Summary of corrections

Addressing General Comments

1. The thesis was clearly written. However, there is a need to proofread for minor errors. Typos, improvements and grammar minor points: Many of those have been annotated in the pdf by the internal examiner and will be sent to Vanisha for her to proceed with the corrections.

- Worked through pdf and corrected spelling, grammar and minor improvements throughout thesis.

2. The study needs to prove the developed ensemble model outperforms the other efficient ensemble models such as Random Forest, Light Gradient boosting and extreme gradient boosting using Precision Recall AUC. For that, the same datasets need to be used for comparison of those approaches.
3. The study should clearly show the contribution to knowledge given that there exist other studies using similar datasets with superior performance (Azhagesan et al)

- My research was to carry out a comparison of models and then investigate the development of an ensemble method. Creating a model that outperforms other models would fall outside the remit of this MPhil. After consulting with Grazziela have added the following information which highlights the novelty of the ensemble model and how it compares with other available models.

Changed research question 4:
After the construction of novel ensemble models it is important to investigate whether they perform better than the individual parts. Two performance methods commonly applied are cross validation and unseen test data.

End of 5.6 Chapter summary:
Our method is able to generalise to unseen data as well as previous studies while requiring less input information than other models. This is an important point for a models application to new organisms for which little to no information is available.

End of 6.2 General discussion:
While the AUCs of our model sit in the middle of other studies that use a large number of training organisms to improve generalisation [REFS] our model requires less input information making it an ideal candidate for the prediction of genes in new or hard to culture bacterial organisms.

Chapter Specific

Introduction: Section 1.2 needs a concluding paragraph reiterating the gaps led to your aims and objectives. The student needs to stress the contributions to knowledge produced.

- Amended paragraph on page 11:
The importance of essential genes and the minimal genome concept has accelerated the advancement of experimental methods within biology which has, in turn, reduced the cost of genome sequencing [24]. The advantage of this is that there are now more genome and essentiality data available for use in computational methods. As a result the application of machine learning to biology has been steadily increasing. While many different pipelines and algorithms are being applied to the important task of gene essentiality prediction [24–27]. Majority of methods are highly specific, being targeted at predictions within the same species or closely related organisms. Or they require additional information such as gene function, this information is available for model organisms and while it can be predicted for unknown genes. The accuracy of this prediction depends relationship between the organisms.

Chapter 2: In the summary section, it would be beneficial to reiterate the related work gaps and opportunities that led to your research in the subsequent chapters.

- Added into chapter summary page 24:
Most current computational methods use small subsets of closely related organisms or require the use of complicated features, as they accuracy of these predictions depend on the organisms being closely related, their accuracy decreases when applied to non-model or unknown organisms.

Chapter 4: It would be interesting to see in the appendix a list of features from Figure 4.2 that tend to appear together, to understand whether there could be synergies. It is also important to discuss the impact of correlated variables to the feature selection methods chosen. For example, LASSO at times will select different sets of features if ran on a same dataset multiple times due to the probabilistic character given by correlated features.

- Added to 4.8 Discussion of feature subsets:
While the approach we chose showed us that for some features, which share biological and or functional similarity, the models selected one of these gene features more frequently than the other, we did not look into feature relationships before carrying out feature selection.

As feature relationships can affect some classifiers it is possible that they will have impacted some feature selection methods. Correlated features can affect classifiers in different ways. For example with recursive feature elimination (RFE) and logistic regression, if two correlated features are both present their importance to the model would be low. But if one feature is removed the importance score of the other would need to increase. This would require feature importance to be recalculated after each removal step.

For RFE with random forest or other tree-based models if the correlated features are both useful for prediction, which one is selected is essentially random choice. In this case the feature selection method might contain highly correlated redundant features. While this may not affect the models prediction accuracy it also does not allow us to gain any information about feature importance. Future work into investigating multicollinearity in our feature set may allow us to gain an insight into which redundant features can be excluded before feature selection.

An interesting line of further investigation from this point would be to look into features which frequently appear together across the different feature selection methods and their impact on the models created.

Chapter 5: It would like to see the results of the proposed ensemble for the same datasets as those used by the literature in Table 5.3. It is hard to judge how effective the ensemble is when different datasets from those used in related work are used. We suggest re-running the experiments for at least one related work and further discussion on the obtained results, if the data are available. (for instance, Deng et al). Similarly, from the results presented in Figures 5.4 - 5.6 and Table 5.2 showing the performance of the ensemble model, it was observed that Random Forest (an ensemble classifier) has similar performance with the developed model by this study across all the experiments performed. Also, when compared to a related study (Azhagesan et al) in Table 5.4, the average AUC of the developed model was inferior, as the datasets used are slightly different from what we understood.

- 5.5.3 Discussion of ensemble validation:
We were unable to validate our models by running them on the same datasets as previous studies as the data within these studies was available to us within the time-frame of the project. While the papers do contain the organism names, they do not contain the NCBI accession or version IDs which allow to use the exact data. Where DEG has been used previous versions of the database were unavailable to us, as was the version history.

Conclusions: It would be interesting to see a discussion on how your methods can aid understanding of feature importance for determining essential genes.

- Added to 6.2 General discussion:
While our methods allowed us to see which features were more frequently selected they do not allow us to gain meaningful insight into feature importance. As understanding this aspect of machine learning could help us improve our ensemble models it is an important line of future research.

Important aspects regarding the experiments to be discussed and further addressed:
The study chose not to address the data imbalance in the training data because this is a natural phenomenon in the target organisms. Unfortunately, ML models used by the study do not learn class distribution in the data when training the classifier which creates a bias for the negative data points (Non-essential genes) in the validated or unseen data. This would have been evident should Precision- Recall AUC or Matthew's correlation coefficient or sensitivity of the analysis was reported. It would be good to see the other metrics reported to appreciate the performance of the developed model. AUROC metric used by this study is not sufficient to evaluate the model's performance given the imbalance nature of the (training and test) data used.

- Added to Chapter 6:
Discussion of evaluation matrix

As covered in section 4.4.1, we used the area under the ROC curve to evaluate our classifier models. This is commonly used for evaluating binary classifiers and is calculated using the false positive and true positive rates. A limitation of this metric is that for imbalanced classes the model may seem more useful for predictions that it actually is. This is because when dealing with imbalanced classes the classifier can predict everything as the larger class and still give good performance when measured using the ROC metric. In these cases precision-recall curves can provide a better insight into performance as it is a measure of how good the model is at predicting the positive class, in our case it would be the essential genes as the minority class.

Our research was based on previous studies in which the ROC curve metric was applied to measure the performance of models with balanced classes. However during the project it was decided that for laboratory experiments identifying the non-essential genes is equally important for targeting genes and as a result we choose not address the class imbalance. This meant our classes had a ratio of 1:6, essential to non-essential genes. Due to the context of the research changing, applying the Precision-Recall AUC metric would have provided a deeper insight into the performance of our models. As this cannot be addressed with the time and funding available it should form the start point of any future work.

Although the study aims to develop a simple classifier for essential genes in non-model organisms, however, the accuracy of a classifier is more important than its simplicity. Recent studies (Campos 2019, Aromolaran 2020) have shown that the application of diverse features improves prediction accuracy for essentiality prediction. The performance of the developed model would be improved if features from other categories such as ontology were considered, given the knowledge that most bacteria have well annotated GO terms (This is a suggestion for discussion with the supervisor).

- We chose not address this point as while GO terms have to be predicted for some organisms, they only experimentally proven for model organisms and depend on evolutionary relationship between organisms. The prediction programs require information such as which organism the sequence is from or which one is it most closely related to. The Gene Ontology resource requires you to input the name of the gene and the organism species.

The problem identified in literature about the use of small training dataset has been addressed by some studies particularly by GEPTOP 2.0 (DOI: 10.3389/fmicb.2019.01236) with 37 organisms and interesting results. This was also confirmed when the author stated that the AUCs of our model sit in the middle of other studies that use a large number of training organisms to improve generalisation (Pg 68).

- GEPTOP was only updated at the end of 2019. Before then they only used 19 organisms from DEG to train their model. My work was carried out on a more up-to-date version of DEG and contains 40 organisms.

Addressed in End of 6.2 General discussion:
While the AUCs of our model sit in the middle of other studies that use a large number of training organisms to improve generalisation [REFS] our model requires less input information making it an ideal candidate for the prediction of genes in new or hard to culture bacterial organisms.

Minor points to be addressed:
The type of scaling used should be described (Min-max or Z-score etc.)

-Added to 5.2.2 Method:
The features were scaled before training using a Min-Max scaler with a range of -1 to 1.

Added to 5.5.1 Method:
The features were scaled before training using a Min-Max scaler with a range of -1 to 1.

Is there any assumption made on the negative samples. Are they experimentally confirmed as OGEE is? These should be clearly stated.

- Covered in chapter 3.3 Database of Essential Genes:
Within DEG are also sub-databases which contain non-essential coding genes, these can be inferred from the set of essential genes or based on the original source. This information can come from transposon mutagenesis studies which determine non-essential genes first while the essential genes are inferred.