Research Project Portfolio

University of Nottingham

Doctorate in Clinical Psychology

June 2025

Forecasting Depressive Symptom Deterioration Using Wearable Sensor Data and LSTM Models: A Longitudinal Analysis from the RADAR-MDD Study.

Fintan James Haley, BSc (Hons), MSc

Thesis submitted in part fulfilment of the requirements for the degree of Doctor of Clinical Psychology
to the University of Nottingham.

Thesis Abstract

Major Depressive Disorder (MDD) is a highly prevalent and disabling mental health condition, with early detection of symptom deterioration offering opportunities for timely intervention and improved outcomes. Advances in wearable technology and machine learning have prompted growing interest in passive monitoring as a scalable means of identifying individuals at risk of worsening depression. This thesis investigates the feasibility of forecasting depressive symptom deterioration using wearable sensor data and Long Short-Term Memory (LSTM) models, leveraging data from the RADAR-MDD study—a large, multi-site, longitudinal cohort of individuals with a history of MDD (N = 623).

The primary aim was to develop a predictive model capable of identifying clinically significant deterioration, defined as a ≥5-point increase on the Patient Health Questionnaire-8 (PHQ-8), using passive data streams (e.g., sleep, step count, heart rate) collected via Fitbit devices. Despite rigorous data preprocessing, normalization, and class-balancing procedures, the LSTM model failed to correctly identify any cases of symptom deterioration, achieving an AUC-ROC of 0.50, F1 score of 0.00, and high specificity driven by substantial class imbalance. A secondary logistic regression analysis using a reduced feature set similarly failed to demonstrate predictive utility, suggesting that the limitations lay not in the modelling technique but in the quality and completeness of the data.

Contributing factors to poor model performance included significant data sparsity—particularly in sleep data—declining participant adherence over time, and the bidirectional nature of certain behavioural indicators (e.g., sleep duration). Furthermore, exploratory analyses revealed significant cultural variation in PHQ-8 response patterns across the UK, Spain, and the Netherlands, raising concerns about the universal applicability of standardised depression measures and global prediction models.

The findings highlight the current limitations of population-level machine learning approaches in forecasting depression and support a shift toward idiographic, context-sensitive modelling strategies. The thesis also examines the psychological, ethical, and practical implications of continuous monitoring technologies in mental health, emphasising the need for user-centred, culturally informed, and ethically robust design. Recommendations for future research include improved data collection strategies, adaptive feedback systems, and the development of person-specific early warning systems that align more closely with clinical realities.

Keywords: Major Depressive Disorder, Wearable Devices, Machine Learning, LSTM, Digital Mental Health, RADAR-MDD, Prediction Models, Cultural Variation

Acknowledgments

I would like to express my sincere gratitude to Jacob Andrews, who saw the potential in this project from my very first email, even before I began the doctorate. This thesis would not have been possible without his support. Jacob facilitated access to the dataset that formed the foundation of this study and provided consistent, timely guidance throughout. His advice was particularly invaluable when I faced challenges in determining the direction of the project.

I am also deeply grateful to Nima Moghaddam, who showed great trust in me and gave me the freedom to pursue my research vision from the outset. Nima offered expert advice on integrating the project within the framework of clinical psychology and aligning it with the broader requirements of the doctorate. His insights into data modelling were especially valuable, and I am very thankful for his support.

Special thanks go to Alexander Turner, whose expertise in data analysis was critical to the success of this project. As someone whose background is not in computer science, I could not have developed the models without his knowledge and support. His dedication, including traveling to collaborate in person, made an immense difference, and I am incredibly appreciative of his generosity and collegiality.

I would also like to acknowledge Alexis Lamb, the librarian who assisted with the development of the search strategy for this review. Her expertise in crafting effective search terms was instrumental in guiding the literature review process. Thank you, Alexis, for your invaluable contribution.

Finally, I wish to thank Nick Cummins for his consultation and thoughtful advice on how best to utilise the dataset. His input was greatly appreciated throughout the project.

Statement of Contribution

1. Project Design
Fintan Haley (with supervision from Nima Moghaddam and Jacob Andrews)

2. Applying for ethical approval
Fintan Haley (with supervision from Nima Moghaddam and Jacob Andrews)

3. Writing of the systematic literature review
Fintan Haley (with supervision from Nima Moghaddam and Jacob Andrews)

4. Data analysis
Fintan Haley and Alexander Turner

5. Writeup
Fintan Haley (with supervision from Nima Moghaddam, Alexander Turner and Jacob Andrews)

The clinical data used in the study were collected as a part of a study with the RADAR-MDD project. The data had ethical approval for secondary data analysis and was kindly released for the current study by Nick Cummins.

Portfolio wordcount: 16981 words

Journal wordcount: 5279 words

Table of Contents

Table of Figures and Tables

1. Journal Paper

Forecasting Depressive Symptom Deterioration Using Wearable Sensor Data and LSTM Models: A Longitudinal Analysis from the RADAR-MDD Study

Authors:

Fintan Haley (Corresponding Author),[1] Jacob Andrews,[2] Nima Moghaddam[3], Alexander Turner[4]

Author affiliations and information:

[1]Trent DClinPsy Programme, University of Nottingham, Nottingham, United Kingdom

Fintan.haley@nottingham.ac.uk

[2]NIHR MindTech Medtech Co-operative, Academic Unit of Mental Health and Clinical Neuroscience, School of Medicine, University of Nottingham, Nottingham, United Kingdom

jacob.andrews@nottingham.ac.uk

ORCID: 0000-0001-8408-5782

[3]College of Health and Science, School of Psychology, Trent DClinPsy Programme, University of Lincoln, Lincoln, United Kingdom

nmoghaddam@lincoln.ac.uk

ORCID: 0000-0002-8657-4341

[4]Department of Computer Science, University of Nottingham, Nottingham, United Kingdom

alexander.turner@nottingham.ac.uk

ORCID: 0000-0002-2392-6549

Intended Paper: Journal of technology in behavioral science

Data Availability Statement:

The data used in this study were obtained from the Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD) study, conducted as part of the RADAR-CNS research programme. Due to ethical restrictions and participant confidentiality agreements, the dataset is not publicly available and cannot be shared.

Competing Interest Declaration:

The authors declare that they have received funding for their respective courses. Jacob Andrews is funded by the National Institute for Health and Care Research (NIHR), Nottingham Biomedical Research Centre, Mental Health and Technology theme. Fintan Haley is funded by the Nottinghamshire Healthcare NHS Foundation Trust. It should be noted that the views presented in this manuscript are those of the authors and do not necessarily reflect the views of the NIHR or Nottinghamshire Healthcare NHS Foundation Trust. Furthermore, the authors wish to clarify that no additional funds, grants, or other support were received specifically for the preparation of this manuscript. The authors have no relevant financial or non-financial interests to disclose.

Author Contribution Declaration:

All authors collectively contributed to the study's conception and design. Material preparation were primarily conducted by Fintan Haley. The data analysis was shared between Alexander Turner and Fintan Haley. The first draft of the manuscript was authored by Fintan Haley, with subsequent revisions and comments provided by all authors. All authors have reviewed and approved the final manuscript.

1.1 Abstract

Major Depressive Disorder (MDD) remains a leading global health concern, with early detection critical to mitigating its profound personal and societal impacts. This study explores the feasibility of using wearable technology and machine learning to predict future depressive episodes, rather than detect current symptoms. Leveraging the Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD) dataset, we developed a Long Short-Term Memory (LSTM) model trained on longitudinal passive data (e.g., sleep, activity, heart rate) from smartwatches and bi-weekly Patient Health Questionnaire-8 (PHQ-8) scores across a diverse sample (N=623) from the UK, Spain, and the Netherlands. Reliable depressive deterioration was defined as a ≥5-point increase on the PHQ-8.

The model failed to detect any positive cases, with precision, recall, and F1 scores of 0.00. Crucially, the AUC-ROC score was 0.50**,** demonstrating that the model performed no better than random chance in distinguishing between deterioration and non-deterioration cases. Class imbalance, data sparsity, particularly in sleep features, and cross-cultural heterogeneity likely contributed to poor model performance. Analyses revealed significant cultural variation in PHQ-8 response patterns, undermining the generalisability of a single predictive model. Our findings highlight the limitations of population-level machine learning approaches in forecasting depression and call for a shift toward person-centred modelling, improved engagement strategies, and culturally adaptive designs. Future research should prioritise dynamic, idiographic models and ethical, user-centred design to enhance the clinical utility of digital mental health tools.

1.2 Introduction

1.2.1 The Prevalence of Depression

Recent estimates indicate that approximately 16% of individuals aged 16 and older in the UK reported moderate-to-severe depressive symptoms in Autumn 2022, based on PHQ-8 assessments (Mullis & Attwell, 2022). This rate has remained elevated since the COVID-19 pandemic. MDD is a severe and persistent mood disorder characterized by prolonged feelings of sadness, hopelessness, and a diminished capacity to experience pleasure (APA, 2022). MDD significantly impairs daily functioning and is often accompanied by disruptions in appetite, sleep patterns, energy levels, and, in severe cases, thoughts of self-harm.

Psychological theories suggest that depression perpetuates itself through reinforcing cycles of behavioural disengagement and reduced environmental reward. Lewinsohn's (1974) model, for instance, posits that a decline in engagement with previously rewarding activities leads to fewer opportunities for positive reinforcement, which in turn sustains depressive symptoms. Research supports this idea, with studies demonstrating that reduced engagement in rewarding activities is associated with increased depressive symptoms (e.g., Vassilopoulos et al., 2017). Furthermore, interventions based on Behavioural Activation (BA), which aim to increase exposure to rewarding experiences, have been shown to effectively alleviate depression, providing additional support for this theoretical framework. This perspective is particularly relevant to research using wearable technology, as passive data on movement, sleep, and heart rate may offer objective markers of behavioural disengagement. Drawing on behavioural activation theory, which posits that reduced engagement in rewarding activities often precedes the onset of depressive symptoms (Jacobson et al., 2001), machine-learning models may be well-suited to detect early warning signs of depression by identifying subtle changes in activity levels and physiological states. Such models could offer a predictive advantage by capturing known prodromal features of depression—such as sleep disturbances

and reduced physical activity—that may emerge before full symptomatic expression (van de Leemput et al., 2014; Cuijpers et al., 2021).

The literature has also established a positive correlation between the duration of untreated depression and symptom severity (Hung et al., 2017). The consequences of untreated MDD can be profound, with suicide being one of the most severe outcomes. Studies indicate that approximately 60% of suicides are linked to MDD (Ng et al., 2017). In 2023, England and Wales recorded 6,069 suicides, equating to an age-standardised mortality rate of 11.4 deaths per 100,000 individuals, the highest rate since 1999. While the relationship between MDD and suicide is complex and multifaceted, studies suggest that individuals with MDD are at a significantly higher risk of suicide compared to the general population. These figures underscore the critical need for improved early detection and intervention strategies for depression to mitigate suicide risks. These findings underscore the urgent need for early identification and intervention to mitigate the risks associated with depression.

Traditional approaches to identifying depression, such as clinical check-ups and self-reported mood diaries, play a crucial role in early intervention. However, these methods are often resource-intensive and difficult to scale effectively (Hopko & Mullane, 2008; Kazdin, 2017). As a result, many individuals experiencing early symptoms of depression do not receive timely support, highlighting the need for innovative, scalable solutions.

1.2.2 The Role of Wearable Devices and Machine Learning

One promising avenue for addressing these challenges is the use of wearable technology and machine learning algorithms. Modern smartwatches, for example, are equipped with sensors that track movement, communication patterns, and other forms of user interaction. The widespread adoption of these devices presents an opportunity for large-scale, passive data collection, with recent figures indicating that 48% of UK adults aged 35–44 own a smartwatch (Laricchia, 2024).

Parallel to the growth of wearable technology, machine learning has seen significant advancements and is increasingly integrated into various domains of healthcare (Jordan & Mitchell, 2015). Machine learning, a subset of artificial intelligence (Bishop, 2006), is particularly useful in analysing complex, non-linear relationships between multiple variables, referred to as features. These features can take various forms, including numerical data, categorical classifications, or more complex representations such as text or images. By associating these features with clinical observations of depression, machine learning models can be trained to detect depressive symptoms based on behavioural patterns alone. Such approaches have the potential to reduce reliance on traditional psychometric assessments and clinical evaluations (Opoku Asare et al., 2021).

1.2.3 The Evidence Base

The application of machine learning in mental health has shown promising results. For instance, Opoku Asare et al. (2021) demonstrated that depression could be predicted with 91.0% accuracy using smartphone sensor data. A systematic review by Haley et al. (2024) further corroborated these findings, highlighting the potential of machine learning models in detecting depression. However, the review also identified critical limitations, including the lack of demographic diversity in training datasets, insufficient external validation, and the absence of models capable of forecasting depression onset rather than simply detecting existing cases.

Current models primarily focus on recognising depression once symptoms have already emerged. However, there is a growing need to develop predictive models that anticipate depressive episodes before they occur. Haley et al. (2024) emphasise the necessity of training models on larger and more diverse datasets to improve their generalizability across different populations. The rationale for this approach is grounded in previous research demonstrating that models trained on homogeneous data perform poorly when applied to novel demographic groups (Dockès et al., 2021). Moreover, predictive models that

can identify individuals at risk of developing depression could enable more efficient and proactive intervention strategies, aligning with the British National Health Service's stepped-care model (Richards et al., 2012).

1.2.4 Aims and Purpose of the Investigation

This study aimed to develop a predictive model for depression using smartwatch-derived data from the RADAR-MDD dataset. Previous research on digital phenotyping for depression has often been limited by homogenous samples, cross-sectional designs, and a focus on detecting rather than forecasting depressive episodes (e.g., Saeb et al., 2016; Rohani et al., 2022). In contrast, this study leverages a diverse, multi-site dataset from the United Kingdom, Spain, and the Netherlands (Matcham et al., 2022), addressing the generalisability issues that have hindered prior models. The study followed 623 participants over two years, capturing both passive behavioral data from smartphone and wearable sensors and active self-reported symptom measures.

By applying machine learning techniques to these longitudinal data streams, this study aimed to develop an algorithm capable of predicting depressive episodes in advance, rather than merely detecting them retrospectively. Given the strong empirical link between behavioral disengagement and worsening depression (Vassilopoulos et al., 2017) and the potential for physiological markers such as sleep disturbances and heart rate variability to indicate early signs of mood deterioration (Ben-Zeev et al., 2015; Teo et al., 2021), integrating these multimodal signals offers a promising path toward real-time, scalable mental health interventions. By improving upon existing limitations in sample diversity, temporal forecasting, and objective behavioral tracking, this study contributes to the development of proactive, data-driven approaches in depression management.

1.3 Methods

1.3.1 Study Design

The study used a large longitudinal data set of an EU research program, RADAR-MDD, which explored the utility of remote measurement technologies in long-term (up to 2 years) depression monitoring (Matcham, et al., 2019). Existing features, including sleep, step count, activity and heartrate were examined as predictors of depression symptom changes, operationalised as an increase of 5 or more points on the PHQ-8. The study received ethical approval from the research ethics committee of the University of Nottingham.

1.3.2 Population

The RADAR-MDD study aimed to detect events of increased depressive symptomatology using fitbit data and estimated that 100 relapses of depression would be required to provide sufficient power for a predictive model with 10 variables (Peduzzi, et al., 1996; Matcham, et al., 2019). It was approximated that 33% of participants would relapse over a year, therefore a minimum of 300 participants would be required for the cohort study. Out of concerns for noisy data and attrition rates the team recruited 623 (Matcham, et al., 2022). The demographics of the population can be seen in Table 1.

Table 1: Socio-demographic and clinical baseline data adapted from the RADAR-MDD study (Matcham, et al., 2022).

	Total Sample
Total, N(%)	623 (100.0)
London, N(%)	350 (56.2)
Barcelona, N(%)	155 (24.9)
Amsterdam, (%)	118 (18.9)

		Total Sample
Socio-demographics		
Age, M (SD)		46.4 (15.3)
Gender, N (%)	Female	471 (75.6)
Marital Status, N (%)	Single/separated/divorced/ widowed	332 (53.3)
	Married/cohabiting/LTR	291 (46.7)
Ethnicity, N(%)	White British/Dutch	369 (78.9)
	White Other	35 (7.5)
	Black ethnic group	14 (3.0)
	Asian ethnic group	16 (3.4)
	Mixed ethnic background	16 (3.4)
	Other	18 (3.9)
Employment Status	Employed/furloughed	260 (41.7)
	Unemployed/sick leave	134 (21.5)
	Student	68 (10.9)
	Retired	123 (19.7)
	Not reported	38 (6.1)
Total years in education, M(SD)		16.4 (6.5)
Benefits Receipt, N(%)	Yes	275 (44.1)

		Total Sample
Accommodation type, N(%)	Own outright/with mortgage	368 (59.1)
	Renting	216 (34.7)
	Living rent-free	29 (4.7)
	Not reported	10 (1.6)
Household income per annum, N(%)	<£/€15,000	154 (24.8)
	£/€15,000 – 55,000	354 (57.0)
	>£€55,000	98 (15.8)
	Prefer not to say	10 (1.6)
	Unknown	5 (0.8)
Clinical Characteristics		
Current depression	IDS-SR total, M(SD)	31.3 (14.5)
	None (0–13), N(%)	61 (10.1)
	Mild (14–25), N(%)	157 (25.9)
	Moderate (26–38), N(%)	206 (33.9)
	Severe (39–48), N(%)	104 (17.1)
	Very severe (49–84), N(%)	79 (13.0)
	Not reported	16 (2.6)
Baseline aRMT PHQ8	PHQ8 total, M(SD)	10.9 (6.0)

		Total Sample
	≥10, N(%)	371 (59.6)

1.3.3 Recruitment

Recruitment for the study was conducted over a period of eighteen months at three international sites, including King's College London (UK), the Netherlands Study of Depression and Anxiety, and other available patient groups at Vrije Universiteit Medisch Centrum (Netherlands), as well as the Centro de Investigacion Biomedica en Red (Spain) (Matcham, et al., 2019). Eligible participants were identified through existing research cohorts or mental health services and contacted by telephone with study information sheets and consent forms sent via email. Eligibility criteria can be seen in Table 2. Enrolment took place either at the research centre or at the participant's home. As an incentive, participants received £15/€20 for enrolling in the study and £5/€10 for every three months of continued participation.

Written consent was obtained during the enrolment session, which included the collection of sociodemographic, social environment, medical history, and technology use questionnaires, as well as baseline data collection of all outcome measures. Subsequently, participants were monitored for a period of 24 months. Upon completion of the study, participants underwent a 60-minute debriefing session, during which their experiences of the study were investigated.

Table 2: Eligibility criteria for participation taken from the RADAR-MDD study (Matcham, et al., 2019).

Inclusion criteria	Exclusion criteria
Meet DSM-5 diagnostic criteria for diagnosis of non-psychotic MDD within the past two years.	Lifetime history of bipolar disorder, schizophrenia, MDD with psychotic features, schizoaffective disorders.
Recurrent MDD (a lifetime history of at least two episodes of depression)	Dementia.
Willing and able to complete self-reported assessments via smartphone.	History of moderate to severe drug or alcohol dependence within the last 6 months.
Able to give informed consent for participation.	History of major medical disease which might impact upon the patient's ability to participate in normal daily activities for more than two weeks (e.g. due to likely hospitalisations or other periods of indisposition). Pregnancy
Fluent in English, Spanish, Catalan or Dutch language.	
Existing ownership of Android smartphone or willingness to use an Android smartphone as their only smartphone.	
Aged 18 or over.	

1.3.5 Data Collection

Passive Data Collection

Each participant in the study installed a passive Remote Measuring Technology (pRMT) application from the RADAR-BASE system for data collection, as described in Matcham et al. (2022). A comprehensive guide to the RADAR-BASE platform for data collection is available in the paper by Ranjan et al. (2019). The pRMT app was connected to their fitbit devices and ran silently in the background of participants' smartphones, collecting data unobtrusively. The passive data collected included sleep, step count, activity and heartrate.

Throughout the process, data was completely anonymized. The researchers appended IDs to each data stream to identify participants. This enabled the research team to contact participants if any data or psychometric information was missing. Access to the visualization dashboard was subject to authentication and authorization. The data used in our analysis is currently stored in a secured communal drive in CSV format. This data has already been pre-processed and is available to members of the Remote Assessment of Disease and Relapse Central Nervous System (RADAR-CNS) consortium, which includes one of the authors of this project.

Active Data Collection

Sociodemographic information was collected during the enrolment procedure of the study (Matcham, et al., 2019). Upon enrolment participants downloaded an active RMT (aRMT) app so that validated measures could be administered remotely to participants at set time intervals. The collection of psychometric data followed the same protocol as the passive data. Numerous metrics were collected during the study for various potential analyses; the primary outcome of our study will be depression severity as measured through the PHQ-8 (Kroenke, et al., 2001).

1.3.6 Primary outcome

The study aimed to investigate reliable deterioration in depression scores, defined as an increase of five points or more on the PHQ-8, a threshold widely recognized in clinical settings as indicating a meaningful worsening of depressive symptoms (Kershaw et al., 2009; Kroenke et al., 2001). Accurately predicting this level of deterioration is clinically valuable, as it can help identify individuals at risk of significant symptom worsening and facilitate timely intervention. The decision to predict reliable change is based on the need to capture the significance of individual changes in depressive symptomology rather than relying on a binary cut-off, such as the presence or absence of depression. By predicting reliable change, the study can identify individuals at risk of deterioration and intervene accordingly.

Bi-weekly administration of the PHQ-8, a validated instrument for measuring depression severity, occurred via the aRMT app in the RADAR-MDD study. The PHQ-8 consists of eight items, each of which corresponds to one of the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V) criteria for major depressive disorder (APA, 2013). The items assess the frequency of specific depressive symptoms over the past two weeks. The PHQ-8 uses a Likert scale to measure the frequency of each symptom, ranging from zero ("not at all") to three ("nearly every day"). The scores for each item are summed to create a total score, which ranges from 0 to 24. Higher scores indicate greater severity of depression. The PHQ-8 can be scored in two ways: as a continuous measure or as a categorical measure. In the categorical measure, specific ranges of scores are used to categorize individuals into different levels of depression severity: minimal (0-4), mild (5-9), moderate (10-14), moderately severe (15-19), and severe (20-24). The PHQ-8 has established reliability, validity, and sensitivity, and is widely used in the NHS's IAPT, making the results of the present study relevant and interpretable to the public (Kroenke, et al., 2001; Levis, et al., 2011). The ease of use of the PHQ-8 via the aRMT app enhances its practicality and accessibility for study participants.

1.3.7 Data Processing[1]

All available data was incorporated into the analysis. To mitigate variations in data availability and scaling, a min-max normalisation technique was implemented, transforming all feature values to a standardized range between 0 and 1. This normalisation is critical in machine learning applications as it prevents features with larger scales from disproportionately influencing the model's performance (Han, Kamber, & Pei, 2011).

To address missing data, imputation using average values was employed. This method was chosen to preserve the structural consistency of the dataset while minimizing the introduction of artificial patterns. Using mean imputation allows for the retention of temporal structure required for sequence-based models, such as the LSTM model utilised in this study, while avoiding the potential distortions that zero-padding may introduce (Che et al., 2018).

The dataset initially exhibited a class imbalance, with 30% of the data representing cases (individuals with depressive episodes) and 70% representing controls. To correct this imbalance, cases were weighted to create a balanced ratio of 1:1. This approach was chosen over other techniques such as SMOTE (Synthetic Minority Over-sampling Technique) due to its simplicity and efficacy in preserving the original data distribution, thus preventing potential biases introduced by synthetic data generation (Chawla, et al., 2002).

1.3.8 Analysis[2, 3]

An LSTM model was selected due to its proven capability to capture temporal dependencies in sequential data, making it particularly suitable for the longitudinal nature of the data collected (Hochreiter & Schmidhuber, 1997). LSTM models are recognized for their capability to effectively manage missing

[1] See 2.1.1 Data Processing for details on the construction of the dataset.
[2] See 2.1.2 Model Selection for a description of the methodology and a comparison of the models evaluated.
[3] See 2.1.3 Logistic Regression for methods on additional analysis comparing the LSTM model with traditional logistic regression.

data within sequences, a frequent challenge encountered in remote monitoring studies (Lipton, Kale, & Wetzel, 2015). However, this robustness typically relies on explicit preprocessing strategies, such as data padding or imputation, to handle missing entries prior to model training. Therefore, although LSTMs demonstrate resilience to incomplete data, careful and transparent handling during preprocessing remains essential for optimal performance.

The dataset was partitioned into three subsets: 70% for training, 15% for validation, and 15% for testing. This split was designed to ensure that the model was trained on a substantial portion of the data while still allowing for validation and testing to prevent overfitting and enhance generalizability.

A feature limit of 300 was established after extensive testing, ensuring that the model could effectively process relevant features without overcomplicating the architecture. The final model architecture incorporated 2048 units and a dropout rate of 0.5 to mitigate overfitting by randomly deactivating a proportion of neurons during training (Srivastava et al., 2014). This configuration was selected following exploratory testing with various unit sizes (32, 64, 256, 1024, and 2048), which yielded similar performance outcomes. These findings suggest that the number of units had minimal impact on model performance. The number of epochs was capped at 1000, with an early stopping mechanism set to halt training after 20 epochs if no improvement was observed in the validation loss, thus safeguarding against overtraining and ensuring optimal model performance.

The performance of the LSTM model was evaluated using multiple metrics, including accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics provided a comprehensive assessment of the model's predictive capabilities, capturing both its overall accuracy and its ability to correctly classify cases versus controls, which is critical in clinical prediction models (Saito & Rehmsmeier, 2015).

1.4 Results[4]

1.4.1 Data Preprocessing and Dataset Description

The dataset used in this study comprised both predicted and actual binary outcomes (0 = negative class, 1 = positive class) structured for input into the LSTM model. A total of 1,975 case-level timepoints were included in the analysis. While approximately 30% of participants experienced at least one depressive deterioration event across the study period, the data were segmented into multiple timepoints per participant. Due to this temporal granularity, the vast majority of individual timepoints reflected periods without deterioration. As a result, 91.4% (n = 1,806) of timepoints were assigned to the negative class (0), indicating no significant change in PHQ-8 score, while only 8.6% (n = 169) were labelled as positive (1), signifying a deterioration defined as a 5-point increase on the PHQ-8 scale. This pronounced imbalance at the timepoint level presented a substantial challenge for the model, which struggled to detect the relatively sparse deterioration events. Descriptive statistics for the dataset are summarised in Table 3.

Table 3: Descriptive Statistics.

Metric	Total (n = 1975)	Negative Class (0)	Positive Class (1)
Count	1975	1806 (91.4%)	169 (8.6%)
Mean Prediction	0.9144	1.0	0.0

1.4.2 Model Performance Metrics [5, 6]

The performance of the LSTM model was evaluated using several key metrics, including accuracy, precision, recall, F1 score, specificity, and AUC-ROC. The model achieved an overall accuracy of 91.44%, indicating a high proportion of

[4] see 2.2.1 Additional Graphs for an extended illustration of results.
[5] See 2.2.2 Comparison of LSTM Models for extended analysis
[6] See 2.2.3 Logistic Regression Comparison for extended analysis

correct classifications. However, this accuracy was likely driven by the class imbalance, as the model demonstrated an inability to correctly classify positive cases. The precision, defined as the proportion of correctly identified positive cases among all predicted positives, was 0.00%, signifying that no positive predictions were made. Similarly, recall (sensitivity) was 0.00%, indicating that the model failed to identify any actual positive cases. The F1 score, which balances precision and recall, was also 0.00, further confirming the absence of correct positive classifications.

Conversely, specificity was 1.00, meaning the model correctly identified all negative cases without misclassification. However, the AUC-ROC score of 0.50 suggests that the model's ability to distinguish between positive and negative cases was no better than random chance, highlighting a critical issue with its predictive capability. The specificity score of 1.00 highlights the model’s correct classification of all negative cases. However, the precision and recall scores of 0.00 demonstrate that it failed to identify any positive cases, underscoring the lack of discriminatory power and bias towards the negative class.

These results highlight a fundamental issue in the model’s ability to distinguish between cases and controls. Future improvements, such as dataset rebalancing, adjusting the decision threshold, or exploring alternative modelling techniques, may be necessary to enhance performance and yield more meaningful predictions.

1.4.3 Implications of Class Imbalance

The substantial class imbalance within the dataset had a marked influence on the performance of the LSTM model. While the high recall and F1 score for the negative class may suggest strong performance, these metrics are misleading given the model’s failure to identify any positive cases. The perfect specificity score and AUC-ROC of 0.50 further confirm the model's inability to effectively distinguish between classes.

1.5 Discussion

1.5.1 Summary of Findings

This study evaluated the feasibility of using passive wearable sensor data to predict future deterioration in depressive symptoms in individuals with MDD. Using an LSTM model trained on data from the RADAR-MDD study, we aimed to detect a reliable worsening of depression, defined as a ≥5-point increase in PHQ-8 scores. Although the model achieved an overall accuracy of 91%, this result was misleading due to the heavily imbalanced dataset, where only 8.6% of timepoints represented deterioration events. Crucially, the model failed to correctly predict any positive cases, with precision, recall, and F1 scores all at 0.00, and an AUC-ROC of 0.50, indicating performance equivalent to random chance. These findings underscore critical challenges in building generalisable, clinically useful machine learning models for mental health prediction[7].

1.5.2 Model Performance and Methodological Considerations

These findings contrast with previous research that has reported high accuracy alongside strong sensitivity and specificity, demonstrating an ability to differentiate between cases and non-cases (Haley et al., 2024). The model's failure to detect deterioration events can be attributed to several interrelated methodological issues. First, class imbalance fundamentally skewed model learning. Despite weighting adjustments during training, the prevalence of negative cases (non-events) led to high specificity but no sensitivity. This imbalance was exacerbated by the timepoint-level segmentation of data, where even participants who experienced deterioration over the study period contributed mostly non-event timepoints.

Second, data sparsity was a major limitation. Although the dataset spanned two years, participant adherence declined markedly over time. On average, participants completed only 40% of scheduled PHQ-8 assessments. Passive

[7] See 2.3.1 Model Performance & Analytical Rigor for extended discussion of analysis.

data availability varied across modalities: step count data had relatively good coverage, while sleep data were often missing, with 78% of participants contributing less than 25% of the expected sleep data. This undermined the temporal continuity and richness necessary for sequence modelling.

Third, the feature engineering process lacked sensitivity to individual baselines. Features such as sleep and activity were standardised across the sample using min-max normalisation and averaged across 24-hour windows. This approach obscures meaningful intra-individual variability. For example, eight hours of sleep may be normative for one participant but signal hypersomnia for another. By failing to capture these personal baselines and directional deviations, the model could not detect the subtle behavioural shifts that often precede depressive relapse (Cornet & Holden, 2018)[8].

Additionally, the chosen operationalisation of depressive deterioration may have influenced outcomes. We defined deterioration as a ≥5-point increase on the PHQ-8, a clinically recognised threshold for reliable change. While valid, this definition introduced sparsity in the outcome variable. A binary diagnostic threshold (e.g., PHQ-8 ≥10) was also considered but abandoned due to insufficient case volume for training. Thus, even exploratory attempts to model symptom presence contemporaneously (rather than prospectively) yielded similarly poor results, suggesting that the issue lay not only in the anticipatory nature of the task but also in data structure and availability.

1.5.3 Cultural and Demographic Heterogeneity

A notable strength of the RADAR-MDD dataset is its multinational composition, comprising participants from the UK, Spain, and the Netherlands. However, this diversity introduces challenges for predictive modelling. Cross-cultural analyses revealed significant differences in PHQ-8 response patterns across sites. Such variation likely reflects underlying cultural differences in symptom expression, mental health literacy, and response styles. This aligns with broader critiques of

[8] See 2.3.2 Limitations of Fitbit Data for a detailed discussion of the data processing phase.

Western psychiatric nosology, which may insufficiently account for how depression manifests in different sociocultural contexts (Kleinman, 1981; Ryder et al., 2002).

These findings suggest that a single, cross-cultural predictive model may struggle to generalise effectively. Although our sample was large in aggregate, there was insufficient statistical power to build and validate separate models for each site. This limitation underscores the need for future studies to adopt stratified or culturally adaptive approaches to modelling, potentially developing site-specific or demographically tailored algorithms[9].

1.5.4 Depression as a Heterogeneous Construct[10, 11].

The limited predictive power of our model also reflects the fundamental complexity of depression as a clinical construct. MDD encompasses a wide range of symptom profiles, severities, and functional impacts (Fried & Nesse, 2015; Zimmerman et al., 2015). Individuals with the same PHQ-8 score may present with entirely different symptom constellations (e.g., anhedonia and fatigue versus guilt and suicidality). Moreover, depression is dynamic and context-sensitive. Even within the same individual, symptom presentation may shift across episodes or in response to life events.

These challenges have led to growing interest in idiographic and person-specific models of mental health. For instance, the WARN-D study (Fried et al., 2023) proposes using dense, individual-level time series data and network analysis to model within-person symptom dynamics and forecast relapse. A static, population-level model based on averaged features is unlikely to capture the nuanced, interactive processes that drive mood changes. Instead, future work should consider hybrid approaches that incorporate both nomothetic trends and idiographic fluctuations.

---

[9] See 3. Cultural Variation in PHQ-8 Responses for additional paper.
[10] See 2.3.3 From data to psychological theory for extended discussion for a broader integration of findings.
[11] See 2.3.4. Depression through a systemic lense for wider concerns around the prediction.

1.5.5 Ethical, Psychological, and Engagement Considerations

Another critical limitation is the potential for selection bias, particularly in relation to participant engagement and data completeness. Individuals who remained compliant with passive and active data collection protocols over extended periods may represent a more motivated, higher-functioning, or health-conscious subgroup of the broader clinical population. As a result, the sample used for model development may be systematically different from the general population of individuals with depression, potentially skewing both the patterns detected and the generalizability of findings (Helgadottir et al., 2021). This bias is especially problematic in digital phenotyping studies, where attrition and differential engagement are known to disproportionately affect more symptomatic or socioeconomically disadvantaged participants (Torous et al., 2016). The resultant dataset may thus underrepresent individuals with more severe, fluctuating, or context-dependent symptom trajectories.

Recent qualitative research (Haley et al., 2025) identified key factors that influence long-term engagement with remote monitoring tools. These include the perceived usefulness of the data, transparency around data use, and trust in the institutions managing the technology. Participants expressed concern about data privacy and the commercialisation of personal health information. Furthermore, a lack of personalised feedback or perceived benefit was cited as a reason for disengagement.

There are also psychological risks associated with passive monitoring. For some individuals, continuous self-tracking can heighten symptom awareness or trigger anxiety. This is particularly problematic in populations with elevated self-critical tendencies, where data may inadvertently reinforce negative cognitive patterns (Harari et al., 2020). It is crucial that future studies incorporate safeguards such as customisable feedback, opt-out mechanisms, and participant consultation to ensure psychological safety.

1.5.6 Implications and Future Directions

The results of this study point to several recommendations for future research and development:

- Adopt person-centred modelling approaches: Future predictive efforts should prioritise individual baselines, behavioural variability, and intra-individual changes over time. Normalised population-level averages are insufficient to capture personal risk trajectories.
- Improve data quality and coverage: Strategies such as real-time engagement tracking, incentivisation, adaptive sampling, and ecological momentary interventions may improve adherence. Passive data quality could be enhanced through improved sensor design and integration with user feedback loops Oetzmann et al. (2022).
- Incorporate cultural adaptation and stratification: Where datasets span multiple countries or demographic groups, stratified modelling or culturally tailored algorithms should be considered. This may involve pre-training models on specific subgroups or incorporating culture-specific behavioural features with later external validation of these models in varied contexts[12].
- Account for construct heterogeneity: Researchers should explore approaches that model depression as a network of interacting symptoms or functional states, rather than a singular latent disorder. Multimodal modelling that includes affect, cognition, and context may better capture the disorder's complexity.
- Prioritise ethical, user-centred design: Engagement can be improved through design strategies that promote trust, transparency, and psychological safety. These may include user dashboards, clear consent processes, and participant co-design methods.

[12] See 2.3.6 External Validation for extended discussion on the topic.

1.5.7 Conclusion

This study highlights the limitations of current population-level, wearable-based machine learning models for forecasting depressive deterioration. Despite leveraging a large, multinational dataset and employing a powerful sequence model, predictive performance was no better than chance. The findings suggest that future progress in this area will require a shift toward person-centred, context-aware, and ethically informed modelling approaches. Addressing issues of data sparsity, class imbalance, cultural variability, and construct heterogeneity is essential for realising the clinical potential of digital mental health tools. Moving forward, interdisciplinary collaboration between clinical psychologists, data scientists, human-computer interaction specialists, and service users will be key to designing systems that are both scientifically valid and practically usable.

1.6 References

American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.). American Psychiatric Publishing.

Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. Psychiatric Rehabilitation Journal, 38(3), 218–226. https://doi.org/10.1037/prj0000130

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. Scientific Reports, 8(1), 6085. https://doi.org/10.1038/s41598-018-24271-9

Cornet, V. P., & Holden, R. J. (2018). Systematic review of smartphone-based passive sensing for health and wellbeing. Journal of Biomedical Informatics, 77, 120–132. https://doi.org/10.1016/j.jbi.2017.12.008

Cuijpers, P., Quero, S., Noma, H., Ciharova, M., Miguel, C., Karyotaki, E., Cipriani, A., Cristea, I. A., & Furukawa, T. A. (2021). Psychotherapies for depression: A network meta-analysis covering efficacy, acceptability and long-term outcomes of all main treatment types. World Psychiatry, 20(2), 283–293. https://doi.org/10.1002/wps.20860

Dockès, J., Varoquaux, G., & Poline, J.-B. (2021). Preventing dataset shift from breaking machine-learning biomarkers. GigaScience, 10(1), giab055. https://doi.org/10.1093/gigascience/giab055

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STARD study. Journal of Affective Disorders, 172, 96–102. https://doi.org/10.1016/j.jad.2014.10.010

Fried, E. I., Proppert, R. K. K., & Rieble, C. L. (2023). Building an Early Warning System for Depression: Rationale, Objectives, and Methods of the WARN-D Study. *Clinical psychology in Europe*, *5*(3), e10075. https://doi.org/10.32872/cpe.10075

Haley, F., Andrews, J., & Moghaddam, N. (2024). Advancements and Limitations: A Systematic Review of Remote-Based Deep Learning Predictive Algorithms for Depression. J. technol. behav. sci. https://doi.org/10.1007/s41347-024-00457-z

Haley, F., Andrews, J., & Moghaddam, N. (2025). Acceptability of Remote Monitoring Technologies for Early Warning of Major Depression. J. technol. behav. sci. https://doi.org/10.1007/s41347-025-00530-1

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2020). *Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges*. Perspectives on Psychological Science, 11(6), 838–854. https://doi.org/10.1177/1745691616650285

Helgadóttir, B., Hallgren, M., Ekblom, Ö., & Forsell, Y. (2016). Training fast or slow? Exercise for depression: A randomized controlled trial. PeerJ, 4, e2321.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hopko, D. R., & Mullane, C. M. (2008). Exploring the relation of depression and overt behavior using daily diaries. Behaviour Research and Therapy, 46(9), 1085–1089. https://doi.org/10.1016/j.brat.2008.05.002

Hung, C.-I., Liu, C.-Y., & Yang, C.-H. (2017). Untreated duration predicted the severity of depression at the two-year follow-up point. PLOS ONE, 12(9), e0185119. https://doi.org/10.1371/journal.pone.0185119

Jacobson, N. S., Martell, C. R., & Dimidjian, S. (2001). Behavioral activation treatment for depression: Returning to contextual roots. Clinical Psychology: Science and Practice, 8(3), 255–270. https://doi.org/10.1093/clipsy.8.3.255

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kazdin, A. E. (2017). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. Behaviour Research and Therapy, 88, 7–18. https://doi.org/10.1016/j.brat.2016.06.004

Kleinman, A. (1981). Patients and healers in the context of culture: An exploration of the borderland between anthropology, medicine, and psychiatry. University of California Press.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. Journal of Affective Disorders, 114(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Laricchia, F. (2024). Number of smartphone users in the United States from 2010 to 2025. Statista. https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/

Levis, B., Benedetti, A., & Thombs, B. D. (2019). Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual

participant data meta-analysis. BMJ, 365, l1476. https://doi.org/10.1136/bmj.l1476

Lewinsohn, P. M. (1974). A behavioral approach to depression. In R. J. Friedman & M. M. Katz (Eds.), *The psychology of depression: Contemporary theory and research.* John Wiley & Sons.

Lipton, Z. C., Kale, D. C., & Wetzel, R. (2016). Modeling missing data in clinical time series with RNNs. In Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), Workshop on Computational Healthcare. https://arxiv.org/abs/1606.04130

Littlewood, R. (2002). Pathologies of the West: An anthropology of mental illness in Europe and America. Cornell University Press.

Matcham, F., Leightley, D., White, K. M., Oetzmann, C., Ivan, A., Lamers, F., Siddi, S., Simblett, S., Rintala, A., Mohr, D. C., Myin-Germeys, I., Wykes, T., Haro, J. M., Penninx, B. W. J. H., Narayan, V. A., Annas, P., Hotopf, M., Dobson, R. J. B., & Folarin, A. A. (2022). Challenges in using mHealth data from smartphones and wearable devices to predict depression symptom severity: Retrospective analysis. medRxiv. https://doi.org/10.1101/2022.12.20.22283760

Matcham, F., Leightley, D., White, K. M., Oetzmann, C., Ivan, A., Lamers, F., Siddi, S., Simblett, S., Rintala, A., Mohr, D. C., Myin-Germeys, I., Wykes, T., Haro, J. M., Pennix, B. W. J. H., Narayan, V. A., Hotopf, M., & Dobson, R. J. B. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. BMC Psychiatry, 19, 72. https://doi.org/10.1186/s12888-019-2049-z

Mullis, R., & Attwell, C. (2022, December 6). Cost of living and depression in adults, Great Britain: 29 September to 23 October 2022. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/m

entalhealth/articles/costoflivinganddepressioninadultsgreatbritain/29septembert o23october2022

Ng, C. W. M., How, C. H., & Ng, Y. P. (2017). Depression in primary care: Assessing suicide risk. Singapore Medical Journal, 58(2), 72–77. https://doi.org/10.11622/smedj.2017006

Oetzmann, C., White, K. M., Ivan, A., Lavelle, G., Simblett, S., Leightley, D., & Matcham, F. (2022). Lessons learned from recruiting into a longitudinal remote measurement study in major depressive disorder. npj Digital Medicine, 5(1), 123. https://doi.org/10.1038/s41746-022-00680-z

Opoku Asare, K., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., & Ferreira, D. (2021). Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: Exploratory study. JMIR mHealth and uHealth, 9(7), e26540. https://doi.org/10.2196/26540

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Ranjan, Y., Rashid, Z., Stewart, C., Conde, P., Begale, M., Verbeeck, D., Boettcher, S., The Hyve, Dobson, R., Folarin, A., & The RADAR-CNS Consortium. (2019). RADAR-base: Open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR mHealth and uHealth, 7(8), e11734. https://doi.org/10.2196/11734

Richards, D. A. (2012). Stepped care: A method to deliver increased access to psychological therapies. The Canadian Journal of Psychiatry, 57(4), 210–215. https://doi.org/10.1177/070674371205700403

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2022). Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: Systematic review. JMIR mHealth and uHealth, 10(1), e14569. https://doi.org/10.2196/14569

Ryder, A. G., Yang, J., & Heine, S. J. (2002). Somatization vs. Psychologization of Emotional Distress: A Paradigmatic Example for Cultural Psychopathology. *Online Readings in Psychology and Culture, 10*(2). https://doi.org/10.9707/2307-0919.1080

Saeb, S., Zhang, M., Kwasny, M. J., Karr, C. J., Kording, K. P., & Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ, 4, e2537. https://doi.org/10.7717/peerj.2537

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Teo, J. T., Newton, P., Hickey, B. A., Lin, C.-T., & Lal, S. (2021). Associations between sleep quality and heart rate variability: Implications for a biological model of stress detection using wearable technology. International Journal of Environmental Research and Public Health, 18(7), 3669. https://doi.org/10.3390/ijerph18073669

Torous, J., Staples, P., Barnett, I., Onnela, J. P., & Keshavan, M. (2018). A crossroad for validating digital tools in schizophrenia and mental health. *NPJ schizophrenia*, *4*(1), 6. https://doi.org/10.1038/s41537-018-0048-6

van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H., Derom, C., Jacobs, N., Kendler, K. S., van der Maas, H. L. J., Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. Proceedings of the National Academy of Sciences, 111(1), 87–92. https://doi.org/10.1073/pnas.1312114110

Vassilopoulos, S. P., Brouzos, A., Moberly, N. J., & Tsiligiannis, G. (2017). Is positive thinking in anticipation of a performance situation better than distraction? An experimental study in preadolescents. Scandinavian Journal of Psychology, 58(6), 510–517. https://doi.org/10.1111/sjop.12355

Wright, A. G. C., & Woods, W. C. (2020). Personalized models of psychopathology. Annual Review of Clinical Psychology, 16, 49–74. https://doi.org/10.1146/annurev-clinpsy-071119-115928

Zimmerman, M., Ellison, W., Young, D., Chelminski, I., & Dalrymple, K. (2015). How many different ways do patients meet the diagnostic criteria for major depressive disorder?. *Comprehensive psychiatry*, *56*, 29–34. https://doi.org/10.1016/j.comppsych.2014.09.007

2. Extended Paper

This extended paper builds upon the accompanying journal article by offering a more comprehensive account of the study's methodological decisions, additional analyses, and expanded theoretical discussion. While the journal article presented the core findings within the constraints of word count and format, this paper provides a fuller exploration of the processes and reasoning behind the research.

A key challenge in the original analysis was the limited quality of the dataset. To address this, a range of analytic strategies were employed to enhance the robustness of the findings. This extended paper includes supplementary analyses not featured in the journal article, aimed at demonstrating how the research team sought to optimise the use of available data and improve the performance of the predictive models.

The discussion section follows the thematic structure of the journal article but is expanded to reflect a broader analytical trajectory. It begins by examining the technical and methodological challenges encountered during the study, before progressing through successive layers of interpretation, culminating in a consideration of the broader societal implications of the technology.

Finally, whereas the journal article operated under the assumption that the PHQ-8 depression scale functions equivalently across different cultural contexts, the additional paper presented here critically examines that assumption. It explores potential cross-cultural variation in responses to the PHQ-8, thereby raising important questions about the universality of this commonly used measure in global mental health research.

2.1 Extended Methods

2.1.1 Data Processing

The dataset for this study was constructed through several stages of data processing to classify participants into two groups: those who experienced an increase in depressive symptoms (cases) and those who did not (non-cases).

Initially, data were collected and stored separately in individual CSV files based on data type. For instance, GPS Homestay data and PHQ-8 questionnaire data were each stored in their own files (see Table 4).

The processing started by calculating the total PHQ-8 scores for each participant at each recorded time point by summing responses across the eight questionnaire items. A clinically significant increase in depressive symptoms was defined as an increase of 5 points or more from the previous PHQ-8 score. Participants meeting this criterion were marked as cases from the first instance of such an increase.

Next, Fitbit data streams were merged based on participant IDs and dates, creating a structured dataset where each row represented a single day's data for one participant, organized chronologically.

This resulted in two primary datasets:

- The PHQ-8 dataset (containing depressive symptom scores)
- The Fitbit dataset (containing wearable device data)

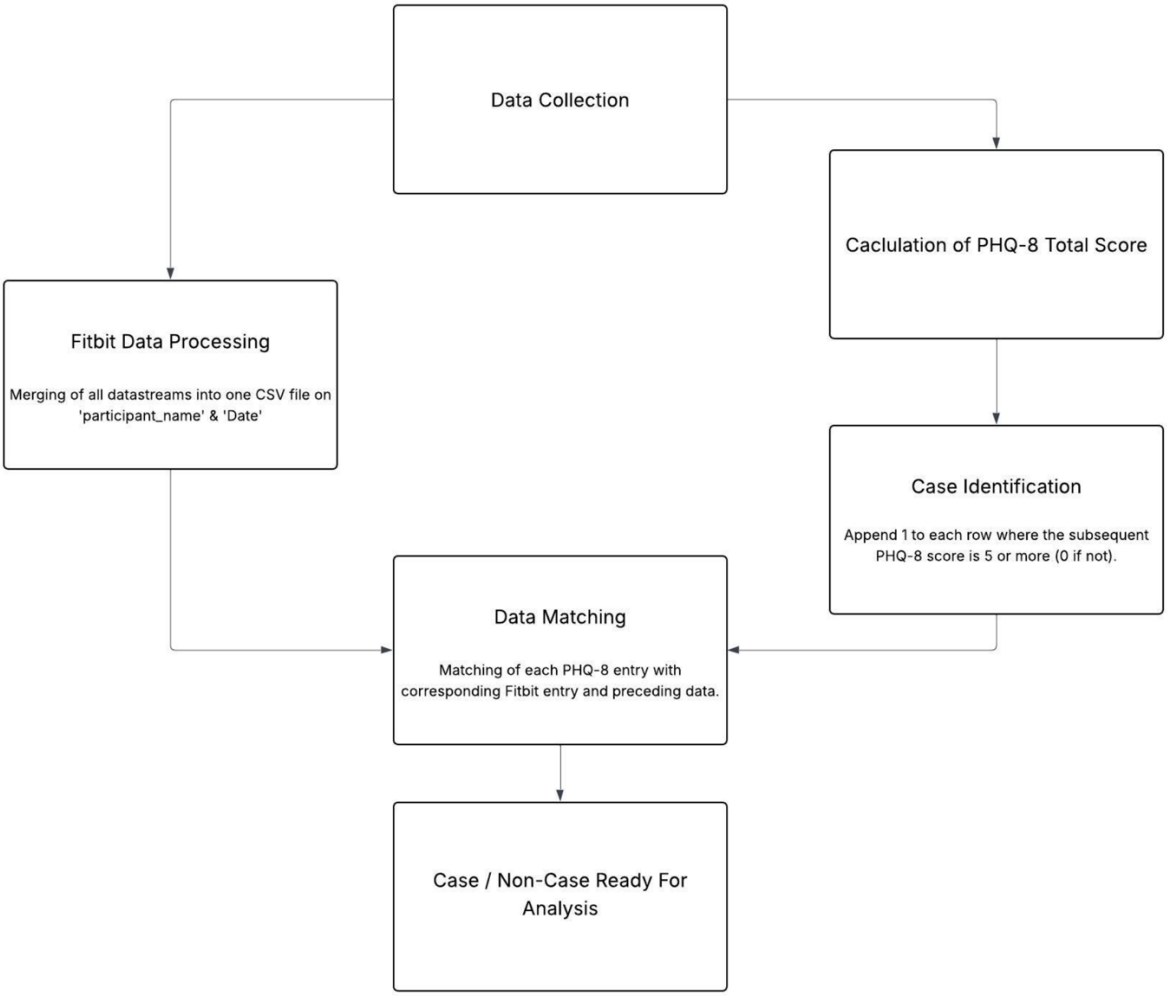
When a participant's PHQ-8 score indicated a case (marked as "1"), the corresponding date and participant ID were matched to the Fitbit dataset. All Fitbit data preceding that point for the participant were then labeled as indicative of a depressive symptom increase. If multiple cases occurred for the same participant, earlier Fitbit data could be reused for subsequent analyses. PHQ-8 entries without an increase (annotated as "0") served as non-case controls for algorithm development.

This approach intentionally allowed flexibility in analyzing the dataset to identify potential predictive signals without preconceived assumptions about the timing of depression onset. If a predictive signal had emerged, further analysis would have refined the dataset by isolating when the signal occurred (e.g., identifying increased sleep activity three weeks before symptom onset). However, due to the absence of such signals, this secondary analytical step was not pursued, concluding the data processing phase.

Table 4: GPS Homestay Data.

dailyFeatures_GPShomestay		
No.	**Column name**	**Short description**
1	file_name	File from which features have been imported
2	participant_name	Foreign key referring to column ‘participant_name’ in table ‘participants’
3	day_timestamp	Timestamp of the day when data was collected
4	day_time	Date on which the data was collected
5	time_interval	Foreign key that points back to the column ‘timeInterval_ID’ in table ‘timeIntervals’
6	homestay_day	How long stayed the person at home?
7	created_at	Date from first database upload
8	updated_at	Date from last data update

Figure 1: Data processing



2.1.2 Model Selection

Data availability posed a significant challenge in this study. This limitation was recognized from the outset, and the primary objective was to strike a balance between maximizing the use of available data and ensuring clinical utility. The decision-making process is illustrated in the decision tree (Figure 2).

Three key factors influenced model selection. The first consideration was the data source. Two passive data streams were available: Fitbit data and smartphone data. While Fitbit data streams were more complete, smartphone data were more clinically relevant due to the widespread use of smartphones. The higher completeness of Fitbit data was likely because participants used the devices solely for monitoring purposes, whereas smartphones also served everyday functions. Prior studies on the same cohort indicated that certain smartphone-based data streams, such as Bluetooth connectivity, had limited availability because participants often disabled Bluetooth or app permissions to conserve battery life (Zhang et al., 2021). Given these constraints, Fitbit data were selected for algorithm development, as they provided the best opportunity for building an accurate predictive model.

The second decision involved defining depressive symptom changes. Three possible approaches were considered: (1) a binary threshold, where a PHQ-8 score above 10 indicated the onset of a depressive episode; (2) an increase of 5 or more points in PHQ-8 scores, which represents a clinically significant change in symptom severity; and (3) an alternative classification based on different cut-off values. While both the binary threshold and the 5-point increase had clinical relevance, the latter was deemed more inclusive. Many individuals experience persistent depressive symptoms without crossing a diagnostic threshold, meaning a jump from 9 to 10 would technically meet criteria for depression but might not reflect a meaningful clinical change. In contrast, a 5-point increase accounts for significant symptom exacerbation regardless of baseline severity. Additionally, this approach had a larger dataset available for model training, whereas the binary threshold approach had fewer than 100 cases—insufficient for algorithm development.

The final decision concerned the timing of prediction: whether to predict depressive episodes before they occurred or at the moment they happened. From a clinical perspective, anticipating symptom deterioration in advance allows for earlier intervention and preventive measures, making this the preferred approach. Although real-time prediction yielded more available data, prioritizing advance detection aligned better with the goal of improving mental health outcomes.

Ultimately, the final model was designed to predict an increase of 5 or more points in PHQ-8 scores, using Fitbit data, and to make predictions ahead of time, meaning the prediction would be made prior to the deterioration between the two PHQ-8 time points. However, to explore the feasibility of alternative approaches, models were also trained on the different classification and prediction strategies. The results of these exploratory models are presented in *section 2.2.2.*

Figure 2: Model Selection

2.1.3 Logistic Regression

To investigate whether the poor performance of the LSTM model was attributable to limitations within the data itself rather than the modelling technique, an additional logistic regression analysis was conducted. Four variables activity, heart rate, step count, and total sleep time were entered into a logistic regression model to assess their predictive value for identifying increases in depression symptoms.

Data Source and Preprocessing

The data used for the logistic regression analysis was derived from the same dataset used for the LSTM model. The original dataset contained multiple daily measurements related to activity, heart rate, step count, and sleep, each with numerous detailed features. Given that logistic regression benefits from a simplified feature set, one representative feature was carefully selected from each category. Each chosen feature provided the clearest summary of daily fluctuations in participant behavior:

- Activity: Total number of active minutes during daytime hours (06:00–23:59).
- Heart Rate: Average daily heart rate across 24 hours.
- Steps: Total daily step count.
- Sleep: Total sleep duration per day, calculated as the sum of all sleep stages excluding wakefulness, recorded in seconds.

Outcome Variable

The binary outcome variable, termed reliable deterioration, was established based on participants' scores from the Patient Health Questionnaire-8 (PHQ-8). Participants experiencing an increase of at least 5 points in their PHQ-8 score were classified as events (positive cases), indicating the potential onset or worsening of depressive symptoms. For each event, the data from the 14 days preceding the PHQ-8 increase were extracted. The average of each

independent variable (activity, heart rate, steps, sleep) was calculated over the 14-day period, forming the predictive dataset for the logistic regression.

A matched control group (non-events) was created by selecting equivalent 14-day periods during which participants showed no significant increase in PHQ-8 scores. These control periods were processed using the same averaging procedure. Both the event and control datasets were combined into a single analytical dataset. Cases containing missing values for any of the variables of interest were excluded prior to analysis.

Statistical Analysis

A logistic regression was conducted to explore whether the selected independent variables—activity, heart rate, steps, and sleep—could predict the onset of depressive episodes (defined as a PHQ-8 increase of ≥5 points). Given the significant imbalance between event cases (fewer cases) and control cases (more numerous), cases were weighted to create a balanced ratio of 1:1. The combined dataset was then randomly divided into a training set (70%) and a test set (30%), stratified by the outcome variable to preserve proportional representation. Model performance was subsequently evaluated on the original, imbalanced test set using several key metrics, including accuracy, precision, recall, F1-score, the confusion matrix, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC). Additionally, model coefficients were examined to assess the relative contribution and significance of each predictor variable.

2.2 Extended Results

2.2.1 Additional Graphs

The graphs presented below were not included in the journal paper, as they do not contribute to the analysis of the null results. However, a description of these results is provided below.

The ROC curve, typically representing the trade-off between the true positive rate (recall) and the false positive rate, follows a diagonal path in this case, resulting in an AUC-ROC value of 0.50. This indicates that the model's performance in distinguishing between classes was no better than random guessing. The Precision-Recall curve, particularly informative in the context of class imbalance, also shows a straight line.

Figure 3: LTSM ROC-Curve

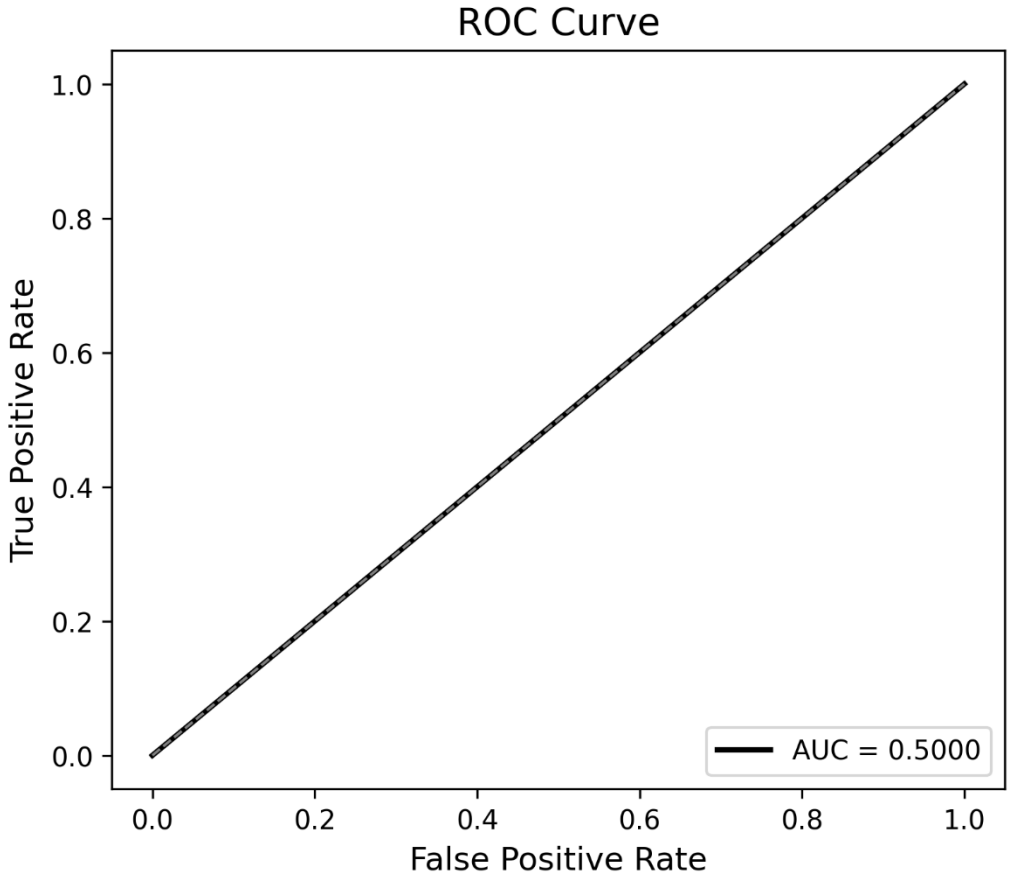
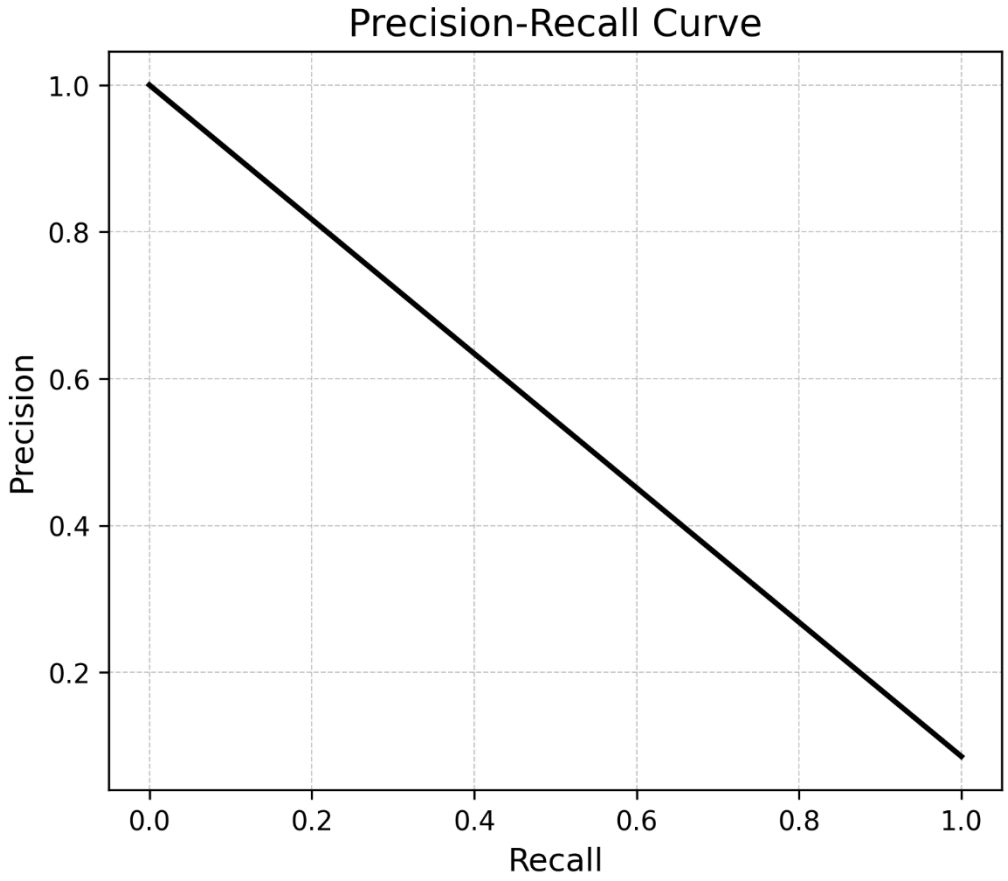


Figure 4: LSTM Precision-Recall Curve



2.2.2 Comparison of LSTM Models

Across all four models, predictive performance was similarly poor. Although the models achieved relatively high accuracy scores and perfect recall , this came at the cost of complete specificity failure. None of the models correctly identified any true negatives, indicating that they predicted all cases as positive. Precision and F1 scores were moderately high, but these metrics were misleading due to the class imbalance and absence of true negative predictions. Overall, all models demonstrated poor prediction power and failed to meaningfully distinguish between classes.

Table 5: LSTM Model Results

Model	Accuracy[a]	Precision	Recall	Specificity	F1	True +[b]	True -[c]	False +	False -
FitBit Prediction	0.92	0.92	1.00	0.00	0.96	4808	0	439	0
FitBit Detection	0.92	0.92	1.00	0.00	0.96	4808	0	439	0
Phone Prediction	0.92	0.92	1.00	0.00	0.96	1975	0	176	0
Phone Detection	0.92	0.92	1.00	0.00	0.96	1975	0	176	0

[a] Not representative of accuracy towards positive cases, due to class imbalance.
[b] + = Positives
[c] - = Negatives

2.2.3 Logistic Regression Comparison

The logistic regression model correctly classified only 48% of cases overall, further highlighting difficulties within the dataset itself. The classification report detailing precision, recall, and F1-scores is summarized in Table 6.

Table 6: Classification Report for Logistic Regression Model

Class	Precision	Recall	F1-Score	Support
0	.93	.47	.62	1410
1	.08	.56	.14	117
Accuracy			.48	1527
Macro Avg.	.50	.52	.38	1527
Weighted Avg.	.86	.48	.59	1527

The ROC AUC score was 0.57, marginally above random guessing (0.50), indicating that the logistic regression model also exhibited limited discriminatory capability. Table 12 presents the logistic regression coefficients for each predictor, revealing negligible effects across the four predictors.

Table 7: Logistic Regression Coefficients for Predicting Depression Symptom Increase

Predictor Variable	Coefficient
Activity	.0012
Heart Rate	-.0148
Step Count	.0000
Total Sleep Time	.0000

When comparing the logistic regression and LSTM models, it becomes evident that the poor predictive capability observed is not unique to the LSTM method. Both models demonstrated substantial limitations in classifying positive cases, underscoring a fundamental difficulty with data predictability rather than inherent

methodological deficiencies. Specifically, while the logistic regression model was marginally better at detecting positive cases (recall of .56 vs. 0 in the LSTM), the low precision (.08) and overall accuracy (.48) further confirmed the challenges inherent in the dataset.

In conclusion, the similarity of poor performance between both analytical approaches reinforces the interpretation that limitations are primarily due to the dataset itself. Future research should therefore consider data collection methodologies, measurement sensitivity, and feature engineering to enhance predictive capabilities and model effectiveness.

2.3 Extended Discussion

2.3.1 Model Performance & Analytical Rigor

The LSTM model achieved an overall classification accuracy of 91.44% but failed to correctly identify any cases of depressive symptom deterioration (precision, recall, F1 score = 0.00). While this discrepancy is initially explained by class imbalance and heterogeneity in symptom expression, a deeper analysis suggests more nuanced analytical limitations. These include potential overfitting or underfitting, limited model calibration, and inadequate handling of data imbalance. Moreover, alternative modelling strategies and interpretability tools were not fully leveraged, which further constrained diagnostic and predictive insights.

Overfitting and Underfitting

The LSTM model likely suffered from a combination of underfitting (insufficient learning of minority class patterns) and overfitting to the majority class. Overfitting is a well-documented challenge in deep learning, particularly when data are noisy or imbalanced (Roelofs et al., 2019). Despite employing dropout regularisation and early stopping, the limited positive cases constrained the model's ability to generalise deterioration patterns.

Hyperparameter tuning was performed via grid search over a limited range due to computational constraints. While this yielded a stable model on the validation set, the absence of recall for the minority class indicates inadequate learning of minority-relevant signal. The lack of temporal signal strength in the features (particularly due to missingness in sleep data) likely exacerbated underfitting in the recurrent layers, which rely on learning time dependencies (Hochreiter & Schmidhuber, 1997).

Additionally, the training loss function used categorical cross-entropy with class weighting, a basic approach that often proves insufficient in highly imbalanced datasets (Lin et al., 2017). More advanced loss functions, such as focal loss, dynamically adjust weightings to penalise hard-to-classify examples more

heavily, and have shown improved performance in skewed datasets across domains (Lin et al., 2017; Khan et al., 2019).

Alternative Analytical Strategies

Although deep learning was the primary modelling approach, a range of alternative strategies could have been considered to improve minority class sensitivity:

- Ensemble Methods: Ensemble models, such as Random Forests or gradient-boosted decision trees (e.g., XGBoost), have demonstrated strong performance in class-imbalanced health datasets (Chen & Guestrin, 2016). Hybrid models combining LSTM feature extraction with tree-based classifiers could exploit both temporal and static information, balancing depth with interpretability (Rajkomar et al., 2018).
- Anomaly Detection: Since depressive symptom deterioration is relatively rare, reframing the problem as an anomaly detection task may be more appropriate. Approaches such as autoencoders or variational autoencoders (VAEs) can learn a compact representation of “normal” behavioural patterns and flag deviations as potential early indicators (Chandola et al., 2009; Xu et al., 2018). This has been applied effectively in wearable sensor data to identify early signals of mental health decline (Barnett et al., 2018).
- Synthetic Minority Oversampling (SMOTE): Though not originally designed for time series, recent adaptations such as SMOTE-ENC or GAN-based oversampling (Douzas & Bacao, 2018) allow for synthetic generation of minority-class time windows. This technique can artificially balance class representation and has improved performance in longitudinal biomedical data (Estabrooks et al., 2004; Wang et al., 2020).
- Cost-Sensitive Learning: Assigning higher costs to false negatives (i.e., missed deteriorations) can drive the model to prioritise recall over precision (Sun et al., 2007). Cost matrices and utility-based evaluation

may better align modelling goals with clinical priorities, where missing a deterioration is more problematic than a false alert.

Feature Importance and Model Interpretability

The high missingness and noise in sleep-related features—due to inconsistent syncing or device non-wear—undermined the model's ability to learn from potentially rich circadian data. An ablation study, in which sleep data were removed from the input sequence, showed no substantial decrease in accuracy, supporting the conclusion that this modality did not contribute meaningfully to the model's learning. This aligns with findings from other wearable studies that note engagement and data quality issues are common barriers to real-world efficacy (Cornet & Holden, 2018).

Importantly, no SHAP (SHapley Additive exPlanations) or Layer-wise Relevance Propagation methods were applied to explain model decisions, as the model did not produce relevant deterioration classifications. In future iterations, these tools could help uncover the model's internal logic and guide feature engineering (Lundberg & Lee, 2017).

The LSTM model's performance limitations stem from more than class imbalance—they reflect challenges in model architecture, and limitations in feature availability and reliability. Future work should consider incorporating ensemble models, anomaly detection, cost-sensitive learning, and SMOTE-based oversampling to address imbalance and improve recall. In parallel, interpretable AI methods should be integrated to identify which features are most clinically meaningful and generalisable. These steps are necessary to advance from proof-of-concept to actionable clinical tools for forecasting depressive deterioration.

2.3.2 Limitations of Fitbit Data

The limited predictive value observed in the wearable metrics used raises important conceptual and methodological concerns about the use of Fitbit-derived data to predict depressive episodes.

Temporal Aggregation and Resolution Loss

One key methodological issue lies in the aggregation of wearable metrics over extended periods (e.g., daily or weekly averages). While this approach simplifies data handling, it may obscure the dynamic, short-term fluctuations that often signal early changes in mood. As Saeb et al. (2015) highlight, momentary deviations in behaviour—such as brief changes in mobility or social interaction—can act as early indicators of mood deterioration. Averaging over longer windows risks masking these subtle but clinically meaningful patterns.

Falkenström et al. (2017) advocate for idiographic, intensive time-series approaches that model within-person variability and capture behavioural anomalies in real time. Supporting this, studies using EMA with high-frequency passive sensing (e.g., Wang et al., 2018) show that even hourly shifts in sleep or movement can correlate with affective instability. These findings suggest that leveraging higher-resolution data could enhance the temporal sensitivity of predictive models.

Temporal Misalignment Between Predictors and Outcomes

Another limitation is the temporal mismatch between predictor and outcome variables. In the present study, Fitbit metrics are averaged daily, while depressive symptoms (measured by the PHQ-8) are assessed biweekly. This misalignment may weaken predictive power, as behavioural precursors to depression may not align neatly with broader, arbitrary measurement intervals. Evidence from Wang et al. (2014) and Barnett et al. (2018) demonstrates improved model performance when behavioural data are temporally synchronised with symptom reporting, particularly when using time-lagged features and frequent EMA-based assessments.

Group-Level Modelling and Individual Variability

Traditional group-level modelling approaches often fail to account for individual baselines, conflating trait-level characteristics (e.g., chronically low physical

activity) with state-level changes that may indicate mental health deterioration. This limitation has been highlighted in several digital phenotyping studies (Torous et al., 2021). In contrast, person-centred approaches—such as control-chart modelling and anomaly detection (Wang et al., 2020)—offer more nuanced detection of within-person deviations, which may be more predictive of state transitions into depressive episodes.

Incorporating baseline-adjusted models and change-point detection methods could improve early identification of depressive episodes by accounting for normative behavioural patterns and their deviations (Huckins et al., 2020).

Contextual Modelling and Composite Indicators

Beyond methodological refinements, future predictive models could benefit from integrating multiple passive metrics into composite indicators that reflect behavioural context. Interactional models, which examine how passive signals (e.g., reduced activity) interact with psychosocial or environmental variables, align with functional-analytic approaches in clinical psychology (Hayes et al., 2012). These models recognise that the same behavioural marker may have different clinical meanings depending on context—a concept further explored in Section 2.3.3.

Integrating Passive and Active Sensing

To address these limitations, hybrid models that combine passive sensing with active data collection methods like EMA offer a promising path forward. EMA enables frequent, real-time reporting of mood, cognition, and context, thereby complementing passive data streams and enhancing interpretability. As shown by Trull and Ebner-Priemer (2020), multimodal models that integrate both data types tend to outperform those relying on a single modality. This integration offers a richer and more temporally precise understanding of how depressive symptoms emerge and evolve over time and in context.

2.3.3 From data to psychological theory

This study was grounded in behavioural models of depression, particularly Lewinsohn's theory (1974), which posits that depression arises from a reduction in positively reinforcing activities. From this perspective, individuals who experience fewer rewarding events gradually reduce their engagement with the environment, leading to a downward spiral of decreased activity, social withdrawal, and worsening mood. This behavioural chain is particularly relevant when considering the use of RMTs, such as wearable sensors, to detect early warning signs of depressive deterioration. The expectation was that reductions in activity as reflected in objective data from devices like Fitbits could signal a shift toward depressive states.

Lewinsohn's behavioural model rests on a foundational assumption within behavioural psychology: that emotional states are shaped by the interaction between environment and behaviour, with emotional responses emerging downstream from these processes (Ferster, 1973; Martell et al., 2001). This view aligns with operant learning theory, suggesting that a lack of environmental reinforcement leads to behavioural disengagement, which in turn fuels low mood. However, this behavioural primacy is situated within a broader theoretical debate in psychology about the causal ordering of emotion, cognition, and behaviour.

Historically, the James-Lange theory argued that emotions follow physiological changes, while the Cannon-Bard theory proposed simultaneous emotional and physiological responses (Cannon, 1927; James, 1884). Cognitive models such as the Schachter-Singer two-factor theory and Lazarus's appraisal theory introduced the idea that cognitive interpretations play a central role in generating emotional responses (Schachter & Singer, 1962; Lazarus, 1991). Contemporary models increasingly recognise the bidirectional interplay between emotion and cognition. Emotion is now understood as both shaping and being shaped by cognitive appraisal, contextual learning, and predictive processing (Barrett, 2017; Pessoa, 2008). This is consistent with neurobiological evidence

indicating that brain regions such as the prefrontal cortex and amygdala are involved in both cognitive evaluation and affective regulation (Etkin et al., 2015).

Cognitive models of depression, particularly Beck's cognitive theory, reinforce this interplay by emphasising the role of distorted thought patterns and core beliefs in the development and maintenance of depression (Beck, 1967; Beck et al., 1979). According to this view, emotional distress arises not from life events per se, but from how individuals interpret those events. Core beliefs, shaped by early experiences, influence how people attend to, interpret, and remember information about themselves and the world (Beck, 2011). These cognitive distortions give rise to maladaptive coping strategies, which may temporarily reduce distress but ultimately reinforce depressive thinking and behaviour (Westbrook et al., 2007; Moghaddam & Dawson, 2015).

From this cognitive-behavioural perspective, interpreting passively sensed data from RMTs without accounting for individual appraisal processes presents a significant limitation highlighting the point made in section 2.3.2. Reductions in physical activity may indeed occur, but their psychological significance cannot be fully understood without knowing how an individual perceives and interprets their situation. Emotional and behavioural responses to life events are inherently context-dependent, and coping strategies vary greatly between individuals marking the importance of person-centered approaches (Folkman & Moskowitz, 2004).

As discussed in Section 2.3.2, hybrid models that integrate passive sensing with active methods such as EMA, mood ratings, or context-rich self-report tools offer a pathway forward. From a theoretical standpoint, such integration does more than improve prediction—it provides a testbed for examining how behavioural, cognitive, and emotional processes interact over time within real-world contexts. For example, if a model can demonstrate that reductions in activity *combined with* negative self-appraisal predict depressive worsening better than either alone, this lends empirical support to co-constructive and dynamic theories of emotion (Scherer, 2009).

2.3.4 Depression through a systemic lense

The concept of predicting depression prior to its clinical onset remains inconsistently defined in the literature, with considerable conceptual and methodological ambiguity (Haley et al., 2024). Most contemporary research in this domain tends to focus on real-time detection of depressive episodes, intervening once symptoms are already manifest, rather than on genuine prediction, which would entail identifying individuals at heightened risk before symptoms arise. This distinction is critical: while real-time detection reflects a reactive model of healthcare, true prediction aligns with a proactive paradigm, enabling timely interventions that might mitigate or prevent the development of MDD.

Early intervention in depression has consistently been associated with better prognostic outcomes (Cuijpers et al., 2008; Kessler et al., 2007). Consequently, the development of robust predictive methodologies has the potential to revolutionize mental healthcare by shifting the emphasis from treatment to prevention. Such a shift not only promises improved individual outcomes but also aligns with broader policy objectives. For example, the NHS Long Term Plan advocates for a preventative, stepped-care model in mental health services, whereby the intensity of intervention is matched to symptom severity, with a strong emphasis on early identification (NHS England, 2019).

The economic rationale for proactive mental healthcare is also compelling. Layard (2006) argued that untreated mental health problems impose substantial costs on both healthcare systems and broader society. These claims have since been corroborated by data from the UK Centre for Mental Health, which estimates that mental ill-health, including depression and anxiety, costs the economy over £105 billion annually, once lost productivity, healthcare expenditure, and social costs are taken into account (Centre for Mental Health, 2020). Notably, common mental disorders such as depression are primary contributors to this economic burden, resulting in high rates of sickness absence, presenteeism, and workforce attrition. The Health and Safety Executive (HSE, 2022) reported that stress, depression, or anxiety were

responsible for over 50% of all work-related ill health in the UK. Moreover, untreated depression is associated with chronic impairment and an increased risk of physical comorbidities, further escalating healthcare costs and compounding cross-sectoral burdens (OECD, 2018). These findings lend support to Layard's proposition that investing in evidence-based psychological interventions yields high economic returns, reducing demand on services while enhancing societal productivity.

Despite these arguments, existing technologies for depression prediction remain underdeveloped. Most studies rely on retrospective self-report measures such as the PHQ-8 (Kroenke et al., 2009), which offer only coarse-grained data relative to the granularity of continuous physiological input from wearable sensors. While clinically valuable, these instruments are ill-suited to identifying the nuanced temporal dynamics preceding symptom onset (Fried & Nesse, 2015). For example, in this study, Fitbit data were collected daily, whereas the PHQ-8 was administered biweekly. This temporal misalignment introduces interpretative ambiguity for predictive models, which must infer depressive transitions without precisely aligned outcome labels (Jacobson & Chung, 2020). Furthermore, there is evidence to show that self-report distress measures has little influence on responses, measuring a mix of current mood-state and response style as oppose to meaningful recall (Sato et al., 2011). This further supports the argument to incorporate additional datastreams into the development of predictive models, since most psychometrics are subject to recall bias, and do not provide insight into moment to moment changes within an individual (Schmier & Halpurn, 2004).

Another underexplored obstacle in predictive modeling is the influence of interpersonal and cultural variability in the experience and expression of depression. As Kleinman (1981) and more recently Adler et al. (2024) argue, depression is not a culturally neutral phenomenon. Sociocultural context shapes symptom expression, illness narratives, and help-seeking behaviors. Predictors validated in one cultural milieu may prove ineffective or misleading in another,

complicating the development of universally applicable models. This issue is explored further in the additional paper of the thesis found in Section 3.

While technological advances in wearable sensors and machine learning models hold considerable promise for enhancing early identification of mental health difficulties, they must be situated within a broader structural and sociocultural framework. Depression is not merely a neurobiological or individual phenomenon, it is deeply embedded in social and economic conditions (Pilgrim, 2019). Indeed, a substantial body of evidence from public health and social epidemiology emphasizes the pivotal role of socioeconomic inequality in shaping population mental health.

Wilkinson and Pickett (2009), in *The Spirit Level*, present compelling cross-national evidence linking higher income inequality with elevated rates of mental illness, including depression. These effects are not confined to the materially deprived but span the socioeconomic gradient, highlighting the psychosocial stress associated with relative status, diminished social cohesion, and eroded collective well-being (Pickett & Wilkinson, 2015; Patel et al., 2018). Inequality thus functions as a societal stressor, undermining protective psychosocial factors such as trust, belonging, and perceived control (Kawachi & Berkman, 2001).

Relative deprivation, a subjective sense of being worse off than others, has been implicated as a key psychosocial mechanism in the development of depression. In highly individualistic societies, where success is equated with personal merit, failure is often internalised as individual inadequacy, reinforcing cognitive distortions such as hopelessness, self-blame, and worthlessness (Marmot, 2004; Beck, 1967; Layte, 2012). These psychological dynamics are deeply intertwined with the structural conditions of inequality.

Childhood experiences further entrench vulnerability. Exposure to economic insecurity, parental stress, and emotionally unsupportive environments, conditions disproportionately borne by lower-income families, are robust predictors of later mental health problems (Lund et al., 2010; Hughes et al.,

2017). Educational systems that prioritize academic attainment over emotional development exacerbate these risks, contributing to a cumulative psychosocial burden across the lifespan (Ecclestone & Hayes, 2009).

Moreover, consumer capitalism, particularly when overlaid with socioeconomic disparity, reinforces maladaptive coping strategies. Individuals are encouraged to seek self-worth through consumption and status acquisition, only to experience dissatisfaction when unable to meet ever-escalating societal standards (Kasser, 2002). This cycle is especially detrimental to young people, who navigate identity formation within highly performative, image-saturated environments (Twenge et al., 2018).

Against this backdrop, the prevailing biomedical framing of depression, as a neurochemical imbalance or personal deficit, appears increasingly insufficient. Such framings risk pathologizing rational responses to systemic adversity, diverting attention from structural determinants such as austerity, insecure employment, and housing precarity (Rose, 2019; Friedli, 2009; White et al., 2017). While predictive technologies may enhance individual-level identification, they also risk reinforcing an individualised discourse that obscures the social roots of mental distress (van Os et al., 2019). Indeed, by emphasizing personal monitoring and self-regulation, these tools may inadvertently intensify individual responsibility for managing distress while ignoring the broader context in which suffering arises (Morozov, 2013).

To achieve meaningful prevention, mental health strategies must extend beyond prediction and detection. Upstream interventions that address the social determinants of mental health, such as reducing inequality, fostering social inclusion, and ensuring access to supportive environments, are essential. Interventions grounded in community engagement, equity, and psychological safety have demonstrated long-term benefits for population mental health (WHO, 2014; Allen et al., 2014). Without structural reform, technological innovations risk becoming sophisticated tools for symptom management, rather than instruments of genuine prevention.

2.3.5 Implications for clinical practice

The integration of predictive models derived from wearable sensor data into routine mental health services such as IAPT or broader stepped care frameworks represents a major frontier in digital mental health. However, this integration must be approached with caution, ethical foresight, and alignment with clinical values and stakeholder expectations.

Ethical Integration into Stepped Care Models

Stepped care models aim to allocate interventions according to clinical need, beginning with the least intensive intervention and stepping up only as necessary (Bower & Gilbody, 2005; Clark, 2011). Within IAPT, this involves progression from low-intensity guided self-help to high-intensity cognitive behavioural therapy or pharmacological treatment based on clinical presentation and risk. Predictive models that detect subtle patterns of symptom deterioration from wearable sensors could theoretically enhance this framework by identifying at-risk individuals earlier and prompting timely escalation or review.

However, the integration of such predictive systems into stepped care raises ethical and procedural questions. Firstly, digital tools must not override or replace clinical judgement. Rather, they should augment the formulation process by providing additional data points that are considered alongside patient narratives, standardised measures, and therapeutic formulations. As Topol (2019) argue, the use of AI in healthcare should preserve and enhance the clinician–patient relationship, not undermine it. This requires clear governance mechanisms, human-in-the-loop systems, and transparency regarding how predictions are generated, interpreted, and actioned. Within IAPT, clinicians may require specific training to interpret passive sensor data and critically assess its relevance within each patient's biopsychosocial context.

Additionally, consent and data transparency are critical. Patients must be informed not only about what data is being collected but also how it will be

used, shared, and potentially acted upon. Ethical integration mandates mechanisms for ongoing consent, opt-out options, and feedback loops that enable patients to challenge or contextualise data-derived insights. Without such provisions, predictive systems risk reinforcing perceptions of surveillance and disempowerment, particularly among individuals already experiencing low mood, anhedonia, or cognitive distortions.

Clinical Risk

Despite their potential, predictive models are inherently probabilistic and imperfect. A major clinical risk lies in false positives, instances where a model inaccurately predicts deterioration, and false negatives, instances where early signs of relapse are missed. In the context of stepped care, a false positive could lead to inappropriate escalation of treatment, unnecessary clinician burden, or anxiety for the patient. Conversely, a false negative may result in missed opportunities for early intervention, prolonged suffering, or avoidable crises. Once more it is important to bear in mind how this sets an expectation on the individual and further re-iterates how the individual is deffective without acknowledging the limitations of the system within which they find themselves.

In clinical terms, such misclassifications have tangible consequences. A flagged deterioration may trigger clinical review or stepped-up intervention, diverting resources from those in more urgent need. Meanwhile, missed detections may delay critical care or increase the risk of relapse. Thus, careful consideration must be given to how model thresholds are set, how predictions are interpreted, and what clinical responses are appropriate. This reinforces the need for ongoing model validation within live service environments and transparent decision-making frameworks that allow clinicians to override or contextualise alerts (to be discussed further in section 2.3.6).

User Perspectives

The successful deployment of predictive models derived from wearable sensor data within clinical services hinges not only on algorithmic accuracy or clinical

utility, but also on their acceptability to end-users and alignment with user values, expectations, and experiences. Recent qualitative work by Haley et al. (2025) offers an important contribution to this area, drawing from service users enrolled in the RADAR-MDD. Their insights illuminate core design and ethical considerations for embedding RMTs into mental health care, and these findings can be conceptually organised using the Behavioral Intervention Technology (BIT) Model (Mohr et al., 2014).

The BIT Model offers a framework for understanding the design and implementation of technology-based interventions, comprising five elements: (1) the clinical aims (why), (2) the behavioural strategies (how), (3) the elements and features (what), (4) the technical workflow (when), and (5) the technological instantiation (how delivered). Haley et al.'s findings align closely with each of these components, providing practical insights for digital tool development that resonates with both therapeutic theory and user expectations.

Clinical Aims (Why): Empowering Self-Management and Early Intervention

Haley et al. (2025) reported that participants perceived RMTs as acceptable primarily when they supported autonomy, enhanced self-awareness, and enabled early intervention. These aims reflect users' preferences for tools that act not merely as passive monitors, but as collaborative partners in care. Participants described RMTs as "empowering," particularly when data visualisation helped them understand patterns of low mood, sleep disruption, or behavioural withdrawal. This aligns with the BIT model's emphasis on specifying therapeutic targets, here, behavioural activation and self-regulation, as central goals for intervention design.

Behavioural Strategies (How): Personalised Feedback and Supportive Prompts

Participants emphasised the importance of personalised, context-sensitive feedback, noting that RMTs were more effective and engaging when they mirrored therapeutic strategies such as behavioural activation or problem-solving. Alerts that posed reflective questions ("What can you do to take care of

yourself today?") rather than deterministic predictions ("You are at risk of relapse") were preferred. This preference maps directly onto the behavioural strategies component of the BIT model, advocating for the use of motivational prompts and psychoeducation rather than algorithmic determinism.

Importantly, users articulated a strong aversion to alarmist messaging that might induce anxiety or create self-fulfilling prophecies. These findings underscore the need for RMTs to be designed with therapeutic sensitivity, echoing principles from Just-In-Time Adaptive Interventions (JITAIs), which tailor delivery to the user's momentary state (Nahum-Shani et al., 2018).

Elements and Features (What): Notifications, Insights, and Human Connection

The core technological elements identified as valuable by Haley et al. (2025) include tailored notifications, self-tracking summaries, and optional social connectivity features (e.g. involving named others in the alert process). These features correspond to the BIT model's "intervention elements" and underscore the importance of balancing automation with human connection. Several participants expressed a desire for RMTs to prompt social check-ins or share warnings with trusted individuals, enabling relational support in moments of vulnerability.

However, users also warned against excessive notifications, which could feel intrusive or burdensome, particularly during low-energy periods. This insight reinforces the value of customisable intensity levels for alerts and notifications—an implementation feature that should be made adjustable at the user level.

Workflow and Timing (When): Passive Sensing with Minimal Burden

Users consistently preferred passive data collection over active self-report, citing that frequent questionnaire prompts diminished over time in acceptability. As one participant observed, "It's fun for a few weeks, but then it slides." The BIT model's "workflow" component, concerned with when and how interventions are triggered, can benefit from such insights. Effective systems may need to

use passive data to identify optimal windows for active engagement—minimising intrusiveness while maximising contextual relevance. This workflow adaptiveness is also central to maintaining engagement over time, a challenge widely acknowledged in digital mental health interventions (Simblett et al., 2019).

Technological Delivery (How): Empathy-Mimicking Interfaces and Ethical Data Practices

The technological form of RMTs—how they are perceived and experienced—was a recurring theme. Participants responded more favourably to systems that simulated empathy through tone, language, and visual design. Haley et al. (2025) note that participants explicitly preferred messages that mimicked supportive, therapeutic language, aligning with research on the digital therapeutic alliance (Lederman & D'Alfonso, 2021). This affirms the BIT model's attention to "instantiation" and the importance of human-centred interface design.

Ethical considerations further shaped perceptions of acceptability. While most participants initially reported few data concerns, deeper probing revealed widespread unease about commercial use or data sharing with insurers. Conversely, participants expressed willingness to share data for altruistic purposes—such as research or peer benefit—highlighting the value of purpose-driven design and transparent governance. These findings support calls from the literature (Huckvale et al., 2019; Peppin, 2022) for privacy-preserving architectures and user control over data flows.

2.3.6 External Validation

Machine learning derives its predictive power from patterns extracted from observed data, enabling reasoning, classification, and decision-making (Mitchell, 1997). However, without rigorous external validation, ML models risk overfitting to specific datasets and failing to generalize effectively across diverse populations (Hastie, Tibshirani, & Friedman, 2009).

Data, although often conceptualized as objective, are not neutral representations of reality. Rather, they are filtered constructs shaped by measurement tools, sampling methods, and human decisions (Floridi, 2011). Philosophically, data analysis has roots in positivism, which assumes an objective reality amenable to systematic empirical study (Popper, 1959). Nevertheless, the assumptions embedded within data collection processes introduce inherent biases that, if unexamined, can propagate through ML models.

The power of machine learning is fundamentally constrained by the quality and representativeness of its training data. Models trained on biased datasets are likely to reinforce those biases rather than ameliorate them (Barocas, Hardt, & Narayanan, 2019). A well-documented example is the COMPAS algorithm used to predict criminal recidivism, which was shown to systematically overestimate risk scores for Black defendants compared to white defendants with similar criminal histories (Angwin et al., 2016). This case highlights the danger of assuming model neutrality when underlying data reflect historical and systemic biases.

In healthcare, similar risks are apparent. Models developed to predict depression using passive sensor data, are often trained on clinical populations that may not reflect broader demographic diversity. In the UK, for instance, middle-aged white women are statistically more likely to access psychological services compared to young Black men (Baker, 2020). Consequently, predictive models disproportionately trained on the former group may exhibit reduced sensitivity and specificity when applied to underrepresented populations, thus

exacerbating existing disparities in mental health outcomes (McDonald et al., 2017).

External validation, the process of evaluating a model's performance on independent, previously unseen datasets, is therefore essential to establishing generalizability and fairness (Varoquaux, 2018). Without such validation, models are vulnerable to overfitting, whereby they perform well on training data but poorly in novel settings (Dietterich, 1995). External validation enables researchers to systematically test the robustness of models, identify and correct biases, and enhance the models' applicability to diverse populations (Gebru et al., 2018).

Moreover, external validation promotes model adaptability. A system trained exclusively on one demographic context may perform inadequately when deployed in different environments, with potentially serious ethical and practical consequences (Buolamwini & Gebru, 2018). By testing models across varied demographic, geographic, and socioeconomic groups, researchers can refine algorithms to ensure greater predictive accuracy and equitable application (Suresh & Guttag, 2021).

While machine learning offers significant potential for advancing predictive mental health interventions, its reliability, fairness, and clinical utility are contingent on rigorous, continuous external validation. Acknowledging the inherent limitations of training data and actively addressing biases during model development and evaluation are critical. Only through such methodological vigilance can machine learning technologies fulfill their promise of supporting ethical, effective, and inclusive outcomes across healthcare and other societal domains.

2.4 Reflections

The development of my thesis project was not a spontaneous endeavour but rather the culmination of a long-standing intellectual trajectory that began during my undergraduate studies. It emerged at the intersection of two foundational interests: the transformative potential of AI and predictive modelling in mental health care (Shatte, Hutchinson & Teague, 2019), and a broader ethical commitment to contributing meaningfully to public wellbeing. Initially, my professional engagement within an IAPT service was shaped by a biomedical understanding of mental illness. Within this framework, depression was conceptualised as a discrete, diagnosable condition, akin to somatic illnesses such as diabetes or cancer, consistent with the dominant nosological approach of the DSM-5 (APA, 2013).

This early clinical context reinforced a positivist epistemology, characterised by the assumption that psychological disorders exist independently of the observer and can be measured objectively (Harper & Speed, 2012). My initial research assumptions mirrored this stance: I believed that depression could be detected through observable behavioural and physiological markers, and predicted using statistical and algorithmic models grounded in machine learning. The idea that data-driven tools could offer 'objective' insights into mental health seemed both plausible and promising at the time, reflecting a broader trend within digital psychiatry (Insel, 2017).

However, my epistemological stance began to shift through the course of my Doctorate in Clinical Psychology. Immersion in diverse clinical settings and exposure to varied conceptual frameworks challenged the rigidity of my earlier assumptions. One pivotal experience involved working with individuals diagnosed with personality disorders. These constructs, while codified in diagnostic systems such as the DSM-5 and ICD-11, have long been critiqued for their conceptual instability and cultural embeddedness (Kendler, 2016). Historical shifts in terminology, from ‘hysteria’ to ‘borderline personality disorder’, for example, underscore the malleability of psychiatric categories in

response to evolving societal norms and professional discourse (Becker, 1997; Paris, 2020). In practice, many clients' difficulties could be better understood as responses to complex trauma rather than as evidence of stable, internal pathology, aligning with trauma-informed frameworks that challenge traditional diagnostic interpretations (Herman, 1992).

This realisation highlighted the socially constructed nature of diagnosis: psychiatric labels are not neutral descriptors of internal dysfunction, but rather products of cultural, historical, and relational processes (Bracken et al., 2012). They reflect power-laden judgments about which behaviours are deemed acceptable or pathological in a given social context.

My evolving perspective was further deepened through work with autistic individuals. Traditionally positioned within the medical model as a disorder requiring intervention, autism has increasingly been reinterpreted through the lens of neurodiversity, a paradigm that affirms cognitive difference as a natural and valuable aspect of human variation (Kapp et al., 2013). This shift draws from the social model of disability, which posits that disability arises not from individual impairments but from societal failure to accommodate difference (Shakespeare, 2013). From this standpoint, the act of diagnosis itself can marginalise individuals by framing variance from neurotypical norms as inherently deficient, again reinforcing the contingent, non-objective nature of psychiatric categorisation.

Another significant turning point occurred during academic inquiry into cross-cultural differences in psychosis. Research has shown that experiences such as auditory hallucinations or delusional beliefs, pathologised within Western psychiatric frameworks, are interpreted very differently across cultural contexts (Luhrmann et al., 2015). In some non-Western societies, such phenomena may be understood as spiritual encounters rather than symptoms of schizophrenia, and may not carry the same distress or stigma (Kirmayer & Swartz, 2013). These findings underscore the cultural relativity of psychiatric diagnosis and the dangers of universalising Western models of mental illness (Watters, 2010).

Even foundational terms in clinical psychology, such as ‘functioning’ or ‘normality’, are shaped by sociocultural expectations. As scholars of critical psychology have argued, concepts of ‘mental health’ often encode normative assumptions about productivity, emotional regulation, and social behaviour (Rose, 1999; Boyle, 2022). What is considered 'functional' is context-dependent, informed by economic structures, cultural values, and policy imperatives, further destabilising the idea that mental health constructs are natural or self-evident.

These epistemological shifts directly impacted the design and interpretation of my thesis. Initially conceived as a study of predictive modelling using wearable technologies to detect episodes of depression, the project gradually evolved into a critical engagement with the assumptions underpinning such models. While quantitative techniques such as factor analysis can demonstrate internal consistency and construct validity within measurement instruments (Costello & Osborne, 2005), they do not resolve the underlying issue: that depression, as a diagnostic category, is itself shaped by sociocultural forces. Research has demonstrated significant cross-cultural variation in both the symptomatology and expression of depression (Kleinman, 1987; Ryder et al., 2008), calling into question the universality of existing measures like the PHQ-9 or HAM-D.

Moreover, even ostensibly objective data, such as daily activity tracked via wearable sensors, are interpreted through the lens of subjective experience. Individuals’ willingness and ability to engage with technology, their interpretation of mood and behaviour, and their understanding of mental health all influence the data that is ultimately captured (Mohr et al., 2017). These insights challenge the assumption that digital biomarkers offer a neutral window into psychopathology. Instead, they highlight the need for contextualised, participatory approaches that foreground lived experience and cultural meaning.

In this light, my project became not only a study of depression prediction but also a reflexive exploration of the epistemological tensions within clinical psychology and mental health research. I began by approaching depression as

a static, measurable construct, but concluded with a more nuanced understanding: that psychological distress is irreducibly shaped by culture, context, and interpersonal meaning. This journey reflects a broader movement within clinical psychology toward integrative, pluralistic, and critical approaches that resist reductive classification and embrace complexity (Johnstone & Boyle, 2018).

2.5 References

Adler, D. A., Stamatis, C. A., Meyerhoff, J., Mohr, D. C., Wang, F., Aranovich, G. J., Sen, S., & Choudhury, T. (2024). Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. npj Mental Health Research, 3(17). https://doi.org/10.1038/s44184-024-00057-y

Allen, J., Balfour, R., Bell, R., & Marmot, M. (2014). Social determinants of mental health. *International review of psychiatry (Abingdon, England)*, *26*(4), 392–407. https://doi.org/10.3109/09540261.2014.928270

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Publishing.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica.

Baker, C. (2020). Mental health statistics for England: Prevalence, services and funding. House of Commons Library.

Barnett, I., Torous, J., Staples, P., Sandoval, L., Keshavan, M., & Onnela, J. P. (2018). Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology, 43(8), 1660–1666. https://doi.org/10.1038/s41386-018-0030-z

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. fairmlbook.org.

Beck, A. T. (1967). Depression: Clinical, experimental, and theoretical aspects. Harper & Row.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). Cognitive therapy of depression. Guilford Press.

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. American Journal of Psychiatry, 165(8), 969–977.

Beck, J. S. (2011). *Cognitive behavior therapy: Basics and beyond* (2nd ed.). Guilford Press.

Becker, D. (1997). Through the looking glass: Women and borderline personality disorder. Basic Books.

Ben-Zeev, D., Brian, R., Wang, R., Wang, W., Campbell, A. T., & Aung, M. H. (2017). CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to monitor symptoms of schizophrenia. Schizophrenia Bulletin, 43(1), 199–209.

Ben-Zeev, D., Young, M. A., & Madsen, J. W. (2009). Retrospective recall of affect in clinically depressed individuals and controls. Cognition and Emotion, 23(5), 1021-1040. https://doi.org/ 10.1080/02699930802607937

Boyle, M. (2022). Schizophrenia: A scientific delusion? (2nd ed.). Routledge.

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: Access, effectiveness and efficiency. *British Journal of Psychiatry*, *186*(1), 11–17. https://doi.org/10.1192/bjp.186.1.11

Bracken, P., Thomas, P., Timimi, S., Asen, E., Behr, G., Beuster, C., ... & Yeomans, D. (2012). Psychiatry beyond the current paradigm. British Journal of Psychiatry, 201(6), 430–434. https://doi.org/10.1192/bjp.bp.112.109447

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 77-91. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American Journal of Psychology, 39*(1/4), 106–124. https://doi.org/10.2307/1415404

Centre for Mental Health. (2020). The economic and social cost of mental health problems in the UK.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 15. https://doi.org/10.1145/1541880.1541882

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry*, *23*(4), 318–327. https://doi.org/10.3109/09540261.2011.606803

Cornet, V. P., & Holden, R. J. (2018). Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics*, *77*, 120–132. https://doi.org/10.1016/j.jbi.2017.12.008

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research & Evaluation, 10(7), 1–9. https://doi.org/10.4135/9781412995627.d8

Cramer, A. O., Waldorp, L. J., van der Maas, H. L., & Borsboom, D. (2010). Comorbidity: a network perspective. *The Behavioral and brain sciences*, *33*(2-3), 137–193. https://doi.org/10.1017/S0140525X09991567

Cuijpers, P., van Straten, A., Smit, F., Mihalopoulos, C., & Beekman, A. (2008). Preventing the onset of depressive disorders: a meta-analytic review of

psychological interventions. *The American journal of psychiatry*, *165*(10), 1272–1280. https://doi.org/10.1176/appi.ajp.2008.07091422

Dietterich, T. G. (1995). Overfitting and undercomputing in machine learning. ACM Computing Surveys (CSUR), 27(3), 326.

Doryab, A., Min, J.-K., Wiese, J., Zimmerman, J., & Hong, J. (2014). Detection of behavior change in people with depression. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 451–460). ACM.

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications, 91*, 464–471. https://doi.org/10.1016/j.eswa.2017.09.030

Ecclestone, K., & Hayes, D. (2009). The dangerous rise of therapeutic education. Routledge.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence, 20*(1), 18–36. https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

Etkin, A., Büchel, C., & Gross, J. J. (2015). The neural bases of emotion regulation. *Nature Reviews Neuroscience, 16*(11), 693–700. https://doi.org/10.1038/nrn4044

Falkenström, F., Finkel, S., Sandell, R., Rubel, J. A., & Holmqvist, R. (2017). Dynamic models of individual change in psychotherapy process research. *Journal of Consulting and Clinical Psychology, 85*(6), 537–549. https://doi.org/10.1037/ccp0000203

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of 'precision psychiatry'. *BMC medicine*, *15*(1), 80. https://doi.org/10.1186/s12916-017-0849-x

Ferster, C. B. (1973). A functional analysis of depression. *American Psychologist, 28*(10), 857–870. https://doi.org/10.1037/h0035605

Floridi, L. (2011). The philosophy of information. Oxford University Press.

Folkman, S., & Moskowitz, J. T. (2004). Coping: Pitfalls and promise. *Annual Review of Psychology, 55*, 745–774. https://doi.org/10.1146/annurev.psych.55.090902.141456

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of affective disorders*, *172*, 96–102. https://doi.org/10.1016/j.jad.2014.10.010

Friedli, L. (2009). Mental health, resilience and inequalities. WHO Europe.

Friston K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews. Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daum√© III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Haley, F., Andrews, J. & Moghaddam, N. Acceptability of Remote Monitoring Technologies for Early Warning of Major Depression. *J. technol. behav. sci.* (2025). https://doi.org/10.1007/s41347-025-00530-1

Haley, F., Andrews, J. & Moghaddam, N. Advancements and Limitations: A Systematic Review of Remote-Based Deep Learning Predictive Algorithms for Depression. *J. technol. behav. sci.* (2024). *https://doi.org/10.1007/s41347-024-00457-z*

Harper, D. J., & Speed, E. (2012). Uncovering recovery: The resistible rise of recovery and resilience*.* Studies in Social Justice, 6*(1), 9–25. https://doi.org/10.26522/ssj.v6i1.1066*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual behavioral science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science, 1*(1-2), 1–16. https://doi.org/10.1016/j.jcbs.2012.09.004

Health and Safety Executive (HSE). (2022). Work-related stress, anxiety or depression statistics in Great Britain, 2022.

Herman, J. L. (1992). Trauma and recovery: The aftermath of violence—from domestic abuse to political terror. Basic Books.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology, 160(1), 106,Äì154.

Huckins, J. F., daSilva, A. W., Wang, W., Hedlund, E., Rogers, C., Nepal, S. K., Wu, J., Obuchi, M., Murphy, E. I., Meyer, M. L., Wagner, D. D., Holtzheimer, P. E., & Campbell, A. T. (2020). Mental Health and Behavior of College Students During the Early Phases of the COVID-19 Pandemic: Longitudinal Smartphone and Ecological Momentary Assessment Study. *Journal of medical Internet research*, *22*(6), e20185. https://doi.org/10.2196/20185

Huckvale, K., Venkatesh, S., & Christensen, H. (2019). Toward clinical digital phenotyping: A timely opportunity to consider purpose, quality, and safety. *NPJ Digital Medicine*, *2*(1), 88. https://doi.org/10.1038/s41746-019-0166-1

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., Jones, L., & Dunne, M. P. (2017). The effect of multiple adverse childhood

experiences on health: a systematic review and meta-analysis. *The Lancet. Public health*, *2*(8), e356–e366. https://doi.org/10.1016/S2468-2667(17)30118-4

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. JAMA, 318(13), 1215–1216. https://doi.org/10.1001/jama.2017.11295

Jacobson, N. C., & Chung, Y. J. (2020). Passive Sensing of Prediction of Moment-To-Moment Depressed Mood among Undergraduates with Clinical Levels of Depression Sample Using Smartphones. *Sensors (Basel, Switzerland)*, *20*(12), 3572. https://doi.org/10.3390/s20123572

Jacobson, N. C., Summers, B., & Wilhelm, S. (2020). Digital Biomarkers of Social Anxiety Severity: Digital Phenotyping Using Passive Smartphone Sensors. *Journal of medical Internet research*, *22*(5), e16875. https://doi.org/10.2196/16875

Jacobson, N. S., Martell, C. R., & Dimidjian, S. (2001). Behavioral activation treatment for depression: Returning to contextual roots. *Clinical Psychology: Science and Practice, 8*(3), 255–270. https://doi.org/10.1093/clipsy.8.3.255

James, W. (1884). What is an emotion? *Mind, 9*(34), 188–205.

Johnstone, L., & Boyle, M. (2018). The Power Threat Meaning Framework: Towards the identification of patterns in emotional distress, unusual experiences and troubled or troubling behaviour, as an alternative to functional psychiatric diagnosis. British Psychological Society.

Kapp, S. K., Gillespie-Lynch, K., Sherman, L. E., & Hutman, T. (2013). Deficit, difference, or both? Autism and neurodiversity. Developmental Psychology, 49(1), 59–71. https://doi.org/10.1037/a0028353

Kasser, T. (2002). The high price of materialism. MIT Press.

Kawachi, I., & Berkman, L. F. (2001). Social ties and mental health. *Journal of urban health : bulletin of the New York Academy of Medicine*, *78*(3), 458–467. https://doi.org/10.1093/jurban/78.3.458

Kendler, K. S. (2016). The phenomenology of major depression and the representativeness and nature of DSM criteria. American Journal of Psychiatry, 173(8), 771–780. https://doi.org/10.1176/appi.ajp.2016.15121509

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry*, *62*(6), 593–602. https://doi.org/10.1001/archpsyc.62.6.593

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2019). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems, 30*(1), 118–131. https://doi.org/10.1109/TNNLS.2018.2801473

Kirmayer, L. J., & Swartz, L. (2013). Culture and global mental health. In V. Patel, H. Minas, A. Cohen, & M. Prince (Eds.), Global mental health: Principles and practice (pp. 41–62). Oxford University Press.

Kleinman, A. (1980). *Patients and healers in the context of culture: An exploration of the borderland between anthropology, medicine, and psychiatry* (Vol. 3). Univ of California Press.

Kleinman, A. (1987). Anthropology and psychiatry. The role of culture in cross-cultural research on illness. The British Journal of Psychiatry, 151(4), 447–454. https://doi.org/10.1192/bjp.151.4.447

Kleinman A. (2004). Culture and depression. *The New England journal of medicine*, *351*(10), 951–953. https://doi.org/10.1056/NEJMp048078

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general

population. *Journal of affective disorders*, *114*(1-3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science, 21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Layard R. (2006). The case for psychological treatment centres. *BMJ (Clinical research ed.)*, *332*(7548), 1030–1032. https://doi.org/10.1136/bmj.332.7548.1030.

Layte, R. (2012). The association between income inequality and mental health: Testing status anxiety, social capital and neo-materialist explanations. European Sociological Review, 28(4), 498-511 https://doi.org./10.2307/23272534

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Lederman, R., & D'Alfonso, S. (2021). The digital therapeutic alliance: Prospects and considerations. *JMIR Mental Health*, *8*(7), e31385. https://doi.org/10.2196/31385

Lewinsohn, P. M. (1974). A behavioral approach to depression. In R. J. Friedman & M. M. Katz (Eds.), *The psychology of depression: Contemporary theory and research.* John Wiley & Sons.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988. https://doi.org/10.1109/ICCV.2017.324

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences, 20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Luhrmann, T. M., Padmavati, R., Tharoor, H., & Osei, A. (2015). Differences in voice-hearing experiences of people with psychosis in the USA, India and Ghana: Interview-based study. The British Journal of Psychiatry, 206(1), 41–44. https://doi.org/10.1192/bjp.bp.113.139048

Lund, C., Breen, A., Flisher, A. J., Kakuma, R., Corrigall, J., Joska, J. A., Swartz, L., & Patel, V. (2010). Poverty and common mental disorders in low and middle income countries: A systematic review. *Social science & medicine (1982)*, *71*(3), 517–528. https://doi.org/10.1016/j.socscimed.2010.04.027

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.48550/arXiv.1705.07874

Marmot, M. (2004). Status syndrome: How your social standing directly affects your health and life expectancy. Bloomsbury Press.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. W. H. Freeman.

Martell, C. R., Addis, M. E., & Jacobson, N. S. (2001). *Depression in context: Strategies for guided action*. Norton.

Matcham, F., Carr, E., White, K. M., Leightley, D., Lamers, F., Siddi, S., Annas, P., de Girolamo, G., Haro, J. M., Horsfall, M., Ivan, A., Lavelle, G., Li, Q., Lombardini, F., Mohr, D. C., Narayan, V. A., Penninx, B. W. H. J., Oetzmann, C., Coromina, M., Simblett, S. K., … RADAR-CNS consortium (2022). Predictors of engagement with remote sensing technologies for symptom measurement in Major Depressive Disorder. *Journal of affective disorders*, *310*, 106–115. https://doi.org/10.1016/j.jad.2022.05.005

McDonald, L., Ramagopalan, S. V., Cox, A. P., & Oguz, M. (2017). Unintended consequences of machine learning in medicine?. *F1000Research*, *6*, 1707. https://doi.org/10.12688/f1000research.12693.1

Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual review of clinical psychology*, *13*, 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

Mohr, D. C., Schueller, S. M., Montague, E., Burns, M. N., & Rashidi, P. (2014). The Behavioral Intervention Technology model: An integrated conceptual and technological framework for eHealth and mHealth interventions. *Journal of Medical Internet Research*, *16*(6), e146. https://doi.org/10.2196/jmir.3077

Moghaddam, N. G., & Dawson, D. L. (2015). The impact of early maladaptive schemas on mental health in adulthood: A systematic review and meta-analysis. *Clinical Psychology Review, 41*, 90–103. https://doi.org/10.1016/j.cpr.2015.06.002

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & van Os, J. (2009). Experience sampling research in psychopathology: opening the black box of daily life. *Psychological medicine*, *39*(9), 1533–1547. https://doi.org/10.1017/S0033291708004947

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, *52*(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8

NHS England. (2019). The NHS Long Term Plan. https://www.longtermplan.nhs.uk/ OECD. (2018). Health at a Glance: Europe 2018. OECD Publishing.

Onnela, J. P., & Rauch, S. L. (2016). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology : official publication of the American College*

of Neuropsychopharmacology, *41*(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Paris, J. (2020). Personality disorders over time: Precursors, course, and outcome. American Psychiatric Association Publishing.

Patel, V., Burns, J. K., Dhingra, M., Tarver, L., Kohrt, B. A., & Lund, C. (2018). Income inequality and depression: A systematic review and meta-analysis of the association and a scoping review of mechanisms. World Psychiatry, 17(1), 76-89. https://doi.org/10.1002/wps.20492

Peppin, A. (2022, May 5). *Who cares what the public think?* Ada Lovelace Institute. https://www.adalovelaceinstitute.org/evidence-review/public-attitudes-data-regulation/

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*(2), 148–158. https://doi.org/10.1038/nrn2317

Pickett, K. E., & Wilkinson, R. G. (2015). Income inequality and health: A causal review. Social Science & Medicine, 128, 316-326. https://doi.org/10.1016/j.socscimed.2014.12.031

Popper, K. (1959). The logic of scientific discovery. Hutchinson & Co.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports, 7*, 13006. https://doi.org/10.1038/s41598-017-12961-9

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., & Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems, 32*.

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018). Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. *JMIR mHealth and uHealth*, *6*(8), e165. https://doi.org/10.2196/mhealth.9691

Rose, N. (1999). Governing the soul: The shaping of the private self (2nd ed.). Free Association Books.

Rose, N. (2019). Our psychiatric future: The politics of mental health. Polity Press. Shiffman, S.,

Russell, M. A., & Gajos, J. M. (2020). Annual Research Review: Ecological momentary assessment studies in child psychology and psychiatry. *Journal of Child Psychology and Psychiatry, 61*(3), 376–394. https://doi.org/10.1111/jcpp.13204

Ryder, A. G., Yang, J., & Heine, S. J. (2008). Somatization vs. psychologization of emotional distress: A paradigmatic example for cultural psychopathology. In S. Choudhury & J. Slaby (Eds.), *Critical neuroscience: A handbook of the social and cultural contexts of neuroscience* (pp. 340–355). Wiley-Blackwell. https://doi.org/10.9707/2307-0919.1080

Ryder, A. G., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S. J., & Bagby, R. M. (2008). The cultural shaping of depression: somatic symptoms in China, psychological symptoms in North America?. *Journal of abnormal psychology*, *117*(2), 300–313. https://doi.org/10.1037/0021-843X.117.2.300

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile Phone Sensor Correlates of Depressive Symptom

Severity in Daily-Life Behavior: An Exploratory Study. *Journal of medical Internet research*, *17*(7), e175. https://doi.org/10.2196/jmir.4273

Sato, H., & Kawahara, J. (2011). Selective bias in retrospective self-reports of negative mood states. *Anxiety, stress, and coping*, *24*(4), 359–367. https://doi.org/10.1080/10615806.2010.543132

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379–399. https://doi.org/10.1037/h0046234

Scherer, K. R. (2009). Emotions are emergent processes: They require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), 3459–3474. https://doi.org/10.1098/rstb.2009.0141

Schmier, J. K., & Halpern, M. T. (2004). Patient recall and recall bias of health state and health status. Expert review of pharmacoeconomics & outcomes research, 4(2), 159-163. doi.org/10.1586/14737167.4.2.159

Shakespeare, T. (2013). Disability rights and wrongs revisited (2nd ed.). Routledge.

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. Psychological Medicine, 49(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual review of clinical psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Simblett, S., Matcham, F., Siddi, S., Bulgari, V., Di San Pietro, C. B., López, J. H., Ferrão, J., Polhemus, A., Haro, J. M., De Girolamo, G., Gamble, P., Eriksson, H., Hotopf, M., & Wykes, T. (2019). Barriers to and facilitators of engagement with mHealth technology for remote measurement and management of depression: Qualitative analysis. *JMIR Mhealth and Uhealth*, *7*(1), e11325. https://doi.org/10.2196/11325

Sun, Y., Wong, A. K., & Kamel, M. S. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 40*(12), 3358–3378. https://doi.org/10.1016/j.patcog.2007.04.009

Suresh, H., & Guttag, J. V. (2021). A framework for understanding unintended consequences of machine learning. Communications of the ACM, 64(5), 62. https://doi.org/10.1145/3442188

Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.

Torous, J., Wisniewski, H., Liu, G., & Keshavan, M. (2018). Mental Health Mobile Phone App Usage, Concerns, and Benefits Among Psychiatric Outpatients: Comparative Survey Study. *JMIR mental health*, *5*(4), e11715. https://doi.org/10.2196/11715

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of abnormal psychology*, *129*(1), 56–63. https://doi.org/10.1037/abn0000473

Twenge, J. M., Martin, G. N., & Spitzberg, B. H. (2018). Trends in US adolescents‚Äô media use, 1976‚Äì2016: The rise of digital media, the decline of TV, and the (near) demise of print. Psychology of Popular Media Culture, 8(4), 329. https://doi.org/10.1037/ppm0000203

van de Leemput, I. A., Wichers, M., Cramer, A. O., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H.,

Derom, C., Jacobs, N., Kendler, K. S., van der Maas, H. L., Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1), 87–92. https://doi.org/10.1073/pnas.1312114110

van Os, J., Guloksuz, S., Vijn, T. W., Hafkenscheid, A., & Delespaul, P. (2019). The evidence-based group-level symptom-reduction model as the organizing principle for mental health care: Time for change? World Psychiatry, 18(1), 88. https://doi.org/10.1002/wps.20609

Varoquaux G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*(Pt A), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Wang, R., Wang, W., Dasilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Campbell, A. T. (2018). Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, *2*(1), 43. https://doi.org/10.1145/3191775

Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*., 1578-1585 https://doi.org/10.1109/IJCNN.2017.7966039

Wang, Y., Kraut, R. E., & Levine, J. M. (2012). To stay or leave? The relationship of emotional and informational support to commitment in online health support groups. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (pp. 833–842). ACM.

Watters, E. (2010). Crazy like us: The globalization of the American psyche. Free Press.

Westbrook, D., Kennerley, H., & Kirk, J. (2007). *An introduction to cognitive behaviour therapy: Skills and applications*. Sage Publications.

White, R. G., Imperiale, M. G., & Perera, E. (2017). Mental health and wellbeing in the Anthropocene: A posthuman rights-based approach. International Journal of Mental Health Systems, 11, 65. https://doi.org/10.1007/978-981-13-3326-2

WHO. (2014). Social determinants of mental health.

Wilkinson, R., & Pickett, K. (2009). The spirit level: Why more equal societies almost always do better. Allen Lane.

Xu, H., Huang, J., & Xu, L. (2018). LSTM-based anomaly detection for time series. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Data Processing* (pp. 1–5). https://doi.org/10.1109/IDAP.2018.8620895

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Laiou, P., Matcham, F., Lamers, F., Siddi, S., Simblett, S., Rintala, A., Mohr, D. C., Myin-Germeys, I., Wykes, T., Haro, J. M., Pennix, B. W. J. H., Narayan, V. A., Annas, P., Hotopf, M., & Dobson, R. J. B. (2021). Predicting depressive symptom severity through individuals' nearby Bluetooth devices count data collected by mobile phones: A preliminary longitudinal study.

3. Additional Paper: Cultural Variation in PHQ8 Responses

Cultural Differences in Depression Symptom Reporting: A Cross-National Analysis Using the RADAR-MDD Dataset

Authors:

Fintan Haley (Corresponding Author),[1] Jacob Andrews,[2] Nima Moghaddam[3], Alexander Turner[4]

Author affiliations and information:

[1]Trent DClinPsy Programme, University of Nottingham, Nottingham, United Kingdom

Fintan.haley@nottingham.ac.uk

[2]NIHR MindTech Medtech Co-operative, Academic Unit of Mental Health and Clinical Neuroscience, School of Medicine, University of Nottingham, Nottingham, United Kingdom

jacob.andrews@nottingham.ac.uk

ORCID: 0000-0001-8408-5782

[3]College of Health and Science, School of Psychology, Trent DClinPsy Programme, University of Lincoln, Lincoln, United Kingdom

nmoghaddam@lincoln.ac.uk

ORCID: 0000-0002-8657-4341

[4]Department of Computer Science, University of Nottingham, Nottingham, United Kingdom

alexander.turner@nottingham.ac.uk

ORCID: 0000-0002-2392-6549

Intended Paper: Journal of technology in behavioral science

Data Availability Statement:

The data used in this study were obtained from the Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD) study, conducted as part of the RADAR-CNS research programme. Due to ethical restrictions and participant confidentiality agreements, the dataset is not publicly available and cannot be shared.

Competing Interest Declaration:

The authors declare that they have received funding for their respective courses. Jacob Andrews is funded by the National Institute for Health and Care Research (NIHR), Nottingham Biomedical Research Centre, Mental Health and Technology theme. Fintan Haley is funded by the Nottinghamshire Healthcare NHS Foundation Trust. It should be noted that the views presented in this manuscript are those of the authors and do not necessarily reflect the views of the NIHR or Nottinghamshire Healthcare NHS Foundation Trust. Furthermore, the authors wish to clarify that no additional funds, grants, or other support were received specifically for the preparation of this manuscript. The authors have no relevant financial or non-financial interests to disclose.

Author Contribution Declaration:

All authors collectively contributed to the study's conception and design. Material preparation and data analysis were primarily conducted by Fintan Haley. The first draft of the manuscript was authored by Fintan Haley, with subsequent revisions and comments provided by all authors. All authors have reviewed and approved the final manuscript.

3.1 Abstract

Depression is a globally prevalent mental health condition, yet its manifestation and measurement may vary significantly across cultural contexts. This study examined cross-national differences in depressive symptom reporting and response style using data from the RADAR-MDD project, a large-scale longitudinal study that monitored individuals with Major Depressive Disorder across the United Kingdom, Spain, and the Netherlands. Utilizing baseline responses to the Patient Health Questionnaire-8 (PHQ-8) from 623 participants, we conducted a series of ANCOVAs to compare total scores, item-level symptom reporting, and response extremity indices between countries, controlling for age, gender, education finish age, and baseline symptom severity.

Results revealed significant differences in overall depression scores, with Spanish participants reporting higher symptom levels than their UK and Dutch counterparts. Analysis of individual PHQ-8 items indicated pronounced cross-national variation in symptoms such as sleep disturbance, fatigue, and psychomotor changes, with the strongest effects observed in Spain. Response style analysis further revealed cultural differences in the tendency to use extreme response categories. These findings suggest that cultural, linguistic, and contextual factors influence how depression is reported and perceived, challenging the universal applicability of standardized diagnostic tools like the PHQ-8.

The study highlights critical implications for cross-cultural mental health assessment, emphasizing the need for culturally sensitive diagnostic frameworks and AI-based tools that accommodate diverse symptom presentations. Future research should explore the mechanisms driving these differences, including sociocultural norms, stigma, and healthcare accessibility, to enhance diagnostic validity and clinical care across global populations.

3.2 Introduction

Depression is a widespread mental health condition, affecting approximately 16% of individuals aged 16 and older in Western countries (Mullis & Attwell, 2022). The American Psychological Association (APA) defines depression as a persistent mood disorder characterized by prolonged sadness, hopelessness, and a reduced ability to experience pleasure (APA, 2022).

In clinical practice, depression is commonly diagnosed using two major classification systems: the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) and the International Classification of Diseases (ICD-10). However, these frameworks were developed within Western contexts and may reflect a Eurocentric and North American conceptualization of psychiatric disorders (Halbreich et al., 2007). This cultural bias raises concerns about the validity and applicability of diagnostic criteria across diverse populations. Even within Europe, significant regional variations in depression prevalence have been observed. Van de Velde et al. (2010) examined depression rates across 25 European countries using the 8-item CES-D scale and found that prevalence tended to cluster by region, with the highest rates in Central and Eastern European countries (e.g., Ukraine, Hungary, Russia) and the lowest in Western and Northern European nations (e.g., Norway, Denmark, Switzerland).

Several factors may contribute to these regional differences. One explanation suggests that environmental and societal influences, such as socioeconomic inequality, impact mental health outcomes. Pickett and Wilkinson (2015) highlight the relationship between societal inequality and overall well-being, suggesting that disparities in income, social support, and access to resources may shape depression rates. However, cross-cultural comparisons of depression present methodological challenges. Bias can occur at multiple levels, including construct bias, where the definition and symptomatology of depression may vary across cultures, and item bias, where specific questionnaire items may be interpreted differently depending on cultural norms and linguistic nuances (Van de Velde et al., 2010).

Another issue is category fallacy, where diagnostic criteria developed in one cultural setting are inappropriately applied to another, leading to inaccurate conclusions (Simon et al., 2002). Standardized diagnostic tools may inadvertently measure different severity levels of depression or fail to capture culturally specific expressions of distress. Therefore, understanding cultural context is crucial when interpreting depression prevalence rates and applying diagnostic frameworks globally.

Depression is not purely a biological condition; it is also influenced by cultural factors, distinguishing it from many physical illnesses with clearly defined biological etiologies and outcomes (Kirmayer & Sartorius, 2007). The classification of mental health disorders is inherently linked to social norms, as diagnoses rely on the identification of distress or functional impairment within the context of societal expectations (Summerfield, 2008).

In Western, individualistic societies, normative functioning is often defined by the ability to maintain consistent employment, sustain personal independence, and focus on tasks for extended periods (Kleinman, 1988). In contrast, collectivist societies may place greater emphasis on communal living and social support networks, which can mitigate psychological distress (Chiao & Blizinsky, 2010). This cultural variation suggests that while depression exists universally, its manifestation, expression, and diagnosis are shaped by social context (Ryder et al., 2008). However, this does not imply that depression is merely a social construct; rather, its presentation and associated challenges differ based on sociocultural expectations. Recognizing these contextual differences is essential for developing culturally sensitive approaches to diagnosis and treatment, ensuring that mental health care reflects the diverse ways distress is experienced and addressed across societies.

Recent advancements in AI have led to growing interest in its potential for diagnosing and predicting mental health conditions such as depression (Haley et al., 2024). Machine learning, a subset of AI, is particularly effective in identifying complex, non-linear relationships between variables and clinical

presentations (Bishop, 2006). However, for machine learning models to function effectively, they require access to a validated reference point, or "ground truth," against which predictions can be assessed. In the context of depression, this ground truth is typically derived from psychometric questionnaires, which classify individuals based on the presence or absence of depressive symptoms (Kessler et al., 2002).

While such diagnostic approaches offer a structured method for identifying depression, they may not fully capture its complexity, particularly across different cultural contexts (Ryder et al., 2008). If machine learning algorithms are trained on datasets that rely primarily on binary classifications of depression, they may fail to generalize across populations where symptom expression varies due to cultural, linguistic, or contextual factors (Kirmayer & Sartorius, 2007). Despite recognition of these challenges, there remains a significant gap in large-scale, longitudinal research examining cultural variations in depression symptomatology. Existing studies often rely on cross-sectional data from single-point surveys with limited sample sizes (Haley et al., 2024), making it difficult to assess how depressive symptoms evolve over time and across different cultural contexts.

This study aimed to explore whether significant differences exist in the symptoms of depression across different countries. To achieve this, the study utilized data from the RADAR-MDD dataset, which includes participants from the United Kingdom, Spain, and the Netherlands (Matcham et al., 2022). The RADAR-MDD study followed 623 participants over two years, collecting self-reported data, including responses to the PHQ-8, a widely used psychometric questionnaire for depression. Statistical analyses were conducted to examine whether specific depressive symptoms varied significantly between countries. By investigating these cross-national differences, this study investigates whether significant cross-cultural differences exist in depression symptom reporting and response style using data from the RADAR-MDD project across the UK, Spain, and the Netherlands.

3.3 Methods

3.3.1 Study Design

The study used a large longitudinal data set of an EU research program, RADAR-MDD, which explored the utility of remote measurement technologies in long-term (up to 2 years) depression monitoring (Matcham, et al., 2019). The study received ethical approval from the research ethics committee of the University of Nottingham.

3.3.2 Population

The RADAR-MDD study aimed to detect events increased depression symptomology through smartphone data and estimated that 100 relapses of depression would be required to provide sufficient power for a predictive model of 10 variables (Peduzzi, et al., 1996; Matcham, et al., 2019). It was approximated that 33% of participants would relapse over a year, therefore a minimum of 300 participants would be required for the cohort study (Trivedi, et al., 2006). Out of concerns for noisy data and attrition rates the team recruited 623 (Matcham, et al., 2022). The demographics of the population can be seen in Table 8.

Table 8: Socio-demographic and clinical baseline data adapted from the RADAR-MDD study (Matcham, et al., 2022).

	Total Sample
Total, N(%)	623 (100.0)
London, N(%)	350 (56.2)
Barcelona, N(%)	155 (24.9)
Amsterdam, (%)	118 (18.9)
Socio-demographics	

		Total Sample
Age, M (SD)		46.4 (15.3)
Gender, N (%)	Female	471 (75.6)
Marital Status, N (%)	Single/separated/divorced/ widowed	332 (53.3)
	Married/cohabiting/LTR	291 (46.7)
Ethnicity, N(%)	White British/Dutch	369 (78.9)
	White Other	35 (7.5)
	Black ethnic group	14 (3.0)
	Asian ethnic group	16 (3.4)
	Mixed ethnic background	16 (3.4)
	Other	18 (3.9)
Employment Status	Employed/furloughed	260 (41.7)
	Unemployed/sick leave	134 (21.5)
	Student	68 (10.9)
	Retired	123 (19.7)
	Not reported	38 (6.1)
Total years in education, M(SD)		16.4 (6.5)
Benefits Receipt, N(%)	Yes	275 (44.1)
Accommodation type, N(%)	Own outright/with mortgage	368 (59.1)

		Total Sample
	Renting	216 (34.7)
	Living rent-free	29 (4.7)
	Not reported	10 (1.6)
Household income per annum, N(%)	<£/€15,000	154 (24.8)
	£/€15,000 – 55,000	354 (57.0)
	>£€55,000	98 (15.8)
	Prefer not to say	10 (1.6)
	Unknown	5 (0.8)
Clinical Characteristics		
Current depression	IDS-SR total, M(SD)	31.3 (14.5)
	None (0–13), N(%)	61 (10.1)
	Mild (14–25), N(%)	157 (25.9)
	Moderate (26–38), N(%)	206 (33.9)
	Severe (39–48), N(%)	104 (17.1)
	Very severe (49–84), N(%)	79 (13.0)
	Not reported	16 (2.6)
Baseline aRMT PHQ8	PHQ8 total, M(SD)	10.9 (6.0)
	≥10, N(%)	371 (59.6)

3.3.3 Recruitment

Recruitment for the study was conducted over a period of eighteen months at three international sites, including King's College London (UK), the Netherlands Study of Depression and Anxiety, and other available patient groups at Vrije Universiteit Medisch Centrum (Netherlands), as well as the Centro de Investigacion Biomedica en Red (Spain) (Matcham, et al., 2019). Eligible participants were identified through existing research cohorts or mental health services and contacted by telephone with study information sheets and consent forms sent via email. Eligibility criteria can be seen in Table 9. Enrolment took place either at the research centre or at the participant's home. As an incentive, participants received £15/€20 for enrolling in the study and £5/€10 for every three months of continued participation.

Written consent was obtained during the enrolment session, which included the collection of sociodemographic, social environment, medical history, and technology use questionnaires, as well as baseline data collection of all outcome measures (Harris, et al., 2009). Subsequently, participants were monitored for a period of 24 months. Upon completion of the study, participants underwent a 60-minute debriefing session, during which their experiences of the study were investigated.

Table 9: Eligibility criteria for participation taken from the RADAR-MDD study (Matcham, et al., 2019).

Inclusion criteria	Exclusion criteria
Meet DSM-5 diagnostic criteria for diagnosis of non-psychotic MDD within the past two years.	Lifetime history of bipolar disorder, schizophrenia, MDD with psychotic features, schizoaffective disorders.
Recurrent MDD (a lifetime history of at least two episodes of depression)	Dementia.

Inclusion criteria	Exclusion criteria
Willing and able to complete self-reported assessments via smartphone.	History of moderate to severe drug or alcohol dependence within the last 6 months.
Able to give informed consent for participation.	History of major medical disease which might impact upon the patient's ability to participate in normal daily activities for more than two weeks (e.g. due to likely hospitalisations or other periods of indisposition). Pregnancy
Fluent in English, Spanish, Catalan or Dutch language.	
Existing ownership of Android smartphone or willingness to use an Android smartphone as their only smartphone.	
Aged 18 or over.	

3.3.4 Data Collection

Sociodemographic information was collected during the enrolment procedure of the study (Matcham, et al., 2019). Upon enrolment participants downloaded an active RMT (aRMT) app so that validated measures could be administered remotely to participants at set time intervals. Throughout the process, data was completely anonymized. The researchers appended IDs to each data stream to identify participants. This enabled the research team to contact participants if any data or psychometric information was missing. Access to the visualization dashboard was subject to authentication and authorization. The data used in our analysis is currently stored in a secured communal drive in CSV format. This data has already been pre-processed and is available to members of the

Remote Assessment of Disease and Relapse Central Nervous System (RADAR-CNS) consortium, which includes one of the authors of this project. Numerous metrics were collected during the study for various potential analyses; the primary outcome of our study will be depression as measured through the PHQ-8 (Kroenke, et al., 2001).

3.3.5 Primary outcomes

The study aimed to examine notable differences in responses to the items on the PHQ-8 questionnaire. Each primary outcome corresponded to one of the eight questions, which assess various symptoms of depression. Participants responded to each question using a four-point scale: "Not at all," "Several days," "More than half the days," or "Nearly every day." The specific questions are outlined below.

1. Little interest or pleasure in doing things
2. Feeling down, depressed, irritable or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down
7. Trouble concentrating on things, such as schoolwork, reading or watching television
8. Moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving

3.3.6 Covariates

To account for potential confounding factors in cross-national comparisons of depression symptomatology and response patterns, several covariates were included in the analysis based on empirical and theoretical justification. Specifically, we controlled for age, gender, education finish age, and baseline PHQ-8 total score.

Age and gender were included due to their established associations with depression prevalence, symptom expression, and reporting behaviours. For instance, women tend to report higher levels of internalising symptoms, while symptom presentation and help-seeking behaviours also vary across age groups (Kuehner, 2017; Salk et al., 2017). Educational attainment, proxied here by age at completion of formal education, was included to account for differences in mental health literacy and cognitive interpretation of questionnaire items, which can influence self-report accuracy (Jorm, 2000; Parker et al., 2001). Baseline depression severity, measured using the PHQ-8 total score, was included to ensure that country-level differences in individual item responses were not confounded by variation in overall symptom burden at study entry.

Other potentially relevant covariates—such as ethnicity, household income, and treatment history—were considered but excluded due to incomplete or inconsistently reported data across the study sites. While these factors are also known to influence depression outcomes and reporting (Williams et al., 2007; Lorant et al., 2003), their exclusion was necessary to maintain analytical robustness and avoid introducing bias through selective missingness.

3.3.7 Analysis

Step 1: PHQ-8 Total Scores

An Analysis of Covariance (ANCOVA) was conducted to assess whether PHQ-8 total scores differed significantly across countries after controlling for potential confounding variables (age, gender, education finish age). Pairwise ANCOVAs with Bonferroni corrections were subsequently performed to identify specific country differences. Effect sizes were calculated using partial eta squared (η^2_p).

Step 2: Response Extremity Index

To evaluate cross-national differences in response styles, a Response Extremity Index (REI) was calculated for each participant, representing the

proportion of extreme responses (scores of 0 or 3) across PHQ-8 items. An ANCOVA was conducted with REI as the dependent variable, country as the between-subject factor, and age, gender, education finish age, and PHQ-8 total score as covariates. Pairwise ANCOVAs with Bonferroni corrections were used to pinpoint differences between specific countries.

Step 3: PHQ-8 Item-by-Item Analysis

For detailed symptom-level analyses, eight separate ANCOVAs were performed, each using a single PHQ-8 item as the dependent variable. Each model included country as the between-subject factor and controlled for age, gender, education finish age, and PHQ-8 total score. Following significant main effects, pairwise ANCOVAs with Bonferroni corrections were conducted. Additionally, Cohen's d effect sizes were calculated to quantify the magnitude of differences in symptom endorsement between countries.

All assumptions for ANCOVA, including normality, linearity, homogeneity of regression slopes, and homogeneity of variances, were assessed prior to analysis. Data cleaning included the exclusion of cases with missing values on relevant variables, ensuring robustness and reliability of the statistical results.

3.4 Results

3.4.1 PHQ-8 Total Scores Across Countries

A one-way ANCOVA was conducted to examine whether PHQ-8 total scores differed across the UK, Spain, and the Netherlands after adjusting for age, gender, and education finish age. The model revealed a significant main effect of country, $F(2, 12361) = 500.26$, $p < .001$, with a moderate effect size ($\eta^2_p = 0.075$). Covariates were also significantly associated with PHQ-8 scores (Table 10).

Table 10: ANCOVA for PHQ-8 Total Scores

Source	SS	DF	F	p-unc	η^2_p
Country	33,585.03	2	500.26	< .001	0.075
Age	17,686.63	1	526.90	< .001	0.041
Gender	565.56	1	16.85	< .001	0.001
Education Finish Age	3,719.56	1	110.81	< .001	0.009
Residual	414,927.04	—	—	—	—

Bonferroni-adjusted pairwise comparisons (Table 11) showed that Spain reported significantly higher total scores than both the Netherlands ($p < .001$) and the UK ($p < .001$). No significant difference was found between the UK and the Netherlands ($p = 1.000$).

Table 11: Pairwise ANCOVAs for PHQ-8 Total (Bonferroni Corrected)

Group 1	Group 2	η^2_p	p-value
Spain	Netherlands	0.083	< .001
Spain	UK	0.078	< .001
Netherlands	UK	0.00001	1.000

3.4.2 Response Style Differences Across Countries

To investigate cross-national response tendencies, a response extremity index was calculated for each participant. ANCOVA showed a significant main effect of country, $F(2, 12360) = 30.04$, $p < .001$, with a small effect size ($\eta^2_p = 0.005$). Age and education were also significant predictors, while gender and PHQ-8 total were not (Table 12).

Table 12: ANCOVA for Response Extremity Index

Source	SS	DF	F	p-value	η^2_p
Country	4.61	2	30.04	< .001	0.005
Age	3.32	1	43.24	< .001	0.0035
Gender	0.04	1	0.56	.453	< .001
Education Finish Age	2.82	1	36.73	< .001	0.0030
PHQ-8 Total Score	0.31	1	3.99	.046	< .001
Residual	948.17	—	—	—	—

Table 13: Pairwise Comparisons for Response Extremity (Bonferroni Corrected)

Group 1	Group 2	η^2_p	p-value
Spain	Netherlands	0.0007	0.152
Spain	UK	0.0034	< .001
Netherlands	UK	0.0053	< .001

3.4.3 Symptom-Specific Differences Across Countries

Eight separate ANCOVAs were run for each PHQ-8 item, adjusting for age, gender, education, and PHQ-8 total score. Results are summarised in Table 14. All models showed significant main effects of PHQ-8 total score, and most showed a significant country effect, particularly for sleep disturbance (Item 3) and motor changes (Item 8).

Table 14: Summary of PHQ-8 Item-Level ANCOVAs (Country Effect Only)

PHQ-8 Item	Symptom	F	p-value	η^2_p
phq8_1	Anhedonia	43.01	< .001	0.0069
phq8_2	Depressed mood	11.66	< .001	0.0019

PHQ-8 Item	Symptom	F	p-value	η^2_p
phq8_3	Sleep problems	115.08	< .001	0.0183
phq8_4	Fatigue	93.30	< .001	0.0149
phq8_5	Appetite changes	4.54	.011	0.0007
phq8_6	Feelings of worthlessness	3.69	.025	0.0006
phq8_7	Trouble concentrating	9.19	< .001	0.0015
phq8_8	Psychomotor changes	254.58	< .001	0.0396

Post-hoc pairwise comparisons indicated that Spain consistently reported higher symptom scores than the UK and Netherlands across most items. The largest effects were observed for Item 8 (motor agitation/retardation) with η^2_p = 0.040 and Item 3 (sleep) with η^2_p = 0.018. Notably, pairwise differences between the UK and the Netherlands were often non-significant.

Unadjusted Cohen's *d* calculations reinforced these findings. Large effect sizes were found for:

- Item 8: *d* = 0.80 (Spain vs UK)
- Item 3: *d* = 0.49 (Spain vs Netherlands)
- Item 4: *d* = 0.36 (Spain vs UK)

3.5 Discussion

The findings of this study underscore notable cross-national variations in depressive symptom reporting and response styles among participants from the United Kingdom, Spain, and the Netherlands. These results highlight the significant role cultural context plays in the manifestation, interpretation, and reporting of depressive symptoms, posing important considerations regarding the universal applicability of diagnostic instruments predominantly developed within Western psychiatric frameworks (Fried et al., 2021; Van de Vijver & Leung, 1997).

Our analysis revealed that there were signficant differences in response styles between countries along with symptom specific differences. These heightened symptom scores may reflect broader socioeconomic factors, including economic instability, higher unemployment rates, and comparatively limited access to mental health services, potentially exacerbating emotional distress and symptom reporting in Spanish populations (Chaves et al., 2018).

An alternative explanation for the observed differences in response patterns is that the PHQ-8 items may not operate equivalently across linguistic and cultural contexts. In other words, what may initially appear to reflect genuine differences in depressive symptomatology could instead result from variation in how individuals interpret, translate, or respond to the items—rather than differences in the underlying construct of depression itself. Although prior research has examined the cross-national measurement invariance of the PHQ-8 (Arias de la Torre et al., 2023), supporting its structural consistency across different populations, the extent to which these findings apply to the current study is uncertain. This is particularly relevant given the unique characteristics of this dataset: repeated app-based self-report measures collected across three countries. As such, questions remain about the generalisability of previous invariance findings to this specific methodological and cultural context.

Nonetheless, these cultural differences highlight critical methodological challenges in the cross-cultural application of standardized instruments such as the PHQ-8. Cultural constructions of distress significantly influence how symptoms are understood and reported, creating risks of misinterpretation and category fallacy when applying diagnostic criteria developed in Anglo-centric contexts globally (Kohrt et al., 2022). Explicitly, our findings suggest a need to reconsider specific diagnostic criteria from DSM and ICD frameworks, which might inadequately capture culturally influenced symptom presentations and severities.

Furthermore, our results have critical implications for the increasing integration of AI into diagnostic practices. Machine learning models built primarily on

Western-centric data risk perpetuating biases or inaccuracies when generalized to culturally diverse populations (Peters & Carman, 2024). Our findings underline the importance of incorporating culturally diverse datasets and nuanced symptom-reporting styles in AI training algorithms, thus promoting accurate, equitable, and culturally sensitive mental health diagnostics (Onnela & Rauch, 2023).

Clinically, practitioners should integrate culturally adapted diagnostic approaches, employing supplementary qualitative assessments and culturally informed guidelines to ensure accurate identification and management of depression. For instance, clinicians might benefit from additional training in recognizing culture-specific expressions of distress, such as somatic complaints or relational difficulties commonly reported in collectivist cultures.

Future research should further investigate mechanisms underpinning observed cultural differences, specifically exploring how socioeconomic conditions, healthcare accessibility, stigma, and social connectedness uniquely shape symptom progression and help-seeking behaviors across different cultural contexts. Qualitative studies capturing subjective experiences can provide deeper insights into culturally distinct depression phenomenologies.

3.5.1 Limitations

Despite notable strengths such as the large longitudinal design and cross-national participant pool, this study has several limitations. First, reliance on self-report measures introduces susceptibility to subjective biases, social desirability effects, and potential inaccuracies in symptom reporting. Objective biomarkers or clinician-administered assessments could complement these data in future research, enhancing the reliability and comprehensiveness of findings.

Additionally, our sample was predominantly female, potentially limiting the generalizability of findings across gender groups. Future studies should aim for

more balanced gender representation to explore gender-specific cultural influences on depression symptomatology.

Finally, although our data are longitudinal, this study primarily focused on cross-sectional analyses of baseline data. Longitudinal analyses tracking symptom trajectories over time could provide more detailed insights into how cultural contexts dynamically influence depression experiences.

3.6 References

American Psychiatric Association. (2022). *What is depression?* https://www.psychiatry.org/patients-families/depression/what-is-depression

Arias de la Torre, J., Vilagut, G., Ronaldson, A., Valderas, J. M., Bakolis, I., Dregan, A., Molina, A. J., Navarro-Mateu, F., Pérez, K., Bartoll-Roca, X., Elices, M., Pérez-Sola, V., Serrano-Blanco, A., Martín, V., & Alonso, J. (2023). Reliability and cross-country equivalence of the 8-item version of the Patient Health Questionnaire (PHQ-8) for the assessment of depression: results from 27 countries in Europe. *The Lancet regional health. Europe*, *31*, 100659. https://doi.org/10.1016/j.lanepe.2023.100659

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Chaves, C., Castellanos, T., Abrams, M., & Vázquez, C. (2018). The impact of economic recessions on depression and individual and social well-being: The case of Spain (2006–2013). *Social Psychiatry and Psychiatric Epidemiology*, 53(9), 977–986. https://doi.org/10.1007/s00127-018-1558-2

Chiao, J. Y., & Blizinsky, K. D. (2010). Culture–gene coevolution of individualism–collectivism and the serotonin transporter gene. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1681), 529–537. https://doi.org/10.1098/rspb.2009.1650

Fried, E. I., Nesse, R. M., Zivin, K., Guille, C., & Sen, S. (2021). Depression is more than the sum score of its parts: Individual DSM symptoms have different risk factors. *Psychological Medicine*, *44*(10), 2067–2076. https://doi.org/10.1017/S0033291713002900

Jorm A. F. (2000). Mental health literacy. Public knowledge and beliefs about mental disorders. *The British journal of psychiatry : the journal of mental science*, *177*, 396–401. https://doi.org/10.1192/bjp.177.5.396

Halbreich, U., Borenstein, J., Pearlstein, T., & Kahn, L. S. (2007). Clinical diagnostic criteria for premenstrual syndrome and guidelines for their quantification for research studies. *Gynecological Endocrinology*, *23*(3), 123–130. https://doi.org/10.1080/09513590601167969

Haley, F., Andrews, J. & Moghaddam, N. Advancements and Limitations: A Systematic Review of Remote-Based Deep Learning Predictive Algorithms for Depression. *J. technol. behav. sci.* (2024). https://doi.org/10.1007/s41347-024-00457-z

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., ... & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/S0033291702006074

Kirmayer, L. J., & Sartorius, N. (2007). Cultural models and somatic syndromes. *Psychosomatic Medicine*, *69*(9), 832–840. https://doi.org/10.1097/PSY.0b013e31815b7635

Kleinman, A. (1988). *Rethinking psychiatry: From cultural category to personal experience*. Free Press.

Kohrt, B. A., Mendenhall, E., & Brown, P. J. (2022). Cultural concepts of distress and psychiatric disorders: Literature review and research recommendations. *International Journal of Epidemiology*, 43(2), 365–406. https://doi.org/10.1093/ije/dyt227

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2001). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Kuehner C. (2017). Why is depression more common among women than among men?. *The lancet. Psychiatry*, *4*(2), 146–158. https://doi.org/10.1016/S2215-0366(16)30263-2

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., & Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. *American journal of epidemiology*, *157*(2), 98–112. https://doi.org/10.1093/aje/kwf182

Matcham, F., Barattieri di San Pietro, C., Bulgari, V., de Girolamo, G., Dobson, R., Eriksson, H., ... & Hotopf, M. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry*, *19*(1), 72. https://doi.org/10.1186/s12888-019-2049-z

Matcham, F., Barattieri di San Pietro, C., Bulgari, V., de Girolamo, G., Dobson, R., Eriksson, H., ... & Hotopf, M. (2022). Longitudinal remote measurement of depression in major depressive disorder: The RADAR-MDD study. *Journal of Affective Disorders*, *296*, 567–575. https://doi.org/10.1016/j.jad.2021.09.056

Mullis, R., & Attwell, C. (2022, December 6). Cost of living and depression in adults, Great Britain: 29 September to 23 October 2022. Office for National Statistics. Retrieved February 5, 2023, from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/mentalhealth/articles/costoflivinganddepressioninadultsgreatbritain/29septembertо23october2022#:~:text=Prevalence%20of%20moderate%20to%20severe%20depressive%20symptoms,-The%20presence%20of&text=The%20estimates%20reported%20in%20this,depressive%20symptoms%20(Figure%201).

Onnela, J. P., & Rauch, S. L. (2023). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, *48*(1), 1–3. https://doi.org/10.1038/s41386-022-01420-0

Parker, G., Gladstone, G., & Chee, K. T. (2001). Depression in the planet's largest ethnic group: the Chinese. *The American journal of psychiatry*, *158*(6), 857–864. https://doi.org/10.1176/appi.ajp.158.6.857

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Peters, U., & Carman, M. (2024). *Cultural Bias in Explainable AI Research: A Systematic Analysis*. *Journal of Artificial Intelligence Research, 79*, 971–1000. https://doi.org/10.1613/jair.1.14888

Pickett, K. E., & Wilkinson, R. G. (2015). Income inequality and health: A causal review. *Social Science & Medicine*, *128*, 316–326. https://doi.org/10.1016/j.socscimed.2014.12.031

Ryder, A. G., Yang, J., & Heine, S. J. (2008). Somatization vs. psychologization of emotional distress: A paradigmatic example for cultural psychopathology. *Online Readings in Psychology and Culture*, *10*(2). https://doi.org/10.9707/2307-0919.1080

Salk, R. H., Hyde, J. S., & Abramson, L. Y. (2017). Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological bulletin*, *143*(8), 783–822. https://doi.org/10.1037/bul0000102

Simon, G. E., Goldberg, D. P., Von Korff, M., & Üstün, T. B. (2002). Understanding cross-national differences in depression prevalence.

Psychological Medicine, *32*(4), 585–594. https://doi.org/10.1017/S0033291702005457

Summerfield, D. (2008). How scientifically valid is the knowledge base of global mental health? *BMJ*, *336*(7651), 992–994. https://doi.org/10.1136/bmj.39513.441030.AD

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., ... & Fava, M. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. *American Journal of Psychiatry*, *163*(1), 28–40. https://doi.org/10.1176/appi.ajp.163.1.28

Van de Velde, S., Bracke, P., Levecque, K., & Meuleman, B. (2010). Gender differences in depression in 25 European countries after eliminating measurement bias in the CES-D 8. *Social Science Research*, *39*(3), 396–404. https://doi.org/10.1016/j.ssresearch.2010.01.002

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Sage Publications, Inc.

Williams, D. R., González, H. M., Neighbors, H., Nesse, R., Abelson, J. M., Sweetman, J., & Jackson, J. S. (2007). Prevalence and distribution of major depressive disorder in African Americans, Caribbean blacks, and non-Hispanic whites: results from the National Survey of American Life. *Archives of general psychiatry*, *64*(3), 305–315. https://doi.org/10.1001/archpsyc.64.3.305

Appendix 1: Journal submission guidelines

https://link.springer.com/journal/41347/submission-guidelines

Appendix 2: Ethical Approval

DPAP Committee: 24/10/2023
Supervisor: Dr Jacob Andrews
Applicant: Mr Fintan Haley
Project ID: 3126

Project Title: "The Development and External Validation of a predictive model of Depression using passive smartphone data."

Dear Fintan,

The committee is pleased to confirm that the above study now has approval on the basis of your application and any subsequent clarifications. You must conduct your research as described in your application, adhere to all conditions under which the ethical approval is granted, and use only materials and documentation specified in your application.

If you need to make any changes (for example to extend your data collection timeframe, change the mode of data collection, or the measures being used), you must create and submit an Amendment Form. To do this, select the 'Create Sub Form' option from the Actions Menu on the left-hand side of the page in the online system and then select 'Amendment Form'.

With best wishes

Katy and Jen

Chairs of the Mental Health and Clinical Neurosciences Research Ethics Sub-committee

Appendix 3: Poster for journal paper

Forecasting Depressive Symptom Deterioration Using Wearable Sensor Data and LSTM Models: A Longitudinal Analysis from the RADAR-MDD Study

Fintan James Haley, Nima Moghaddam | Trent Doctorate in Clinical Psychology
Alexander Turner | Department of Computer Science, University of Nottingham
Jacob Andrews | NIHR MindTech Medtech Co-operative, University of Nottingham

Introduction

Major Depressive Disorder is a prevalent and debilitating mental illness affecting millions worldwide. Early detection of symptom deterioration is critical for timely intervention, yet traditional clinical approaches are often reactive and resource-intensive.

Recent advances in wearable technology and machine learning (ML) offer promising avenues for scalable, passive mental health monitoring. However, most models focus on detecting current symptoms rather than predicting future depressive episodes.

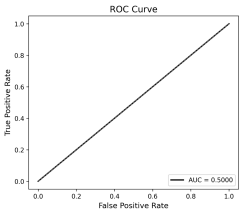
Aims:
This study leverages the RADAR-MDD dataset to evaluate whether an LSTM neural network can forecast clinically meaningful symptom deterioration using passive data from wearable sensors.

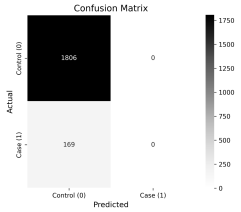
Results

Dataset:
- 1,975 timepoints.
- 91.4% negative class, 8.6% positive class.

Model Performance:
- Accuracy: 91.4%
- Precision/Recall/F1 (Positive Class): 0.00
- AUC-ROC: 0.50 (no better than chance)
- Specificity: 1.00

Key Issue:
Model failed to detect any deterioration events (zero true positives).





Methods

Data Source:
RADAR-MDD longitudinal study (N = 623).
Data Collected:
- Passive data: Sleep, step count, activity, and heart rate via Fitbit.
- Active data: Bi-weekly PHQ-8 questionnaires.

Outcome Definition:
Reliable deterioration = ≥5-point increase in PHQ-8.
Preprocessing:
- Min-max normalization of features.
- Mean imputation for missing values.
- Class weighting to address imbalance (8.6% positive cases).

Model:
LSTM model (2048 units, dropout = 0.5, early stopping after 20 epochs).
Train/validation/test split: 70/15/15.
Evaluation Metrics:
Accuracy, Precision, Recall, F1 Score, AUC-ROC.

Discussion

Causes for lack of predictability: Severe class imbalance and sparse data. Cross-cultural variability in PHQ-8 symptom reporting. Features modeled without accounting for individual baselines.
Future directions : Consistent data collection, idiographic modeling, deviation-from-baseline feature engineering, Culturally adaptive designs.

Reference: Haley, F.; Andrews, J.; Moghaddam, N. Advancements and Limitations: A Systematic Review of Remote-Based Deep Learning Predictive Algorithms for Depression. J. Technol. Behav. Sci. 2024, 1–14, https://doi.org/10.1007/s41347-024-00457-z.