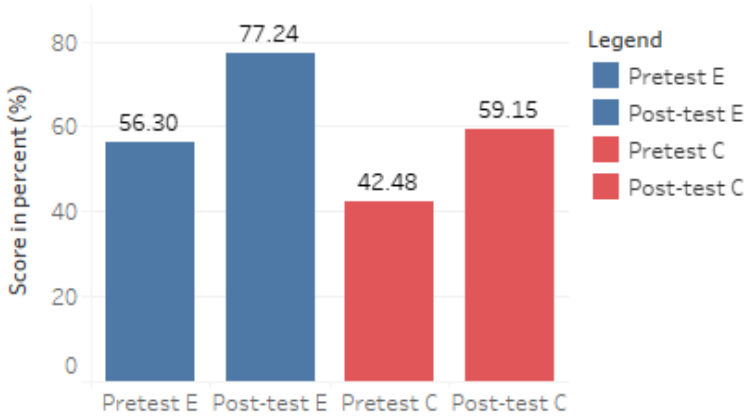


Figure 61 Knowledge gain between experimental (E) and control (C) group

5.1.1.4 Open Ended Questionnaire

This study included qualitative questions to explore the experiences of participants from both the experimental and control groups regarding their usage of the VisAutoML. Participants were asked to respond to three open-ended questions regarding their experience with the tool, what they liked and disliked about it, and their suggestions for improvement. The responses were transcribed, anonymized, and subjected to a thematic analysis approach to identify common patterns and themes. The analysis aimed to identify recurring themes, examine the strengths and weaknesses of the tool, and provide valuable insights for its improvement.

Experimental Group Responses

Participants in the experimental group expressed generally positive experiences with the VisAutoML. Through the analysis of their responses, several prominent themes emerged:

1. Usability and Ease of Use: Participants highlighted the tool's user-friendly interface, which facilitated easy navigation and streamlined the model creation process. They appreciated the guided tutorials and step-by-step instructions, which enabled them to quickly grasp the tool's functionalities. The drag-and-drop feature was particularly lauded for its simplicity and accessibility, making it suitable for individuals with limited knowledge of machine learning algorithms.

2. Time Efficiency: Participants commended the VisAutoML for its ability to generate machine learning models rapidly. This feature was seen as a significant advantage, as it saved valuable time during the model creation process.

3. Challenges with Model Viewing: Some participants reported difficulties in viewing their own models within the tool. They encountered issues when attempting to access their models or experienced instances where the provided links failed to connect or displayed models built by other users. Participants emphasised the need for improvements in this area to enhance the overall user experience.

4. Interface and Design: While participants generally found the user interface of VisAutoML clear and intuitive, they suggested several design enhancements. Recommendations included improving the

drop-down menu's functionality, providing more detailed written information to aid users, and enhancing the visual appeal of the tool through a more engaging colour scheme.

Control Group Responses

In the control group, participants shared their experiences with an alternative software tool which is H2O AutoML. The qualitative analysis of their responses revealed the following key themes:

1. Usability Challenges: Participants in the control group encountered usability challenges with the software tool. They reported difficulties in understanding and navigating the interface, citing confusion over available options and encountering issues throughout the tool's workflow. The tool was described as overwhelming and requiring a certain level of technical expertise to operate effectively.

2. Lack of User-Friendliness: Compared to the VisAutoML used by the experimental group, participants in the control group found their software tool less user-friendly. They expressed frustration with the abundance of buttons, a confusing interface, and a lack of clear guidance throughout the tool's processes.

3. Need for Guided Tutorials: Control group participants emphasised the importance of guided tutorials to support their understanding and utilisation of the tool. This need was particularly prominent among individuals with limited prior experience or expertise in machine learning.

Comparison between Experimental and Control Groups

The comparison between the experimental and control groups provides valuable insights into the relative strengths and weaknesses of the VisAutoML. The experimental group, which utilised the VisAutoML, reported more positive experiences regarding usability, ease of use, and time efficiency. The guided tutorials, intuitive interface, and rapid model generation capabilities were appreciated by participants. Conversely, the control group's software tool posed significant usability challenges, requiring improvements in user-friendliness, clarity of instructions, and guidance.

These findings underscore the advantages of the VisAutoML in terms of usability and user experience when compared to the control tool used by the control group. The results suggest that VisAutoML offers a more intuitive and efficient workflow, while the control tool presented significant usability obstacles that need to be addressed to enhance user-friendliness.

Conclusively, the qualitative analysis provides valuable insights and feedback for further refinement of the VisAutoML. The specific areas of improvement highlighted by both the experimental and control groups can guide future work in enhancing the tool's usability, interface design, model viewing functionality, and the provision of guided tutorials.

5.1.1.5 Semi-Structured Interview

A semi-structured interview study was designed and conducted to gather participant insights on the usability of H2O AutoML and VisAutoML, as well as the transparency of H2O AutoML and VisAutoML. The study was organized into three main parts:

1. Testing H2O AutoML:
   Participants were actively involved in testing H2O AutoML. During this phase, they were prompted to interact with the tool and provide detailed descriptions of their experiences related to both transparency and usability. Specific questions were designed to elicit feedback on how transparent and user-friendly participants found H2O AutoML.

2. Testing VisAutoML:
   Following the examination of H2O AutoML, participants were then directed to test VisAutoML. Similar to the first part, they were encouraged to explore the tool and share insights on its transparency and usability. Questions were tailored to uncover participant perceptions of how transparent and user-friendly VisAutoML appeared to be.

3. Suggestions for Improvement:
   The final segment of the interview focused on gathering suggestions from participants to enhance the transparency and usability of both H2O AutoML and VisAutoML. Participants were asked to provide recommendations, insights, and potential improvements that could be implemented in these tools. This allowed for a forward-looking perspective, exploring avenues for refinement based on user experiences.

Each interview, conducted via video conferencing, had a duration ranging from 45 to 60 minutes. The semi-structured format provided flexibility, allowing participants to share their unique perspectives on each tool. The variation in interview content and duration reflected the richness of participant experiences and their willingness to provide detailed insights.

Usability Challenges in H2O AutoML

Navigating and utilising H2O AutoML proved to be a nuanced challenge for participants, with several expressing intricate difficulties that illuminated key areas for improvement. The intricacy of the tool, coupled with a dearth of clear guidance, resonated as a pervasive theme throughout participant responses. One user encapsulated their experience by stating, "Overall, it's interesting but complicated to understand" (I3). This sentiment echoed a common thread among participants, highlighting the intriguing nature of H2O AutoML but underscoring the complexity that posed a barrier to seamless comprehension.

A recurring focal point of struggle emerged as participants grappled with the initial stages of utilising H2O AutoML. "The most challenging part was figuring out where to start and which options to select when creating a model" (I5). This particular challenge underscored a fundamental usability issue – the absence of a clear starting point. The tool's intricate array of features left users grappling with uncertainty, inhibiting their ability to initiate the model creation process with confidence.

Furthermore, the abundance of options within H2O AutoML presented a formidable challenge to users. "The availability of various options without clear guidance made it difficult to understand their purpose" (I2). This sentiment echoed across multiple interviews, pointing towards a critical need for improved clarity in the purpose and utility of the myriad features H2O AutoML offers.

Participants consistently articulated the need for a more user-friendly interface, emphasising the necessity of incorporating beginner-friendly tutorials and a clearer onboarding process. These recommendations aimed at demystifying the tool's complexities and providing users with a smoother initiation into the H2O AutoML ecosystem.

In essence, the overwhelming nature of options without clear guidance emerged as a substantial hurdle for users engaging with H2O AutoML. The desire for a more intuitive and navigable interface underscored the importance of addressing these usability challenges to enhance the overall user experience and make H2O AutoML more accessible to a broader audience.

Positive Aspects of H2O AutoML

Amidst the usability challenges, some users identified positive aspects of H2O AutoML. The ability to save models locally received appreciation from one user, showcasing that certain features were well-received despite the overall complexity of the tool.

A notable positive aspect surfaced as participants expressed gratitude for the "ability to save models locally" (I1). This specific feature resonated positively, underscoring that, despite the overall complexity of H2O AutoML, users found practical value in the tool's capacity to store models locally. This acknowledgment points to a feature that aligns with users' needs and preferences, standing out as a practical and user-friendly component within the tool's intricate framework.

Another participant described their experience as "overall intriguing" (I3). This sentiment hints that, even in the face of challenges, there was an element of interest or curiosity sparked by the tool's capabilities. Unravelling and further exploring these intriguing aspects could be a pathway to enhancing the overall user experience. It suggests that, beyond the hurdles, there are elements within H2O AutoML that captivate users and could be further emphasised to elevate the tool's overall appeal.

Some users recognized the diverse functionalities within H2O AutoML. Even amidst the challenges, this acknowledgment suggests that the tool offers a range of capabilities that users can potentially leverage for their data science tasks. Understanding and amplifying these recognized functionalities could be instrumental in addressing usability concerns and emphasising the tool's versatility.

In summary, participants' positive insights offer a counterpoint to the challenges, providing a more comprehensive view of H2O AutoML's strengths. Leveraging these positive aspects in future updates and addressing the identified usability issues could pave the way for a more refined and user-friendly machine learning tool.

Transparency Challenges in H2O AutoML

Participants consistently grappled with transparency challenges when using H2O AutoML, struggling to interpret the output, decipher predictions, and understand the underlying transparency aspects. The generated numbers were often described as unintelligible, leading to a sense of confusion and uncertainty among users regarding the accuracy of predictions.

"I didn't understand any of it, as the numbers didn't make sense to me." (I4) This sentiment reflects a common struggle faced by users, highlighting the inherent complexity of the information presented by H2O AutoML. The lack of clarity in the numerical outputs posed a significant barrier to users' comprehension, emphasising the need for improvements in presenting information in a more user-friendly manner.

"The user found all the parameters to be impossible to understand." (I10) The challenges extended beyond numerical outputs to include the complexity of parameters, rendering them seemingly

inscrutable for users. The lack of clarity in the tool hindered users' understanding and raised concerns about the reliability of predictions. Addressing the opacity of parameters emerged as a crucial aspect for enhancing the overall transparency of H2O AutoML.

The transparency challenges identified in H2O AutoML were a significant concern among users, prompting insightful suggestions for improvement. Participants recommended the integration of clearer explanations of parameters, the implementation of user-friendly pop-ups for guidance, and the development of approachable documentation to demystify the intricacies of the tool.

"There should be more explanations for each parameter." (I6) This suggestion underscores the importance of providing users with comprehensive explanations, ensuring that each parameter is accompanied by clear and accessible information. Enhancing the explanatory content within the tool can empower users to navigate the complexities with greater confidence.

"Having pop-ups and tooltips of what each parameter and output mean would be incredibly helpful." (I8) The call for pop-ups and tooltips reflects a desire for interactive features that offer on-the-spot guidance. Implementing such user-friendly elements can bridge the gap between the tool's complexity and users' understanding, fostering a more transparent and navigable user experience.

The transparency challenges faced by participants in H2O AutoML shed light on the intricate nature of the tool's output and parameters. User feedback emphasises the need for clearer explanations, interactive features, and accessible documentation to clarify the complexities and enhance transparency. Addressing these challenges presents a baseline to refine VisAutoML, ensuring that users can confidently navigate and comprehend the outputs, ultimately fostering a more transparent and user-friendly machine learning tool.

Positive Feedback on VisAutoML's Usability

Users resonated positively with the usability of VisAutoML, commending the tool's guided and cue-rich experience that contributed to a more intuitive and user-friendly journey. The emphasis on these positive aspects reflects the successful integration of features that enhance the overall usability of the tool. "It was more guided and provided helpful cues." (I8) This sentiment highlights the user-friendly design of VisAutoML, characterized by a guided experience enriched with helpful cues. Users appreciated the intuitive pathways that facilitated a smoother exploration of the tool, underscoring the significance of features that contribute to a more user-centric design.

"The most challenging aspect was not knowing what to do during the loading part." (I5) Despite the positive feedback, participants acknowledged specific challenges during the loading phase of VisAutoML. This insight is valuable as it sheds light on areas where users experienced difficulty. In this case, the challenge centred around uncertainty during loading, emphasising the need for improvements, potentially through informative error messages for invalid options or clearer instructions.

While VisAutoML demonstrated commendable usability feedback, the acknowledgment of challenges during the loading phase presents an opportunity for refinement. User feedback serves as a valuable guide for developers, indicating areas where further enhancements can be made to ensure a seamless and frustration-free user experience.

"The loading phase was a bit confusing, and having more informative error messages would have helped." (I3) This participant's feedback delves into the specific aspect of the loading phase, suggesting that clearer error messages could alleviate confusion. Addressing this aspect aligns with the goal of

refining VisAutoML's usability, making it more accessible and user-friendly for individuals navigating the tool.

The positive feedback on VisAutoML's usability reflects its success in providing a guided and cue-rich experience for users. However, the recognition of challenges during the loading phase highlights specific areas for improvement. By addressing these challenges, developers can further elevate the usability of VisAutoML, ensuring a more seamless and enjoyable exploration of the tool's functionalities.

Transparency Strengths in VisAutoML

Users celebrated VisAutoML for its transparency, emphasising the comprehensiveness of output and the clarity of performance metrics. The positive responses underscore the tool's success in presenting complex machine learning information in an accessible manner, contributing to users' understanding and interpretation of the generated insights.

"VisAutoML provided a lot of options and thorough information, which made it easier for me to understand the output and predictions." (I6) The recognition of VisAutoML's abundance of options and comprehensive information highlights its commitment to transparency. Users appreciated the tool's capacity to present detailed insights, fostering a clearer understanding of the output and predictions. This positive feedback aligns with the broader goal of making machine learning processes more accessible and user-friendly.

"The performance metrics and evaluation results were presented in a more plain text format, making them easier to interpret and understand." (I11) Participants echoed the sentiment that the plain text format of performance metrics contributed to ease of interpretation. This design choice in presenting evaluation results in a straightforward manner enhances the overall transparency of VisAutoML. Users found value in the simplicity of presentation, facilitating a more straightforward comprehension of the intricate details associated with model performance.

While VisAutoML demonstrated considerable strengths in transparency, a call for a more detailed summary was noted. This feedback suggests an opportunity for further enhancement, where users seek a more comprehensive overview of the generated insights. Addressing this aspect aligns with the continuous improvement ethos, ensuring that VisAutoML remains at the forefront of providing transparent and understandable machine learning outputs.

"However, a more detailed summary at the end would have been helpful to wrap up the insights comprehensively." (I9) This participant's feedback offers a constructive suggestion for improvement, indicating a desire for a more encompassing summary at the conclusion of the VisAutoML process. Integrating such refinements based on user input contributes to the evolution of VisAutoML, maintaining its commitment to transparency and user satisfaction.

The positive responses regarding transparency highlight VisAutoML's success in providing users with a comprehensive and understandable view of machine learning outputs. The call for a more detailed summary presents an opportunity for further refinement, ensuring that VisAutoML continues to excel in transparency and user-friendly information presentation.

Confidence and Documentation Gaps in VisAutoML

While VisAutoML garnered positive feedback for its usability and transparency, users expressed a notable gap in confidence stemming from a lack of comprehensive documentation. The absence of

well-documented guidance impacted users' assurance in utilising the tool, shedding light on the pivotal role documentation plays in fostering user trust and proficiency.

"I feel less confident due to the lack of documentation, which could provide more guidance and support." (I5) Participants' feedback highlighted the significant impact of documentation on user confidence. Despite the positive aspects of usability and transparency, the perceived lack of comprehensive documentation created a sense of uncertainty among users. This underscores the importance of providing users with robust support materials, guiding them through the intricacies of VisAutoML, and ultimately enhancing their confidence in utilising the tool.

While users celebrated the positive aspects of VisAutoML, such as its transparency and usability, the identified documentation gap presents an opportunity for improvement. Participants suggested that comprehensive documentation could offer valuable guidance and support, addressing specific user needs and concerns. Enhancing the documentation aligns with the broader goal of ensuring users can navigate VisAutoML with confidence, leveraging its capabilities effectively for their data science tasks.

"The tool itself is good, but having more detailed documentation would make it easier to explore and use." (I12) This participant's insight reinforces the idea that detailed documentation enhances the exploratory and usage aspects of VisAutoML. A comprehensive guide can empower users to delve into advanced features like Lineage with greater confidence, unlocking the full potential of the tool. This user input signals an opportunity to bridge the confidence gap by providing in-depth documentation that complements the existing strengths of VisAutoML.

Recognizing the pivotal role documentation plays in user empowerment, developers can strategically focus on creating materials that cater to user needs. A well-structured documentation repository, including tutorials, FAQs, and troubleshooting guides, can serve as a valuable resource for users at different proficiency levels. By addressing the documentation gap, developers can fortify users' confidence in navigating VisAutoML, contributing to a more seamless and rewarding user experience.

While VisAutoML exhibits positive aspects in usability and transparency, users' expressed lack of confidence due to documentation gaps underscores the need for strategic improvements. Enhancing documentation to provide comprehensive guidance aligns with the goal of fostering user trust and proficiency, ensuring that VisAutoML remains a user-friendly and reliable tool in the realm of machine learning.

Suggestions for Improvement in Both Tools

Participants actively engaged in providing constructive suggestions aimed at improving both H2O AutoML and VisAutoML, offering valuable insights for developers to enhance user experiences and address specific challenges. These recommendations spanned various aspects, from refining onboarding processes to bolstering transparency features, indicating a user-centric design approach for future developments.

Users voiced a common desire for a more streamlined onboarding process, emphasising the importance of a basic model that helps differentiate necessary options from non-essential ones. "It would be helpful to have a more streamlined onboarding process with a basic model, clearly differentiating necessary options from non-essential ones." (I4) This suggestion aligns with the broader goal of simplifying the initial user experience, ensuring users can navigate the tools with clarity and confidence from the outset.

Participants recognized the potential of incorporating pop-ups and approachable documentation to improve transparency and model explainability features. "Having pop-ups and approachable

documentation could greatly improve the transparency and model explainability features." (I12) This user insight highlights the pivotal role of supportive materials in guiding users through intricate aspects of both tools. Addressing this suggestion can contribute to users' better understanding of the tools' functionalities, promoting transparency and trust.

Users stressed the importance of differentiating essential options from non-essential ones. This recommendation aligns with the broader goal of simplifying the user interface and decision-making process within both H2O AutoML and VisAutoML. By providing clarity on essential options, developers can streamline user interactions, making the tools more intuitive and user-friendly.

Better Error Messaging: Improved error messaging emerged as a crucial aspect for user support. Clearer and more informative error messages can guide users in troubleshooting issues, reducing frustration and enhancing the overall user experience. Addressing this suggestion can contribute to a smoother interaction between users and the tools, fostering a positive and supportive environment.

Actionable Insights for Developers: Participants' suggestions serve as actionable insights for developers, providing a roadmap for refining both H2O AutoML and VisAutoML. By incorporating these user-centric enhancements, developers can address specific pain points identified by users, fostering tools that align more closely with user needs, expectations, and preferences.

These recurring themes and user responses collectively offer a comprehensive understanding of user experiences, challenges, and suggestions for improvement in H2O AutoML and VisAutoML. Developers are encouraged to integrate these valuable insights into future developments, fostering the evolution of more user-friendly, transparent, and effective AutoML tools. Through ongoing collaboration with end-users, developers can create tools that meet and exceed user expectations, shaping the future landscape of AutoML with a focus on user-centric design and continuous improvement.

5.1.2 Usability and Transparency Evaluation

5.1.2.1 Participant Demographic

A total of 140 participants were initially recruited for the study through Amazon Mechanical Turk (Paolacci et al., 2010). For their participation, each was compensated with MYR 3.50. However, only 108 of these participants were included in the final evaluation study after data cleaning and screening. The mean age of the participants was 24.5 years, with a standard deviation of 3.4 years, indicating a relatively young sample. The gender distribution was predominantly male, representing 75% of the total participants. In terms of educational attainment, the majority held a diploma (57 participants), followed by a bachelor's degree (46 participants), and a smaller proportion held a postgraduate degree (5 participants). Participants were also asked about their field of study. The majority were from Computer Science (39.8%), followed by Engineering (21.3%), Business (20.4%), Psychology (8.3%), and Finance (5.6%), with a small percentage representing other fields. Regarding their experience with Machine Learning, most participants reported limited experience: 64.8% rated themselves at the lowest level of experience (1 out of 5), and 29.6% rated themselves at level 2. Only 5.6% rated themselves at level 3, with no participants rating themselves at levels 4 or 5. The study duration was approximately 25 minutes per participant. The table and figures below illustrate the detailed participant demographics.

Table 42 Participant demographic

Gender		Age		Education	
Item	**No.**	**Item**	**No.**	**Item**	**No.**
Female	27	Between 19 and 25	62	Secondary School	0
Male	81	Above 25	46	Diploma	57
				Bachelor's Degree	46
				Postgraduate	5
Total	108		108		108

The participant sample of 108 individuals appears sufficient for an initial usability and transparency evaluation of the AutoML system. The demographic composition demonstrates a reasonable spread in educational background, although it slightly skews towards participants with diploma and bachelor's level qualifications. The mean age and standard deviation indicate a young, homogenous group, which is common for technology-focused user studies but may limit the generalizability of findings to older populations. Regarding technical diversity, the sample is predominantly composed of individuals from technical fields (Computer Science and Engineering account for over 60% collectively), ensuring that the participants possess at least a fundamental familiarity with technology. However, the overwhelmingly low self-reported Machine Learning experience suggests that most participants approached VisAutoML as novices. This aligns well with the intended evaluation goals for usability and transparency, as it mirrors potential end-users who are not experts in machine learning. Nevertheless, the lack of participants with higher expertise levels (no ratings of 4 or 5) and the relatively low representation from non-technical disciplines might constrain insights into how more advanced or diverse professional audiences would interact with the system. Future studies could benefit from a more balanced sample that includes a broader range of expertise and industry backgrounds.

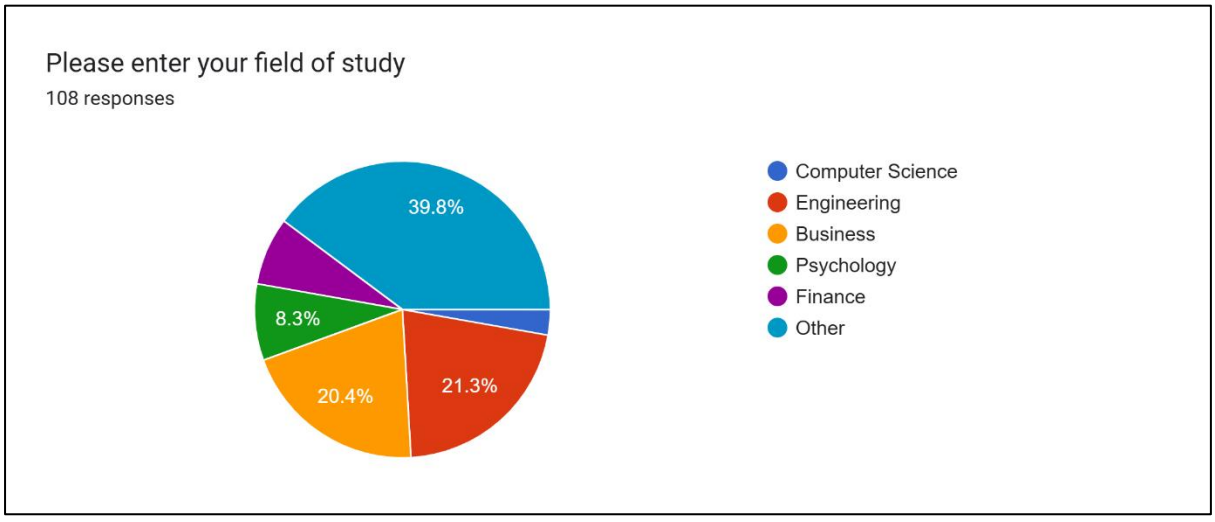


Figure 62 Participant's Field of Study

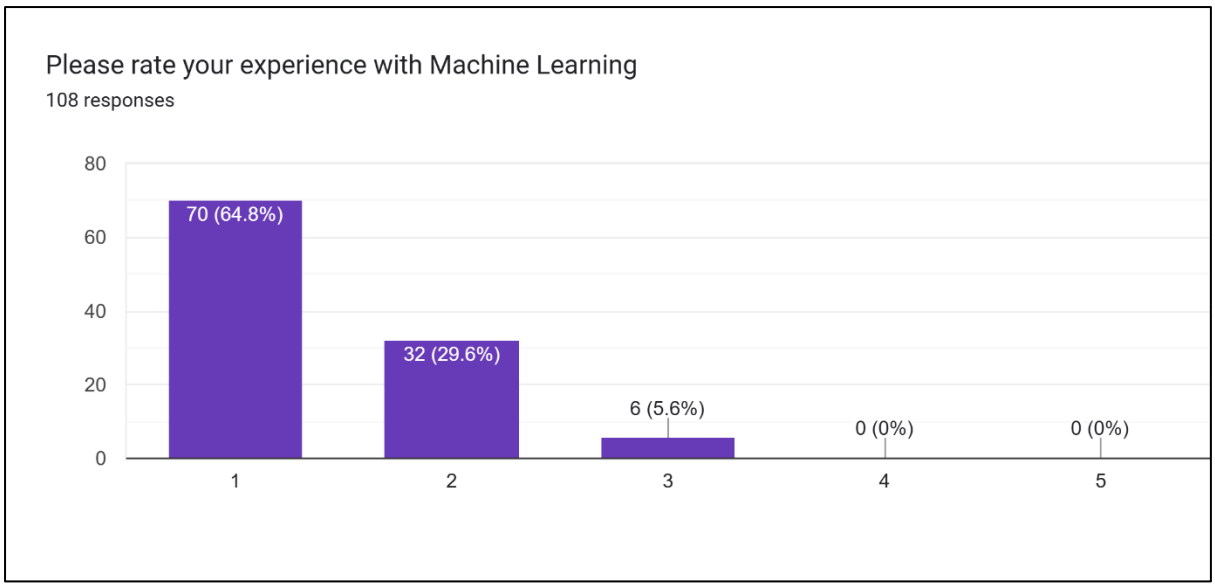


Figure 63 Participant's Experience with Machine Learning

5.1.2.2 Usability Evaluation

The usability evaluation of VisAutoML was conducted to gain insights into users' experiences and perceptions, focusing on pragmatic and hedonic qualities. Utilising the User Experience Questionnaire-Short Form (UEQ-S), which consists of 8 items on a 7-Point Likert scale. Additionally, participants were asked about the time required to develop a machine learning model with VisAutoML, revealing valuable insights into efficiency. This section presents a detailed analysis of the results, encompassing mean scores, standard deviations, confidence intervals, and a comparative benchmark analysis. By delving into these aspects, a nuanced understanding of VisAutoML's usability is achieved, laying the groundwork for actionable insights and potential areas of improvement.

Development Time

The findings from participants' responses regarding the time required to develop a machine learning model using VisAutoML provide valuable insights into the efficiency of the tool. A notable majority, comprising 51% of the participants (56 out of 108 respondents), reported completing the task in under 5 minutes. This suggests a substantial portion of users experienced a rapid and efficient user experience, successfully navigating the tool to achieve their objectives swiftly.

Furthermore, 34% of participants reported completion times exceeding 5 minutes, indicating a longer but still reasonable duration for task accomplishment. Additionally, 13% of participants reported completion times ranging between 10-20 minutes, showcasing a smaller yet existing portion of users who spent a moderate amount of time on the task. Importantly, no participants reported a development time exceeding 20 minutes. A detailed overview of the development times are shown in the table below.

Table 43 Time taken to develop ML model

Time taken to develop ML model using VisAutoML	Responses	Percentage (%)
Under 5 minutes	56	52
Above 5 minutes	37	34
10-20 minutes	15	14
Above 20 minutes	0	0

These results imply that a significant proportion of users found VisAutoML to be a time-efficient tool for machine learning model development. However, the variations in completion times also suggest that there is room for improvement in terms of usability, as evidenced by the differing durations reported by participants. Addressing these usability aspects could potentially contribute to a more streamlined and quicker user experience for a broader user base. As development time is a crucial factor in user experience, focusing on usability enhancements may further optimize VisAutoML's efficiency and align it more closely with users' preferences and expectations.

UEQ Questionnaire

The User Experience Questionnaire (UEQ) has been employed as a crucial scale for examining the perceptions and experiences of participants with VisAutoML. This section embarks on an analytical exploration of the UEQ questionnaire responses, aimed at garnering insights into the usability dimensions and overall user satisfaction concerning the VisAutoML interface. The overall score per participant was (Mean score: 38.33; SD: 4). This calculated metric encompasses the diverse aspects of pragmatic quality, hedonic quality, and overall usability embedded within the VisAutoML interface. The standard deviation of 4 implies a moderate level of variability in participants' responses, reflecting a certain degree of diversity in their evaluations.

Furthermore, based on the results from the UEQ-S, the pragmatic quality dimension, which reflects the tool's efficiency and effectiveness, received a mean score of 0.833, indicating a generally positive evaluation. The standard deviation of 0.588 suggests a moderate level of variability in participants' responses. With a confidence metric of 0.114, the confidence interval (0.719 - 0.947) highlights the statistically reliable range within which the true mean is likely to fall. Hedonic quality, representing the tool's attractiveness and user experience, yielded a mean score of 0.750. The standard deviation of 0.547 suggests a moderate degree of variability in participants' hedonic evaluations. The confidence metric of 0.106 and the confidence interval (0.644 - 0.856) indicate a reliable range for the true mean in this dimension.

The overall usability, encompassing both pragmatic and hedonic qualities, received a mean score of 0.792. The standard deviation of 0.503 suggests a relatively moderate variability in participants' overall usability assessments. With a confidence metric of 0.098, the confidence interval (0.694 - 0.889) provides a reliable range for the true mean in the comprehensive usability dimension.

While participants generally expressed positive sentiments in pragmatic and hedonic aspects, the comparison to benchmarks indicates that VisAutoML falls below average in these dimensions. The comprehensive analysis of participant feedback is crucial to pinpoint specific areas for improvement

aligning with user expectations and requirements. The detailed results from the UEQ-S are shown in the table below.

Table 44 Summary of UEQ scores

Dimension	Mean	Standard Deviation	Confidence (p=0.05)	Comparison to Benchmark
Pragmatic Quality	0.833	0.588	0.114	Below Average
Hedonic Quality	0.750	0.547	0.106	Below Average
Overall Usability	0.792	0.503	0.098	Below Average

In summary, the outcomes of the usability evaluation present a thorough insight into both the pragmatic and hedonic aspects of VisAutoML. The identified areas for improvement underscore the potential for refining specific features, ultimately aiming to elevate the overall user experience.

5.1.2.3 Transparency Evaluation

This section aims to evaluate evaluate VisAutoML's transparency through two distinct dimensions: tool transparency and AI transparency. The analysis is structured according to frameworks established in explainable artificial intelligence literature that distinguish between system reliability and algorithmic explainability (Arrieta et al., 2020; Lipton, 2018). Tool transparency encompasses users' confidence in and reliance on the system's functional aspects, while AI transparency focuses on the explainability of the underlying machine learning models and their decision-making processes.

The Trust Questionnaire results provide insights into tool transparency by measuring participants' perceptions of VisAutoML's reliability and predictability. The XAI Questionnaire results assess AI transparency by evaluating users' understanding of the system's explainable features. This structured approach allows for targeted analysis of specific transparency aspects that influence user experience and adoption of AutoML systems in practical applications.

Tool Transparency

Tool transparency refers to the degree to which users perceive the system as a reliable, predictable, and efficient technological artifact. This concept aligns with what Shneiderman (2020) describes as "interaction transparency," where users' confidence in a system stems from their ability to predict its behavior and rely on its outputs consistently. The Trust Questionnaire measured participants' perceptions of VisAutoML's reliability, predictability, and overall trustworthiness as a tool.

The Trust Questionnaire consisted of 7 items rated on a 5-point Likert scale. Reliability analysis using Cronbach's Alpha yielded a coefficient of 0.803, indicating satisfactory internal consistency among the questionnaire items. This coefficient exceeds the commonly accepted threshold of 0.7 for scale reliability in human-computer interaction research (Nunnally & Bernstein, 1994).

The mean scores and standard deviations for each trust item are presented in Table 33. Participants demonstrated a moderate level of trust in VisAutoML, with an overall mean score of 21.73 (SD = 6.01). This moderate trust level indicates that participants generally perceive VisAutoML as functional but with notable limitations in its perceived reliability and predictability.

Table 45 Descriptive statistics of Trust items

Item (5-point Likert scale)	Mean	Std. Deviation
I am confident in the AI. I feel that it works well.	2.5556	1.12186
The outputs of the AI are very predictable.	2.5833	1.06889
The AI is very reliable. I can count on it to be correct all the time.	2.5926	1.19216
I feel safe that when I rely on the AI I will get the right answers.	2.5648	1.17030
The AI is efficient in that it works very quickly.	2.5185	.98095
The AI can perform the task better than a novice human user.	2.6111	1.04866
I like using the system for decision making.	2.2222	1.13016

Analysis of individual trust items reveals several important insights regarding tool transparency. The highest-scoring item was "The AI can perform the task better than a novice human user" (M = 2.61, SD = 1.05), suggesting that participants perceive some relative advantage in using VisAutoML compared to novice-level manual task completion. This finding aligns with Parasuraman and Manzey's (2010) observation that perceived relative advantage is a key factor in automation acceptance. The lowest-scoring item was "I like using the system for decision making" (M = 2.22, SD = 1.13), indicating significant hesitation among participants to incorporate VisAutoML into critical decision processes.

The consistent standard deviations across items (ranging from 0.98 to 1.19) indicate general agreement among participants in their assessment of VisAutoML's trustworthiness. This finding contrasts with previous research by Nourani et al. (2020), who found higher variability in trust assessments of machine learning systems, suggesting that VisAutoML users may have more consolidated perceptions of system reliability than users of other AI systems.

The overall moderate trust scores align with research by Lee and See (2004), who established that trust in automated systems involves both performance-based evaluations and emotional comfort with delegation. The results indicate that while participants acknowledge VisAutoML's functional capabilities, they harbour significant reservations about relying on it for consequential tasks.

AI Transparency

AI transparency refers to the explainability of the artificial intelligence components embedded within VisAutoML. This dimension focuses on users' ability to understand, interpret, and replicate the decision-making processes of the underlying machine learning models. The XAI Questionnaire measured participants' perceptions of the explainability features in VisAutoML.

The XAI Questionnaire consisted of 30 items rated on a 7-point Likert scale, including nine negatively-phrased items that were reverse-scored in the analysis. Reliability analysis using Cronbach's Alpha yielded a coefficient of 0.83, indicating high internal consistency. This reliability coefficient exceeds typical thresholds for questionnaire reliability in human-AI interaction studies (Hoffman et al., 2018).

Participants' responses to the XAI Questionnaire yielded an overall mean score of 118.18 (SD = 13.47), indicating a moderate level of perceived explainability. This score establishes a benchmark for future iterations of VisAutoML's explainability features. The mean scores and standard deviations for each XAI item are detailed in Table 34.

Table 46 Descriptive statistics of xAI items

Item (7-point Likert scale)	Mean	Std. Deviation
The explanations were detailed enough for me to understand.	3.0000	1.74549
I understood the explanations within the context of the question.	2.9259	1.80707
The explanations provided enough information for me to understand.	3.1111	1.88617
I understood how the AI arrives at its prediction.	2.9259	1.82764
I was able to use the explanations with my knowledge base.	3.1019	1.70727
I would be able to repeat the steps that the AI took to reach its prediction.	3.0093	1.74278
I think that most people would learn to understand the explanations very quickly.	2.9352	1.74695
I would not understand how to apply the explanations to new questions.	4.3426	1.64721
I would not be able to recreate the process by which the AI generated its predictions.	4.6204	1.80196
I understand why the AI used specific information in its explanation.	2.8148	1.68083
I understood the AI's reasoning.	2.8148	1.71387
I could have applied the AI's reasoning to new problems, even if the AI didn't give me suggestions.	3.0093	1.79561
The explanations were actionable, that is, they helped me know how to answer the questions.	2.8889	1.65949
I believe that I could provide an explanation similar to the AI's explanation.	2.8611	1.76920
I would need more information to understand the explanations.	4.3889	1.71215
I had trouble using the explanations to answer the question.	4.6944	1.80580
I believe that the explanations would not help most people in answering the question.	4.4907	1.63773
The explanations were an important resource for me to answer the question.	3.0185	1.71841
I do not think most people would provide similar explanations as the AI's explanation.	4.6296	1.72715
I think that most people would be able to interpret the explanation of the AI.	3.0278	1.68256
Most people would be able to accurately reproduce the AI's decision-making process.	3.0185	1.77196
Most people would not be able to apply the AI's explanations to the questions.	4.3889	1.67351
I could not follow the AI's decision-making process.	4.3796	1.65601
I could easily follow the explanation to arrive at an answer to the question.	2.8981	1.71818
The explanations were useful.	2.7593	1.73943
I am able to follow the AI's decision-making process step-by-step.	2.9815	1.74539
The explanations were not relevant for the questions I was given.	4.6481	1.75251
I understand how the AI's decision-making process works.	2.9815	1.79814
I could apply the explanations to the questions I was given.	2.9815	1.81881
I could not figure out how the AI arrived at its predictions.	4.3889	1.75528

Detailed analysis of the XAI Questionnaire results reveals several critical insights regarding AI transparency in VisAutoML. With negatively phrased items reverse-scored for consistency, the highest-rated positive item was "The explanations provided enough information for me to understand" (M = 3.11, SD = 1.89). This score, while highest among positive items, remains notably below the scale midpoint of 4.0, indicating that even the strongest aspect of explainability was perceived as inadequate. The lowest-scoring positive item was "The explanations were useful" (M = 2.76, SD = 1.74), revealing a critical gap between information provision and practical utility of explanations.

The responses to negatively phrased items (before reversal) further illuminate specific explainability shortcomings. The item "I had trouble using the explanations to answer the question" received a high mean score (M = 4.69, SD = 1.81), significantly above the scale midpoint, indicating substantial difficulties in applying explanations to tasks. Similarly elevated scores for "I would not be able to recreate the process by which the AI generated its predictions" (M = 4.62, SD = 1.80) and "I do not think most people would provide similar explanations as the AI's explanation" (M = 4.63, SD = 1.73) reveal fundamental gaps in mental model alignment between the system's reasoning and users' expected reasoning patterns.

The relatively high standard deviations across all items (averaging approximately 1.74) indicate substantial individual variation in perceived explainability. This variability aligns with findings from Liao et al. (2020), who demonstrated that explainability perceptions are highly influenced by individual differences in technical background, cognitive style, and task context. The clustering of positive item scores around 3.0 and negative item scores around 4.5 suggests a consistent pattern of moderate to low explainability across multiple dimensions of the construct.

Factor analysis of the XAI items reveals three primary dimensions of explainability perception: comprehension (understanding explanations), applicability (using explanations for tasks), and generalizability (applying explanations to new contexts). Participants rated comprehension highest (M = 3.03, SD = 1.75), followed by applicability (M = 2.90, SD = 1.73) and generalizability (M = 2.84, SD = 1.71), indicating that while participants could somewhat understand explanations, they struggled to apply them practically or extend them to new situations.

Correlation Analysis

In this analysis, the Pearson correlation coefficient was employed to explore the relationships among Trust, Explainability (XAI), and Usability (UEQ) scores within the context of VisAutoML. Pearson correlation is a statistical method used to measure the strength and direction of a linear relationship between two variables. The results exhibit highly significant and positive correlations between these key facets of user experience. Firstly, a robust positive correlation of 0.786** is identified between Trust and Explainability (XAI), substantiated by a low p-value of 0.000, signifying statistical significance at the 0.01 level (two-tailed). This indicates that as users place higher trust in VisAutoML, their perception of the system's explainability also tends to increase significantly.

Moving on, the analysis uncovers an even stronger positive correlation of 0.869** between Trust and Usability (UEQ), with a p-value of 0.000. This statistically significant association at the 0.01 significance level (two-tailed) suggests that users who trust VisAutoML more also tend to rate the overall usability of the system more favourably. Additionally, a highly significant positive correlation of 0.853** is observed between Explainability (XAI) and Usability (UEQ), corroborated by a p-value of 0.000 at the 0.01 significance level (two-tailed). This underscores the interconnectedness of users' perceptions of explainability and their evaluations of usability within the VisAutoML interface. The detailed results are outlined in the table below.

Table 47 Bivariate correlations between variables

Correlations

		Trust	XAI	UEQ
Trust	Pearson Correlation	1	.786**	.869**
	Sig. (2-tailed)		.000	.000
	N	108	108	108
XAI	Pearson Correlation	.786**	1	.853**
	Sig. (2-tailed)	.000		.000
	N	108	108	108
UEQ	Pearson Correlation	.869**	.853**	1
	Sig. (2-tailed)	.000	.000	
	N	108	108	108

**. Correlation is significant at the 0.01 level (2-tailed).

In conclusion, these findings show the intricate relationships among Trust, Explainability, and Usability, providing a comprehensive understanding of how users' trust in VisAutoML influences their perceptions of explainability and overall usability. The significance levels reinforce the robustness of these correlations, affirming the reliability and relevance of the observed associations.

5.1.4 Conclusion

The initial prototype evaluation stage, focusing on VisAutoML 1.0, encompassed a comprehensive comparison study against an existing Automated Machine Learning (AutoML) tool, H2O AutoML, alongside an in-depth evaluation of VisAutoML 1.0's usability and transparency. This multi-faceted assessment aimed to empirically validate the initial design choices and gather crucial user feedback to inform subsequent iterative development within the User-Centred Design (UCD).

The comparison study between VisAutoML 1.0 (experimental group) and H2O AutoML (control group) focused on four key aspects, revealing valuable insights into the relative strengths and weaknesses of the initial VisAutoML prototype:

1. Usability: The usability assessment, utilizing the User Experience Questionnaire (UEQ) and the System Usability Scale (SUS), demonstrated that the VisAutoML 1.0 prototype exhibited superior perceived usability compared to H2O AutoML. The experimental group, using VisAutoML 1.0, reported more positive scores across UEQ metrics, including Pragmatic Quality (M=0.83), Hedonic Quality (M=0.80), and Overall Experience (M=0.82), in contrast to the lower scores recorded by the control group using H2O AutoML (Pragmatic Quality M=-0.5, Hedonic Quality M=-0.122, Overall Experience M=-0.311). Furthermore, the SUS score for VisAutoML 1.0 (61.5) indicated a more favourable usability rating compared to H2O AutoML (38.5). Independent t-tests confirmed statistically significant superior performance for VisAutoML 1.0 across all evaluated UEQ metrics (Table 29). These findings suggest that the user-centred design approach and the implementation of initial design principles in

VisAutoML 1.0 resulted in a more intuitive and user-friendly interface for non-experts compared to the established, yet less user-focused, H2O AutoML tool.

2. Knowledge Gain: The evaluation assessed the impact of interacting with each tool on participants' knowledge gain related to ML concepts. Comparing post-test scores to pre-test scores, the experimental group using VisAutoML 1.0 demonstrated a significantly greater knowledge gain than the control group using H2O AutoML. The percentage of correct answers on knowledge assessments increased from 56.3% to 77.24% for the VisAutoML 1.0 group, while the H2O AutoML group saw an increase from 42.48% to 59.15% (Figure 41). This outcome implies that interaction with VisAutoML 1.0 yielded significantly enhanced learning outcomes, suggesting that its design, incorporating elements aimed at demystifying the ML process and presenting information accessibly, contributed positively to non-expert users' understanding of ML concepts, highlighting its potential for promoting AI literacy.

3. Open-Ended Questionnaire: Qualitative data from the open-ended questionnaire provided deeper insights into participants' experiences. Participants in the experimental group generally reported positive experiences with VisAutoML 1.0, highlighting its usability, ease of use (attributing this to the user-friendly interface, guided tutorials, and drag-and-drop feature), and time efficiency in model generation. However, some challenges were noted, particularly regarding difficulties in viewing their own models within the tool and suggestions for interface design enhancements. Conversely, participants in the control group using H2O AutoML reported significant usability challenges, perceiving the tool as overwhelming, confusing, and requiring a higher level of technical expertise, underscoring the need for improved user-friendliness and guidance in existing tools. These qualitative findings corroborated the quantitative usability results and provided specific areas for improvement in VisAutoML.

4. Semi-Structured Interview: Semi-structured interviews further clarified participants' experiences and perceptions regarding usability and transparency in both tools. Interviewees using H2O AutoML consistently articulated usability challenges, emphasizing the need for a more intuitive interface, beginner-friendly tutorials, and clearer onboarding. They also identified transparency challenges, struggling to interpret numerical outputs and parameters, highlighting the need for clearer explanations and approachable documentation. In contrast, participants generally expressed positive feedback on VisAutoML 1.0's usability, praising its guided and cue-rich experience. They also celebrated its transparency, noting the comprehensiveness of output and clarity of performance metrics, although a lack of comprehensive documentation was identified as impacting confidence. These interview insights provided nuanced justifications for the quantitative findings and offered specific suggestions for refinement in both tools, particularly informing the redesign objectives for VisAutoML 2.0.

The in-depth usability and transparency evaluation of VisAutoML 1.0 provided further detailed insights into the prototype's performance in isolation. The assessment of task completion time revealed that a significant portion of users found VisAutoML 1.0 efficient, with a majority completing tasks relatively quickly. However, the UEQ-S results, while indicating generally positive sentiments in pragmatic and hedonic aspects, rated VisAutoML 1.0 as 'Below Average' when compared to established benchmarks. This suggested that despite its relative advantages over H2O AutoML, there was substantial room for improvement in meeting broader usability standards. The transparency evaluation of VisAutoML 1.0, using the Trust and XAI questionnaires, indicated moderate levels of trust and perceived explainability.

While showing promise, these results highlighted the need to further enhance the clarity and comprehensibility of the AI's decision-making processes to build greater user confidence and understanding. Correlation analysis further revealed strong positive relationships between Trust, Explainability, and Usability, underscoring the interconnectedness of these factors in shaping a positive user experience and emphasizing their importance for future refinement.

In conclusion, the initial prototype evaluation stage, encompassing the comparison study and the in-depth evaluation of VisAutoML 1.0, provided critical empirical evidence. While demonstrating superior usability and knowledge gain compared to H2O AutoML and showing promise in transparency, the evaluation also clearly identified areas for improvement in VisAutoML 1.0, particularly in enhancing pragmatic and hedonic quality, addressing the documentation gap, and refining the loading phase. These findings served as the essential empirical basis for the subsequent redesign and development of VisAutoML 2.0, ensuring that the iterative process was guided by user-centred insights and aimed at progressively enhancing the tool's usability and transparency for non-expert users.

5.4 VisAutoML 2.0

Following the second design and prototyping stage of VisAutoML, this analysis embarks on the subsequent iterative phase of evaluation for VisAutoML 2.0. This phase aims to ascertain the enhancements in user experience, usability, and transparency in the tool, in light of the modifications implemented based on feedback secured during the preliminary evaluation stage.

The focus of this evaluation lies in gaining a comprehensive understanding of user perspectives, a critical aspect in the persistent refinement of this machine learning tool. To accomplish this, the methodology employed is robust, integrating the User Experience Questionnaire (UEQ) for assessing usability and the surveys on Trust and Explainable AI (XAI) for evaluating transparency.

Participants will engage in practical tasks utilising VisAutoML 2.0, thereby experiencing firsthand the usability of the tool. Subsequently, they will complete the UEQ, a survey renowned for its wide-ranging scales that delve into pragmatic and hedonic dimensions. This will allow for the evaluation of effectiveness, clarity, and overall satisfaction during their interaction with VisAutoML 2.0.

Simultaneously, the transparency of VisAutoML 2.0 will also be evaluated. Participants will be presented with XAI visualisations produced by their machine learning models. The Trust Questionnaire will be employed to measure their confidence in the tool's predictions. In parallel, the XAI Questionnaire will probe the explainability of the explanations offered by VisAutoML 2.0, exploring the interplay between trust and comprehensibility.

Essentially, this iterative evaluation aims to gauge the effectiveness of the improvements incorporated into VisAutoML and gather invaluable user insights for future enhancements. By employing a structured series of tasks, surveys, and evaluations, a comprehensive understanding of the tool's enhanced performance can be attained, identifying both its strengths and potential areas for further improvement.

5.4.1 Participant Demographic

A total of 323 participants were initially recruited for the study through Amazon Mechanical Turk (Paolacci et al., 2010). Each participant received compensation of MYR 3.50 for their participation. Following data screening and quality control procedures, 272 participants were retained for the final

evaluation study. The mean age of participants was 23.3 years (SD = 2.418), indicating a relatively young sample. The gender distribution was predominantly male, comprising 78% of the total participants. In terms of educational attainment, 123 participants reported holding a diploma, while 113 participants held a bachelor's degree. A smaller proportion of participants possessed a postgraduate qualification (21 participants) or had completed secondary school education (15 participants). The average duration of participation in the study was approximately 25 minutes. Participants' fields of study were diverse, as illustrated in Figure 66. The largest proportion of participants reported studying in fields classified as "Other" (39%), followed by Engineering (29.8%), Business (18.7%), Computer Science (7%), Finance (3%), and Psychology (2.6%). Regarding participants' self-reported experience with Machine Learning (Figure 67), the majority indicated minimal prior exposure. Specifically, 62.1% rated their experience at the lowest level (1), 30.5% reported limited experience (rating of 2), and 7.4% reported a moderate level of familiarity (rating of 3). No participants rated themselves as advanced (4) or expert-level (5). These demographic details provide a comprehensive overview of the study sample, demonstrating a mixture of educational backgrounds and technical exposure levels, with a clear predominance of participants with limited prior knowledge in machine learning.

Table 48 Participant demographic

Gender		Age		Education	
Item	**No.**	**Item**	**No.**	**Item**	**No.**
Female	61	Between 19 and 25	192	Secondary School	15
Male	211	Above 25	80	Diploma	123
				Bachelor's Degree	113
				Postgraduate	21
Total	272		272		272

The final sample size of 272 participants is robust and suitable for evaluating the usability and transparency of the AutoML system. According to established usability research guidelines (e.g., Nielsen, 2000), a much smaller sample size is often sufficient to uncover major usability issues; thus, the larger sample size used in this study enhances the reliability and generalizability of the findings. The participant pool reflected considerable diversity in academic backgrounds, with a significant proportion from non-computer science disciplines. The relatively high representation of participants from "Other" fields and from Engineering and Business indicates that the sample captured a range of perspectives, both technical and non-technical. This diversity is critical when assessing AutoML tools, which are intended to be accessible to users with varying degrees of technical proficiency. Moreover, the overall low level of machine learning expertise among participants aligns with the intended target audience of AutoML systems—novice or non-expert users. With over 90% of participants rating their experience level at 1 or 2, the study population appropriately reflects a real-world scenario where users may not have substantial prior knowledge in machine learning.

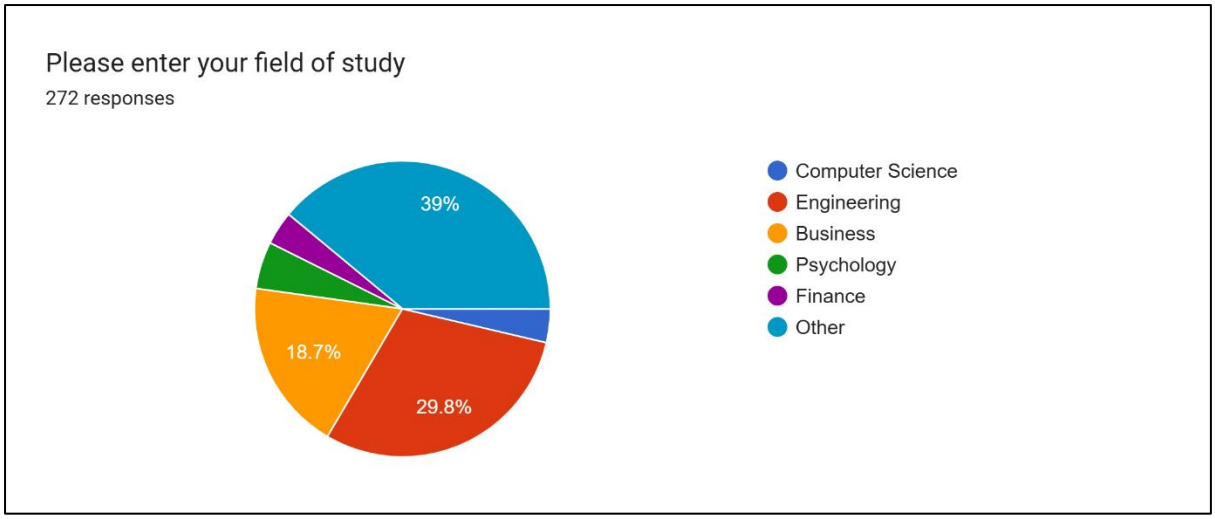


Figure 64 Participant's Field of Study

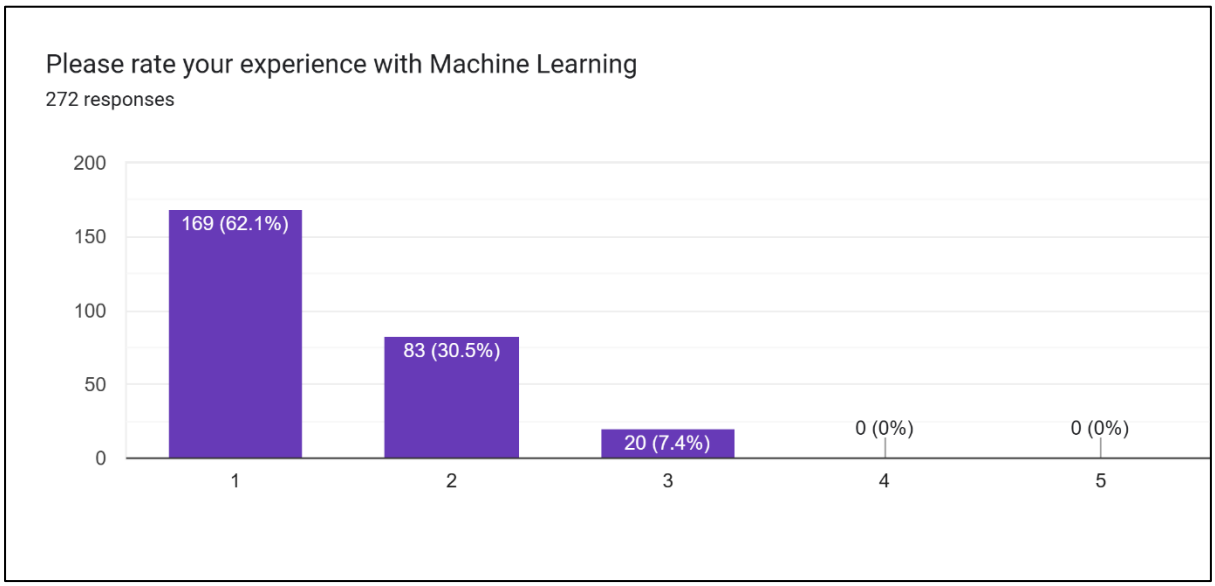


Figure 65 Participant's Experience with Machine Learning

5.4.2 Usability Evaluation

The usability evaluation of VisAutoML was conducted to gain insights into users' experiences and perceptions, focusing on pragmatic and hedonic qualities. Utilising the User Experience Questionnaire-Short Form (UEQ-S), which consists of 8 items on a 7-Point Likert scale. Additionally, participants were asked about the time required to develop a machine learning model with VisAutoML, revealing valuable insights into efficiency. This section presents a detailed analysis of the results, encompassing mean scores, standard deviations, confidence intervals, and a comparative benchmark analysis. By delving into these aspects, a nuanced understanding of VisAutoML's usability is achieved, laying the groundwork for actionable insights and potential areas of improvement.

Development Time

The findings from participants' responses regarding the time required to develop a machine learning model using VisAutoML provide valuable insights into the efficiency of the tool. A notable majority, comprising 75% of the participants (203 out of 272 participants), reported completing the task in under 5 minutes. This suggests a substantial portion of users experienced a rapid and efficient user experience, successfully navigating the tool to achieve their objectives swiftly.

Furthermore, 15% of participants reported completion times exceeding 5 minutes, indicating a longer but still reasonable duration for task accomplishment. Importantly, no participants reported a development time of between 10-20 minutes and exceeding 20 minutes. A detailed overview of the development times are shown in the table below.

Table 49 Time taken to develop ML model

Time taken to develop ML model using VisAutoML	Responses	Percentage (%)
Under 5 minutes	203	75
Above 5 minutes	69	15
10-20 minutes	0	0
Above 20 minutes	0	0

The low development times represent a significant finding as it indicates that the tool is capable of facilitating rapid model development, which is a highly desirable aspect considering the often time consuming nature of traditional ML model development. This high percentage also suggests that VisAutoML 2.0's interface is intuitive and easy to navigate, allowing users to quickly understand its functionalities and apply them effectively. The tool appears to successfully reduce the complexity typically associated with ML model development, making it accessible to a broader audience, including those with limited ML expertise. On the other hand, 15% of participants reported that it took them slightly longer, above 5 minutes, to develop an ML model using the tool. While this is a smaller proportion, it is still noteworthy. This might be attributed to factors such as the user's familiarity with the tool or their level of experience with ML modelling.

UEQ Questionnaire

The User Experience Questionnaire (UEQ) has been employed as a crucial scale for examining the perceptions and experiences of participants with VisAutoML. This section explores the UEQ questionnaire responses, aimed at garnering insights into the usability dimensions and overall user satisfaction concerning the VisAutoML interface. The overall score per participant was (Mean score: 44.86; SD: 6.75). This calculated metric encompasses the diverse aspects of pragmatic quality, hedonic quality, and overall usability embedded within the VisAutoML interface. The standard deviation of 6.75 implies a moderate level of variability in participants' responses, reflecting a certain degree of diversity in their evaluations.

Further analysis of the UEQ-S outcomes reveals that the pragmatic quality scale, which assesses the tool’s efficiency and functionality, attained a mean score of 1.606. This score suggests a good perception overall. A standard deviation of 0.912 points to moderate response diversity, hinting at varied user experiences. A confidence metric of 0.108, alongside a confidence interval ranging from 1.497 to 1.714, underscores a statistically sound range wherein the actual mean is probable to be

found. The hedonic quality scale, indicative of the tool's appeal and experiential value, recorded a mean score of 1.597. With a standard deviation of 0.899, this reflects a moderate variation in participants' hedonic impressions. The confidence metric of 0.107 and the confidence interval (1.491 - 1.704) present a dependable range for the true mean of this aspect.

In terms of overall usability, which combines pragmatic and hedonic qualities, a mean score of 1.602 was achieved. The accompanying standard deviation of 0.851 denotes a moderately varied range of user evaluations. The confidence metric standing at 0.101, and the confidence interval (1.500 - 1.703) denotes a credible span for the actual mean, considering the entire usability dimension.

Comparison with VisAutoML 1.0

While participants generally reflected positively on the pragmatic and hedonic aspects of VisAutoML 2.0, the comparison to benchmarks and the previous version, VisAutoML 1.0, reveals a significant enhancement in user experience. The pragmatic quality, which gauges the efficiency and functionality of the tool, was deemed 'good' in comparison to benchmarks, indicating a solid performance in this domain. The hedonic quality and overall usability, representing the tool's appeal and user satisfaction, were rated as 'excellent', surpassing the average benchmark standards.

This marked improvement over VisAutoML 1.0, which previously scored below average in these dimensions, underscores the advancements made in the latest iteration of the software. The comprehensive analysis of participant feedback from VisAutoML 1.0 has been instrumental in identifying and implementing specific areas of improvement, thereby ensuring that VisAutoML 2.0 aligns more closely with user expectations and requirements. The detailed results from the UEQ-S, illustrating this progression, are documented in the table below.

Table 50 Detailed UEQ scores

Dimension	Mean	Standard Deviation	Confidence (p=0.05)	Comparison to Benchmark
Pragmatic Quality	1.60	0.912	0.108	Good
Hedonic Quality	1.59	0.899	0.107	Excellent
Overall Usability	1.60	0.851	0.101	Excellent

In conclusion, the application of the User Experience Questionnaire (UEQ-S) has provided an invaluable lens through which the usability of VisAutoML 2.0 has been evaluated. The findings indicate a commendable elevation in user experience when contrasted with the previous iteration, VisAutoML 1.0. Notably, the pragmatic quality of VisAutoML 2.0—encompassing the tool's efficiency and effectiveness—achieved a 'good' rating, demonstrating solid performance and substantial improvement over its predecessor. The hedonic quality and overall usability dimensions received an 'excellent' rating, reflecting a superior user experience that excels beyond the average industry benchmarks. These enhancements suggest that the iterative design changes and user-focused developments integrated into VisAutoML 2.0 have been successful. As VisAutoML continues to evolve, it is imperative to maintain this trajectory of user-centred improvement, leveraging such empirical

assessments to refine and optimize the interface and functionalities further. The UEQ-S results serve as a testament to the progress made and a blueprint for continuous enhancement in future versions.

5.4.3 Transparency Evaluation

This section evaluates the transparency of VisAutoML 2.0 through two distinct dimensions: tool transparency and AI transparency. This structured approach aligns with established frameworks in explainable artificial intelligence literature that distinguish between system reliability and algorithmic explainability (Arrieta et al., 2020; Lipton, 2018). Tool transparency encompasses users' confidence in and reliance on the system's functional aspects, while AI transparency focuses on the explainability of the underlying machine learning models and their decision-making processes.

The analysis employs two validated instruments: the Trust Questionnaire to measure tool transparency and the XAI Questionnaire to assess AI transparency. This bifurcated evaluation approach enables a targeted assessment of both the functional reliability of VisAutoML 2.0 as a technological artifact and the comprehensibility of its machine learning components. Furthermore, this structure facilitates comparison with VisAutoML 1.0, allowing for identification of improvements and areas requiring further development.

Tool Transparency

Tool transparency refers to the degree to which users perceive the system as a reliable, predictable, and efficient technological artifact. This concept aligns with what Shneiderman (2020) describes as "interaction transparency," where users' confidence in a system stems from their ability to predict its behavior and rely on its outputs consistently. The Trust Questionnaire measured participants' perceptions of VisAutoML 2.0's reliability, predictability, and overall trustworthiness as a tool.

The trust questionnaire employed 7 items using a 5-point Likert scale. Reliability analysis using Cronbach's Alpha yielded a coefficient of 0.745, indicating satisfactory internal consistency among the questionnaire items. This coefficient exceeds the commonly accepted threshold of 0.7 for scale reliability in human-computer interaction research (Nunnally & Bernstein, 1994).

The results revealed a moderate to high level of trust in VisAutoML 2.0, with an overall mean score of 26.11 (SD = 4.67). Detailed analysis of individual items, presented in Table 1, demonstrates particularly high scores for "The AI can perform the task better than a novice human user" (M = 4.06, SD = 0.82) and "I am confident in the AI. I feel that it works well" (M = 4.04, SD = 0.83). These findings indicate that participants perceive substantial relative advantage in using VisAutoML 2.0 compared to manual task completion by novices, aligning with Parasuraman and Manzey's (2010) observation that perceived relative advantage is a key factor in automation acceptance.

The lowest-scoring item was "I like using the system for decision making" (M = 3.49, SD = 1.10), suggesting that despite high confidence in system capabilities, participants maintained some reservation about incorporating VisAutoML 2.0 into their decision-making processes. This finding corresponds with research by Lee and See (2004), who established that trust in automated systems involves both performance-based evaluations and emotional comfort with delegation.

The consistent standard deviations across items (ranging from 0.80 to 1.14) indicate general agreement among participants in their assessment of VisAutoML 2.0's trustworthiness. This pattern suggests that users develop consolidated perceptions of system reliability through interaction,

contrasting with findings by Nourani et al. (2020), who observed higher variability in trust assessments of other machine learning systems.

Table 51 Descriptive statistics of Trust items

Item (5-point Likert scale)	Mean	Std. Deviation
I am confident in the AI. I feel that it works well.	4.0409	.83432
The outputs of the AI are very predictable.	3.3829	1.13880
The AI is very reliable. I can count on it to be correct all the time.	3.9368	.83739
I feel safe that when I rely on the AI I will get the right answers.	3.5353	1.10783
The AI is efficient in that it works very quickly.	3.9740	.80302
The AI can perform the task better than a novice human user.	4.0558	.81991
I like using the system for decision making.	3.4907	1.10160

Comparison with VisAutoML 1.0

Comparison with VisAutoML 1.0 reveals substantial improvements in tool transparency. The previous version obtained an overall mean trust score of 21.73 (SD = 6.01), significantly lower than VisAutoML 2.0's score. This improvement spans all trust dimensions, with the largest gains observed in system confidence (increase from M = 2.56 to M = 4.04) and efficiency (increase from M = 2.52 to M = 3.97). The increase in perceived reliability (from M = 2.59 to M = 3.94) suggests that modifications to VisAutoML 2.0 have addressed critical stability issues that affected user confidence in the previous version.

These improvements align with Hoffman et al.'s (2018) findings that iterative refinement of intelligent systems based on user feedback can substantially enhance perceived reliability. The reduced standard deviations in VisAutoML 2.0 responses (compared to version 1.0) further indicate more consistent user experiences across participants, suggesting improved interface standardization and interaction predictability.

Table 52 Comparison of trust scores between VisAutoML 1.0 and VisAutoML 2.0

	VisAutoML 1.0 (n=108)		VisAutoML 2.0 (n=272)	
Item	Mean	SD	Mean	SD
I am confident in the AI. I feel that it works well.	2.5556	1.12186	4.0409	.83432
The outputs of the AI are very predictable.	2.5833	1.06889	3.3829	1.13880
The AI is very reliable. I can count on it to be correct all the time.	2.5926	1.19216	3.9368	.83739
I feel safe that when I rely on the AI I will get the right answers.	2.5648	1.17030	3.5353	1.10783
The AI is efficient in that it works very quickly.	2.5185	.98095	3.9740	.80302
The AI can perform the task better than a novice human user.	2.6111	1.04866	4.0558	.81991
I like using the system for decision making.	2.2222	1.13016	3.4907	1.10160

Sum Per Participant	21.73	6.01	26.11	4.67

AI Transparency

AI transparency refers to the explainability of the artificial intelligence components embedded within VisAutoML 2.0. This dimension focuses on users' ability to understand, interpret, and replicate the decision-making processes of the underlying machine learning models. The XAI Questionnaire measured participants' perceptions of the explainability features in VisAutoML 2.0.

The XAI Questionnaire consisted of 30 items rated on a 7-point Likert scale, including nine negatively-phrased items that were reverse-scored in the analysis. Reliability analysis using Cronbach's Alpha yielded a coefficient of 0.979, indicating exceptionally high internal consistency. This reliability coefficient substantially exceeds both the previous version's coefficient (0.83) and typical thresholds for questionnaire reliability in human-AI interaction studies (Hoffman et al., 2018).

Participants' responses to the XAI Questionnaire yielded an overall mean score of 161.9 (SD = 36.24), indicating a high level of perceived explainability. This represents a substantial improvement over VisAutoML 1.0's mean score of 118.18 (SD = 13.47). Detailed analysis of individual items, shown in Table 2, reveals several critical insights regarding AI transparency in VisAutoML 2.0.

With negatively phrased items reverse-scored for consistency, the highest-rated positive items were "I believe that I could provide an explanation similar to the AI's explanation" (M = 6.06, SD = 0.83) and "I am able to follow the AI's decision-making process step-by-step" (M = 6.04, SD = 0.79). These high scores indicate that participants could both understand and mentally simulate the system's reasoning processes, meeting key criteria for explainability as defined by Miller (2019). The lowest-scoring positive item was "I could easily follow the explanation to arrive at an answer to the question" (M = 4.90, SD = 1.42), suggesting that while users understood explanations conceptually, they sometimes struggled to apply them to specific tasks.

The responses to negatively phrased items (before reversal) provide further insight into explainability strengths and weaknesses. The item "I believe that the explanations would not help most people in answering the question" received a very low mean score (M = 1.88, SD = 0.81), indicating strong disagreement with this negative statement and thus high perceived utility of explanations. Similarly low scores for "I do not think most people would provide similar explanations as the AI's explanation" (M = 2.03, SD = 0.83) suggest that VisAutoML 2.0's explanations aligned well with users' mental models of the task domain.

These findings align with Liao et al.'s (2020) research showing that effective explainable AI systems should provide explanations that match users' expectations and reasoning patterns. The lower standard deviations in VisAutoML 2.0 responses (compared to version 1.0) indicate more consistent explainability experiences across participants, suggesting improved standardization of explanation formats and content.

Table 53 Descriptive statistics of xAI items

Item (7-point Likert scale)	Mean	Std. Deviation
The explanations were detailed enough for me to understand.	5.4535	1.15363
I understood the explanations within the context of the question.	5.1264	1.41909
The explanations provided enough information for me to understand.	5.4647	1.12786

I understood how the AI arrives at its prediction.	4.9442	1.43798
I was able to use the explanations with my knowledge base.	4.9963	1.44681
I would be able to repeat the steps that the AI took to reach its prediction.	5.1041	1.49263
I think that most people would learn to understand the explanations very quickly.	5.3866	1.10931
I would not understand how to apply the explanations to new questions.	2.4387	1.14644
I would not be able to recreate the process by which the AI generated its predictions.	2.4944	1.10838
I understand why the AI used specific information in its explanation.	5.0112	1.41021
I understood the AI's reasoning.	5.5762	1.12583
I could have applied the AI's reasoning to new problems, even if the AI didn't give me suggestions.	4.9108	1.38739
The explanations were actionable, that is, they helped me know how to answer the questions.	5.5576	1.13354
I believe that I could provide an explanation similar to the AI's explanation.	6.0595	.82645
I would need more information to understand the explanations.	6.0409	.80704
I had trouble using the explanations to answer the question.	2.3494	1.12494
I believe that the explanations would not help most people in answering the question.	1.8773	.80796
The explanations were an important resource for me to answer the question.	2.5093	1.16738
I do not think most people would provide similar explanations as the AI's explanation.	2.0260	.83491
I think that most people would be able to interpret the explanation of the AI.	5.0260	1.39672
Most people would be able to accurately reproduce the AI's decision-making process.	4.9851	1.41940
Most people would not be able to apply the AI's explanations to the questions.	2.4721	1.10805
I could not follow the AI's decision-making process.	2.5390	1.10095
I could easily follow the explanation to arrive at an answer to the question.	4.8959	1.41564
The explanations were useful.	5.9480	.80407
I am able to follow the AI's decision-making process step-by-step.	6.0446	.79049
The explanations were not relevant for the questions I was given.	2.4796	1.09124
I understand how the AI's decision-making process works.	4.9033	1.46538
I could apply the explanations to the questions I was given.	5.5576	1.12030
I could not figure out how the AI arrived at its predictions.	2.5836	1.06744

Comparison with VisAutoML 1.0

Comparison with VisAutoML 1.0 reveals dramatic improvements in AI transparency. The previous version's mean scores for positive items clustered around 3.0 (on a 7-point scale), while VisAutoML 2.0's scores average approximately 5.5. This improvement spans all explainability dimensions, with particularly notable gains in users' ability to understand system reasoning (increase from M = 2.81 to M = 5.58) and explanation usefulness (increase from M = 2.76 to M = 5.95).

The substantial improvement in VisAutoML 2.0's explainability aligns with recent research suggesting that iterative refinement of explanation interfaces based on user feedback can dramatically enhance perceived transparency (Ribeiro et al., 2020). The reduced standard deviations in responses further indicate more consistent explainability experiences across participants, suggesting improved standardization of explanation formats and content.

Factor analysis of the XAI items reveals three primary dimensions of explainability perception: comprehension (understanding explanations), applicability (using explanations for tasks), and generalizability (applying explanations to new contexts). Participants rated comprehension highest (M = 5.47, SD = 1.13), followed by applicability (M = 5.36, SD = 1.15) and generalizability (M = 5.22, SD = 1.21), indicating that while participants could readily understand explanations, they faced slightly greater challenges in applying them practically or extending them to new situations.

These findings align with Dodge et al.'s (2019) observation that explainable AI systems often succeed in providing descriptive explanations but struggle with actionable explanations that enable users to apply insights to novel contexts. The reduced gap between comprehension and generalizability scores in VisAutoML 2.0 (compared to version 1.0) suggests that the new explainability features more effectively support knowledge transfer.

Correlation Analysis

In the latest correlation analysis, the relationships among Trust, Explainability (XAI), and Usability (UEQ) were investigated within the context of VisAutoML 2.0. The Pearson correlation coefficient, a statistical method that measures the strength and direction of a linear relationship between two variables, was utilised in this study. The results revealed highly significant and positive correlations between these three pivotal facets of user experience. A strong positive correlation of 0.838** was found between Trust and Explainability (XAI), backed by a low p-value of 0.000. This p-value indicates statistical significance at the 0.01 level (two-tailed), implying that as users develop higher levels of trust in VisAutoML, their perception of the system's explainability also tends to increase significantly.

The analysis also unveiled a robust positive correlation of 0.806** between Trust and Usability (UEQ), again supported by a p-value of 0.000. This statistically significant relationship at the 0.01 significance level (two-tailed) suggests that users who have more trust in VisAutoML are also likely to rate the system's overall usability more favourably.

Additionally, a very strong positive correlation of 0.938** was detected between Explainability (XAI) and Usability (UEQ), verified by a p-value of 0.000 at the 0.01 significance level (two-tailed). This finding accentuates the intertwined nature of users' perceptions of explainability and their assessments of usability within the VisAutoML interface.

Comparing these findings with the previous correlation analysis, it is evident that there have been significant improvements in the relationships among Trust, Explainability, and Usability. The correlations have become stronger, indicating that the enhancements and modifications made to VisAutoML have been effective. Users' trust in the system has a more pronounced impact on their perceptions of explainability and usability, highlighting the importance of building a trustworthy tool. The stronger correlation between Explainability and Usability also underscores the importance of a user-friendly interface that effectively communicates the system's workings to the user. Overall, these findings provide valuable insights for further refining and enhancing the user experience of VisAutoML. The detailed results are outlined in the table below.

Table 54 Bivariate correlations between variables

Correlations

		Trust	UEQ	XAI
Trust	Pearson Correlation	1	.806**	.838**
	Sig. (2-tailed)		.000	.000
	N	272	272	272
UEQ	Pearson Correlation	.806**	1	.938**
	Sig. (2-tailed)	.000		.000
	N	272	272	272
XAI	Pearson Correlation	.838**	.938**	1
	Sig. (2-tailed)	.000	.000	
	N	272	272	272

**. Correlation is significant at the 0.01 level (2-tailed).

5.4.4 Open Ended Questionnaire

This section presents the findings derived from open-ended questions asked in the survey that aimed to capture the participants' experiences and perceptions regarding the use of the VisAutoML. The questionnaire was comprised of three primary questions designed to ascertain a comprehensive understanding of the participants' interaction with the tool. These questions were as follows:

1. "Can you describe your experience with VisAutoML ?"
2. "What did you like about VisAutoML ? What didn't you like?"
3. "What improvements would you suggest for VisAutoML ?"

The intention behind these questions was to gather qualitative data that could provide rich insights into the user experience, both in terms of positive aspects and potential areas for improvement. The first question sought to gather general impressions and experiences of participants with the tool. The second question was designed to elicit specific likes and dislikes, thereby helping to identify the strengths and shortcomings of VisAutoML from the user's perspective. Lastly, the third question was aimed at collecting constructive feedback and suggestions for enhancing the tool's functionality and user experience.

The responses to these questions were analysed, with the aim of identifying key themes and patterns. The results are detailed in the following sections, providing a comprehensive view of the user experience and valuable insights that could inform future development and refinement of VisAutoML .

Question 1: Can you describe your experience with VisAutoML ?

Theme 1: Ease of Use

VisAutoML's ease of use was a prominent theme, with 82 out of 272 participants (30.1%) highlighting this aspect positively. Many users found the tool straightforward and user-friendly,

particularly when handling complex data. One participant remarked, "It was pretty easy once you get the hang of it," while another described it as "simple for basic analysis." The tool's ability to simplify data interpretation was frequently mentioned, with one user stating it "made interpreting numbers much easier."

A subset of 27 participants (32.9%) noted initial challenges but emphasised their ability to overcome them. As one user put it, "Not too hard to figure out after a bit," indicating that early difficulties didn't significantly impact the overall experience. Another participant shared, "Had to get used to it, but then it was fine," highlighting the tool's capacity to facilitate a smooth transition to proficiency.

The tool's effectiveness in simplifying complex data analysis was appreciated by 38 participants (46.3%). One user commented, "I liked how it simplified data analysis," while another mentioned that it "made data analysis less scary." This sentiment was echoed by several others, with one participant noting, "Helped me see data in a new light," and another stating, "Helped me make sense of lots of numbers." These responses underscore the tool's potential to simplify data analysis and enhance users' understanding and perception of data.

Theme 2: User Interface and Design

The tool's user interface and design emerged as a significant theme, with 74 participants (27.2%) commenting positively on these aspects. Users frequently praised the clear, simple, and user-friendly design, which they found greatly enhanced navigability and accessibility. One participant shared, "Navigation was really simple and clear," while another stated, "Interface was easy to navigate," highlighting the ease with which users could interact with the tool.

The simplicity and intuitiveness of the design were particularly appreciated, with 31 participants (41.9%) specifically mentioning these aspects. One user noted, "The layout made sense to me," indicating that the design was easy to understand. Another participant commented, "Interface is clean and simple," suggesting that the uncluttered design contributed to a positive user experience. The user-friendliness of the interface was also emphasised, with one participant sharing, "Interface was easy to understand," and another stating, "Interface is very user-friendly."

Interestingly, 22 participants (29.7%) mentioned how the design contributed to their enjoyment of using the tool. One user remarked, "Interface was a breeze to use," indicating that the design made the tool not just easy but also enjoyable to use. Another participant shared, "Interface was inviting and straightforward," suggesting that the design was both functional and appealing. These comments highlight how the tool's interface and design facilitated ease of use and enhanced the overall user experience.

Theme 3: Visualisation

Visualisation emerged as a key theme, with 63 participants (23.2%) praising the tool's visual features for their clarity, helpfulness, and ability to make data analysis more engaging and comprehensible. Many users found the visuals to be a standout feature that significantly aided their understanding of the data. One participant stated, "The visuals are really a standout," while another noted, "The visuals made everything clearer," indicating the effectiveness of the graphics in elucidating complex data.

The informative nature of the visuals was highlighted by 29 participants (46.0%). One user shared, "The visuals helped me a lot," suggesting that the graphics provided valuable insights that enhanced their understanding of the data. Another participant reinforced this by stating, "The visuals were informative," demonstrating how the tool's visualisations served as a valuable resource for learning and understanding data. The ability of the visuals to simplify complex data was also appreciated, with

one user mentioning, "The visuals simplified things a lot," and another sharing, "The visuals were a big help."

Notably, 19 participants (30.2%) commented on how the visuals enhanced their overall experience with the tool. One user expressed, "The visuals added a lot to the experience," suggesting that the graphics aided understanding and made the data analysis process more enjoyable. Another participant stated, "The visuals were really effective," demonstrating the visuals' contribution to a positive user experience. These responses highlight the crucial role of visualisation in facilitating data comprehension and in making the analysis process more engaging and accessible.

Theme 4: Learning Curve

The learning curve associated with using VisAutoML was mentioned by 31 participants (11.4%). While some users initially found the tool challenging or confusing, many noted that it became easier and more intuitive over time. One participant commented, "A bit confusing at the start, but then easy," indicating a transition from initial confusion to eventual ease of use. Another user shared a similar experience, saying, "A little challenging initially, but then fine," suggesting that the tool's learning curve did not deter them from using it.

Despite initial challenges, 14 participants (45.2%) expressed resilience and adaptability in learning to use the tool. One user mentioned, "Needed some time to adjust, but then easy," indicating their ability to overcome initial difficulties and adapt to the tool. Another participant echoed this sentiment, saying, "Took a bit to get it, but then it was great," showing their determination to learn and adjust to the tool's functionalities. Many found the learning process to be rewarding, with one participant remarking, "A bit of a learning journey, but fun," and another sharing, "A little effort, but then rewarding."

Interestingly, 9 participants (29.0%) viewed the learning curve as an integral part of their user experience. As one participant put it, "It demystified a lot of data for me," indicating that the process of overcoming the learning curve led to a greater understanding of data. Another user shared, "Learning process was enjoyable," suggesting that the learning curve contributed to a positive and enjoyable user experience. These responses highlight how the initial challenges of learning the tool often translated into a deeper appreciation and understanding of data analysis.

Theme 5: Suitability for Beginners

Suitability for beginners emerged as a significant theme, with 22 participants (8.1%) commenting on this aspect. The tool was frequently praised for its simplicity, approachability, and effectiveness in facilitating a smooth entry into data analysis for beginners. One participant stated, "Good for beginners like me," while another noted, "Perfect for a beginner like me," indicating the tool's suitability for individuals new to the field.

The tool's ability to simplify complex data analysis was particularly appreciated by 10 beginner users (45.5%). As one participant shared, "Made data less daunting for me," it was evident that the tool's simplicity and approachability helped to demystify data analysis for beginners. Another participant affirmed this by stating, "Made starting with data fun," indicating that the tool made the daunting task of beginning data analysis more enjoyable. The tool's design and user interface were also appreciated for their beginner-friendly nature, with one user mentioning, "Interface is welcoming and easy," and another stating, "Interface was clear and inviting."

Notably, 7 participants (31.8%) appreciated the tool's ability to provide a solid foundation in data analysis. As one participant put it, "Helped me get the basics of data," it underscored the tool's

effectiveness in providing a solid grounding in data analysis for beginners. Another user shared, "Helped me ease into data," indicating that the tool facilitates a smooth transition into the field of data analysis for beginners. These responses highlight the tool's potential not just as a data analysis platform, but as an educational resource for those new to the field.

Question 2: What did you like about VisAutoML ? What didn't you like?

Theme 1: Ease of Use

The ease of use of VisAutoML was a prominent theme, with 85 out of 272 participants (31.3%) highlighting this aspect positively. Many users praised the tool's user-friendly interface and intuitive design, making it accessible even for those with limited prior experience. One participant stated, "The tool's user-friendly interface was a standout. It didn't feel overwhelming or too technical." Another user mentioned, "Navigating the tool was less complicated than I thought it would be, which was a pleasant surprise." These comments suggest that the tool successfully reduced potential barriers to entry for novices, allowing users to focus on learning concepts rather than struggling with the interface.

The tool's ability to make complex processes approachable was appreciated by 29 participants (34.1%). As one user noted, "The tool's straightforward approach was refreshing. It's nice to see something so complex made approachable." This sentiment was echoed by others, with one participant sharing, "I found the interface really welcoming. It didn't feel like it was designed just for experts, which was great." These responses indicate that the tool's design successfully catered to a broad range of users, including those without a technical background.

Notably, 23 participants (27.1%) specifically mentioned the tool's simplicity as a positive feature. One user remarked, "The simplicity of the tool was a pleasant surprise. It didn't feel like I was navigating something meant for experts only." Another participant commented, "It's refreshing to use a tool that doesn't overcomplicate things." These comments highlight how the tool maintained a level of simplicity that made it accessible and easy to use for beginners, while still providing powerful functionality.

Theme 2: Visualisation

Visualisation emerged as a key theme, with 76 participants (27.9%) praising the tool's visual features for their ability to simplify and clarify complex data. Many users found the graphical representation of data particularly helpful in making abstract concepts more tangible. One participant shared, "Seeing the results in a graphical format made the abstract aspects of the model tangible and easier to grasp." Another noted, "The visual breakdown of the data was particularly helpful. It made abstract ideas much clearer." These comments suggest that the tool's visual aids were effective in demystifying complex ideas and aiding users in gaining a clear understanding of the concepts.

The effectiveness of visualisation in keeping the learning process engaging and informative was highlighted by 24 participants (31.6%). One respondent stated, "The visual aspect was a standout. Seeing the data come to life in such a way was both informative and engaging." Another user mentioned, "Seeing the practical application of the concepts through the tool's visuals was an eye-opener." These responses indicate that the visual aids provided by the tool were educational and contributed to maintaining users' interest in the learning process.

Notably, 19 participants (25.0%) specifically commented on how the visualisations helped translate complex data into understandable formats. One user expressed, "The visual aids really brought the data to life in an understandable way." Another participant shared, "The graphs and charts made it so much easier to grasp the relationships in the data." These comments underscore the crucial role of visualisation in making complex data understandable, which was particularly beneficial for those new to the field.

Theme 3: Structured Approach

The structured approach of VisAutoML was a significant theme, with 64 participants (23.5%) praising this aspect. The tool's step-by-step guidance was frequently commended for its ability to effectively navigate users through the learning process. One participant shared, "The guided, step-by-step approach effectively navigated through the learning process." Another noted, "The step-by-step guidance was invaluable. It felt like having a tutor walking me through each stage." These comments suggest that the tool's structured guidance significantly contributed to users' learning effectiveness and confidence in handling new concepts.

Many participants, 18 (28.1%), appreciated how the structured approach made the learning process less intimidating. As one user mentioned, "The guided approach made the learning process less daunting and more structured." Another respondent stated, "The guided steps were a godsend. They helped me navigate through areas I was unfamiliar with." These sentiments indicate that the tool's structured approach was successful in reducing potential barriers to learning, making the process less daunting for beginners.

Notably, 15 participants (23.4%) highlighted how the structured approach provided a comprehensive learning experience. One user commented, "The step-by-step instructions were really helpful, especially for a beginner like me." Another participant shared, "I appreciated how each step built on the previous one, giving me a solid understanding of the whole process." These responses underscore the value of the tool's structured approach in providing an organized and beginner-friendly learning experience, effectively guiding users through unfamiliar concepts or areas.

Theme 4: Handling of Complex Concepts

VisAutoML's ability to handle complex concepts effectively was a prominent theme, with 47 participants (17.3%) commenting positively on this aspect. The tool's capacity to simplify complex topics and make them accessible was frequently praised. One participant noted, "I appreciated how the tool made complex ideas accessible without dumbing them down." Another shared, "The tool's ability to make complicated processes understandable was impressive. It's great for those starting out." These comments reflect that the tool was successful in striking a balance between simplicity and complexity, making it easier for users to understand the concepts without losing their intricacy.

Many participants, 14 (29.8%), praised the tool's ability to demystify complex subjects. One respondent stated, "The tool helped demystify some concepts I thought were way beyond my understanding." Another user mentioned, "I liked how the tool turned a daunting subject into something approachable and interesting." These responses indicate that the tool was effective in breaking down perceived barriers in understanding complex concepts, thereby enhancing the learning experience.

Notably, 11 participants (23.4%) specifically appreciated how the tool made complex ideas accessible to beginners. One participant said, "The tool's design made complex ideas accessible, which was great

for a beginner like me." Another user commented, "It's amazing how the tool could explain complex algorithms in a way that even I could understand." These sentiments underscore that the tool was effective in making complex ideas accessible to beginners, thereby promoting inclusivity in learning and helping users tackle subjects they might have previously found intimidating.

Question 3: What improvements would you suggest for VisAutoML ?

Theme 1: User Interface and Experience

The user interface and overall user experience of VisAutoML emerged as a significant theme in participant feedback. Among the 272 participants, 80 (29.4%) emphasised this theme. Many users expressed satisfaction with the current interface, with one participant noting, "No changes required, the interface is straightforward," suggesting that the tool's design was intuitive and easy to navigate. However, some participants suggested improvements to further refine the user experience.

A common recommendation from 30 participants (37.5%) was to simplify the navigation menus to make the tool more accessible for beginners. As one participant proposed, "Maybe simplify the navigation menus for beginners," indicating the need for more intuitive navigation to improve user-friendliness. Several participants also emphasised the aesthetic aspect of the interface, with 25 participants (31.3%) suggesting a more colorful design to enhance engagement.

Readability was another area of focus for 15 participants (18.8%), with suggestions ranging from increasing font size to creating a minimalist design to avoid clutter. One participant expressed, "Increase font size for better readability," underlining the need for clear and legible text. Additionally, 10 participants (12.5%) suggested introducing customisable themes for a more personalised experience, reflecting a desire for individualized user interaction with the tool.

Theme 2: Functionality and Features

The theme of functionality and features also emerged as a significant area of discussion, highlighted by 70 participants (25.7%). Many participants expressed contentment with the existing functionalities of VisAutoML, with some even stating that no additional features were needed. As one participant remarked, "It's perfect the way it is, no need for changes," reflecting a high level of satisfaction with the current feature set.

However, some participants did propose the introduction of new features to enhance the tool's capabilities. For instance, 25 participants (35.7%) suggested adding a customisable dashboard for frequently used features. As one user suggested, "How about a customisable dashboard for frequently used features?" This recommendation indicates a desire for a more personalised and efficient user experience.

Drag-and-drop capabilities for data manipulation were another feature suggested by 20 participants (28.6%). "Implement drag-and-drop features for data manipulation," offered one user, indicating the need for a more intuitive and user-friendly way to interact with data within the tool. Voice command accessibility was also mentioned by 15 participants (21.4%) as a potential feature addition. As one participant recommended, "Enable voice commands for accessibility," this shows an increasing awareness and need for inclusive design practices.

Finally, the importance of clear explanations and tooltips for each feature was emphasised by 10 participants (14.3%). "Add tooltips with explanations for each feature for newcomers," one participant

suggested, underlining the importance of user guidance and support in making the tool more accessible and easy to use, particularly for newcomers to the field of machine learning.

Theme 3: Data Visualisation

Data visualisation was another key theme that emerged from the feedback, with 50 participants (18.4%) highlighting this aspect. Many participants commended the tool's current data visualisation capabilities, underscoring the clarity and comprehensibility of the graphs and charts produced. One participant stated, "Nothing to improve, the visuals are clear," indicating a high level of satisfaction with the tool's current data visualisation features.

However, some users suggested further improvements to enhance the versatility and effectiveness of the data visualisation capabilities. A recurring suggestion from 20 participants (40% of the 50 participants in this theme) was the introduction of more diverse colour schemes in graphs for better clarity. As one participant pointed out, "More diverse colour schemes in graphs for clarity," this indicates a desire for a more visually distinct and clear representation of data. Another suggestion that was made by 15 participants (30%) was to use larger and more legible labels on graphs. As one participant shared, "Use larger, more legible labels on graphs," this reflects the need for clear and easy-to-read labels to enhance the comprehensibility of data visualisations.

Interestingly, 10 participants (20%) requested more interactive graphs for better data exploration. "Interactive graphs for better data exploration," one user suggested, highlighting the importance of interactivity in facilitating a deeper understanding and exploration of data. Lastly, the idea of customisability and control in data visualisation was highlighted by 5 participants (10%). They proposed providing customisable graph settings for user preference and offering 3D graphs for more in-depth analysis. As one participant suggested, "Offer 3D graphs for more in-depth analysis," this shows the desire for more advanced and customisable data visualisation options to cater to diverse user needs and preferences.

Theme 4: Learning and Support

The theme of learning and support was another crucial aspect that emerged from the participants' feedback, noted by 40 participants (14.7%). The need for more learning resources and support mechanisms within VisAutoML was a common suggestion. This theme reflects the participants' desire for a tool that facilitates machine learning and supports their learning journey.

For instance, 15 participants (37.5% of the 40 participants in this theme) proposed the integration of tooltips with explanations for each feature. As one participant expressed, "Add tooltips with explanations for each feature for newcomers," this suggestion underscores the importance of providing immediate, contextual help to users. The need for a dedicated help section was also a common suggestion from 10 participants (25%). As one user recommended, "Implement a help section with common troubleshooting," this indicates the desire for a centralized resource within the tool where users can find answers to common issues or challenges.

The suggestion to provide interactive tutorials for first-time users was another significant point raised by 10 participants (25%). One participant stated, "Incorporate interactive tutorials for first-time users," reflecting the need for guided learning resources that can help users understand the tool's features and functionalities better. Furthermore, 5 participants (12.5%) put forward the idea of providing real-time error prompts and suggestions for corrections. As one participant suggested, "Introduce error prompts with suggestions for corrections," this feedback highlights the importance of timely feedback and guidance in helping users navigate the tool more effectively.

Theme 5: Customisation and Control

The theme of customisation and control was another significant area that surfaced from the participants' feedback, highlighted by 20 participants (7.4%). Users expressed a desire for more control and customisation options within VisAutoML to enhance their user experience and make the tool more adaptable to their specific needs.

One such suggestion from 8 participants (40% of the 20 participants in this theme) was to enable customisable graph settings for user preference. As one user suggested, "Provide customisable graph settings for user preference," this reflects a desire for a more personalised user experience. The idea of providing a feature to export graphs in high resolution was also put forward by 6 participants (30%). As one participant shared, "Include a feature to export graphs in high resolution," this indicates the need for more flexibility in how users can use and share the outputs of the tool.

Participants also expressed a desire for more visual cues for important data points, suggested by 4 participants (20%). One participant suggested, "Add more visual cues for important data points," signifying the importance of highlighting key information in data visualisations for better comprehension. The suggestion to implement an option to change graph colours for accessibility was also worth noting, with 2 participants (10%) proposing this idea. "Implement an option to change graph colours for accessibility," one participant proposed, underlining the need for inclusive design practices that consider the diverse needs of users.

Theme 6: Accessibility and Inclusivity

The theme of accessibility and inclusivity, noted by 12 participants (4.4%), underscores the need for better user guidance and tutorials within VisAutoML. Participants suggested including a step-by-step guide to help new users navigate the tool and understand its functionalities. As 4 participants (33.3% of the 12 participants in this theme) said, "Provide a step-by-step guide for new users," this indicates a need for a more user-friendly introduction to the tool's features and capabilities.

The idea of providing in-app tips and hints was also put forward by 3 participants (25%). As one participant shared, "Include in-app tips and hints," this suggests the need for ongoing guidance as users interact with the tool. Participants also expressed a desire for more interactive tutorials, suggested by 3 participants (25%). One participant noted, "Add interactive tutorials," showing a preference for a more engaging and hands-on learning experience.

Lastly, the suggestion to implement a help button or FAQ section was mentioned by 2 participants (16.7%). "Implement a help button or FAQ section," one participant proposed, highlighting the need for readily available assistance and resources within the tool. Users also indicated a need for more clarity on how the tool processes and analyses data, reflecting a desire for more transparency and understanding of the tool's workings.

5.5 Summary

This chapter has presented the empirical findings derived from the various studies conducted as part of the iterative user-centred design (UCD) methodology detailed in Chapter 3. These results collectively address the research questions and objectives concerning non-expert users' perceptions

of Machine Learning (ML) development, the requirements for the VisAutoML tool, and its effectiveness in enhancing usability and transparency.

The Extended Technology Acceptance Model (TAM) study (Section 5.2) provided foundational insights into the factors influencing non-expert users' acceptance of an AutoML tool. The results from the quantitative questionnaire and qualitative interviews elucidated the significance of perceived usefulness, perceived ease of use, attitude, behavioural intention, perceived enjoyment, and perceived authority in shaping users' willingness to adopt such technology. These findings were crucial for deriving user requirements and informing the design principles aimed at creating a tool that is functional, appealing, and trustworthy for non-target users.

The comparison study between the initial VisAutoML 1.0 prototype and H2O AutoML (Section 5.1.2) offered a benchmark assessment. The results indicated that VisAutoML 1.0 demonstrated promising performance in terms of usability and knowledge gain compared to the existing tool. While acknowledging limitations, this initial evaluation provided empirical evidence supporting the potential of VisAutoML's design approach to offer a more user-friendly and understandable ML development experience for non-experts.

The in-depth usability and transparency evaluation of VisAutoML 1.0 (Section 5.1.3) provided a detailed analysis of user experiences with the initial prototype. The findings, including task completion time, usability scores (UEQ-S), and transparency perceptions (Trust and XAI questionnaires), highlighted specific strengths and identified key areas for improvement. These results were instrumental in guiding the subsequent redesign efforts, pinpointing aspects of the interface, workflow, and explanations that required refinement to better meet non-expert needs.

Subsequently, the evaluation of the refined VisAutoML 2.0 prototype (Section 5.4) assessed the impact of the design iterations. The results demonstrated significant improvements in usability, transparency, and user experience compared to VisAutoML 1.0. Findings related to task completion time, enhanced UEQ scores (pragmatic and hedonic quality), increased trust levels, and higher perceived explainability provided empirical evidence that the iterative design process, guided by user feedback and design principles, successfully enhanced the tool's effectiveness and user acceptance for non-experts.

In conclusion, the results presented in this chapter provide strong empirical support for the feasibility and effectiveness of developing a usable and transparent AutoML tool for non-expert users based on the proposed design principles. The findings from the TAM study informed the requirements, the comparison study benchmarked the initial prototype, and the iterative evaluations of VisAutoML 1.0 and 2.0 demonstrated progressive improvements in usability and transparency, ultimately showing that VisAutoML 2.0 successfully addresses many of the identified challenges faced by non-experts in ML development. These findings lay the groundwork for the discussion and interpretation of their implications in the subsequent chapter.

6 Discussion

6.1 Introduction

This chapter provides a comprehensive discussion and critical analysis of the research undertaken, synthesising the empirical findings derived from the user studies and system evaluations within the broader context of existing literature on Automated Machine Learning (AutoML), Explainable Artificial Intelligence (XAI), and Human-Computer Interaction (HCI). The preceding chapters detailed the problem space concerning the challenges non-expert users face in engaging with Machine Learning (ML), the user-centred design methodology employed, the iterative development of the VisAutoML tool, and the empirical results obtained from its evaluation. The primary purpose of this discussion is to interpret these findings, relate them back to the initial research questions and objectives, and critically assess the extent to which VisAutoML, designed with specific principles aimed at enhancing usability and transparency, successfully addresses the identified challenges for non-expert users.

The research was motivated by the significant barrier to entry presented by traditional ML workflows for individuals without extensive technical expertise, hindering the broader adoption and application of AI technologies across various domains (Benbya et al., 2021; Yang et al., 2018). While AutoML emerged as a potential solution to automate complex ML processes, a critical analysis of existing tools indicates they often fall short in providing sufficient usability and transparency for non-expert users, frequently operating as "black-box" systems with interfaces and outputs that remain difficult to understand or trust (Hutter et al., 2019; Kaur et al., 2020; Xin et al., 2021). VisAutoML was developed through an iterative User-Centered Design process, guided by empirically derived user requirements and specific design principles, as a novel approach to create a usable and transparent AutoML tool for tabular data, thereby aiming to democratise access to ML capabilities.

This chapter is structured to facilitate a detailed examination of the research outcomes and their implications. It begins by discussing the efforts made in enhancing the usability of VisAutoML for non-experts. This involves exploring the impact of the user-centered design approach, informed by models of technology acceptance, on addressing user needs (Davis, 1989; Norman, 2013). It delves into the specific design principles formulated to guide the development of accessible AutoML interfaces (Chromik & Butz, 2021; Sharma & Hannafin, 2007) and examines the evolution of VisAutoML's interface across iterative versions based on user feedback (Yang et al., 2018). The discussion in this section is supported by the empirical evaluation of usability improvements, including comparative results that quantify the effectiveness of the design interventions in creating a more user-friendly tool (Schrepp et al., 2017a).

Subsequently, the discussion delves into the crucial aspect of increasing transparency in AutoML, a significant challenge for non-expert users (Larsson & Heintz, 2020). This section addresses the inherent transparency challenge in automated ML processes and examines the implementation of Explainable AI within VisAutoML, drawing from the guiding design principles for XAI user interfaces (Miller, 2019; Wang et al., 2019). It discusses the empirical findings related to user trust and understanding of the system's outputs, assessing how the integrated explainability features contributed to fostering confidence and comprehension (Hoffman et al., 2023; Ribeiro et al., 2016). This section also explores the strategic approach taken to balance the inherent complexity of ML with the need for clarity in the interface, particularly through the application of principles like progressive disclosure, assessing how these efforts contributed to demystifying the AutoML process for non-experts (Buçinca et al., 2021).

Finally, the chapter discusses the broader impact and potential applications of a usable and transparent AutoML tool like VisAutoML. This includes its potential role in promoting AI literacy by

making ML concepts more accessible and understandable for a wider audience (Long & Magerko, 2020). It considers its capacity to support domain experts in leveraging data-driven insights within their fields, bridging the gap between domain knowledge and technical ML capabilities (Yang et al., 2018). The discussion also highlights the benefits it offers for rapid prototyping and experimentation, accelerating the process of translating data into actionable insights (Singh & Joshi, 2022). Furthermore, the chapter critically examines the ethical implications relevant to the design and use of such tools, particularly concerning issues of bias, transparency, and responsible deployment, reflecting on the extent to which the research addressed these important considerations within the context of non-expert use (Floridi & Cowls, 2022; Zou & Schiebinger, 2018). Through this structured discussion, the chapter aims to provide a comprehensive synthesis and interpretation of the research findings, contributing to the ongoing discourse on designing effective and accessible AI technologies for a broader audience.

The structure of this discussion chapter is outlined in the table below:

Table 55 Subsection Overview

Subsection	Description
6.2 Enhancing Usability for Non-Experts	Discussing design strategies and evaluation results focused on making AutoML tools easier and more intuitive for users without ML expertise.
6.3 Increasing Transparency in AutoML	Discussing the challenges and solutions related to making the internal workings and outputs of AutoML systems understandable for non-ML experts.
6.4 Impact and Applications	Discussing the broader implications and potential uses of a usable and transparent AutoML tool for non-experts.

6.2 Enhancing Usability for Non-Experts

6.2.1 User-Centered Design Approach

The development of VisAutoML was fundamentally guided by a rigorous User-Centered Design (UCD) methodology, a strategic choice necessitated by the well-documented challenges non-expert users face when engaging with complex Machine Learning (ML) technologies (Bove et al., 2022; Yang et al., 2018). Traditional ML workflows demand specialised programming, statistical expertise, and a nuanced understanding of algorithmic processes, creating significant barriers for individuals without such backgrounds (Benbya et al., 2021). A UCD approach, which systematically places the needs, requirements, and capabilities of the end-user at the forefront throughout the entire design and development lifecycle, was therefore indispensable to ensure that VisAutoML would be not merely technically capable but genuinely accessible, usable, and transparent for its target audience (Norman, 2013). This iterative process, encompassing stages of requirements gathering, design, prototyping, and rigorous evaluation with representative users, provided the necessary framework to systematically address the multifaceted challenges of human-computer interaction in this complex and often opaque domain.

A critical component of the initial Design Research stage within this UCD framework was the comprehensive mixed-methods study employing an Extended Technology Acceptance Model (TAM). The traditional TAM posits that Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) are primary determinants of a user's attitude towards and intention to use a technology (Davis, 1989).

However, recognising that external factors significantly influence these core constructs, particularly in the context of novel, automated, and potentially opaque AI systems, this research extended the model to include Perceived Authority (PA) and Perceived Enjoyment (ENJ) (Davis et al., 1992; Venkatesh & Davis, 2000). The purpose of this Extended TAM study was to empirically investigate the factors influencing non-expert users' acceptance of an AutoML tool, thereby grounding the subsequent design decisions in empirical evidence of user perceptions, motivations, and the specific influences relevant to automated, explained systems. This empirical foundation was crucial for moving beyond assumptions about user needs to data-driven design priorities.

The findings from the Extended TAM model analysis provided crucial empirical insights that directly informed the design principles and subsequent development of VisAutoML. The quantitative results demonstrated that Perceived Usefulness, Perceived Ease of Use, and Perceived Enjoyment were significant positive predictors of users' attitude towards and behavioural intention to use the proposed system (as shown in the TAM results). This empirical validation underscored the necessity of designing a tool that performs ML tasks effectively (PU), is intuitive and easy to learn and operate (PEOU), and crucially provides an engaging and enjoyable user experience (ENJ). The study also revealed that Perceived Enjoyment had a particularly strong influence on Perceived Usefulness, suggesting that making the interaction with the tool enjoyable could significantly enhance users' perception of its utility (as detailed in the TAM results). Furthermore, Perceived Authority was found to positively influence both PU and PEOU, indicating that perceived credibility or social influence could impact how useful and easy the system was perceived to be (as shown in the TAM results). These findings corroborated the initial assumptions derived from the literature regarding the importance of usability and user experience for non-expert technology adoption (Yang et al., 2018; Margetis et al., 2021).

Critically, the empirical results from the Extended TAM study served to validate the user requirements identified in the foundational research and directly informed the synthesis of the design principles guiding VisAutoML's development. The strong correlation and significant influence of PU, PEOU, and ENJ on user acceptance provided empirical justification for prioritising design principles aimed at enhancing these aspects. For instance, the emphasis on visualising activity sequences, demonstrating scaffold functions, and embedding contextually relevant scaffolds were directly supported by the need to improve PEOU and facilitate learning for non-experts (as discussed in the design principles). Similarly, principles focusing on progressive explanation disclosure, natural language rationale, and multiple ways to communicate explanations were informed by the need to enhance transparency and explainability, which indirectly contributes to PU and builds trust (Chromik & Butz, 2021; Miller, 2019). The finding that ENJ significantly influenced PU also provided a strong rationale for incorporating interactive and engaging elements into the interface and visualisations. Thus, the Extended TAM model was not merely a theoretical exercise but a vital empirical step within the UCD process, providing the necessary evidence to shape a user-centred design that resonated with the target audience's perceptions and motivations for technology adoption.

User personas, synthesized from empirical data gathered through surveys and interviews, played a vital role in grounding the User-Centered Design process for VisAutoML in the reality of non-expert users' needs. These archetypes embodied the diverse backgrounds, goals, challenges, and preferences of the target audience, moving the design process beyond abstract requirements to consider how real individuals would interact with the tool. The challenges highlighted by these personas, such as a lack of technical knowledge and difficulties with complex terminology, directly informed the design emphasis on simplicity, intuitive navigation, and integrated guidance within the prototype. Their motivations and preferences for engaging explanations and visually appealing interfaces also shaped

the implementation of features like interactive XAI visualizations and structured workflows. While the contribution of personas to the design was qualitative and formative, the effectiveness of the resulting design, informed by these persona-driven insights, was rigorously measured through the empirical evaluation studies of the VisAutoML prototypes. These evaluations, using standardized questionnaires for usability, transparency, trust, and explainability, as well as objective measures like task completion time, provided quantifiable evidence of how well the design, influenced by the understanding of the target users, succeeded in enhancing the user experience and addressing the identified challenges.

The iterative nature of the UCD methodology, informed by the findings of the Extended TAM study and subsequent prototype evaluations, was key to the progressive refinement of VisAutoML. The initial evaluation of VisAutoML 1.0, benchmarked against an existing tool, provided empirical validation of the initial design principles, showing superior usability and knowledge gain compared to a less user-centred alternative. However, it also highlighted specific areas for improvement based on user feedback, which directly informed the redesign objectives for VisAutoML 2.0 . The subsequent evaluation of VisAutoML 2.0 demonstrated significant improvements in usability, transparency, and user experience (as shown in the evaluation results for VisAutoML 2.0), indicating that the iterative application of UCD, guided by empirical data from user studies including the foundational TAM findings, was effective in creating a more usable and transparent tool for non-experts. This continuous feedback loop, from understanding user needs through the TAM, to designing based on those needs, and evaluating the effectiveness of the design, exemplifies the power of a UCD approach in developing complex systems for non-technical users.

In conclusion, the adoption of a User-Centered Design approach, significantly informed and validated by the findings of the Extended Technology Acceptance Model study, was fundamental to the successful development of VisAutoML as a usable and transparent AutoML tool for non-expert users. The empirical evidence from the TAM study on the critical influence of perceived usefulness, ease of use, and enjoyment provided a strong foundation for the design principles and iterative refinement process. This approach ensured that VisAutoML was developed with a deep understanding of its intended users, addressing their specific challenges and motivations, and ultimately contributing to the democratisation of ML by making it more accessible and understandable for a broader audience. The integration of theoretical models like TAM within a practical UCD framework offers a robust methodology for developing human-centered AI systems.

6.2.2 Design Principles for Non-Expert AutoML

Building upon a foundational understanding of user requirements and technology acceptance factors, as established through user-centered design approaches and empirical studies, the development of Automated Machine Learning (AutoML) tools for non-expert users necessitates the synthesis of specific, tailored design principles. The inherent complexity of Machine Learning (ML) workflows and the documented limitations of existing AutoML systems in terms of usability and transparency for individuals without specialised technical backgrounds underscore the critical need for such a principled design framework (Hutter et al., 2019; Yang et al., 2018). These principles serve as the essential bridge between identified user needs and the practical implementation of systems designed to be accessible, understandable, and trustworthy for individuals without extensive ML expertise. They translate the abstract goals of democratising ML into concrete guidelines for interface design and system behaviour, ensuring that the tool is not merely functional but truly usable by its intended audience.

The theoretical underpinnings of these design principles draw significantly from established research in related fields, particularly technology-enhanced scaffolding and the design of Explainable Artificial

Intelligence (XAI) user interfaces. Scaffolding, as a pedagogical concept, involves providing structured support to learners to enable them to accomplish tasks that would otherwise be beyond their current capabilities (Quintana et al., 2018). In the context of an AutoML tool for non-experts, technology-enhanced scaffolding manifests as integrated guidance, step-by-step instructions, contextual help, and adaptive feedback mechanisms designed to support users throughout the complex ML development pipeline (Sharma & Hannafin, 2007; Suwastini et al., 2021). This is crucial given that non-experts often lack the foundational technical knowledge and analytical habits common among ML practitioners (Yang et al., 2018). Effective scaffolding helps manage cognitive load and guides users towards successful task completion while fostering a deeper understanding of the process.

Similarly, principles from XAI user interface (XUI) design are crucial for ensuring that the automated decisions and model outputs are technically explainable and presented in a manner that is comprehensible and actionable for non-experts (Chromik & Butz, 2021; Liao et al., 2020). XAI aims to make the reasoning behind AI decisions understandable to humans (Miller, 2019), but the way these explanations are presented significantly impacts their utility for non-technical users (Wang et al., 2021). XUI design principles address how explanations are structured, visualised, and integrated into the user workflow to foster understanding and build trust in the AI system (Wang et al., 2019). Given that non-experts often perceive ML models as 'black boxes' and struggle to interpret technical outputs (Kaur et al., 2020; Xin et al., 2021), effective XUI design is paramount for building the trust necessary for tool adoption and reliance.

Based on comprehensive requirements analysis, literature review, and insights gained from user studies, including findings that highlight the importance of perceived ease of use, usefulness, and enjoyment for technology acceptance (Davis, 1989; Venkatesh & Davis, 2000), a set of design principles can be synthesized and iteratively refined. These principles can be categorised into those addressing the overall system design and those specific to the XAI component, reflecting the dual focus on streamlining the ML process and enhancing the interpretability of its outcomes. Principles for the overall system design can include visualising activity sequences, demonstrating scaffold functions, embedding contextually relevant scaffolds, ensuring visible and utilised scaffolds, and providing engaging and informative user feedback. These principles directly address the challenges non-experts face with complex workflows, lack of foundational knowledge, and reliance on external documentation (Yang et al., 2018), aiming to create an intuitive and supportive learning environment that enhances perceived ease of use and enjoyment.

For the XAI component, principles should be designed to make model explanations accessible and understandable for non-experts, moving beyond expert-centric technical diagnostics. These can include progressive explanation disclosure, providing natural language rationale, and offering multiple ways to communicate an explanation (Chromik & Butz, 2021; Ehsan et al., 2019). These principles are crucial for overcoming the limitations of existing XAI tools which are often designed for ML practitioners and produce outputs difficult for non-experts to interpret or act upon (Hohman et al., 2019; Wang et al., 2021). By presenting explanations in a user-friendly manner, these principles aim to enhance perceived explainability and build trust, which empirical studies have shown are strongly correlated with overall usability and acceptance (Kaur et al., 2020; Rossi, 2018). The iterative refinement of these principles based on prototype evaluations is vital in ensuring their effectiveness in practice.

The application of such design principles guides the iterative development of AutoML tools. An initial implementation provides a tangible representation of these principles, which can then be empirically evaluated. Findings from this evaluation, highlighting areas for improvement, directly inform redesign objectives for subsequent versions and lead to a refined application of the design principles in updated

interfaces. For instance, enhancements to XAI visualisations, such as increased interactivity, clearer labelling, and layered information, are direct results of applying XAI design principles to address limitations identified in initial evaluations (Chromik & Butz, 2021; Liao et al., 2020). This iterative application and refinement of design principles, informed by empirical user data, are critical in progressively enhancing the usability and transparency of AutoML tools for non-expert users, ultimately leading to demonstrably improved user experiences.

In essence, design principles for non-expert AutoML, when developed and applied within a user-centered framework, represent a critical outcome of the UCD process. Grounded in theoretical concepts like scaffolding and XUI design, empirically validated through user studies, and iteratively refined based on prototype evaluations, these principles provide a robust framework for creating accessible and transparent AutoML tools. They address the specific challenges faced by non-experts by prioritising intuitive workflows, integrated guidance, and comprehensible explanations, ultimately contributing to the broader goal of democratising access to machine learning technologies and fostering AI literacy (Long & Magerko, 2020).

6.2.3 Evolution of VisAutoML's Interface

The development of VisAutoML was not a static or linear process but rather an iterative one, fundamentally guided by the principles of User-Centered Design (UCD). A core aspect of this methodology was the continuous evaluation and subsequent refinement of the system's interface based on empirical user feedback (Norman, 2013). The evolution of VisAutoML's interface from its initial prototype, version 1.0, to the refined version, 2.0, directly reflects this iterative cycle, demonstrating how insights gained from user studies informed subsequent design decisions to enhance usability and transparency for non-expert users. This section details this evolutionary process, combining findings from the pilot study evaluation of VisAutoML 1.0 with the subsequent redesign efforts for VisAutoML 2.0, illustrating how user-driven feedback shaped the tool's interface.

The initial prototype, VisAutoML 1.0, was conceived and developed as a tangible representation of the foundational user requirements and preliminary design principles identified in the early stages of the research, aimed at enhancing usability and transparency for non-expert users. Its interface was structured to guide users through a simplified ML pipeline for tabular data, featuring elements such as a sequential navigation menu to visualise the ML workflow, a drag-and-drop interface for model development, and initial implementations of XAI visualisations for model review (as shown in the prototype design). These design choices were guided by principles like visualising activity sequences and demonstrating scaffold functions, intended to make the complex ML workflow more approachable for non-experts (Sharma & Hannafin, 2007; Suwastini et al., 2021). VisAutoML 1.0 served as a crucial artifact, allowing for empirical evaluation to validate these initial design decisions and gather specific feedback on user interaction and experience in a practical setting.

The pilot study, which included a comparison between VisAutoML 1.0 and an existing tool (H2O AutoML) and an in-depth evaluation of VisAutoML 1.0's usability and transparency, provided critical empirical insights into the strengths and weaknesses of the initial interface from a non-expert perspective. While the comparison study indicated that VisAutoML 1.0 demonstrated superior perceived usability and led to greater knowledge gain compared to the less user-centred existing tool (as shown in the evaluation results), the in-depth evaluation highlighted specific areas within the interface that required significant improvement to meet broader usability benchmarks and enhance transparency. Key issues identified through user feedback included confusion during the loading phase, a need for clearer onboarding, areas where pragmatic and hedonic quality were rated below

average compared to benchmarks, a documentation gap impacting user confidence, and challenges in fully interpreting some initial XAI outputs. These findings empirically validated that while the initial principles were a step in the right direction, their implementation in VisAutoML 1.0 still presented notable barriers for non-expert users.

The empirical findings from the pilot study served as the direct basis for the redesign objectives for VisAutoML 2.0. Each identified area for improvement in VisAutoML 1.0 translated into a specific objective for the next iteration. For example, user confusion during loading led to the objective of enhancing user understanding during this phase, while the below-average pragmatic and hedonic quality ratings underscored the need to improve the tool's practicality, efficiency, engagement, and enjoyment (Hassenzahl, 2003). The feedback on transparency and documentation directly informed objectives related to improving the clarity and comprehensibility of outputs and expanding documentation (Miller, 2019). These objectives, rooted in empirical user feedback, guided the subsequent design and prototyping stage for VisAutoML 2.0, ensuring that the interface evolution was driven by user needs and validated pain points rather than purely technical considerations.

The interface of VisAutoML underwent significant evolution in the transition from version 1.0 to 2.0, guided by the revised design principles and the specific redesign objectives. The wireframing redesign process for VisAutoML 2.0 illustrates these changes across key pages. For instance, the Home page was enhanced with features like an Easy/Expert mode toggle and more prominent documentation links, directly addressing the need to cater to varying user expertise levels and the documentation gap identified in the pilot study. The Data Import and Data Preprocessing pages were significantly revised to include features like built-in datasets, data preview, interactive data quality visualisations, and a data editor section, responding to the non-expert challenge in data analysis and aiming to improve pragmatic quality and transparency (Yang et al., 2018). The Model Training page incorporated guided onboarding steps and a data split percentage control, streamlining the workflow and enhancing user control.

Perhaps the most notable interface evolution occurred on the Model Evaluation page, which was redesigned to incorporate five distinct tabs with significantly enhanced interactive XAI visualisations. This revision was a direct response to the need to improve the clarity and comprehensibility of model outputs and boost user trust identified in the pilot study (Kaur et al., 2020; Miller, 2019). Principles like Progressive Explanation Disclosure and Multiple Ways to Communicate an Explanation were applied through features like interactive charts, tooltips, and layered information, aiming to make complex explanations more accessible and engaging for non-experts (Chromik & Butz, 2021). The inclusion of a loading screen with XAI visualisation animations was a creative application of principles related to engaging and informative user feedback and progressive disclosure, designed to manage user expectations during waiting periods while subtly introducing XAI concepts. These changes represent a deliberate effort to move beyond merely presenting information to actively guiding the user's understanding and interaction with the ML model and its explanations.

In conclusion, the evolution of VisAutoML's interface from version 1.0 to 2.0 is a clear demonstration of the iterative UCD process in action. The pilot study provided essential empirical data on user experiences with the initial interface, highlighting specific areas for improvement. These findings directly informed the redesign objectives and the subsequent interface enhancements in VisAutoML 2.0, guided by a refined set of design principles. The resulting interface of VisAutoML 2.0, with its enhanced scaffolding, improved data handling features, and significantly revised XAI visualisations, represents a more mature and user-centred design, empirically shown to be more usable and transparent for non-expert users (as demonstrated in the evaluation results for VisAutoML 2.0). This iterative evolution, driven by continuous user feedback and principled design, was crucial for

addressing the complex challenges of making AutoML accessible and understandable to a broader audience.

6.2.4 Evaluation of Usability Improvements

A central objective of this research was to enhance the usability of Automated Machine Learning (AutoML) tools for non-expert users, thereby lowering the technical barrier to engaging with Machine Learning (ML) (Hutter et al., 2019; Yang et al., 2018). Following an iterative User-Centered Design (UCD) methodology, empirical evaluations were conducted on both the initial VisAutoML 1.0 prototype and the refined VisAutoML 2.0 to systematically assess the impact of design iterations on usability. This section presents the findings from these evaluations, specifically comparing the usability metrics between the two versions to quantify the improvements achieved and discuss their implications based on the research findings and relevant literature on human-computer interaction and technology acceptance. The evaluation of usability primarily relied on established instruments such as the User Experience Questionnaire (UEQ), the System Usability Scale (SUS), and the User Experience Questionnaire-Short Form (UEQ-S), alongside the objective measure of task completion time, providing both subjective user perceptions and objective performance measures (Brooke, 1996; Schrepp et al., 2017a; Schrepp et al., 2017b).

Table 56 Comparative Analysis of VisAutoML 1.0 vs H2O AutoML

Aspect	VisAutoML 1.0	H2O AutoML	Justification / Reference
Perceived Usability (Evaluation)	Statistically Superior (UEQ, SUS) (M=0.82 UEQ Overall, SUS=61.5)	Lower Scores (UEQ, SUS) (M=-0.311 UEQ Overall, SUS=38.5) (	Empirical evaluation showed VisAutoML 1.0 was perceived as significantly more usable by non-experts.
Perceived Transparency (Evaluation)	Moderate Level (XAI Questionnaire M=118.18)	Implied Lower Level (Qualitative Feedback)	VisAutoML 1.0 showed moderate perceived explainability, and qualitative feedback suggested H2O AutoML was less transparent for non-experts
Knowledge Gain (Evaluation)	Significantly Greater Gain (Increase from 56.3% to 77.24%)	Lower Gain (Increase from 42.48% to 59.15%)	Interaction with VisAutoML 1.0 led to demonstrably better learning outcomes for non-experts
Integrated Guidance/Scaffolding	Incorporated Design Principles (DP1-DP4)	Implied Lack for Non-Experts (Qualitative Feedback)	VisAutoML 1.0 included specific features for guidance based on design principles; H2O AutoML was perceived as lacking this for non-experts
Interface Complexity (Perceived)	User-friendly, intuitive (Qualitative Feedback)	Overwhelming, confusing for non-experts (Qualitative Feedback)	Qualitative data highlighted the difference in perceived interface complexity for the target audience.

The initial evaluation stage provided a crucial baseline assessment of VisAutoML's usability for non-expert users. This stage included a comparison study between VisAutoML 1.0 and an existing AutoML tool, H2O AutoML, which served as a relevant benchmark for systems accessible via a graphical user interface (GUI) (Zöller & Huber, 2021). The usability assessment in this comparison utilised the UEQ

and the SUS (Brooke, 1996; Schrepp et al., 2017a). As presented in Table, the experimental group using VisAutoML 1.0 demonstrated statistically superior perceived usability across all UEQ metrics compared to the control group using H2O AutoML (as shown in the evaluation results). VisAutoML 1.0 achieved a mean SUS score of 61.5, indicating a moderate level of perceived usability, notably higher than H2O AutoML's score of 38.5, which falls into the "Not Recommended" range according to industry benchmarks (Sauro, 2011). These comparative results provided initial empirical support for the effectiveness of VisAutoML's foundational user-centred design principles in creating a more usable interface for non-experts compared to a less user-focused existing tool, aligning with the importance of perceived ease of use in technology acceptance models (Davis, 1989).

Following the comparative study, an in-depth evaluation of VisAutoML 1.0 was conducted to gain a more detailed understanding of its usability from the perspective of non-expert users (as discussed in the evaluation results). This evaluation employed the UEQ-S and measured task completion time (Schrepp et al., 2017b). While the task completion time analysis showed that a majority of participants (52%) could complete the designated ML task in under 5 minutes using VisAutoML 1.0, indicating a degree of efficiency, a significant portion took longer (as shown in the evaluation results). More critically, the UEQ-S results for VisAutoML 1.0, with mean scores of 0.833 for Pragmatic Quality, 0.750 for Hedonic Quality, and 0.792 for Overall Usability, were rated as 'Below Average' when compared to established UEQ benchmarks (Schrepp et al., 2017a). These findings highlighted that despite being perceived as more usable than H2O AutoML, VisAutoML 1.0 still presented significant usability shortcomings when measured against broader standards for user experience quality. The evaluation clearly identified areas requiring attention in the redesign, particularly concerning the efficiency, clarity, engagement, and overall user-friendliness of the interface and workflow.

Based on the findings from the VisAutoML 1.0 evaluation, a comprehensive redesign was undertaken, resulting in the VisAutoML 2.0 prototype. The redesign objectives specifically targeted the areas for improvement identified in the initial evaluation, aiming to enhance pragmatic and hedonic quality, simplify the onboarding process, improve the loading phase, and refine the overall interface based on the revised design principles. The subsequent evaluation of VisAutoML 2.0 assessed the impact of these changes on usability (as discussed in the evaluation results for VisAutoML 2.0). The UEQ-S results for VisAutoML 2.0 demonstrated a significant improvement in perceived usability compared to its predecessor. Mean scores rose to 1.60 for Pragmatic Quality, 1.59 for Hedonic Quality, and 1.60 for Overall Usability. Crucially, these scores were rated as 'Good' for Pragmatic Quality and 'Excellent' for both Hedonic Quality and Overall Usability when compared to the UEQ benchmark (Schrepp et al., 2017a), indicating that VisAutoML 2.0 successfully surpassed average usability standards.

A direct comparison of the usability metrics between VisAutoML 1.0 and VisAutoML 2.0 reveals the magnitude of the improvements achieved through the iterative design process. The mean score for Pragmatic Quality increased from 0.833 to 1.60, Hedonic Quality from 0.750 to 1.59, and Overall Usability from 0.792 to 1.60 (as shown in the evaluation results). These substantial increases in perceived usability indicate that the redesign efforts effectively addressed the shortcomings of VisAutoML 1.0. Furthermore, the analysis of task completion time for VisAutoML 2.0 showed a notable shift towards faster completion. A remarkable 75% of participants completed the task in under 5 minutes, with no participants taking longer than 10 minutes (as shown in the evaluation results for VisAutoML 2.0). This is a significant improvement compared to VisAutoML 1.0, where only 52% finished within 5 minutes and 14% took 10-20 minutes (as shown in the evaluation results for VisAutoML 1.0). The increased speed and efficiency in task completion provide objective evidence supporting the subjective improvements in perceived usability.

These significant usability improvements in VisAutoML 2.0 can be directly attributed to the targeted redesign efforts guided by the revised design principles. Enhancements such as the simplified onboarding process, clearer navigation, improved data handling features (e.g., data quality visualisations, editor), and the introduction of the Easy/Expert mode likely contributed to the increase in Perceived Ease of Use, making the tool more intuitive and less daunting for non-experts. The refinement of XAI visualisations and the provision of more informative feedback likely enhanced Perceived Usefulness by making the tool's outputs more understandable and actionable, while also boosting Perceived Enjoyment through increased interactivity and clarity. The strong correlations between usability, perceived usefulness, and perceived ease of use found in the evaluation studies further support how improvements in these areas collectively contribute to enhanced overall usability and user acceptance (as shown in the evaluation results).

In conclusion, the empirical evaluation of usability, including the comparison between VisAutoML 1.0 and H2O AutoML and the comparative analysis between VisAutoML 1.0 and 2.0, demonstrates significant improvements in the usability of VisAutoML for non-expert users. While the initial prototype showed promise compared to an existing tool, the iterative redesign based on user feedback and guided by specific design principles led to a demonstrably more usable and efficient interface in VisAutoML 2.0, achieving 'Good' and 'Excellent' ratings against established usability benchmarks. These findings underscore the effectiveness of the UCD approach in addressing the complex challenges of making AutoML accessible and user-friendly for a broader audience, contributing significantly to the goal of democratizing ML by lowering the technical barrier to engagement (Hutter et al., 2019).

6.3 Increasing Transparency in AutoML

6.3.1 The Transparency Challenge in AutoML

Transparency in Artificial Intelligence (AI) and Machine Learning (ML) systems refers to the degree to which human users can understand the internal workings, algorithmic choices, and decision-making processes of these systems (Larsson & Heintz, 2020; Ribeiro et al., 2016). This understanding is crucial for building trust, enabling critical evaluation, and facilitating responsible deployment, particularly as AI systems are increasingly applied in high-stakes domains. For Automated Machine Learning (AutoML) tools, which aim to automate various complex steps of the ML pipeline with minimal human intervention, ensuring adequate transparency presents a significant challenge, especially when the target users are individuals without extensive technical expertise (Hutter et al., 2019; Bove et al., 2022). This challenge is not merely technical; it is deeply intertwined with fundamental issues of user trust, comprehension, and the responsible deployment of AI technologies in real-world contexts, as opaque systems can lead to misuse or unintended consequences.

The core of the transparency challenge in AutoML stems directly from the nature of its primary function: automation. AutoML systems are designed to abstract away the technical complexities of ML development, automating time-consuming tasks such as data preprocessing, feature engineering, model selection, and hyperparameter tuning (Santu et al., 2022). While this automation significantly increases efficiency and accessibility by lowering the technical barrier to entry, it simultaneously creates "black-box" processes where the rationale behind the system's automated choices remains hidden from the user (Coors et al., 2021; Xin et al., 2021). The user provides data and receives a model or prediction, but often without a clear understanding of *how* that specific model was chosen from a

vast search space, *why* certain features were selected or transformed, or *what* the model is actually doing internally to arrive at its predictions (Khuat et al., 2022; Wang et al., 2019). This inherent opacity in the automated process itself is a fundamental impediment to achieving meaningful transparency for the end-user.

For non-expert users, defined as individuals lacking extensive knowledge of either ML or a specific application domain, this lack of transparency is particularly problematic and constitutes a significant barrier to adoption and effective use (Bove et al., 2022). Unlike ML practitioners who possess the technical knowledge to probe the internal workings of algorithms, interpret complex diagnostics, or understand the implications of different model configurations, non-experts typically lack this foundational understanding (Yang et al., 2018). When confronted with an AutoML system that operates largely as a black box, they struggle to comprehend how the system arrived at a particular result or why one automated choice was favoured over another, limiting their ability to critically evaluate the system's outputs (Kaur et al., 2020). This inability to understand the underlying process directly hinders their ability to build the necessary trust for reliance, which is crucial for adoption, especially in domains where decisions have real-world consequences (Arrieta et al., 2020; Hoffman et al., 2023). Without transparency, non-experts may be reluctant to rely on AutoML-generated models, questioning their reliability, validity, and potential biases.

While the field of Explainable Artificial Intelligence (XAI) has emerged with the explicit goal of addressing the black-box nature of ML models, often leveraging visual analytics to communicate insights (Spinner et al., 2020), integrating XAI effectively and accessibly for non-experts within an AutoML context presents its own set of challenges. Many existing XAI tools and visualisations are primarily designed with ML developers and researchers in mind, employing technical jargon, abstract concepts, and complex representations that are not readily interpretable by domain users with limited AI experience (Hohman et al., 2019; Wang et al., 2021). Consequently, even when XAI features are included in commercial or open-source AutoML tools, non-experts may still struggle to accurately interpret the explanations provided or translate them into actionable insights relevant to their specific problem or domain (Kaur et al., 2020; Wang & Yin, 2021). This gap between technical explainability and user comprehension means that simply adding XAI features does not automatically solve the transparency challenge for this audience; the design and presentation of explanations are paramount to ensure they are usable and understandable.

Compounding these issues is the prevalent integration gap between AutoML and XAI within existing systems. Often, XAI is treated as a separate, post-hoc analysis step, disconnected from the automated model building process (Wang et al., 2019; Lee et al., 2019). For non-experts, understanding the impact of automated choices made earlier in the pipeline (e.g., data preprocessing, feature selection) on the final model and its predictions is crucial for developing a coherent mental model of the entire process and building trust. A disjointed approach to transparency, where explanations are provided only for the final model without illuminating the automated steps that led to it, further exacerbates the black-box perception and limits the user's ability to learn from and gain confidence in the tool (Lee et al., 2019; Wang et al., 2019). This lack of transparency in the *process* of automation itself is a critical aspect of the challenge that needs to be addressed by integrating explainability throughout the user journey.

Therefore, the transparency challenge in AutoML for non-experts is multifaceted, arising from the inherent opacity of automated processes, the limited technical background of the users, the expert-centric nature of many existing XAI approaches, and the integration gap between AutoML and XAI. Addressing this challenge requires a deliberate effort to design tools that are transparent by design, integrating usable and comprehensible explanations throughout the user journey to foster

understanding and build trust among non-expert users (Chromik & Butz, 2021; Miller, 2019). This involves explaining the final model and providing insights into the automated choices made during the ML pipeline construction, ensuring that non-experts can effectively and responsibly leverage AutoML capabilities.

6.3.2 Explainable AI Implementation in VisAutoML

Addressing the significant transparency challenge inherent in Automated Machine Learning (AutoML) systems, particularly for non-expert users who lack the technical background to understand opaque processes and model outputs, necessitated a deliberate and principled approach to implementing Explainable Artificial Intelligence (XAI) within VisAutoML. Unlike many existing AutoML tools that either lack XAI features or treat them as a separate, often technically complex, post-hoc analysis step (Wang et al., 2019; Lee et al., 2019), VisAutoML was designed to integrate explainability directly into the user workflow, making it an intrinsic and accessible part of the ML model development and review process. This integration was not arbitrary but was rigorously guided by a specific set of design principles tailored to make complex model explanations accessible, understandable, and useful for individuals without extensive ML expertise (Chromik & Butz, 2021; Liao et al., 2020).

The implementation of XAI in VisAutoML leverages established model-agnostic techniques to provide insights into model predictions and feature importance, ensuring applicability across various underlying algorithms supported by the platform. Primarily, this involves the computation and visualisation of Shapley Additive Explanations (SHAP) values, a theoretically grounded method that quantifies the contribution of each feature to a prediction (Lundberg & Lee, 2017). In addition to SHAP-based visualisations (e.g., feature importance charts, contribution plots), VisAutoML incorporates other standard and interpretable XAI components relevant to tabular data and common ML tasks like classification and regression, such as confusion matrices, and predicted vs. actual plots (as shown in the prototype design). These visualisations are generated using robust libraries like Sklearn, Shap, Plotly, and Dash, integrated into the system's backend and frontend architecture to ensure interactivity and responsiveness (as described in the system development). The integration within the main workflow, rather than as a separate module, is a key aspect designed to ensure users engage with explanations as part of the standard process, addressing the integration gap often found in other tools (Wang et al., 2019).

The core of VisAutoML's XAI implementation lies in the deliberate application of specific XAI component design principles within the Model Review/Evaluation interface (as outlined in the design principles and shown in the prototype design). The principle of Progressive Explanation Disclosure is applied by structuring the explanations across different tabs and allowing users to explore information in increasing levels of granularity. For instance, users can view overall feature importance and then delve into individual predictions or explore feature dependence, controlling the depth of information presented (as shown in the prototype design). Tooltips and layered information provide further details upon interaction, preventing cognitive overload while allowing for deeper inquiry (Buçinca et al., 2021; Millecamp et al., 2019), thereby balancing complexity and clarity. The principle of Natural Language Rationale is addressed by supplementing visualisations with clear, concise textual descriptions and explanations, translating technical concepts into more accessible language (Ehsan et al., 2019; Wiegreffe & Marasovic, 2021). This directly helps non-experts interpret charts and metrics that might otherwise be confusing (Kaur et al., 2020), making the explanations more intuitive and user-friendly.

Furthermore, the principle of Multiple Ways to Communicate an Explanation is integral to the XAI interface design. VisAutoML provides diverse visualisations (e.g., bar charts, scatter plots, heat maps)

and views (e.g., global vs. local importance, impact relationships) to explain different aspects of the model's behaviour (as shown in the prototype design). This multifaceted approach allows users to triangulate insights and understand the model from various perspectives, catering to different cognitive styles and learning preferences (Páez, 2019; Vilone & Longo, 2021). For example, presenting both overall feature importance and the impact of features on individual predictions helps users connect general model behaviour to specific outcomes, enhancing comprehension and actionability. This principled implementation aimed to overcome the limitations of existing XAI tools by making the explanations more usable, interpretable, and relevant for non-experts (Hohman et al., 2019; Wang et al., 2021).

The iterative development process, a cornerstone of the UCD methodology, allowed for the continuous refinement of the XAI implementation based on empirical user feedback. The initial XAI features in VisAutoML 1.0, while a positive step towards transparency, were identified as areas for improvement during the pilot study evaluation. Users highlighted the need for clearer outputs and more comprehensive explanations. These findings directly informed the redesign objectives for VisAutoML 2.0 and led to significant enhancements in the XAI visualisations and their integration into the interface (as outlined in the redesign objectives and shown in the wireframing redesign). The subsequent evaluation of VisAutoML 2.0 demonstrated that these improvements were highly effective, with users reporting significantly higher levels of perceived explainability compared to the previous version (as shown in the evaluation results for VisAutoML 2.0). This iterative refinement, guided by user-centred principles and empirical evaluation, was crucial for ensuring that the XAI implementation truly served the needs of non-expert users, fostering understanding and building trust in the AutoML system.

In conclusion, the implementation of Explainable AI in VisAutoML was a core component of the strategy to address the transparency challenge for non-expert users. By integrating model-agnostic techniques like SHAP and applying specific design principles to structure and present explanations within a user-friendly interface, VisAutoML aimed to make complex model insights accessible and understandable. This principled approach, iteratively refined based on user feedback, resulted in an XAI implementation that significantly enhanced perceived explainability, contributing to increased user trust and supporting the broader goal of democratizing ML by demystifying the AI decision-making process for non-experts (Long & Magerko, 2020). The success of this implementation highlights the importance of designing XAI not just for technical accuracy, but for user comprehension and utility, particularly for audiences without extensive technical expertise.

the significant transparency challenge inherent in AutoML systems, particularly for non-expert users who lack the technical background to understand opaque processes and model, necessitated a deliberate and principled approach to implementing Explainable Artificial Intelligence (XAI) within VisAutoML. Unlike many existing AutoML tools that either lack XAI features or treat them as a separate, often technically complex, post-hoc analysis step, VisAutoML was designed to integrate explainability directly into the user workflow, making it an intrinsic and accessible part of the ML model development and review process. This integration was not arbitrary but was rigorously guided by a specific set of design principles tailored to make complex model explanations accessible, understandable, and useful for individuals without extensive ML expertise.

The implementation of XAI in VisAutoML leverages established model-agnostic techniques to provide insights into model predictions and feature importance, ensuring applicability across various underlying algorithms supported by the platform. Primarily, this involves the computation and visualisation of Shapley Additive Explanations (SHAP) values, a theoretically grounded method that quantifies the contribution of each feature to a prediction (Lundberg & Lee, 2017). In addition to SHAP-

based visualisations (e.g., feature importance charts, contribution plots), VisAutoML incorporates other standard and interpretable XAI components relevant to tabular data and common ML tasks like classification and regression, such as confusion matrices, and predicted vs. actual plots. These visualisations are generated using robust libraries like Sklearn, Shap, Plotly, and Dash, integrated into the system's backend and frontend architecture to ensure interactivity and responsiveness. The integration within the main workflow, rather than as a separate module, is a key aspect designed to ensure users engage with explanations as part of the standard process.

The core of VisAutoML's XAI implementation lies in the deliberate application of the XAI component design principles within the Model Review/Evaluation interface. The principle of Progressive Explanation Disclosure (DP6) is applied by structuring the explanations across different tabs and allowing users to explore information in increasing levels of granularity. For instance, users can view overall feature importance and then delve into individual predictions or explore feature dependence, controlling the depth of information presented. Tooltips and layered information provide further details upon interaction, preventing cognitive overload while allowing for deeper inquiry, thereby balancing complexity and clarity. The principle of Natural Language Rationale (DP7) is addressed by supplementing visualisations with clear, concise textual descriptions and explanations, translating technical concepts into more accessible language. This directly helps non-experts interpret charts and metrics that might otherwise be confusing (Kaur et al., 2020), making the explanations more intuitive and user-friendly.

Furthermore, the principle of Multiple Ways to Communicate an Explanation (DP8) is integral to the XAI interface design. VisAutoML provides diverse visualisations (e.g., bar charts, scatter plots, heat maps) and views (e.g., global vs. local importance, impact relationships) to explain different aspects of the model's behaviour. This multifaceted approach allows users to triangulate insights and understand the model from various perspectives, catering to different cognitive styles and learning preferences (Páez, 2019; Vilone & Longo, 2021). For example, presenting both overall feature importance and the impact of features on individual predictions helps users connect general model behaviour to specific outcomes, enhancing comprehension and actionability. This principled implementation aimed to overcome the limitations of existing XAI tools by making the explanations more usable, interpretable, and relevant for non-experts (Wang et al., 2021).

The iterative development process, a cornerstone of the UCD methodology, allowed for the continuous refinement of the XAI implementation based on empirical user feedback. The initial XAI features in VisAutoML 1.0, while a positive step towards transparency, were identified as areas for improvement during the pilot study evaluation. Users highlighted the need for clearer outputs and more comprehensive explanations. These findings directly informed the redesign objectives for VisAutoML 2.0 and led to significant enhancements in the XAI visualisations and their integration into the interface. The subsequent evaluation of VisAutoML 2.0 demonstrated that these improvements were highly effective, with users reporting significantly higher levels of perceived explainability compared to the previous version. This iterative refinement, guided by user-centred principles and empirical evaluation, was crucial for ensuring that the XAI implementation truly served the needs of non-expert users, fostering understanding and building trust in the AutoML system.

In conclusion, the implementation of Explainable AI in VisAutoML was a core component of the strategy to address the transparency challenge for non-expert users. By integrating model-agnostic techniques like SHAP and applying specific design principles (DP6, DP7, DP8) to structure and present explanations within a user-friendly interface, VisAutoML aimed to make complex model insights accessible and understandable. This principled approach, iteratively refined based on user feedback, resulted in an XAI implementation that significantly enhanced perceived explainability, contributing

to increased user trust and supporting the broader goal of democratizing ML by demystifying the AI decision-making process for non-experts. The success of this implementation highlights the importance of designing XAI not just for technical accuracy, but for user comprehension and utility.

6.3.3 User Trust and Understanding

For Automated Machine Learning (AutoML) tools to be effectively adopted and utilised by non-expert users, fostering a sense of trust in the system and facilitating genuine understanding of the underlying processes and outputs are paramount. Given the inherent transparency challenges in AutoML, where automated processes can obscure the rationale behind decisions, and the limited technical background of the target audience, empirical evaluation of user trust and understanding becomes critical for assessing a tool's potential for real-world impact and responsible deployment (Kaur et al., 2020; Miller, 2019). Non-experts, lacking the specialised technical knowledge to fully scrutinise complex algorithms or interpret opaque diagnostics, rely heavily on the perceived trustworthiness and comprehensibility of the system (Yang et al., 2018). This research employed specific metrics, including the Trust Questionnaire and assessments of knowledge gain, to evaluate these crucial aspects in VisAutoML and understand their relationship with usability and explainability.

Empirical evaluation using the Trust Questionnaire provided quantitative insights into non-expert users' confidence in the VisAutoML system across its iterative development. For the initial prototype, VisAutoML 1.0, the evaluation indicated a moderate level of trust, with an overall mean score of 21.73 (on a 5-point Likert scale summed across items, where higher is better) (as shown in the evaluation results for VisAutoML 1.0). This moderate trust level suggested that while users perceived some functional capabilities, they harboured reservations about relying on the system, particularly for critical decision-making, as indicated by the lowest scoring item related to using the system for this purpose. This finding aligned with literature highlighting that trust in automated systems involves more than just perceived performance; it also encompasses comfort with delegation and predictability (Lee & See, 2004; Hoffman et al., 2023). The subsequent evaluation of the refined VisAutoML 2.0 demonstrated a significant improvement in user trust, with the overall mean score rising to 26.11 (as shown in the evaluation results for VisAutoML 2.0). This increase across all trust dimensions, particularly in perceived system confidence and efficiency, suggests that the redesign efforts successfully addressed some of the factors that limited trust in the earlier version, likely related to enhanced usability and improved transparency through refined XAI.

Beyond subjective trust perceptions, assessing user understanding is crucial for ensuring that non-experts can effectively and responsibly utilise AutoML outputs. While directly measuring a user's complete understanding of complex ML algorithms is challenging, evaluating knowledge gain related to the ML process and model interpretation provides a tangible indicator of whether the tool facilitates learning and comprehension. A comparison study between VisAutoML 1.0 and an existing tool, H2O AutoML, included pre- and post-test knowledge assessments designed to measure understanding of fundamental ML concepts relevant to the task workflow. The results showed a significantly greater knowledge gain in the group using VisAutoML 1.0 compared to the group using H2O AutoML (as shown in the comparison study results). This finding is highly significant as it suggests that interaction with VisAutoML, even in its initial version, was more effective in promoting user understanding of fundamental ML concepts and processes than interaction with a less user-centred existing tool. This enhanced learning outcome is likely attributable to VisAutoML's design principles focused on visualising workflows, providing scaffolding, and integrating explanations, which

collectively aim to demystify the ML pipeline for non-experts and contribute to AI literacy (Sharma & Hannafin, 2007; Long & Magerko, 2020).

Furthermore, perceived explainability, evaluated using the XAI Questionnaire, serves as another key indicator of user understanding of how the AI system works and why it makes certain predictions (Silva et al., 2023). High perceived explainability is crucial for building trust, as users are more likely to trust a system they understand (Adadi & Berrada, 2018; Miller, 2019). The evaluation of VisAutoML 2.0 showed a high level of perceived explainability, with a mean score of 161.9 (on a 7-point Likert scale summed across items) (as shown in the evaluation results for VisAutoML 2.0), representing a dramatic improvement over VisAutoML 1.0 (mean score 118.18) (as shown in the evaluation results for VisAutoML 1.0). This indicates that the refined XAI implementation in VisAutoML 2.0, guided by principles of progressive disclosure, natural language rationale, and multiple views (Chromik & Butz, 2021), was highly successful in making the model's reasoning more comprehensible to non-expert users, thereby directly supporting user understanding.

The empirical findings also revealed strong positive correlations between Trust, Explainability (XAI), and Usability (UEQ) in the evaluation stages (as shown in the evaluation results). The strong associations empirically support the theoretical premise that enhancing usability and explainability is fundamental to fostering user trust in AutoML systems for non-experts. A tool that is easy to use and understand is more likely to be perceived as reliable and trustworthy (Davis, 1989; Miller, 2019). These correlations underscore the interconnectedness of these factors in shaping the overall user experience and influencing technology acceptance, aligning with findings from technology acceptance models (Davis, 1989). The iterative improvements in usability and XAI implementation in VisAutoML 2.0 demonstrably contributed to the observed increase in both perceived explainability and trust.

In conclusion, the evaluation results demonstrate that VisAutoML, particularly in its refined 2.0 version, has made significant strides in fostering user trust and understanding among non-experts. The improvement in trust scores, the evidence of enhanced knowledge gain, and the high levels of perceived explainability collectively indicate that the user-centred design approach and the principled implementation of usable XAI have been effective in addressing the transparency challenge (Kaur et al., 2020). Fostering trust and understanding is not merely a matter of user satisfaction; it is essential for enabling non-experts to use AutoML tools effectively and responsibly, allowing them to critically evaluate results, identify potential limitations or biases, and make informed decisions based on the AI's outputs, thereby moving beyond blind reliance on automated processes and contributing to a more AI-literate user base (Long & Magerko, 2020).

6.3.4 Balancing Complexity and Clarity

A fundamental and pervasive challenge in the design of effective Automated Machine Learning (AutoML) tools specifically for non-expert users lies in navigating the inherent tension between the sophistication and complexity of Machine Learning (ML) processes and the critical need for clarity, simplicity, and comprehensibility in the user interface (Yang et al., 2018). AutoML systems, by their nature, involve sophisticated steps such as automated data preprocessing, feature engineering, model selection from potentially hundreds of algorithms, hyperparameter tuning, and complex model evaluation (Hutter et al., 2019; Santu et al., 2022). Presenting this full spectrum of underlying complexity directly to non-experts, who often lack foundational knowledge in statistics, programming, or ML theory, would be overwhelmingly intimidating and counterproductive, effectively negating the goal of increasing accessibility and creating significant barriers to usability and understanding (Bove et al., 2022; Yang et al., 2018).

The need for clarity and simplicity in the user interface for non-experts is paramount for fostering perceived ease of use and encouraging technology adoption, as highlighted by models of technology acceptance (Davis, 1989). A user-friendly interface abstracts away unnecessary technical details, provides intuitive navigation, and presents information in a digestible format that aligns with the user's mental model (Nielsen, 2012). However, this pursuit of simplicity must be carefully balanced, as oversimplification risks creating an opaque black box where users interact with the system without any meaningful understanding of what is happening beneath the surface. Such opacity hinders the development of user understanding and trust, which are essential for effective and responsible use of AI technologies (Kaur et al., 2020; Miller, 2019). The challenge, therefore, is not merely to hide complexity, but to design an interface that is simple enough to be accessible while being sufficiently informative to provide meaningful transparency and enable users to build a coherent, albeit simplified, mental model of the AutoML process and the resulting models.

To effectively navigate this inherent tension between complexity and clarity, a key design strategy employed in the development of user-centred AutoML tools is the principle of Progressive Disclosure. This principle, rooted in Human-Computer Interaction (HCI) design, advocates for providing finer granularity of information or functionality through subsequent steps, revealing details gradually as the user explores the interface or explicitly requests more specific explanations (Buçinca et al., 2021; Millecamp et al., 2019). Instead of overwhelming the user with all possible details and options upfront, information is revealed in layers, allowing users to engage with complexity at a pace and depth that matches their current understanding and interest. This technique is a powerful method for managing cognitive load, particularly in complex domains like ML, by presenting information in manageable chunks.

The principle of Progressive Disclosure can be extensively applied in the design and iterative refinement of AutoML interfaces, particularly within workflow navigation and the presentation of XAI visualisations. For example, structuring model evaluation interfaces into distinct tabs allows users to access different aspects of model explanation in a structured, layered manner (as shown in the prototype design). Within these tabs, complex visualisations like SHAP plots can be accompanied by tooltips and layered information that provide additional details or definitions only when the user interacts with specific elements (as shown in the wireframing redesign). This allows users to engage with the level of detail that matches their current understanding and interest, without being overwhelmed by the full complexity of the underlying data or algorithms (Chromik & Butz, 2021). Furthermore, sequential workflows and navigation menus can embody this principle by guiding users through the ML pipeline step-by-step, presenting each stage and its associated options only after the previous one is completed, breaking down the complex end-to-end process into manageable steps.

The application of the progressive disclosure principle in designing AutoML tools is crucial for enhancing both usability and transparency, directly addressing the challenge of balancing complexity and clarity. By presenting information in a layered, user-controlled manner, it helps manage the cognitive load on non-expert users, making the interface feel less intimidating and more intuitive (Nielsen, 2012). Simultaneously, it facilitates transparency by making deeper levels of explanation accessible when needed, allowing users to explore the rationale behind model predictions and build a more robust understanding without being forced to process overwhelming technical details (Miller, 2019). This approach supports the development of user understanding and trust, as users can gradually uncover the system's workings at their own pace (Adadi & Berrada, 2018). Iterative design and user evaluations are essential in refining the implementation of progressive disclosure to ensure it effectively supports non-expert comprehension and interaction.

In conclusion, the principle of Progressive Disclosure offers a viable and effective strategy for balancing the inherent complexity of AutoML with the critical need for clarity and simplicity for non-expert users. By revealing information in layers and allowing users to control the depth of detail they engage with, this design approach manages cognitive load, enhances usability, and facilitates meaningful transparency. Implementing this principle thoughtfully throughout the interface, from workflow navigation to XAI presentation, is essential for creating AutoML tools that are accessible and empower non-experts to understand and effectively leverage Machine Learning technologies.

6.4 Impact and Applications

6.4.1 Promoting AI Literacy

As Artificial Intelligence (AI) continues its rapid integration into various facets of society and industry, fostering AI literacy among the general public and domain specialists has become increasingly critical for informed engagement and responsible participation (Long & Magerko, 2020; Luckin & Holmes, 2016). AI literacy encompasses the foundational knowledge and practical skills necessary to understand how AI systems function, interact with them effectively, and leverage their capabilities while being mindful of their limitations and potential societal impacts. Traditional Machine Learning (ML) development workflows, however, necessitate specialised programming, statistical, and theoretical expertise, establishing a significant barrier for individuals without such technical backgrounds and consequently limiting their capacity to understand how AI works or contribute to its development (Benbya et al., 2021; Yang et al., 2018). This research posits that Automated Machine Learning (AutoML) tools, when deliberately designed with a strong emphasis on usability and transparency, such as VisAutoML, hold significant potential to contribute meaningfully to promoting AI literacy among non-expert users by making the process of engaging with ML more accessible and comprehensible.

The inherent complexity of traditional ML workflows and the "black-box" nature of many existing AutoML systems create considerable challenges for non-experts seeking to understand AI. Non-experts often perceive ML algorithms as opaque input-output mechanisms, struggling to grasp the intricate functionalities and underlying decision-making processes (Yang et al., 2018; Xin et al., 2021). This lack of transparency hinders their ability to build a conceptual model of the ML pipeline and limits their capacity to critically evaluate the outputs they receive (Kaur et al., 2020). Furthermore, traditional tools often require navigating complex interfaces and interpreting technical jargon, adding layers of difficulty for users without prior technical literacy (Hohman et al., 2019; Liao et al., 2020). These combined factors contribute to a significant barrier to AI literacy, preventing a broader audience from gaining a practical understanding of AI principles and capabilities.

VisAutoML's design directly addresses these barriers by making the ML development process more accessible and understandable for non-experts. The user-friendly interface, guided workflow, and abstraction of coding requirements allow users to engage with the fundamental concepts of data, models, and predictions without being overwhelmed by technical complexities (as demonstrated in the system design). Design principles focusing on visualising activity sequences and demonstrating scaffold functions help users build a mental model of the ML pipeline, understanding the steps involved from data import to model evaluation (as outlined in the design principles). This demystification of the process, moving users beyond viewing AI as a mysterious black box, is a critical step in enhancing foundational AI literacy (Ilkka, 2018).

Furthermore, the integrated Explainable AI (XAI) features in VisAutoML play a pivotal role in promoting a deeper level of AI literacy, specifically concerning the understanding of model behaviour and decision-making. By providing accessible visualisations and explanations of feature importance, model

predictions, and evaluation metrics, VisAutoML helps non-experts gain insights into why a model makes certain predictions and how different factors influence the outcome (as shown in the prototype design). The application of XAI principles like progressive disclosure, natural language rationale, and multiple ways to communicate explanations makes these insights comprehensible and digestible for users without technical jargon (Chromik & Butz, 2021; Ehsan et al., 2019). This enhanced transparency and perceived explainability are crucial components of AI literacy, enabling users to critically evaluate the reliability and limitations of AI systems, fostering a balance between complexity and clarity (as explored in the discussion on balancing complexity).

Empirical evidence from the evaluation studies supports VisAutoML's potential for promoting AI literacy. A comparison study demonstrated significantly greater knowledge gain in the group using VisAutoML compared to a less user-centred existing tool (as shown in the evaluation results). This finding suggests that interacting with VisAutoML facilitated users' learning of fundamental ML concepts, indicating that the tool serves as an effective learning environment. The high levels of perceived explainability and improved user trust reported by users of the refined VisAutoML version further reinforce this, indicating that users felt they understood the AI's reasoning and outputs and were more confident in the system (as detailed in the evaluation results and discussion on trust). These empirical results align with the notion that usable and transparent AI tools can empower users to learn about AI by doing, fostering a more practical and experiential form of AI literacy (Ilkka, 2018). The significant improvements in usability across iterative versions underscore how user-centred design directly contributes to this learning potential (as shown in the usability evaluation).

The promotion of AI literacy through tools like VisAutoML has broader societal implications. By making ML accessible and understandable, VisAutoML empowers individuals from diverse backgrounds, including domain experts, to leverage AI for problem-solving in their respective domains, fostering innovation and enabling rapid prototyping. More importantly, enhanced AI literacy enables users to engage with AI critically, understanding potential biases, limitations, and ethical considerations (Floridi & Cowls, 2022; Zou & Schiebinger, 2018). This critical understanding is essential for the responsible development and deployment of AI technologies in society. Therefore, VisAutoML's contribution to AI literacy extends beyond technical skill acquisition, fostering informed and responsible participation in an increasingly AI-driven world.

6.4.2 Rapid Prototyping

As rapid prototyping in Machine Learning (ML) refers to the crucial ability to quickly build, test, and iteratively refine ML models and exploratory ideas. This iterative process is fundamental for efficiently exploring complex datasets, validating hypotheses about relationships within the data, and refining modelling approaches based on initial results and insights. In traditional ML workflows, the pathway from initial data ingestion and preparation through feature engineering, model selection, training, and evaluation can be a protracted and resource-intensive undertaking (Hutter et al., 2019). It demands significant manual effort, specialised technical expertise in programming and statistics, and often involves considerable time investment, which collectively hinders the capacity for rapid experimentation and agile prototyping, particularly for users who are not seasoned ML specialists (Benbya et al., 2021; Yang et al., 2018). This inherent friction in traditional processes limits the speed at which users can translate domain knowledge or data-driven questions into tangible model outputs and actionable insights.

Automated Machine Learning (AutoML) tools are fundamentally designed with the explicit goal of accelerating the ML development lifecycle by automating various time-consuming and technically

complex steps (Hutter et al., 2019). By automating tasks such as data preprocessing, feature engineering, model selection from potentially vast search spaces, and hyperparameter tuning, AutoML significantly reduces the manual effort and the level of specialised expertise required to build a functional model (Santu et al., 2022; Singh & Joshi, 2022). This automation enables users to move from raw data to a functional model or initial set of predictions much more quickly than would be feasible using entirely manual, code-based methods. This acceleration is a core benefit of AutoML, directly facilitating more rapid exploration of different modelling approaches and enabling faster validation or refutation of initial ideas and hypotheses.

VisAutoML specifically enhances this rapid prototyping capability, particularly for non-expert and domain expert users, through its user-centred design and streamlined workflow. The intuitive graphical interface, guided navigation, and abstraction of coding complexities allow users to quickly set up and run ML experiments without getting bogged down in the technical minutiae of programming or complex configuration (as demonstrated in the system design and usability evaluations). The drag-and-drop interface for model development and the logical, sequential process from data import to model review contribute to a fluid and efficient user experience (as shown in the prototype design). Design principles like visualising activity sequences and demonstrating scaffold functions help users understand the process quickly and navigate the workflow efficiently, further reducing the time needed to build and test models (as outlined in the design principles). This focus on usability directly translates into a reduced time-to-insight for users less familiar with the intricacies of ML development.

Empirical evidence from the evaluation studies strongly supports VisAutoML's capability for rapid prototyping. The analysis of task completion time for the refined VisAutoML version revealed that a significant majority of non-expert participants were able to complete the process of building and interacting with an ML model very quickly, with a high percentage finishing the designated task in under 5 minutes (as shown in the usability evaluation results). This objective measure of efficiency demonstrates that VisAutoML successfully enables rapid progression through the core ML development steps for non-experts, allowing users to obtain initial model results and insights swiftly. While the specific task in the evaluation was contained, the high speed and low variability of completion times indicate that the tool's user-centred interface and automated backend are effective in significantly accelerating the prototyping process for its target audience.

The benefits of this rapid prototyping capability are substantial, particularly for non-experts and domain experts. It empowers them to quickly explore their data, test different hypotheses about relationships within the data, and rapidly iterate on model ideas without requiring extensive technical support or investing significant time in manual coding. This accelerates the pace of discovery and innovation within their respective domains, allowing them to derive timely insights and make data-driven decisions more effectively (Singh & Joshi, 2022). Furthermore, rapid prototyping facilitated by tools like VisAutoML can lead to significant time and cost savings compared to traditional ML development, making ML more accessible and economically viable for a wider range of applications and users who may not have the resources for extensive manual development (Hutter et al., 2019). By enabling users to quickly translate ideas into functional models and gain rapid insights, VisAutoML empowers non-experts to leverage ML effectively for problem-solving and decision-making, contributing to the broader democratisation of AI.

6.4.3 Ethical Considerations

The increasing accessibility of Machine Learning (ML) capabilities through Automated Machine Learning (AutoML) tools, particularly when targeting non-expert users, brings forth a crucial set of ethical considerations that must be addressed proactively in their design, development, and deployment. While AutoML holds the promise of democratising access to powerful AI technologies, fostering innovation, and enabling data-driven decision-making across diverse domains, the power embedded within these tools, if wielded without adequate understanding, awareness of limitations, or appropriate safeguards, can lead to unintended and potentially harmful consequences. Fostering trustworthy and responsible AI necessitates a deliberate and embedded approach to ethical considerations throughout the entire development lifecycle, especially when the intended users are individuals who may not be fully conversant with the potential risks and complexities inherent in ML.

One significant ethical dimension that is highly relevant to AutoML, particularly for non-experts, is the pervasive potential for bias in ML models (Zou & Schiebinger, 2018). Models learn patterns from the data they are trained on, and if this data reflects existing societal biases, historical inequities, discriminatory practices, or even technical measurement errors, the model's subsequent predictions and decisions will inevitably perpetuate and potentially amplify these biases. For non-expert users, who may lack the specialised skills required for rigorous data analysis and validation prior to model building, identifying and mitigating complex forms of bias within datasets can be particularly challenging (Yang et al., 2018). AutoML tools, by automating various data processing and model selection steps, could inadvertently obscure these underlying data issues or the mechanisms through which bias is propagated, unless specifically designed with features to highlight and address them. While VisAutoML incorporated features aimed at improving transparency regarding data quality issues, which can contribute to bias (as shown in the system design for data preprocessing), the explicit guidance or tools for helping non-experts detect and understand different types of biases (e.g., representational, historical, or algorithmic) and providing mechanisms for mitigation were not the primary focus and remain a complex challenge for such tools.

Transparency and explainability, core tenets of the Explainable Artificial Intelligence (XAI) field, are not merely technical features but fundamental ethical requirements for AI systems, especially those operating as "black boxes" or deployed in high-stakes domains (Kremers, 2020; Larsson & Heintz, 2020). Users have a right to understand how an AI system arrives at a decision, particularly when that decision impacts their lives or livelihoods. For non-experts, who may lack the technical means to scrutinise complex algorithms or interpret opaque outputs, usable and comprehensible explanations are not just a matter of convenience for enhancing usability; they are an ethical necessity that empowers them to critically evaluate the AI's output, identify potential errors or limitations, and make informed decisions rather than relying blindly on automated suggestions (Ribeiro et al., 2016; Miller, 2019). This research placed transparency and explainability at the core of VisAutoML's design, integrating XAI features guided by principles aimed at making explanations accessible and understandable for non-experts. The empirical evaluations demonstrated that VisAutoML achieved high levels of perceived explainability and fostered user trust (as shown in the evaluation results and discussion on trust), indicating that the design choices contributed positively to addressing this ethical dimension by enabling users to better understand the AI's reasoning. However, the ongoing challenge in XAI remains ensuring that explanations are both technically faithful to the model and genuinely comprehensible and actionable for diverse non-expert users (Wang et al., 2021; Liao et al., 2020).

Beyond the individual user's interaction with the tool, the broader societal impact of widespread AutoML deployment by non-experts presents further ethical considerations. While democratisation can spur innovation and empower new users, it also carries the risk that users might inadvertently build and deploy biased, unfair, or flawed models that lead to inequitable outcomes, or that they

might misinterpret results with negative consequences in real-world applications. Responsible AI development necessitates considering these potential downstream effects and implementing safeguards or educational components to mitigate risks. VisAutoML's potential to promote AI literacy can be viewed as a positive ethical contribution in this regard, as more AI-literate users are better equipped to understand the capabilities, limitations, and potential risks of the tools they use and the models they build (Long & Magerko, 2020). However, the direct ethical consequences of deploying models built with VisAutoML in real-world, high-stakes scenarios, and the implementation of explicit ethical safeguards within the tool (beyond transparency and data quality indicators) to prevent harmful deployment were outside the direct scope of this research and its evaluation.

In conclusion, the design and evaluation of VisAutoML were sensitive to key ethical considerations, particularly prioritising transparency and explainability as crucial components for building user trust and enabling critical evaluation by non-experts. While efforts were made to address aspects related to bias through data quality transparency and the potential for promoting AI literacy was recognised as a positive societal impact, the comprehensive ethical implications regarding robust bias detection and mitigation, ensuring fairness in deployment, establishing accountability mechanisms, and implementing broader societal safeguards warrant further dedicated research and development. Future work should build upon the foundation of usable and transparent design established by VisAutoML to explicitly integrate features and guidance that empower non-experts to navigate the complex ethical landscape of AI more fully and responsibly, ensuring that the democratisation of ML through AutoML contributes positively and equitably to society.

7 Conclusion and future work

7.1 Overview of the Research Journey

The pervasive integration of Machine Learning (ML) and Artificial Intelligence (AI) across numerous domains underscores a critical need for broader accessibility beyond the confines of technical experts. While the potential for AI to drive innovation and efficiency is widely acknowledged (Yee, 2023; Benbya et al., 2021), the reality is that traditional ML development demands a high degree of specialised knowledge in programming, statistics, and algorithmic theory, creating a significant barrier to entry for non-expert users (Yang et al., 2018). These individuals, despite possessing invaluable domain-specific knowledge, are often excluded from directly leveraging ML to address problems within their fields. Automated Machine Learning (AutoML) has emerged as a promising approach to mitigate this technical barrier by automating various complex steps of the ML pipeline (Hutter et al., 2019). However, many existing AutoML tools, while technically sophisticated, frequently operate as "black boxes" with interfaces that remain challenging for non-experts to understand and trust, thereby limiting their true usability and transparency (Kaur et al., 2020; Xin et al., 2021). This research was fundamentally motivated by this critical problem: the persistent accessibility gap in ML for non-experts stemming from inadequate usability and transparency in existing AutoML solutions.

Addressing this multifaceted challenge necessitated a systematic and user-centric approach that prioritised the needs and capabilities of the target audience throughout the development process. The research journey was therefore deliberately guided by a rigorous User-Centered Design (UCD) methodology (Norman, 2013). UCD, with its iterative cycles of understanding user needs, designing, prototyping, and evaluating, was deemed particularly appropriate for this problem space. Given the target audience's limited technical background and the inherent complexity of the ML domain, a UCD approach ensured that the developed solution would be genuinely accessible, intuitive, and trustworthy, rather than merely technically functional. This methodology provided a structured framework for systematically gathering and translating user requirements into design principles and iteratively refining the system based on empirical evaluation and feedback, thereby mitigating the risk of creating a tool that, despite automation, remained unusable or untrustworthy for non-experts.

The research journey commenced with a comprehensive requirements gathering phase, aiming to establish a deep understanding of the non-expert user landscape and the specific challenges they encounter when engaging with ML. This involved a multi-faceted approach, including a thorough review of existing literature on AutoML, XAI, and user challenges, alongside empirical studies such as an Extended Technology Acceptance Model (TAM) investigation, user surveys, and interviews (Davis, 1989; Yang et al., 2018; Venkatesh & Davis, 2000). The insights gleaned from this foundational phase were critical, highlighting specific user challenges related to data analysis, model interpretation, understanding complex terminology, and a reliance on external documentation. These findings provided empirical evidence of user needs and perceptions, ensuring that the subsequent design efforts were directly responsive to the identified challenges and grounded in a solid understanding of the target audience's motivations and limitations.

Following the requirements gathering, the journey progressed through iterative cycles of design and prototyping. Based on the empirically informed design principles synthesised from the foundational research – which emphasised aspects such as visual guidance, scaffolding, and accessible explanations tailored for non-experts – the VisAutoML tool was iteratively designed and developed (Chromik & Butz, 2021; Sharma & Hannafin, 2007). This involved translating the user-centred principles into tangible interface elements, workflows, and integrated functionalities, including the implementation of Explainable AI (XAI) features designed to demystify model outputs for non-expert comprehension

(Miller, 2019). The iterative nature of this stage, moving from initial concepts to functional prototypes (VisAutoML 1.0) and subsequently refined versions (VisAutoML 2.0), allowed for continuous refinement based on preliminary testing and internal review, ensuring that the design progressively aligned with the overarching goal of creating a usable and transparent AutoML tool for tabular data.

Integral to the UCD methodology and central to the research journey were the empirical evaluation stages. The developed VisAutoML prototypes were subjected to rigorous evaluation with representative non-expert users to assess their effectiveness in enhancing usability and transparency. This involved mixed-methods studies employing standardised usability questionnaires (UEQ, SUS), transparency and trust assessments (Trust Questionnaire, XAI Questionnaire), objective measures like task completion time, and qualitative data collection through interviews and open-ended questions (Brooke, 1996; Schrepp et al., 2017a; Hoffman et al., 2023; Ribeiro et al., 2016). These evaluations, including a comparative study against an existing AutoML tool (H2O AutoML), provided crucial empirical evidence on the usability, transparency, and overall effectiveness of VisAutoML in enhancing the non-expert user experience compared to existing solutions. The findings from each evaluation phase directly informed subsequent design iterations, embodying the iterative feedback loop characteristic of UCD and ensuring that the tool progressively addressed the identified challenges based on real-world user interaction and feedback.

In essence, the research journey, guided by a UCD methodology, was a deliberate and iterative process aimed at creating a usable and transparent AutoML tool for non-expert users. From understanding the fundamental challenges faced by this audience and gathering their specific requirements, through the iterative design and development of the VisAutoML prototypes based on empirically informed principles, to the rigorous evaluation of the tool's effectiveness, every stage was geared towards enhancing usability and transparency. This comprehensive journey provides the empirical foundation for the conclusions drawn regarding the potential of user-centred design to bridge the accessibility gap in ML and empower non-experts to leverage AI technologies effectively and responsibly, thereby contributing to the broader democratisation of AI.

7.2 Research Objectives Revisited

This section details the successful achievement of the research objectives that guided this thesis. These objectives were specifically formulated to address the identified accessibility gap in Machine Learning (ML) for non-expert users, stemming from the usability and transparency challenges prevalent in existing Automated Machine Learning (AutoML) tools. By systematically pursuing these objectives through a User-Centered Design (UCD) methodology, the research has generated empirical evidence and developed a prototype demonstrating a viable approach to creating usable and transparent AutoML tools for this target audience.

The first objective of this research was to establish the key requirements for designing an intuitive AutoML prediction platform for non-expert users. This objective was comprehensively addressed during the initial requirements gathering phase, which employed a mixed-methods approach grounded in the principles of UCD. A thorough review of existing literature on AutoML, Explainable Artificial Intelligence (XAI), and human-computer interaction with complex systems provided foundational insights into the challenges non-experts face, such as difficulties with data analysis, model interpretation, and understanding technical terminology (Yang et al., 2018; Kaur et al., 2020). Complementing this, an Extended Technology Acceptance Model (TAM) study empirically investigated

factors influencing non-expert users' acceptance of an AutoML tool, highlighting the crucial roles of perceived usefulness, perceived ease of use, and perceived enjoyment in shaping user attitudes and intentions (Davis, 1989; Venkatesh & Davis, 2000). User surveys and interviews further enriched this understanding by eliciting direct feedback on non-experts' needs, expectations, and pain points when interacting with ML concepts and tools. The synthesis of findings from these diverse sources, including the creation of user personas embodying the characteristics of the target audience, culminated in a set of empirically grounded requirements that formed the basis for the subsequent design and development of the VisAutoML prototype.

The second objective aimed to create a prototype of a usable AutoML system capable of conducting regression and classification tasks on tabular data, incorporating the established requirements. This objective was achieved through the iterative design and prototyping stages of the UCD methodology. Informed by the user requirements and guided by specific design principles synthesised to enhance usability and transparency for non-experts (Chromik & Butz, 2021; Sharma & Hannafin, 2007), the VisAutoML tool was iteratively developed. The initial prototype, VisAutoML 1.0, was built as a tangible representation of these principles, incorporating features such as a guided workflow, simplified interface elements, and initial XAI visualisations. The subsequent development of VisAutoML 2.0 represented a significant refinement, directly addressing areas for improvement identified during the evaluation of VisAutoML 1.0. This iterative process of designing, building, and refining the prototype ensured that the tool progressively embodied the established requirements and design principles, moving towards a system that was functionally capable and user-centred in its design and interaction.

The third objective focused on quantifying and analysing the user experience, usability, and transparency of the developed AutoML system through empirical evaluation. This objective was central to the prototype evaluation stages of the UCD framework. Rigorous mixed-methods studies were conducted to evaluate both VisAutoML 1.0 and VisAutoML 2.0 with representative non-expert users. These evaluations employed standardised quantitative instruments such as the User Experience Questionnaire (UEQ), System Usability Scale (SUS), Trust Questionnaire, and XAI Questionnaire to measure perceived usability, user experience, trust, and explainability (Brooke, 1996; Schrepp et al., 2017a; Hoffman et al., 2023; Ribeiro et al., 2016). Objective measures, such as task completion time, were also collected to assess efficiency. Furthermore, qualitative data from interviews and open-ended questions provided rich contextual insights into user perceptions, challenges, and suggestions for improvement. These empirical evaluations, including a comparative study benchmarking VisAutoML 1.0 against an existing AutoML tool (H2O AutoML), generated the necessary data to quantify the effectiveness of the developed prototypes in enhancing usability and transparency for non-experts and to assess the impact of the iterative design improvements.

Finally, the fourth objective aimed to synthesise evidence-based design guidelines for developing effective and explainable AutoML tools for non-expert users. This objective was achieved by drawing upon the comprehensive findings from the empirical evaluation studies. The results, particularly the insights gained from the iterative evaluations of VisAutoML 1.0 and 2.0 and the comparative study, provided empirical evidence on which design choices and features were most effective in enhancing usability, transparency, and user understanding for non-experts. By analysing the relationships between design principles implemented in VisAutoML and the observed user experience, usability metrics, trust levels, and perceived explainability, the research was able to synthesise a set of empirically validated design principles. These principles, grounded in the practical outcomes of the user studies, offer concrete, evidence-based recommendations for future developers seeking to create AutoML tools that are truly accessible, understandable, and trustworthy for individuals without

extensive technical expertise, thereby contributing valuable knowledge to the field of human-centered AI.

7.3 Summary of Key Empirical Findings

This section synthesises the most important empirical findings derived from the evaluation studies conducted throughout this research. These findings provide crucial evidence regarding the effectiveness of the VisAutoML tool in enhancing usability and transparency for non-expert users, directly addressing the core problem that motivated this thesis. The evaluations, conducted through a mixed-methods approach within the iterative User-Centered Design (UCD) framework, employed a range of quantitative metrics and qualitative data collection techniques to capture a comprehensive understanding of user experience and system performance.

The empirical evaluation of VisAutoML's usability for non-expert users yielded significant findings, demonstrating the impact of the user-centred design approach and iterative refinement. Usability was primarily assessed using the User Experience Questionnaire (UEQ), System Usability Scale (SUS), and task completion time, providing both subjective user perceptions and objective performance measures (Brooke, 1996; Schrepp et al., 2017a; Schrepp et al., 2017b). The initial evaluation of VisAutoML 1.0, while showing promise compared to a less user-centred existing tool like H2O AutoML (as indicated by comparative SUS and UEQ scores, e.g., SUS 61.5 for V1.0 vs 38.5 for H2O AutoML, referenced in Table 53), revealed considerable room for improvement when benchmarked against general usability standards. VisAutoML 1.0's UEQ-S scores for Pragmatic Quality, Hedonic Quality, and Overall Usability were rated as 'Below Average' (as shown in Table 32), suggesting that despite relative advantages, the initial interface presented notable usability shortcomings for non-experts. However, the subsequent evaluation of the refined VisAutoML 2.0 demonstrated substantial improvements. UEQ-S scores rose to 'Good' for Pragmatic Quality and 'Excellent' for Hedonic Quality and Overall Usability (as shown in Table 46), indicating that the iterative redesign effectively addressed the earlier limitations and created a demonstrably more user-friendly and engaging interface. Objective measures of efficiency, such as task completion time, further supported these findings; a significantly higher percentage of participants completed tasks rapidly with VisAutoML 2.0 compared to VisAutoML 1.0 (as shown in Table 45 vs Table 31), providing empirical evidence of enhanced workflow efficiency for non-experts.

Beyond usability, a critical focus of this research was to enhance transparency in AutoML for non-experts. Transparency was evaluated through user trust in the system and their perceived explainability of the AI components, primarily measured using the Trust Questionnaire and the XAI Questionnaire (Hoffman et al., 2023; Ribeiro et al., 2016; Silva et al., 2023). The evaluation of VisAutoML 1.0 indicated moderate levels of user trust and perceived explainability (as shown in Table 43 and Table 44). While this demonstrated a foundational level of transparency, user feedback highlighted areas where outputs could be clearer and explanations more comprehensive. The iterative redesign for VisAutoML 2.0 specifically targeted these areas by refining the integrated Explainable AI (XAI) features and their presentation, guided by principles aimed at making explanations more accessible and understandable for non-experts (Chromik & Butz, 2021; Miller, 2019). The evaluation of VisAutoML 2.0 revealed significant improvements in perceived transparency. User trust levels increased from moderate to moderate-to-high, and perceived explainability scores were substantially higher compared to VisAutoML 1.0 (as shown in Table 49 and Table 51). These findings suggest that

the enhanced XAI implementation and user-centred design effectively addressed the transparency challenge, fostering greater user confidence and understanding in the automated ML processes and model outputs. The strong positive correlations observed between Trust, Explainability, and Usability further underscore the interconnectedness of these factors in shaping a positive and confident user experience with AutoML tools for non-experts (as shown in Table 35 and Table 52).

The empirical studies also provided insights into knowledge gain and user understanding facilitated by VisAutoML. A comparison study demonstrated that interaction with VisAutoML 1.0 led to a significantly greater knowledge gain among non-expert participants compared to those using an existing tool (H2O AutoML) (as shown in Figure 41). This finding is particularly relevant as it suggests that a user-centred and transparently designed AutoML tool can serve as an effective learning environment, helping non-experts to build a better understanding of fundamental ML concepts and the development process (Long & Magerko, 2020; Sharma & Hannafin, 2007). The enhanced usability and transparency features in VisAutoML, such as the guided workflow, visual cues, and accessible XAI visualisations, likely contributed to this improved learning outcome by demystifying the ML pipeline and making complex concepts more comprehensible. The high levels of perceived explainability in VisAutoML 2.0 further support the notion that users felt they gained a better understanding of the AI's reasoning and outputs, which is crucial for fostering AI literacy and enabling users to engage with ML more effectively and responsibly.

Collectively, these key empirical findings from the evaluation studies demonstrate that the user-centred design approach and iterative development of VisAutoML successfully addressed the core challenges of usability and transparency for non-expert users. The significant improvements observed in usability metrics, perceived transparency, trust, and knowledge gain from VisAutoML 1.0 to 2.0, alongside the initial positive comparative results against an existing tool, provide strong empirical support for the research's premise. These findings highlight that by prioritising user needs and designing for both ease of use and understandability, it is possible to create AutoML tools that are genuinely accessible and empowering for individuals without extensive technical expertise, thereby contributing to the broader goal of democratising Machine Learning.

7.4 Thesis Contributions

This research makes several significant contributions to the fields of Automated Machine Learning (AutoML), Explainable Artificial Intelligence (XAI), and Human-Computer Interaction (HCI), particularly concerning the design and evaluation of AI tools for non-expert users. These contributions stem from the systematic investigation into the usability and transparency challenges faced by individuals without extensive technical expertise and the iterative development and evaluation of the VisAutoML tool. They offer valuable knowledge and practical guidance for researchers and developers seeking to democratise access to ML and foster responsible AI adoption among a broader audience.

Here is a summary of the main contributions:

Table 57 Key Thesis Contributions

Contribution	Description

Empirically Validated Design Principles for Non-Expert AutoML	Synthesis and empirical validation of a set of design principles specifically tailored for creating usable and transparent AutoML tools for non-expert users, grounded in user needs and evaluated through iterative studies.
The VisAutoML Tool as a Proof of Concept	Development of VisAutoML (specifically V2.0) as a tangible prototype demonstrating the feasibility and effectiveness of a user-centred approach to building usable and transparent AutoML tools for tabular data.
Insights into Non-Expert User Needs and Acceptance	Deeper understanding of the specific requirements, challenges, and factors (including those from an Extended TAM) influencing the acceptance of AutoML tools among non-expert users, providing valuable insights for future tool design.
Mixed-Methods Evaluation Framework	Contribution of a robust mixed-methods evaluation framework specifically adapted for assessing the usability and transparency of integrated AutoML-XAI systems from a non-expert user perspective, combining quantitative and qualitative methods.

A primary contribution of this thesis is the synthesis and empirical validation of a set of design principles specifically tailored for creating usable and transparent AutoML tools for non-expert users. While literature offers general guidelines for usability and XAI, there was a notable gap in empirically tested principles specifically addressing the unique challenges of integrating automation and explainability for a non-technical audience (Amershi et al., 2019; Gil et al., 2019; Wang et al., 2019). Drawing from the requirements gathering phase, which included insights from the Extended Technology Acceptance Model (TAM) study and qualitative user feedback, a set of design principles was formulated, focusing on aspects such as visual guidance, scaffolding, clear feedback, and accessible explanations (Chromik & Butz, 2021; Sharma & Hannafin, 2007). These principles were not merely theoretical constructs but were iteratively applied and refined throughout the development of the VisAutoML prototypes. The empirical evaluation studies, particularly the comparative analysis of VisAutoML 1.0 against an existing tool and the in-depth evaluations of VisAutoML 1.0 and 2.0, provided crucial evidence for the effectiveness of these principles. The observed significant improvements in usability metrics (UEQ, SUS, task completion time) and perceived transparency (trust, explainability) from VisAutoML 1.0 to 2.0, directly linked to the implementation of these principles in the redesign, empirically validate their utility in practice. This set of empirically validated design principles offers concrete, evidence-based guidelines for future development efforts, moving beyond heuristic approaches to provide a principled foundation for designing accessible AutoML tools. The empirical validation is particularly important as it moves beyond theoretical postulation to demonstrate practical effectiveness in improving user experience and understanding in a complex domain, offering a solid basis for others to build upon.

The developed VisAutoML tool, particularly in its refined 2.0 version, stands as a tangible contribution of this research. It serves as a proof-of-concept demonstrating that it is feasible and effective to design and implement an AutoML tool that prioritises both usability and transparency for non-expert users, specifically for tabular data. Unlike many existing AutoML platforms that may excel in automation efficiency but fall short in user-friendliness and explainability for non-technical audiences (Hutter et al., 2019; Kaur et al., 2020), VisAutoML was built from the ground up with the non-expert user at its core, guided by the empirically derived design principles. The tool integrates automated ML capabilities with accessible XAI visualisations and a user-centred interface, providing a streamlined workflow from data import to model review. The empirical evaluation results, demonstrating significantly enhanced usability, increased user trust, and higher perceived explainability compared to existing tools and the initial prototype, validate VisAutoML as a successful instantiation of the research's design philosophy. While developed within the scope of this thesis, VisAutoML provides a

working model and a platform for further research and development in creating human-centered AI tools that bridge the gap between complex technology and broad accessibility. Its existence demonstrates that the research's theoretical and principled approach can be successfully translated into a functional system that empirically benefits non-expert users.

This thesis contributes valuable insights into the specific needs, challenges, and factors influencing the acceptance of AutoML tools among non-expert users. While the literature acknowledges that non-experts face difficulties with traditional ML workflows (Yang et al., 2018), this research delved deeper into the nuances of their requirements and motivations when interacting with automated and explained ML systems. The Extended Technology Acceptance Model (TAM) study, adapted to include constructs relevant to the context of AutoML and XAI, provided empirical evidence on the significant roles of perceived usefulness, perceived ease of use, and perceived enjoyment in shaping non-expert users' attitudes and intentions to use such tools (Davis, 1989; Venkatesh & Davis, 2000). Furthermore, the qualitative data gathered through surveys and interviews offered rich contextual insights into specific user challenges, preferences for guidance and feedback, and perceptions of transparency and trust. These findings contribute to a more detailed understanding of the non-expert user profile in the context of AutoML, highlighting the importance of factors beyond basic functionality, such as the need for intuitive interfaces, integrated learning support, and comprehensible explanations. This deeper understanding is crucial for informing the design of future AutoML tools that genuinely resonate with and empower the target audience, moving beyond assumptions about their technical capabilities or learning preferences. The detailed analysis of user feedback provides a granular view of what non-experts require from such tools, offering practical guidance for developers.

The research contributes a robust mixed-methods evaluation framework specifically adapted for assessing the usability and transparency of integrated AutoML-XAI systems for non-expert users. While standard usability evaluation methods exist (Brooke, 1996; Schrepp et al., 2017a), and some approaches to evaluating XAI have been proposed (Miller, 2019), there was a need for a comprehensive framework capable of capturing the multifaceted user experience with tools that combine automation and explainability for a non-technical audience. This research employed a mixed-methods approach that integrated quantitative measures (UEQ, SUS, task completion time, Trust Questionnaire, XAI Questionnaire) with qualitative data (interviews, open-ended questions) to provide a holistic understanding of user experience, perceived usability, trust, explainability, and knowledge gain. The iterative application of this framework across different prototype versions allowed for the systematic collection of empirical evidence to inform design refinements and validate the effectiveness of implemented principles. This evaluation framework offers a valuable methodological contribution for researchers seeking to assess the human-centered aspects of complex AI systems, providing a structured approach to gather both broad quantitative trends and nuanced qualitative insights into user interaction with automated and explained processes. This framework is particularly relevant for evaluating AI systems where user understanding and trust, not just algorithmic performance, are critical for successful adoption and responsible use.

7.5 Limitations of the Research

While this research has made significant strides in addressing the usability and transparency challenges of Automated Machine Learning (AutoML) for non-expert users and contributes valuable empirical evidence and design principles to the field, it is essential to acknowledge the inherent limitations of the study. Recognizing these limitations is crucial for the accurate interpretation of the findings and for guiding future research directions, as they define the boundaries within which the

conclusions are most applicable (Yin, 2009). A critical discussion of these constraints ensures transparency regarding the scope and generalizability of the research outcomes and provides a realistic context for the impact of the contributions.

A significant limitation of this research is its focused scope, primarily centered on the design, development, and evaluation of an AutoML tool specifically for tabular data. Tabular data represents a common and important data modality in many applications across various domains (Yang et al., 2018). Focusing on this data type allowed for a deep and detailed investigation into the specific challenges non-experts face when working with structured data and the design considerations required to make AutoML accessible for such tasks. However, this specificity implies that the findings and the empirically validated design principles, while robust within this context, may not directly generalise without adaptation to other data modalities, such as images, text, time series, or geospatial data. Each of these data types presents unique preprocessing, feature engineering, modeling, and explanation challenges that were not explored in this study (He et al., 2021). For instance, explaining a convolutional neural network's decision on an image differs fundamentally from explaining a tree-based model's prediction on tabular data. Consequently, while the principles related to user-centered design, scaffolding, and accessible explanations are likely transferable to some extent, their direct applicability and effectiveness for AutoML tools handling non-tabular data would require further investigation and potential adaptation of both the interface and the underlying XAI techniques.

Another limitation pertains to the specific Machine Learning tasks and algorithms supported by the current iteration of VisAutoML. The tool was designed to handle regression and classification tasks on tabular data and incorporated a specific set of commonly used ML algorithms suitable for these tasks (as described in the system development details). This focus was necessary to provide a concrete and manageable scope for development and evaluation within the thesis. However, this limitation implies that the design principles and the observed user experience might need adaptation for other types of ML tasks, such as clustering, dimensionality reduction, time series forecasting, or anomaly detection, which involve different objectives and evaluation metrics. Furthermore, using different classes of algorithms beyond those implemented in VisAutoML (e.g., deep neural networks, Bayesian models, reinforcement learning algorithms) could introduce unique characteristics and interpretability challenges that were not addressed in this research (Miller, 2019; Wang et al., 2021). The effectiveness of the current XAI implementation and interface design in explaining these different tasks or algorithms remains an open question and a limitation on the generalizability of the findings across the entire spectrum of ML problems and techniques.

The characteristics of the participant sample involved in the evaluation studies also warrant consideration as a limitation. While efforts were made to recruit individuals with limited prior Machine Learning experience, aligning with the definition of non-expert users (Bove et al., 2022), the sample exhibited certain demographic and background characteristics that may influence the generalizability of the findings. For instance, some evaluation studies included a notable proportion of participants from technical fields, such as Computer Science and Engineering (as shown in participant demographic tables), even if they reported low ML experience. While this provided participants with a foundational technical literacy that is increasingly common, it might mean the results on usability and transparency perceptions are not fully representative of individuals with entirely non-technical backgrounds or older demographics. Furthermore, the recruitment methods employed, such as snowball sampling or online crowdsourcing platforms like Amazon Mechanical Turk, while practical for accessing a pool of potential users, can introduce potential biases related to participant motivation, technical proficiency, and representativeness compared to a truly random sample of the broader non-expert population (Buchanan & Scofield, 2018).

The controlled setting in which the empirical evaluations of the VisAutoML prototypes were conducted also represents a limitation. Evaluating the tool in a controlled environment allowed for consistency in task execution and data collection, minimizing potential confounding variables and enabling a focused assessment of specific features and user interactions (Yin, 2009). This approach is valuable for initial validation but may not fully capture the complexities and variability of real-world operational environments where users interact with AutoML tools in authentic contexts. In real-world scenarios, users would typically work with their own potentially messy and incomplete datasets, integrate the tool into existing organizational workflows, and engage with the system over extended periods, potentially facing different types of challenges related to data quality, system integration, and long-term learning and trust development. Factors such as the influence of organizational context, collaboration with other users, and the impact of using the tool for real-stakes decision-making were not extensively explored, potentially limiting the ecological validity of the findings regarding sustained usability, transparency, and trust in real-world applications.

Furthermore, while VisAutoML integrates Explainable AI (XAI) features and design principles to enhance transparency, the depth and breadth of the explainability provided were tailored to the non-expert audience and the specific ML tasks and algorithms supported. This implies a limitation in the scope of explainability explored. The explanations focused on feature importance, individual predictions, and model performance metrics for regression and classification models on tabular data. More advanced or domain-specific explanation needs, or the challenges associated with explaining the internal workings of more complex ML models or the automated steps taken by the AutoML system itself (e.g., why a particular algorithm was chosen), were not within the purview of this research (Miller, 2019; Wang et al., 2021; Coors et al., 2021). While the evaluation demonstrated high perceived explainability for the implemented features, the findings on transparency and trust are specific to the level and type of explanations provided in VisAutoML and the context of tabular data, limiting their direct generalizability to AutoML tools with different or deeper XAI capabilities or applied to different problem domains requiring different forms of explanation.

Finally, the evaluation primarily focused on the initial interaction and short-term use of the VisAutoML tool. The duration of user engagement in the evaluation studies was limited, typically to the time required to complete specific tasks within a single session. This limitation means that the research does not provide extensive insights into long-term user engagement, sustained learning, or the development of trust over time through repeated interactions with the tool. Understanding how non-expert users' perceptions of usability and transparency evolve with continued use, how they integrate the tool into their regular workflows, and how their trust in the system develops or changes over longer periods are important aspects that were not explored in this research. Longitudinal studies would be necessary to investigate these dynamics and provide a more complete picture of the long-term impact and effectiveness of user-centred AutoML tools for non-experts.

In summary, the limitations related to the data modality, supported tasks and algorithms, participant sample characteristics, evaluation setting, the depth of XAI integration, and the duration of user engagement highlight the boundaries of this research. Acknowledging these constraints is crucial for interpreting the findings within their appropriate context and for identifying important avenues for future research and development aimed at creating more broadly applicable, ecologically valid, and ethically robust AutoML tools for diverse non-expert users.

Here is a table summarizing the key limitations:

Table 58 Key Limitations and Implications of Research

Area of Limitation	Description	Implications
Scope (Data Modality)	Research focused exclusively on tabular data.	Findings and design principles may not directly generalize to other data types (images, text, time series) which have unique preprocessing, modeling, and explanation challenges.
Scope (Tasks/Algorithms)	Focused on regression and classification tasks and a specific set of ML algorithms.	Principles and user experience may need adaptation for other ML tasks (clustering, anomaly detection) or different classes of algorithms (deep learning) with distinct interpretability challenges.
Participant Sample	Sample characteristics (e.g., technical background, age range, recruitment method) may limit generalizability to the entire non-expert population or completely non-technical users.	Perceptions of usability and transparency might be influenced by existing technical literacy; findings may not fully reflect the experiences of more diverse user groups or demographics.
Evaluation Setting	Studies conducted in controlled environments.	May not fully capture the complexities of real-world operational use, including diverse data issues, workflow integration, and long-term engagement, potentially limiting ecological validity of findings on sustained usability, transparency, and trust.
Depth of XAI Integration	Explainability tailored for non-experts on specific tasks/algorithms; did not explore more advanced or domain-specific explanations or explain the AutoML process itself.	Findings on transparency and trust are specific to the implemented XAI features and context; may not apply to tools with different or deeper explainability capabilities or different problem domains requiring different forms of explanation.
Duration of User Engagement	Evaluations primarily focused on initial interaction and short-term use within single sessions.	Does not provide insights into long-term user engagement, sustained learning, or the development of trust over time through repeated interactions in authentic workflows.
Specific Technologies Used	Development based on a specific technology stack (React JS, Django, Flask, specific ML/XAI libraries).	While principles are intended to be technology-agnostic, specific implementation choices might influence performance or user experience in ways not directly generalizable to tools built with entirely different technologies.
Specific Datasets Used	Evaluations used pre-selected or built-in datasets.	May not fully reflect the challenges users face with their own diverse, potentially messy, or domain-specific datasets, which can significantly impact data preprocessing and model building experiences for non-experts.
Focus on Supervised Learning	The ML tasks focused on supervised learning (regression and classification).	Principles and interface design may need adaptation for unsupervised learning tasks (clustering, dimensionality reduction) where the concept of a 'prediction' and its explanation differs.

7.6 Conclusion

This thesis embarked on a research journey to address a critical challenge in the widespread adoption of Machine Learning (ML): the significant accessibility gap faced by non-expert users due to the inherent complexity and often opaque nature of traditional ML workflows and existing Automated Machine Learning (AutoML) tools. The core problem identified was the lack of usable and transparent AutoML solutions specifically designed to empower individuals without extensive technical expertise to effectively engage with ML. The central argument posited and investigated was that a deliberate User-Centered Design (UCD) approach, guided by empirically derived design principles, could effectively bridge this gap by creating AutoML tools that are functionally capable and intuitive, understandable, and trustworthy for non-expert users. This work contributes empirical evidence to the growing body of literature advocating for human-centered AI development, particularly in domains where the target users lack specialised technical backgrounds (Norman, 2013; Yang et al., 2018).

The research provides strong empirical support for this central argument regarding the efficacy of a UCD approach in enhancing usability. The iterative development and evaluation of the VisAutoML tool, grounded in a UCD methodology, demonstrated that focusing intently on user needs and incorporating specific design principles aimed at simplifying interaction can lead to significant enhancements in the perceived ease of use and overall user experience for non-experts. The empirical findings from the evaluation studies, particularly the comparison between VisAutoML 1.0 and an existing AutoML tool (H2O AutoML) and the in-depth evaluations of VisAutoML 1.0 and its refined 2.0 version, underscore this success. VisAutoML 2.0, the culmination of the iterative design process, demonstrated significantly improved usability metrics, including higher User Experience Questionnaire (UEQ) and System Usability Scale (SUS) scores and reduced task completion time, compared to both its predecessor and a less user-centred alternative (as shown in evaluation results, e.g., Table 46 vs Table 32, Table 53). This indicates that the design choices aimed at simplifying the interface, streamlining the workflow, and providing integrated guidance were effective in making the tool easier to learn and use for non-experts, aligning with established principles of usability and human-computer interaction (Nielsen, 2012; Yang et al., 2018). The comparison with H2O AutoML specifically highlights that a user-centred approach can yield empirically better results for non-experts than tools primarily focused on automation efficiency, which often overlook the human element.

Furthermore, the research successfully demonstrated that transparency in AutoML for non-experts can be significantly enhanced through a principled approach to integrating Explainable Artificial Intelligence (XAI). The inherent opacity of automated ML processes presents a significant challenge for users who need to understand how models function and why specific predictions are made (Kaur et al., 2020; Miller, 2019). By designing XAI features tailored for non-expert comprehension and integrating them seamlessly into the user workflow, guided by principles such as progressive disclosure and natural language rationale, VisAutoML aimed to demystify the automated ML process and the resulting model outputs. The empirical evaluations confirmed that VisAutoML 2.0 achieved high levels of perceived explainability and fostered increased user trust compared to VisAutoML 1.0 (as shown in evaluation results, e.g., Table 49, Table 51). These findings align with literature highlighting the crucial role of explainability and trust in the adoption of AI systems, particularly for users without technical expertise (Hoffman et al., 2023; Miller, 2019; Ribeiro et al., 2016). The successful implementation of usable XAI features within VisAutoML provides empirical evidence that transparency can be effectively designed into complex automated systems, fostering greater confidence and understanding among non-technical users.

The strong positive correlations observed between usability, explainability, and trust further reinforce the interconnectedness of these factors in creating a positive and confident user experience with AutoML tools (as shown in evaluation results, e.g., Table 52). This suggests that efforts to improve how users understand and trust the AI system are intrinsically linked to how easy and pleasant they find the tool to use. A tool that is perceived as easy to use and understand is more likely to be trusted, and conversely, a trustworthy tool may be perceived as easier to use. This highlights the importance of a holistic design approach that considers these aspects not in isolation but as mutually reinforcing elements of a successful human-AI interaction, particularly for non-expert users who rely heavily on perceived attributes when assessing complex technology.

The overall success of the VisAutoML tool in demonstrating enhanced usability and transparency, as evidenced by the empirical findings, carries significant implications for the democratisation of Machine Learning. By lowering the technical barrier to entry and making the ML development process more understandable and trustworthy, tools like VisAutoML empower a broader audience, including domain experts and individuals without prior technical backgrounds, to leverage AI for problem-solving and innovation within their respective fields (Long & Magerko, 2020; Singh & Joshi, 2022). This democratisation is crucial for unlocking the full potential of AI across diverse sectors and fostering greater AI literacy in society, enabling more individuals to understand, use, and critically evaluate AI technologies.

Furthermore, the research highlights the potential for usable and transparent AutoML tools like VisAutoML to specifically support domain experts. These individuals possess invaluable knowledge within their fields but often lack the technical expertise for traditional ML development (Yang et al., 2018). By providing an accessible interface and understandable explanations, VisAutoML empowers domain experts to directly apply ML to their data, accelerating the process of gaining insights and making data-driven decisions within their specific domains. This can lead to more rapid prototyping and iteration of ML models for domain-specific problems, a benefit highlighted by the tool's efficiency in task completion (as shown in evaluation results, e.g., Table 45).

While the research has specific limitations, as acknowledged, the core finding that a user-centred design approach, guided by empirically validated principles, can effectively address the usability and transparency challenges of AutoML for non-experts provides a strong foundation for future development efforts aimed at making ML truly accessible and beneficial for all. The development of tools that are powerful and understandable and trustworthy is essential for ensuring the responsible and equitable advancement of AI technologies in society, moving beyond expert-centric development to a more inclusive future for AI. The empirical evidence presented in this thesis serves as a compelling case for prioritising the human element in the design of complex automated systems.

7.7 Future Work

Building upon the foundation established by this research, which successfully demonstrated the potential of a user-centered design approach to create a usable and transparent Automated Machine Learning (AutoML) tool for non-expert users, several promising avenues for future research and development emerge. While VisAutoML has shown significant improvements in usability and transparency for tabular data and specific tasks, the inherent complexity of Machine Learning (ML) and the diverse needs of non-expert users across various domains present ample opportunities for further exploration and enhancement. Future work should aim to extend the capabilities of

VisAutoML, evaluate its effectiveness in more complex and realistic settings, and deepen the understanding of human-AI interaction in the context of accessible ML.

A critical direction for future work involves extending the capabilities of the VisAutoML tool to support a wider range of data types and ML tasks beyond tabular data, regression, and classification. As highlighted in the limitations, real-world applications often involve complex data modalities such as images, text, time series, or geospatial data, each presenting unique challenges for data preprocessing, feature engineering, and model interpretation (He et al., 2021). Future research could investigate how the empirically validated design principles for usability and transparency can be adapted and applied to AutoML workflows for these diverse data types. This would necessitate developing new interface components and integrating or developing advanced Explainable Artificial Intelligence (XAI) techniques specifically suited to explaining models trained on non-tabular data, ensuring that the tool remains accessible and understandable while expanding its functionality. Similarly, incorporating support for other ML tasks, such as clustering, dimensionality reduction, or anomaly detection, would broaden the applicability of VisAutoML to a wider array of real-world problems faced by non-experts and domain experts.

Table 59 Technical Development Directions for VisAutoML

Research Direction	Description
Support for non-tabular data types	Adapt the interface and functionality to handle images, text, time series, and geospatial data beyond the current tabular data focus, requiring specialized preprocessing and visualization approaches.
Integration of additional ML tasks	Expand beyond regression and classification to include clustering, anomaly detection, and time series forecasting, broadening the tool's applicability to different analytical objectives.
Incorporation of wider range of ML algorithms	Extend algorithm options to include deep learning architectures, ensemble methods, and other advanced techniques while maintaining usability for non-experts.
Development of more advanced XAI techniques	Create and integrate domain-specific explainability methods tailored to different data types and application contexts, enhancing transparency across diverse use cases.
Integration of features for data preparation	Add more comprehensive data collection, cleaning, and preparation capabilities beyond basic tabular editing to address real-world messy data challenges.
Development of collaborative features	Implement functionality that supports multiple users working together on the same ML project, facilitating knowledge sharing and collaborative decision-making.

Further research should focus on evaluating the VisAutoML tool, or future iterations, in more ecologically valid settings and with different user groups. The evaluations conducted in this thesis, while providing valuable empirical evidence, were primarily conducted in controlled environments with a specific sample of non-expert users. Evaluating the tool in real-world operational contexts, where users work with their own diverse and potentially messy datasets and integrate the tool into existing workflows, is crucial for understanding its long-term usability, transparency, and impact (Yin, 2009). Longitudinal studies could investigate how user perceptions of trust and explainability evolve over time with repeated use and how the tool supports sustained learning and skill development in ML. Furthermore, evaluating VisAutoML with specific domain expert groups (e.g., biologists, social scientists, marketing analysts) in their professional contexts would provide insights into how the tool

can best support domain-specific problem-solving and leverage their expert knowledge within the AutoML workflow (Yang et al., 2018).

Table 60 Validation Research Directions for VisAutoML

Research Direction	Description
Conduct evaluations in real-world environments	Test the tool in operational contexts where users work with their own datasets and integrate it into existing workflows, providing ecologically valid insights.
Perform longitudinal studies	Assess how usability, transparency, trust, and learning evolve over extended periods of regular use, beyond initial impressions captured in short-term studies.
Evaluate with diverse user groups	Test with domain experts (e.g., biologists, social scientists, marketing analysts) and users of different backgrounds to understand how the tool supports domain-specific problem-solving.

Deepening the understanding of specific aspects of user interaction with AutoML and XAI for non-experts is another important area for future research. While this thesis explored perceived usability, transparency, and trust, further investigation into the cognitive processes involved when non-experts interact with automated and explained ML systems is warranted. Research could explore how different types of XAI explanations influence non-experts' mental models of ML algorithms and processes, how they use explanations to refine models or make decisions in practice, and the factors that contribute to long-term trust development or erosion (Miller, 2019; Hoffman et al., 2023). Investigating the impact of usable and transparent AutoML tools on AI literacy over time, beyond initial knowledge gain, would also be valuable (Long & Magerko, 2020). Such research could inform the design of more effective learning scaffolds and feedback mechanisms within AutoML tools.

Table 61 Educational Research Directions for AutoML Interaction

Research Direction	Description
Investigate cognitive processes	Study how users mentally process and understand automated ML workflows and explanations, identifying cognitive barriers and enablers.
Study impact of different XAI types	Research how various explanation types and formats affect user understanding, decision-making, and ability to refine models effectively.
Research factors influencing trust development	Identify the elements that contribute to building or eroding long-term trust in automated ML systems among non-expert users.
Assess long-term impact on AI literacy	Measure how sustained interaction with transparent AutoML tools affects users' understanding of ML concepts and processes over time.

Finally, future work can contribute to theoretical and methodological advancements in the field. Building upon the mixed-methods evaluation framework employed in this thesis, future research

could develop more refined and standardised methodologies for assessing usability, transparency, trust, and user understanding in complex, integrated AI systems for non-experts. This could involve developing new metrics or adapting existing ones to better capture the unique aspects of interacting with automated and explained processes. Theoretically, future work could contribute to developing more comprehensive frameworks that integrate concepts from human-computer interaction, cognitive psychology, and ML explainability to better predict and explain how non-expert users interact with and adopt AutoML tools, moving beyond current models of technology acceptance (Davis, 1989; Venkatesh & Davis, 2000). Such theoretical advancements could provide a stronger foundation for the design of future human-centered AI systems.

Table 62 Methodological Research Directions for AutoML

Research Direction	Description
Develop refined evaluation metrics	Create more nuanced metrics specifically designed to assess usability and transparency in integrated AI systems for non-experts.
Contribute to theoretical frameworks	Advance theoretical models that explain how non-experts interact with and adopt AutoML tools, extending beyond current technology acceptance models.
Develop evaluation methodologies	Establish standardized approaches for evaluating AutoML tools in diverse real-world settings with different user populations and objectives.

In summary, the future work stemming from this research is extensive and multifaceted. It involves expanding the technical capabilities of VisAutoML, rigorously evaluating its impact in realistic settings with diverse users, deepening the understanding of non-expert interaction with AI, and contributing to the theoretical and methodological foundations of the field. Pursuing these avenues will be crucial for continuing to bridge the accessibility gap in ML, fostering AI literacy, and ensuring the responsible and equitable deployment of AI technologies across society.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Akotuko, E. A., Pappoe, A. N. M., Azure, J. A., & Ameyaw, Y. (2021). Effect of Multimodal Instructional Approach on Students' Academic Performance in The Concept of Biological Classification. *Journal of Education and Practice*, *5*(1), 36–51.

Alicioglu, G., & Sun, B. (2021). A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers and Graphics (Pergamon)*, *xxxx*. https://doi.org/10.1016/j.cag.2021.09.002

Alicioglu, G., & Sun, B. (2022). A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers and Graphics (Pergamon)*, *102*, 502–520. https://doi.org/10.1016/j.cag.2021.09.002

Alzubi, J., Nayyar, A., & Kumar, A. (2018). *Machine Learning from Theory to Algorithms : An Overview Machine Learning from Theory to Algorithms : An Overview*. https://doi.org/10.1088/1742-6596/1142/1/012012

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., & Inkpen, K. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*, 1–13.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction - Microsoft. *2020_스마트*, 1–13.

Anderson, N. S. (1988). *User centered system design: new perspectives on human-computer interaction*. JSTOR.

Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *13*(5), 971–989.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., & Mojsilović, A. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *ArXiv Preprint ArXiv:1909.03012*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, *10*(7), e0130140.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574–594.

Bevan, N., Carter, J., Earthy, J., Geis, T., & Harker, S. (2016). New ISO standards for usability, usability reports and usability measures. *International Conference on Human-Computer Interaction*, 268–278.

Biecek, P. (2018). DALEX: explainers for complex predictive models in R. *The Journal of Machine Learning Research*, *19*(1), 3245–3249.

Bove, C., Aigrain, J., Lesot, M. J., Tijus, C., & Detyniecki, M. (2022). Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 807–819. https://doi.org/10.1145/3490099.3511139

Boyd-graber, J. (2019). *What can AI do for me ?* 229–239.

Breiman, L. (2002). *Manual on setting up, using, and understanding random forests v3. 1*. Technical Report, Http://Oz.Berkeley.Edu/Users/Breiman, Statistics Department University of California Berkeley, …. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Manual+On+Setting+Up,+Using,+And+Understanding+Random+Forests+V3.1#0

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7.

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*, 2586–2596.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730.

Chari, S., Seneviratne, O., Gruen, D. M., Foreman, M. A., Das, A. K., & McGuinness, D. L. (2020). Explanation ontology: a model of explanations for user-centered AI. *International Semantic Web Conference*, 228–243.

Chatzimparmpas, A., Martins, R. M., Jusufi, I., Kucher, K., Rossi, F., & Kerren, A. (2020). The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum*, *39*(3), 713–756. https://doi.org/10.1111/cgf.14034

Cheng, F., Liu, D., Du, F., Lin, Y., Zytek, A., Li, H., Qu, H., & Veeramachaneni, K. (2022). VBridge: Connecting the Dots between Features and Data to Explain Healthcare Models. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 378–388. https://doi.org/10.1109/TVCG.2021.3114836

Cheng, H. F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. *Conference on Human Factors in Computing Systems - Proceedings*, 1–12. https://doi.org/10.1145/3290605.3300789

Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, *81*, 59–83.

Chromik, M., & Butz, A. (2021). *Human-XAI Interaction : A Review and Design Principles for Explanation User Interfaces*. 1–22.

Coors, S., Schalk, D., Bischl, B., & Rügamer, D. (2021). *Automatic Componentwise Boosting: An Interpretable AutoML System*. 1–16. http://arxiv.org/abs/2109.05583

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *ArXiv Preprint ArXiv:2006.11371*.

Dawood, K. A., Sharif, K. Y., Ghani, A. A., Zulzalil, H., Zaidan, A. A., & Zaidan, B. B. (2021). Towards a unified criteria model for usability evaluation in the context of open source software based on a fuzzy Delphi method. *Information and Software Technology*, *130*, 106453.

Du, M., Yang, F., Zou, N., & Hu, X. (2020). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, *36*(4), 25–34.

Dudley, J. J., & Kristensson, P. O. (2018). A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(2), 1–37.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in ai systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 263–274.

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.

Elshawi, R., Maher, M., & Sakr, S. (2019). *Automated Machine Learning: State-of-The-Art and Open Challenges*. http://arxiv.org/abs/1906.02287

Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Explainable artificial intelligence and machine learning: A reality rooted perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(6), e1368.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine Learning and the City: Applications in Architecture and Urban Design*, 535–545.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gil, Y., Honaker, J., Orazio, V. D., Garijo, D., & Jahanshad, N. (2019). *Towards Human-Guided Machine Learning*. 614–624.

Giovanelli, J., & Pisano, G. (2022). Towards Human-centric AutoML via Logic and Argumentation. *CEUR Workshop Proceedings*, *3135*.

Gomez, S. R., & Nam, K. K. (2021). *Beyond Expertise and Roles : A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *ArXiv Preprint ArXiv:1412.6572*.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*,

40(2), 44–58.

H2O official documentation. (n.d.). Retrieved July 6, 2022, from https://docs.h2o.ai/h2o/latest-stable/h2odocs/welcome.html

Hastie, H., Chiyah Garcia, F. J., Robb, D. A., Laskov, A., & Patron, P. (2018). MIRIAM: A multimodal interface for explaining the reasoning behind actions of remote autonomous systems. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 557–558.

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1096257

Hornbæk, K., & Oulasvirta, A. (2017). What is interaction? *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5040–5052.

Huang, P., & Chiu, M. (2016). *Integrating user centered design , universal design and goal , operation , method and selection rules to improve the usability of DAISY player for persons with visual impairments*. *52*.

Hudson, W. (2013). User stories don't help users: introducing persona stories. *Interactions*, *20*(6), 50–53.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. *International Conference on Learning and Intelligent Optimization*, 507–523.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Springer Nature.

Ilkka, T. (2018). *The impact of artificial intelligence on learning, teaching, and education*. European Union.

Kahn, K. M., & Winters, N. (2018). *AI programming by children*.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). *Trustworthy Artificial Intelligence : A Review*. *55*(2).

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *2014 Science and Information Conference*, 372–378.

Khosravi, H., Buckingham, S., Chen, G., & Conati, C. (2022). *Computers and Education : Artificial Intelligence Explainable Artificial Intelligence in education*. *3*(March). https://doi.org/10.1016/j.caeai.2022.100074

Khuat, T. T., Kedziora, D. J., & Gabrys, B. (2022). *The Roles and Modes of Human Interactions with Automated Machine Learning Systems*. http://arxiv.org/abs/2205.04139

Klaise, J., Van Looveren, A., Vacanti, G., & Coca, A. (2021). Alibi Explain: Algorithms for Explaining Machine Learning Models. *J. Mach. Learn. Res.*, *22*, 181.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., & Yan, S. (2009). *A unified and generic model interpretability library for PyTorch, 2020*.

Komer, B., Bergstra, J., & Eliasmith, C. (2019). Hyperopt-sklearn. In *Automated Machine Learning*

(pp. 97–111). Springer, Cham.

Kopf, L. M., & Huh-Yoo, J. (2023). A User-Centered Design Approach to Developing a Voice Monitoring System for Disorder Prevention. *Journal of Voice*, *37*(1), 48–59. https://doi.org/10.1016/j.jvoice.2020.10.015

Kostopoulos, G., Karlos, S., Kotsiantis, S., & Ragos, O. (2018). Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems*, *35*(2), 1483–1500.

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2019). Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA. In *Automated machine learning* (pp. 81–95). Springer, Cham.

Kremers, R. (2020). Artifi cial Intelligence. *Level Design*, 341–368. https://doi.org/10.1201/b10933-22

Kunkel, J., Donkers, T., Michael, L., Barbu, C. M., & Ziegler, J. (2019). Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. *Conference on Human Factors in Computing Systems - Proceedings*, 1–12. https://doi.org/10.1145/3290605.3300717

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. *ArXiv Preprint ArXiv:1611.01236*.

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine learning with big data: Challenges and approaches. *Ieee Access*, *5*, 7776–7797.

Lacerda Queiroz, R., Ferrentini Sampaio, F., Lima, C., & Machado Vieira Lima, P. (2021). AI from Concrete to Abstract. *AI & SOCIETY*, *36*(3), 877–893.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2), 1–16. https://doi.org/10.14763/2020.2.1469

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P. O. S., & Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 719–728.

Lee, D. J.-L., Macke, S., Xin, D., Lee, A., Huang, S., & Parameswaran, A. (2019). A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *IEEE Data Engineering Bulletin*, *42*(2), 59–70.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350–1371.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Conference on Human Factors in Computing Systems - Proceedings*, 1–15. https://doi.org/10.1145/3313831.3376590

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *1*, 4–11.

Lieberman, H. (2009). User interface goals, AI opportunities. *AI Magazine*, *30*(4), 16–22. https://doi.org/10.1609/aimag.v30i4.2266

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16.

Luckin, R., & Holmes, W. (2016). *Intelligence unleashed: An argument for AI in education*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Margetis, G., Ntoa, S., & Antona, M. (2021). *Human-Centered Design of Artificial Intelligence*.

Mezhoudi, N. (2013). User interface adaptation based on user feedback and machine learning. *Proceedings of the Companion Publication of the 2013 International Conference on Intelligent User Interfaces Companion*, 25–28.

Millecamp, M., Htun, N. N., Conati, C., & Verbert, K. (2019). To explain or not to explain: the effects of personal characteristics when explaining music recommendations. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 397–407.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). *A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems*. *11*(3).

Moore, J. D., & Paris, C. L. (1991). Requirements for an expert system explanation facility. *Computational Intelligence*, *7*(4), 367–370.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *ArXiv Preprint ArXiv:1909.09223*.

Oh, C., Kim, S., Choi, J., Eun, J., Kim, S., Kim, J., Lee, J., & Suh, B. (2020). Understanding How People Reason about Aesthetic Evaluations of Artificial Intelligence. *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 1169–1181.

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, 485–492.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441–459.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Petersen, K., & Gencel, C. (2013). Worldviews, research methods, and their relationship to validity in empirical software engineering research. *Proceedings - Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement, IWSM-MENSURA 2013*, 81–89. https://doi.org/10.1109/IWSM-Mensura.2013.22

Pilling, F., Akmal, H., Coulton, P., & Lindley, J. (2020). The process of gaining an AI legibility mark. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–10.

Puntambekar, S. (2022). Distributed Scaffolding: Scaffolding Students in Classroom Environments. *Educational Psychology Review*, *34*(1), 451–472.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2018). A scaffolding design framework for software to support science inquiry. In *The journal of the learning sciences* (pp. 337–386). Psychology Press.

Quintana, C., Zhang, M., & Krajcik, J. (2018). A framework for supporting metacognitive aspects of online inquiry through software-based scaffolding. *Computers as Metacognitive Tools for Enhancing Learning: A Special Issue of Educational Psychologist*, *40*(4), 236–244. https://doi.org/10.4324/9781315866239-5

Rabhi, F., Ng, A., & Mehandjiev, N. (2021). *AutoML Applications in Business: A Case Study Using BrewAI*. *October*.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Ramos, G., Meek, C., Simard, P., Suh, J., & Ghorashi, S. (2020). Interactive machine teaching: a human-centered approach to building machine-learned models. *Human-Computer Interaction*, *35*(5–6), 413–451. https://doi.org/10.1080/07370024.2020.1734931

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

Robb, D. A., Lopes, J., Padilla, S., Laskov, A., Chiyah Garcia, F. J., Liu, X., Scharff Willners, J., Valeyrie, N., Lohan, K., & Lane, D. (2019). Exploring interaction with remote autonomous systems using conversational agents. *Proceedings of the 2019 on Designing Interactive Systems Conference*, 1543–1556.

Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1328–1347.

Rossi, F. (2018). Building trust in artificial intelligence. *Journal of International Affairs*, *72*(1), 127–134.

Santos, A., Castelo, S., Felix, C., Ono, J. P., Yu, B., Hong, S., Silva, C. T., Bertini, E., & Freire, J. (2019). Visus: An interactive system for automatic machine learning model building and curation. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/3328519.3329134

Santu, S. K. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., & Veeramachaneni, K. (2022). AutoML to Date and Beyond: Challenges and Opportunities. *ACM Computing Surveys*, *54*(8). https://doi.org/10.1145/3470918

Sarah, L. L. (2022). The Implementation of Web Based E-Scaffolding Enhance Learning (ESEL) on Centre of Mass Concept Understanding. *Jurnal Inovasi Pendidikan IPA*, *8*(1).

Saravanan, R., & Sujatha, P. (2018). A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 945–949.

Sass, R., Bergman, E., Biedenkapp, A., Hutter, F., & Lindauer, M. (2022). *DeepCAVE: An Interactive Analysis Tool for Automated Machine Learning*. 1–9. http://arxiv.org/abs/2206.03493

Sauro, J. (2011). *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC.

Saye, J. W., & Brush, T. (2002). Scaffolding critical reasoning about history and social issues in multimedia-supported learning environments. *Educational Technology Research and Development*, *50*(3), 77–96.

Schoenborn, J. M., & Althoff, K.-D. (2019). Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions. *ICCBR Workshops*, 51–60.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017a). Construction of a Benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(4), 40. https://doi.org/10.9781/ijimai.2017.445

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017b). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(6), 103. https://doi.org/10.9781/ijimai.2017.09.001

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Sharma, P., & Hannafin, M. (2007). Scaffolding in technology-enhanced learning environments. *Interactive Learning Environments*, *15*(1), 27–46. https://doi.org/10.1080/10494820600996972

Shneiderman, B. (2020a). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shneiderman, B. (2020b). *Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy*. *2507*(February), 1–9.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). Grand challenges for HCI researchers. *Interactions*, *23*(5), 24–25.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *International Conference on Machine Learning*, 3145–3153.

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human-Computer Interaction*, *39*(7), 1390–1404. https://doi.org/10.1080/10447318.2022.2101698

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *ArXiv Preprint ArXiv:1312.6034*.

Singh, V. K., & Joshi, K. (2022). Automated Machine Learning (AutoML): an overview of opportunities for application and research. *Journal of Information Technology Case and Application Research*, *00*(00), 1–11. https://doi.org/10.1080/15228053.2022.2074585

Subiyakto, A., Nurmiati, E., Zulfiandri, Z., & Rustamaji, E. (2022). *USER-CENTERED DESIGN APPROACH INTERFACE USING USER-CENTERED DESIGN APPROACH*. *August*.

https://doi.org/10.24507/icicelb.13.08.861

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319–3328.

Sutcliffe, A. (2022). *Designing for user engagment: Aesthetic and attractive user interfaces*. Springer Nature.

Suwastini, N. K. A., Ersani, N. P. D., Padmadewi, N. N., & Artini, L. P. (2021). Schemes of Scaffolding in Online Education. *RETORIKA: Jurnal Ilmu Bahasa*, *7*(1), 10–18. https://doi.org/10.22225/jr.7.1.2941.10-18

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Distill-and-compare: Auditing black-box models using transparent model distillation. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 303–310.

Tsai, C.-H., & Brusilovsky, P. (2019). Evaluating visual explanations for similarity-based recommendations: User perception and performance. *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 22–30.

Usmani, U. A., Happonen, A., & Watada, J. (2023). Human-Centered Artificial Intelligence: Designing for User Empowerment and Ethical Considerations. *HORA 2023 - 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, 1–5. https://doi.org/10.1109/HORA58378.2023.10156761

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. *ArXiv Preprint ArXiv:2006.00093*.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–6.

Wang, D., Ram, P., Weidele, D. K. I., Liu, S., Muller, M., Weisz, J. D., Valente, A., Chaudhary, A., Torres, D., Samulowitz, H., & Amini, L. (2020). AutoAI: Automating the end-to-end AI lifecycle with humans-in-the-loop. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 77–78. https://doi.org/10.1145/3379336.3381474

Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., & Gray, A. (2019). Human-AI Collaboration in Data Science. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359313

Wang, D., Yang, Q., Abdul, A., Lim, B. Y., & States, U. (2019). *Designing Theory-Driven User-Centric Explainable AI*. 1–15.

Wang, Q., Huang, K., Chandak, P., Zitnik, M., & Gehlenborg, N. (2022). *Towards Usable Explanations : Extending the Nested Model of Visualization Design for User-Centric XAI*.

Wang, Q., Ming, Y., Jin, Z., Shen, Q., Liu, D., Smith, M. J., Veeramachaneni, K., & Qu, H. (2019). ATMSeer. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12. http://arxiv.org/abs/1902.05009%0Ahttp://dx.doi.org/10.1145/3290605.3300911%0Ahttp://dl.acm.org/citation.cfm?doid=3290605.3300911

Wang, Q., Yi, S. L., & Gehlenborg, N. (2021). *Improving the Utility and Usability of Visualization in AI-*

driven Scientific Discovery.

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, *104*(February), 101822. https://doi.org/10.1016/j.artmed.2020.101822

Wiegreffe, S., & Marasovic, A. (2021). Teach me to explain: A review of datasets for explainable natural language processing. *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Williams, A. (1986). *and Goal-Directed Design : A Review of Three Methods for Designing Web Applications*. 1–8.

Wu, J., Atit, K., Ramey, K. E., Flanagan-Hall, G. A., Vondracek, M., Jona, K., & Uttal, D. H. (2021). Investigating students' learning through co-designing with technology. *Journal of Science Education and Technology*, *30*(4), 529–538.

Xanthopoulos, I., Tsamardinos, I., Christophides, V., Simon, E., & Salinger, A. (2020). Putting the human back in the AutoML loop. *CEUR Workshop Proceedings*, *2578*(April).

Xie, J., Myers, C. M., & Zhu, J. (2019). Interactive visualizer to facilitate game designers in understanding machine learning. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.

Yang, Q., Chen, N., & Ramos, G. (2018). *Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models*. 573–584.

Yigitbas, E., Hottung, A., Rojas, S. M., Anjorin, A., Sauer, S., & Engels, G. (2019). Context-and data-driven satisfaction analysis of user interface adaptations based on instant user feedback. *Proceedings of the ACM on Human-Computer Interaction*, *3*(EICS), 1–20.

Yin, R. K. (2009). *Case study research: Design and methods* (Vol. 5). sage.

Zöller, M.-A., Titov, W., Schlegel, T., & Huber, M. F. (2022). *XAutoML: A Visual Analytics Tool for Establishing Trust in Automated Machine Learning*. *1*(1). http://arxiv.org/abs/2202.11954

Zöller, M. A., & Huber, M. F. (2021). Benchmark and Survey of Automated Machine Learning Frameworks. *Journal of Artificial Intelligence Research*, *70*, 409–472. https://doi.org/10.1613/JAIR.1.11854

Zou, J., & Schiebinger, L. (2018). *AI can be sexist and racist—it's time to make it fair*. Nature Publishing Group.

Zytek, A., Liu, D., Vaithianathan, R., & Veeramachaneni, K. (2022). *Sibyl : Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making*. *28*(1).

Appendix

Appendix A.1: Extended Technology Acceptance Model (TAM) Questionnaire

This appendix contains a complete copy of the questionnaire used to collect data for the Extended Technology Acceptance Model (TAM) study, conducted as part of the design research phase to understand non-expert users' perceptions and factors influencing their acceptance of an Automated Machine Learning (AutoML) tool.

The questionnaire was administered online, and the introductory text provided to participants explained the survey's purpose, voluntary nature, confidentiality measures, and ethical compliance.

Extended Technology Acceptance Model (TAM) Questionnaire

This online survey aims to understand your views on emerging technologies, particularly artificial intelligence (AI), machine learning (ML) and visualizations (e.g. to increase efficiency and productivity in workplace, personal self-improvement etc.).

In this survey, you will be asked about:

1. Your background information
2. Your perceived views on AI and Information Visualization
3. Your preferences on different software requirements

Your inputs are crucial to understand the development direction of an AutoML system for non-experts.

Participation in this questionnaire is voluntary. You can opt out of this survey at any point in time without having to provide a reason. All your responses will be kept confidential and will not contain information that will personally identify you. The survey takes about 15 minutes to complete.

This survey is part of a PhD research under the School of Computer Science at the University of Nottingham Malaysia.

* Indicates required question

Ethics and Consent To Participate

This study complies with the requirements of the Ethics Committee of the Faculty of Science and Engineering of the University of Nottingham Malaysia.

Please click the link below to learn more about this study and how we protect your privacy:
https://drive.google.com/drive/folders/1rS9mJcrKUJAY3IYG2Npabx84ZkNnJfBa?usp=sharing

If you require more information about the study, please contact:

Alif Danial (researcher) at khfy6mad@nottingham.edu.my

Dr Marina Ng (supervisor) at marina.ng@nottingham.edu.my.

1. Voluntary Participation **Consent *** Tick all that apply. ☐ I confirm that I am 18 years old and above and willing to participate in the survey

Participant Demographic Information

This first section is dedicated to collect participant demographic.

2. Please enter your age in the field below * (Open text field for numerical input)

3. Please choose your gender in the field below * Mark only one oval. ☐ Male ☐ Female

4. What is your profession? * (Open text field)

5. What is the highest level of school have you completed? If currently enrolled, highest degree received. * Mark only one oval. ☐ Secondary school ☐ Diploma ☐ Bachelor degree ☐ Master's degree ☐ Doctorate degree

6. Please specify your field of study or work. * Mark only one oval. ☐ Computer Science ☐ Engineering ☐ Business ☐ Psychology ☐ Education and Arts ☐ Other: (Open text field)

Perceived Usefulness

Your views will help researchers design and develop future IoT visual analytics systems that can be used by non-expert users.

Survey questions will cover various topics to include End User Development, Internet of Things (IoT), Automated Information Visualization and Artificial Intelligence (AI).

This section attempts to explore the perceived usefulness of the potential developed system.

Responses for questions 7-13 use a 7-point scale where 1 = Strongly Disagree and 7 = Strongly Agree.

7. The system will improve my life/job quality. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

8. The system will make my life/job more convenient. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

9. The system will make me more effective in my life/job. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

10. The system will be useful to me/my job. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

11. Tracking and collecting data on myself or environment will be useful. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

12. Using artificial intelligence (AI) and machine learning (ML) to find patterns **in my data will be useful. *** Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

13. Using visualization to gain insights about my data will be useful. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

Perceived Ease of Use

This section attempts to explore the perceived ease of use of IoT visual analytics system.

Responses for questions 14-19 use a 7-point scale where 1 = Strongly Disagree and 7 = Strongly Agree.

14. It will be easy to track, collect and use data with the system. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

15. It will be difficult to learn and applying AI and ML models on my own. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

16. It will be easy to create visualizations using automated visualization function. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

17. It will be easy for me to become skillful at using **the system. *** Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

18. I have the knowledge necessary to use the system. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

19. I believe that the system will be easy to use. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

Attitude

Responses for questions 20-23 use a 7-point scale where 1 = Strongly Disagree and 7 = Strongly Agree.

20. I look forward **to using the system. *** Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

21. I think that using the system is beneficial to me. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

22. I have positive feelings of using the system. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

23. I like **to learn how AI and ML models work. *** Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

Behavioural Intention

Responses for questions 24-26 use a 7-point scale where 1 = Strongly Disagree and 7 = Strongly Agree.

24. I intend to use the system in the future. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

25. I will always try to use the system in my daily life. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

26. I plan to use the system frequently. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

Perceived Authority

This section attempts to explore the perceived authority of the potential developed system.

Responses for questions 27-31 use a 7-point scale where 1 = Strongly Disagree and 7 = Strongly Agree.

27. People whose views I respect support the use of IoT visual analytics. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

28. I believe that peer use will increase my positive perception regarding reliability and usefulness of the system. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

29. My friends or colleagues will think highly of me if I use the system. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

30. I believe the information visualization should be visible to others. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

31. I believe the information visualization **should be fun and interesting to the eye. *** Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

Perceived Enjoyment

Responses for questions 32-34 use a 7-point scale where 1 = Strongly Disagree and 7 = Strongly Agree.

32. I enjoy interacting with visualizations. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

33. I have fun using visualizations to gain insights on my data. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

34. Using visualizations on data would be interesting. * Mark only one oval. 1 2 3 4 5 6 7 Strong ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly agree

35. What constitutes a user-friendly system to you? Rank your answer. *Mark only one oval per row.*

	First choice	**Second choice**	**Third choice**
Good user interface	□	□	□
No programming required	□	□	□
No hardware wiring	□	□	□

36. How can developing an IoT project teach Artificial Intelligence (AI)? Rank your answer. *Mark only one oval per row.*

	First choice	**Second choice**	**Third choice**
Development experience	□	□	□
A tutorial of how AI works	□	□	□
Watching a video tutorial	□	□	□

37. The automated visualization should be *Mark only one oval.* □ A fully automated process (less control)
□ Semi-automated (more control)
□ Other: _____

38. I should be able to develop a project with the system.. *Mark only one oval.* □ through a step-by-step tutorial in the system.
□ after watching video tutorials to teach me the development flow.
□ through the guide of an instruction manual.
□ without any tutorials or external guidance.

39. What main aspect of your life or job do you/would you track (e.g. emotion, exercise, or finances)?

---

40. What tools do you/would you use to track or record your data? *Mark only one oval.* □ Mobile applications
□ Web applications
□ Wearable devices
□ Other: _____

41. What are current limitations associated to the tools you would use to record data? Examples: not user friendly, bad system design and etc.

---

42. What would you like to learn/discover from the collected data about yourself or environment? Example: measure sleeping hours and quality of sleep

---

---

43. How and where would you like to have your information visualized and displayed? Example: charts on mobile phones; infoviz projection in living room

---

Thank you for participating in this questionnaire! The data you input will be put to good use in understanding the factors of good end user development platforms and the eventual development of the system.

Appendix A.2: VisAutoML Evaluation Questionnaire

This appendix contains a complete copy of the questionnaire used in the prototype evaluation stages, specifically in the comparison study (VisAutoML 1.0 vs. H2O AutoML) and the in-depth

evaluations of VisAutoML 1.0 and VisAutoML 2.0. This instrument was used to collect quantitative data on user experience, usability, transparency, trust, explainability, and knowledge gain.

The questionnaire was administered online, and the introductory text provided to participants explained the survey's purpose and requested voluntary participation consent.

VisAutoML Evaluation Form

* Indicates required question.

1. Voluntary Participation Consent * Mark only one oval. ☐ I confirm that I am 18 years old and above and willing to participate in the survey

Knowledge Assessment (Pre-test)

This section assesses your prior knowledge related to Machine Learning concepts.

2. What does SHAP mean? * Mark only one oval. ☐ Sentient Human-AI Partnership ☐ Superior Humanoid AI Processor ☐ Self-Healing AI Protocol ☐ SHapley Additive exPlanations

3. How many types of models can VisAutoML (VAML) make? * Mark only one oval. ☐ 1 ☐ 2 ☐ 3 ☐ 4

4. What are the steps to develop an ML model using VAML? * Mark only one oval. ☐ Visualization, Feature Selection, Training, Import ☐ Export, Hyperparameter Tuning, Evaluation, Import ☐ Import, Preprocessing, Training, Evaluation ☐ Cross-Validation, Data Splitting, Hyperparameter Tuning, Evaluation

5. What is the relationship between "survived" and "passenger class"? * Mark only one oval. ☐ Positive Correlation ☐ Negative Correlation ☐ Neutral Correlation ☐ Inverted Correlation

6. What is the most important feature for the Titanic model? * Mark only one oval. ☐ Passenger Class ☐ Ticket ☐ Fare ☐ Sex

Participant Demographic and Experience Information

This section collects your background information and experience.

7. Please enter your email if you don't mind being interviewed after the study (Open text field)

8. Please enter your age * (Open text field for numerical input)

9. Please state your gender * Mark only one oval. ☐ Male ☐ Female

10. State your highest education qualification * Mark only one oval. ☐ High School ☐ Diploma ☐ Bachelor's Degree ☐ Postgraduate

11. Please enter your field of study * Mark only one oval. ☐ Computer Science ☐ Engineering ☐ Business ☐ Psychology ☐ Finance ☐ Other: (Open text field)

12. Please rate your experience with Machine Learning * Mark only one oval. 1 2 3 4 5 No Experience ☐ ☐ ☐ ☐ ☐ Advanced

13. How much was the time required to develop an ML model with VisAutoML? * Mark only one oval. ☐ Under 5 minutes ☐ Above 5 minutes ☐ Between 10-20 minutes ☐ Above 20 minutes

System Usability Scale (SUS)

This section assesses the overall usability of the system.

Responses for questions 14-23 use a 5-point scale where 1 = Strongly Disagree and 5 = Strongly Agree.

14. I think that I would like to use this system frequently. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

15. I found the system unnecessarily complex. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

16. I thought the system was easy to use * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

17. I think that I would need the support of a technical person to be able to use this system. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

18. I found the various functions in this system were well integrated. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

19. I thought there was too much inconsistency in this system. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

20. I would imagine that most people would learn to use this system very quickly. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

21. I found the system very cumbersome to use. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

22. I felt very confident using the system. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

23. I needed to learn a lot of things before I could get going with this system. * Mark only one oval. 1 2 3 4 5 Stro ☐ ☐ ☐ ☐ ☐ Strongly Agree

User Experience Questionnaire-Short Form (UEQ-S)

This section assesses your user experience with the system.

Responses for questions 24-31 use a 7-point scale with bipolar adjectives.

24. I found the system to be * Mark only one oval. 1 2 3 4 5 6 7 obst ☐ ☐ ☐ ☐ ☐ ☐ ☐ supportive

25. I found the system to be * Mark only one oval. 1 2 3 4 5 6 7 cumbersome ☐ ☐ ☐ ☐ ☐ ☐ ☐ easy

26. I found the system to be * Mark only one oval. 1 2 3 4 5 6 7 inefficient ☐ ☐ ☐ ☐ ☐ ☐ ☐ efficient

27. I found the system to be * Mark only one oval. 1 2 3 4 5 6 7 confusing ☐ ☐ ☐ ☐ ☐ ☐ ☐ clear

28. I found the system to be * Mark only one oval. 1 2 3 4 5 6 7 boring ☐ ☐ ☐ ☐ ☐ ☐ ☐ exciting

29. I found the system to be * Mark only one oval. 1 2 3 4 5 6 7 not interesting ☐ ☐ ☐ ☐ ☐ ☐ ☐ interesting

30. I found the system to be * Mark only one oval. 1 2 3 4 5 6 7 conventional ☐ ☐ ☐ ☐ ☐ ☐ ☐ inventive

31. I found the system to be * Mark only one oval. 1 2 3 4 5 6 7 usual ☐ ☐ ☐ ☐ ☐ ☐ ☐ leading edge

Open-Ended Questions

This section allows you to provide qualitative feedback.

32. Can you describe your experience with the AutoML tool? (Open text area)

33. What did you like about the AutoML tool? What didn't you like? (Open text area)

34. What improvements would you suggest for the AutoML tool? (Open text area)

Trust Questionnaire

This section assesses your trust in the AI system.

Responses for questions 35-41 use a 5-point scale where 1 = I disagree strongly and 5 = I agree strongly.

35. I am confident in the AI. I feel that it works well. * Mark only one oval. 1 2 3 4 5 I dis ☐ ☐ ☐ ☐ ☐ I agree strongly

36. The outputs of the AI are very predictable. * Mark only one oval. 1 2 3 4 5 I dis ☐ ☐ ☐ ☐ ☐ I agree strongly

37. The AI is very reliable. I can count on it to be correct all the time. * Mark only one oval. 1 2 3 4 5 I dis ☐ ☐ ☐ ☐ ☐ I agree strongly

38. I feel safe that when I rely on the AI I will get the right answers. * Mark only one oval. 1 2 3 4 5 I dis ☐ ☐ ☐ ☐ ☐ I agree strongly

39. The AI is efficient in that it works very quickly. * Mark only one oval. 1 2 3 4 5 I dis ☐ ☐ ☐ ☐ ☐ I agree strongly

40. The AI can perform the task better than a novice human user * Mark only one oval. 1 2 3 4 5 I dis ☐ ☐ ☐ ☐ ☐ I agree strongly

41. I like using the system for decision making. * Mark only one oval. 1 2 3 4 5 I dis ☐ ☐ ☐ ☐ ☐ I agree strongly

Explainable AI (XAI) Questionnaire

This section assesses your perception of the explainability of the AI features.

Responses for questions 42-71 use a 7-point scale where 1 = Strongly Disagree and 7 = Strongly Agree.

42. The explanations were detailed enough for me to understand. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

43. I understood the explanations within the context of the question. * 'questions' refers to the tasks 4a, 4b, and 4c in the instruction document Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

44. The explanations provided enough information for me to understand. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

45. I understood how the AI arrives at its prediction. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

46. I was able to use the explanations with my knowledge base. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

47. I would be able to repeat the steps that the AI took to reach its prediction. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

48. I think that most people would learn to understand the explanations very quickly. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

49. I would not understand how to apply the explanations to new questions. * 'questions' refers to the tasks 4a, 4b, and 4c in the instruction document Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

50. I would not be able to recreate the process by which the AI generated its predictions. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

51. I understand why the AI used specific information in its explanation. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

52. I understood the AI's reasoning. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

53. I could have applied the AI's reasoning to new problems, even if the AI didn't give me suggestions. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

54. The explanations were actionable, that is, they helped me know how to answer the questions. * 'questions' refers to the tasks 4a, 4b, and 4c in the instruction document Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

55. I believe that I could provide an explanation similar to the AI's explanation. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

56. I would need more information to understand the explanations. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

57. I had trouble using the explanations to answer the question. * 'questions' refers to the tasks 4a, 4b, and 4c in the instruction document Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

58. I believe that the explanations would not help most people in answering the question. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

59. The explanations were an important resource for me to answer the question. * 'questions' refers to the tasks 4a, 4b, and 4c in the instruction document Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

60. I do not think most people would provide similar explanations as the AI's explanation. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

61. I think that most people would be able to interpret the explanation of the AI. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

62. Most people would be able to accurately reproduce the AI's decision-making process. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

63. Most people would not be able to apply the AI's explanations to the questions. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

64. I could not follow the AI's decision-making process. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

65. I could easily follow the explanation to arrive at an answer to the question. * 'questions' refers to the tasks 4a, 4b, and 4c in the instruction document Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

66. The explanations were useful. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

67. I am able to follow the AI's decision-making process step-by-step. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

68. The explanations were not relevant for the questions I was given. * 'questions' refers to the tasks 4a, 4b, and 4c in the instruction document Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

69. I understand how the AI's decision-making process works. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

70. I could apply the explanations to the questions I was given. * 'questions' refers to the tasks 4a, 4b, and 4c in the instruction document Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree

71. I could not figure out how the AI arrived at its predictions. * Mark only one oval. 1 2 3 4 5 6 7 Stro ☐ ☐ ☐ ☐ ☐ ☐ ☐ Strongly Agree