**GROUP REPORT**

Table

Description automatically generated

# ***AGREED STRUCTURE***

**INTRODUCTION**

Project Objectives (50 words)

Data Description (50 words)

Approach and Contributions (100 words) *Summarise your approach and contribution*

**METHODS**

- *Explain the methods and why we choose with references*

1. Sampling (100 words)
2. Feature Selection (100 words)
3. Machine Learning Models (100 words) *summarise different ML models, then for Strengths and Limitations we can add more details*

(500 words) *Put results in appendix, summarise it here and explain everyone’s approach, more details in individual report*

1. Logistic Regression Approach (166 words)
2. SVM Approach (166 words)
3. DT Approach (166 words)

**DISCUSSION**

Comparison (300 words) *Comparison TABLE between outputs from different approaches*

Strengths and Limitations (200 words) *Identify limitations and strengths of the tested methods*

Responsible use of methods (300 words) *Considerations on explainability/ responsible use of methods*

**CONCLUSION**

Summary (100 words)

Improvements (100 words)

**WORK TO COMPLETE**

# INTRODUCTION (200w)

*Objective*

Wideband absorbance (WBA) tests for hearing conditions by measuring the middle ear’s absorbance across multiple frequencies, where this was previously measured at only one frequency. This change produces effective results but more complex output, making it difficult for audiologists to process in a clinical setting. The primary purpose of this project is to address this difficulty. We have applied statistical and machine learning techniques to WBA test results to determine the most ‘influential’ frequencies for diagnosing hearing conditions. This may provide audiologists with a starting point for analysing WBA test output.

*Dataset*

The dataset contained 239 observations of wideband absorbance test results that span across 107 different frequencies, being the data fed into our models (*n* = 239; *k* = 107). The output variable was whether the hearing test passed or failed, designated by the binary *OverallPoF* field (1 = fail; 0 = pass). Other features such as Y-admittance, Phase and demographics were also received, but the focus of our project was on the absorbance.

*Contribution*

The key contribution from this project is to provide audiologists (and other end-users of WBA tests) the best frequency, and a small range of frequencies, that are the most effective for diagnosing hearing conditions. This will hopefully allow for a faster, directed approach to analysing the test output, or at least provide a starting point.

# METHODS

**\*\*ADD MORE HIGH IMPACT REFERENCES!!**

## SELECTION OF APPROACHES (3 MARKS) – REFERENCES TO EXPLANATIONS APPENDIX B

### 1- SAMPLING

Considering our data is imbalanced, we applied sampling methods discussed in [6] to create a balanced distribution of the classes. One method, NearMiss, is an undersampling method that removes the instances of the majority class closest to the minority. An opposite approach is SMOTE, an oversampling technique that creates artificial data based on the similarities between the minority class. Combined methods, such as SMOTETomek and SMOTEEN, applies undersampling after SMOTE to clean out overlapping data. These sampling methods tend to produce balanced samples which increases classifier’s performance.

2 - FEATURE SELECTION

In order to fulfill the objective of choosing top frequencies, we experimented with various feature selection methods. According to [8], these methods are broadly categorised into filter, wrapper and embedded. Filter methods rank feature importance using statistical criteria, such as variance or chi2 and therefore independent of any learning algorithms. In contrast, wrapper methods such as RFE and SFS rely on learning algorithms to iteratively evaluate feature subsets and choose the best subset. Embedded methods similarly utilise learning algorithms, but specifically those that intrinsically performs feature selection, eliminating the need of evaluating one subset at a time. Regularization models and decision trees are popular approaches of embedded methods.

### 3- MODELS(Everyone dot points) (500 words) (SVM 150 words; RFE; logistic regression 150 words; DT + RF 200 words)

WHAT CLASSIFIERS WERE CHOSEN INITIALLY AND WHY?

* One sentence describing the model in general. （what）
  + - Why we select xxx model (why)

Our task is to select optimal frequencies in predicting pass or fails of a WBA test. We are provided with a relatively small dataset containing high number of features. Keeping our task and the data in mind, we researched for suitable binary classification algorithms.

#### Decision trees

Decision trees (DTs) are interpretable machine learning models that make predictions from decision-based rules and information learned from features (in our project, frequencies) in the dataset [DT\_REF]. We chose DTs primarily for their interpretability and presentable visualisation of the choices made by the model. DTs also work well with non-linear datasets (which our project is addressing). Our best-performing DT classifier used SMOTE over-sampling to balance the dataset, and GridSearchCV was used to optimise the parameters of the model. The top features, accuracy and loss are shown in the table below.

|  |  |  |
| --- | --- | --- |
| Top influential frequencies for measuring absorbance *(Hz)* | Test recall | Loss |
| **1296**, 1155, 4117, 1943, 2519, 3363, 280, 1455 | 0.89 | 0.06 |

#### Random Forests

Random Forests (RFs) are an ensemble of decision trees, aggregating the predictions from decision trees to make a new prediction. RFs typically provide a more accurate and stable prediction than a decision tree, however they can be more difficult to interpret. As with DTs, RFs can be used for both classification and regression problems and can detect non-linearity. To improve the interpretability of our RF model, we used SHAP – a method for clearly explaining the impact of features on predictions. Figure xxx shows that the top 3 frequencies on test set are 1189 Hz, 771Hz, and 1155Hz, and indicate that for 1189Hz, most of the low values (blue dots) contribute positively to the predicted output. In other words, if the 1189Hz of a single data instance is low, its chances of failing hearing tests are greatly increased, while high values (red dots) will decrease the probability of being diagnosed with hearing loss.

#### LOGISTIC REGRESSION

Logistic regression is a widely used method to predict the probability of a binary event. This method is selected as they are one of the easiest and to implement and modify, in addition to being computationally cheap. Other than this, Logistic regression is a Whitebox method, meaning the results are interpretable: Coefficients can be used to understand the importance of each feature towards the dependent variable [LR1]. Along with the regression, several regularization methods can be used to prevent over-fitting of the data [LR2].

Two approaches were used for Logistic regression, first being multinomial logistic regression using all frequencies for model creation and the second being simple logistic regression to create individual models per frequency. Regularization parameters, namely L1, L2 and Elastic Net, were then tested on the models for both approaches. For the multinomial approach, an additional variance thresholding feature selection was also attempted.

Training sets with SMOTEENN samples results in the highest overall performance for both approaches. Multinomial LR approach returns 88% and 85% recall on the training and test set, using MinMax scaling and L1 or lasso regression, while the Simple LR approach returns 87.51% and 92.78% recall on the training and test set, using L2 or ridge regression as the regularization parameter.

#### SUPPORT VECTOR MACHINES

SVM maps data in a hyper-dimensional feature space and tries to maximise the distance between the classes. Burges [7] proved that this mechanism allows SVM to avoid overfitting when features are numerous, which suits our case.

Several feature selection methods were conducted in combination with the SVM classifier. These included RFE+Linear estimator, RFE+Random Forest, Sequential feature selection + KNN (forward and backward), ANOVA F-test, and embedded methods with L1 and L2 regularisation.

On average, training sets with SMOTEENN and SMOTETomek samples performed the highest in accuracy [see table]. In terms of selection methods, wrapper and embedded methods tended to perform better where the RFE-SVM performed the highest in the wrapper methods category with 91% and 89% recall on the train and test set respectively (C2.1).

For embedded feature selection, both L1 lasso and L2 ridge regularisations were tested across all samples. Using SMOTETomek samples, ridge penalty was able to effectively reduce the features to 13 frequencies. This combination was found in appendix C2.2 to result in the best performing SVM classifier, scoring 91.67% recall on the validation set. The C2.2 table also shows the model generalises well, with minimal difference in performance between training and validation. The selected 13 frequencies were shown to be less correlated with each other in C2.3.

# DISCUSSION (8 MARKS) (800 words)

## COMPARISON OF MODELS (3 MARKS) (Jessie) (300 words)

* Choice of metrics for scoring models?
* Use recall.... because …...
* 1. based on XXX 2. for the same xxx, …...

BIG SUMMARY TABLE OF RESULTS

We choose the recall score on the validation set as the metric for comparing the models in each machine learning algorithm. Since the project aims to select effective frequencies to diagnose ears with conductive conditions, we propose the recall score that measures the performance of our model in correctly identifying hearing loss among all true hearing loss children should be used to select the model. We are particularly concerned about hearing-impaired children who pass the hearing tests.

Firstly, we select the best model in each machine learning method using the recall on the validation set. If several models provide the same validation recall score, we will pick the model with the smallest absolute difference (diff = training recall – validation recall) to avoid overfitting. After that, we compare models from different algorithms based on the recall on the test set.

The model results are shown in table xx. According to table xxx, we find that most algorithms perform well on the SMOTEENN sampling method, which is consistent with Estabrooks, Jo and Japkowicz’s [reference number] finding that combining both oversampling and undersampling helps to deal with class-imbalance problem.

According to table xxx, five out of six models indicate fa1296 as an important frequency. In addition to that, at least 50% of the models show that fa1334, fa1090, fa1155 are efficient frequencies. We find that most of the important frequencies are derived from the 1000hz to 2000hz interval, which is consistent with the box plot of the data distribution (graph XXXX). Within this interval, the pass and fail groups are more easily distinguishable. After comprehensive consideration of recall scores, top frequencies results, and model interpretability, we believe that the decision tree is our final best model.

## LIMITATIONS AND STRENGTHS OF MODELS (2 MARKS) (Everyone) (200 words)

* What are the limitations of your models? (Tip google pros and cons of your models)
* Does the model overfit or underfit at higher frequencies?

DTClassifier, in general:

|  |  |
| --- | --- |
| **Pros** | **Cons** |
| Clear visualisation of models’ thought process, improving the explainability of the model. | Decision trees can tend to overfit data. |
| Decision trees also handle non-linear datasets efficiently. | Decision trees can be heavily affected by small amounts of noise. |
| Decision trees are robust against outliers in data. | Decision trees are generally not suitable for larger datasets. |

Decision trees have the benefit of presenting a clear and visual thought process of the decisions made by the model. This enhances the explainability of the model, increasing the confidence of end users (Robyn; other audiologists) in the model

RF: (remember to reword below)

[It is a difficult tradeoff between the training time (and space) and increased number of trees.

Random forest is like a black box that we have little control over.

it is sensitive to parameters, noise, environmental changes and other factors.]

Decision trees tend to be overfitted in prediction, random forest reduces the degree of overfitting through voting, but its prediction is still overfitting compared to linear model, which is characterized by good matching of existing data. But very conservative with unknown data, and high probability of false negative error.

Logistic Regression:

|  |  |
| --- | --- |
| **Pros** | **Cons** |
| Easy and less time-consuming to implement, train, and modify | **Logistic regression requires no to low multicollinearity among features** |
| **Coefficients can be used to measure importance of each feature** | **Requires sufficiently large sample size.** |
| Coefficients also have positive or negative direction of association | Logistic regression can't describe complex relationships between dependent and independent variables |
| Results are highly interpretable, as coefficients provide quantitative measure for each feature | Often outperformed by more powerful algorithms |
| **Regularisation techniques could be used to avoid overfitting** | Assumption of linearity between dependent and independent variables |

Logistic regression assigns coefficients to features, allowing interpretation of the importance of each feature. However, logistic regression usually requires a large amount of data to avoid overfitting, but built-in regularization parameters could be used to tackle this issue. Further

SVM

|  |  |
| --- | --- |
| **Pros** | **Cons** |
| Highly effective in high dimensional spaces | Does not work well when there are large overlaps in target classes |
| Works well when the number of dimensions is greater than the number of samples | Requires extensive training which may not be suitable when the number of observations becomes large |
| SVM is computationally efficient since it uses less memory than traditional methods | Performs poorly when the number of features is greater than the number of training samples |
| Works well with unstructured data or when there are insufficient information about the data | Difficult to choose an appropriate kernel function for the problem |
|  |  |
|  |  |

To ensure transparency, we have limited our SVM Classifiers to the linear kernel. This is ineffective when there is high overlap between target classes, that is, when our data is not linearly separable. If this is true, non-linear but less-explainable kernels might lead to SVM classifiers with better performance.

## MODEL EXPLAINABILITY (3 MARKS) (David) (300 words)

DISCUSSION ON HOW EXPLAINABLE EACH MODEL IS

The goal of the project is to aid with decision-making in a clinical setting while fulfilling the responsible building and deployment of AI. The preliminary choice of classifiers was weighted on the potential explainability of the models for our clients. In general, if it is difficult to explain the processes and results of our models, it would be equally difficult for our models to be adopted by our clients.

In a recent survey on explainable AI, A. Barredo discusses the trade-off between interpretability and performance of models. He proposes that both *Decision Trees* and *Logistic Regression* are high in model interpretability, however it compromises on model accuracy. On the contrary, *SVM* models held relatively high model accuracy though fell short on model interpretability [ref XAI].

As an example, decision tree models create an appealing tree diagram that shows the decision made at each node, and the ranked importance of the features used in that tree. Similarly, logistic regression produces coefficients that can be used to interpret the impact of its features on the outcome.

To justify using SVM, a linear estimator was adopted instead of the non-linear alternatives. Unlike logistic regression and decision trees, Linear SVM produces coefficients that define the boundary of separation (hyperplane) as opposed to the importance of features. However, it can be interpreted that the relative sizes of the coefficients represent the importance of the feature. Where models were less interpretable, such as the Random Forest Classifier, we calculated and presented the SHAP values. For example, we used force plot and decision plot to show how a single instant is predicted and adopted a summary plot to illustrate global interpretation (appendix C.3.2).

# CONCLUSION (2 MARKS) (200 words) (ANDY + … + … +)

* Sampling methods are they effective?
* Overall key frequencies (maybe
* Frequency mostly in 1000-1400 does it makes sense, yes based on robyn
* Explainability – DT over LR and SVM
* Caveat -

## SUMMARY OF RESEARCH (ALL)

## FUTURE IMPROVEMENTS (David)

The next steps of this project would be to produce a tool for end-users based on our best-performing model. The tool would automatically pull data from the output of WBA/WBT testing machines, run the data through the model, and produce an output of confidence scores for a range of conductive conditions.

We would also continually seek more datapoints to re-run model testing and update the live model as necessary, as the number of observations (n=239) makes it difficult to rely on results.

# REFERENCES

[MyBib](https://www.mybib.com/#/projects/0XNmxp/citations)

DT reference: <https://core.ac.uk/download/pdf/42982626.pdf>

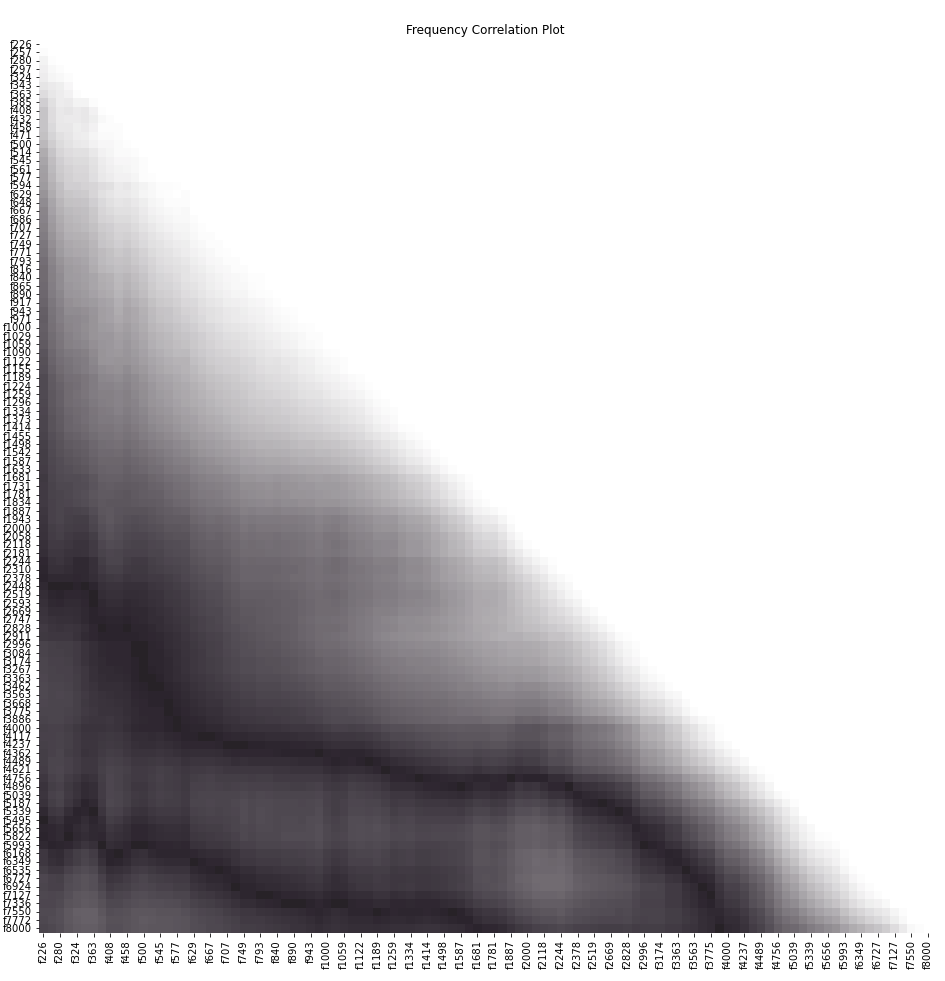
|  |  |
| --- | --- |
| [DT\_REF] | ,. N. A. A. B. F. S. M. S. S. S. H. Md. Rajib Hasan, “Single decision tree classifiers' accuracy on medical data,” in *Proceedings of the 5th International Conference on Computing and Informatics*, Istanbul, 2015. |

[LR1] Christoph Molnar, “Interpretable Machine Learning,” *Github.io*, Oct. 13, 2022. <https://christophm.github.io/interpretable-ml-book/logistic.html>

[LR2] J. M, “Understanding Regularization in Logistic Regression,” *Medium*, Jan. 03, 2020. <https://towardsdatascience.com/understanding-regularization-in-machine-learning-5a0369ac73b9>

# APPENDIX

## **A1 – Data Visualisations**



Chart

Description automatically generated

Chart, bar chart

Description automatically generated

## **B1 – APPROACHES AND METHODS**

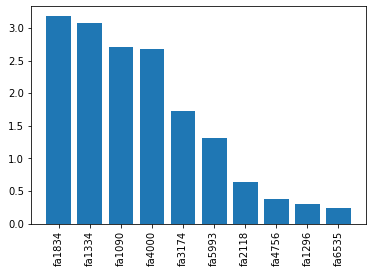
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table XXX: | | | | | |
| **Model** | **Sample** | **Top Frequencies (Hz)** | **Recall (Test set)** | **Diff** | **Note** |
| Logistic regression - Dennis | SMOTEENN | 'fa1834', 'fa1334', 'fa1090', 'fa4000', 'fa3174', 'fa5993', 'fa2118', 'fa4756', 'fa1296', 'fa6535' | 0.93 | 0.05 | Performed MinMax Scaling on the dataset and used L1 (Lasso) regularization to remove unnecessary features. |
| Logistic regression - Andrew | SMOTEENN | 865, 890, 1000, 1029, 1059, 1122, 1155, 1189, 1224, 1296, 1334, 1373, 1414, 1681, 1731, 1781, 1834, 1887, 1943 | 0.927 | 0.052 | Simple logistic regression to create a model for each frequency. Then used the frequencies from best performing models to create a single model. Used L2 (ridge) regularization without scaling to obtain final model. |
| SVM | SMOTEENN | 'fa1090' 'fa1259' 'fa1296' 'fa1334' 'fa2118' 'fa2181' 'fa2911' 'fa3174'  'fa4000' 'fa4896' 'fa6349' | 0.927 | 0.0137 |  |
| SVM | SMOTETomek | 5656  5495  5993  4237  4117  1296  771  1334  4000  727  2996  1498  280 | 0.95 | 0.06 | Selection Method is embedded. |
| Decision Tree | SMOTE | Features and importance ranking:  1296 --> 0.503  1155 --> 0.108  4117 --> 0.1  1943 --> 0.089  2519 --> 0.062  3363 --> 0.058  280 --> 0.05  1455 --> 0.031 | 0.889 | 0.0573 | {'ccp\_alpha': 0.0,  'class\_weight': None,  'criterion': 'entropy',  'max\_depth': 4,  'max\_features': None,  'max\_leaf\_nodes': None,  'min\_impurity\_decrease': 0.0,  'min\_impurity\_split': None,  'min\_samples\_leaf': 3,  'min\_samples\_split': 2,  'min\_weight\_fraction\_leaf': 0.0,  'random\_state': None,  'splitter': 'best'} |
| Random Forest | SMOTEENN | fa1189, fa771, fa1155, fa971, fa1090, fa793, fa1029, fa1122, fa1000, fa943 | 0.903 | 0.0021 | {'criterion': 'gini', 'max\_depth': 3, 'n\_estimators': 100} |
| Note: Diff is calculated as the absolute recall difference between the training set and the test set. | | | | | |

|  |  |  |
| --- | --- | --- |
| **Table xxx:** | | |
| **Frequency** | **The number of occurences** | **Model** |
| Fa1296 | 5 | Logistic regression1, Logistic regression 2, SVM1, SVM2, Decision tree |
| Fa1334 | 4 | Logistic regression1, Logistic regression 2, SVM1, SVM2, |
| Fa1090 | 3 | Logistic regression 1, SVM1, Random forest |
| Fa1155 | 3 | Logistic regression 2, Decision tree, Random forest |
| Fa1000 | 2 | Logistic regression 2, Random forest |
| Fa1029 | 2 | Logistic regression 2, Random forest |
| Fa1122 | 2 | Logistic regression 2, Random forest |
| Fa1189 | 2 | Logistic regression 2, Random forest |
| Fa1834 | 2 | Logistic regression 1, Logistic regression 2 |
| Fa1943 | 2 | Logistic regression 2, Decision tree |
| Fa280 | 2 | SVM2, Decision tree |
| Fa4117 | 2 | SVM2, Decision tree |

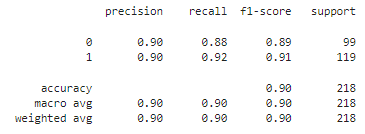
## **C1 – LOGISTIC REGRESSION RESULTS**

**C1.1 - All-Frequencies used approach (Dennis)**

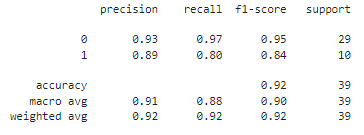
* Features selected by the lasso regularization and their importance (coefficient) ranked



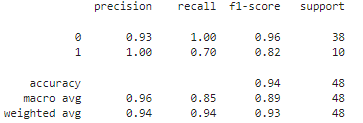
* Classification report on the training set



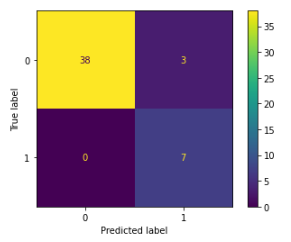
* Classification report on the validation set



* Classification report on the test set

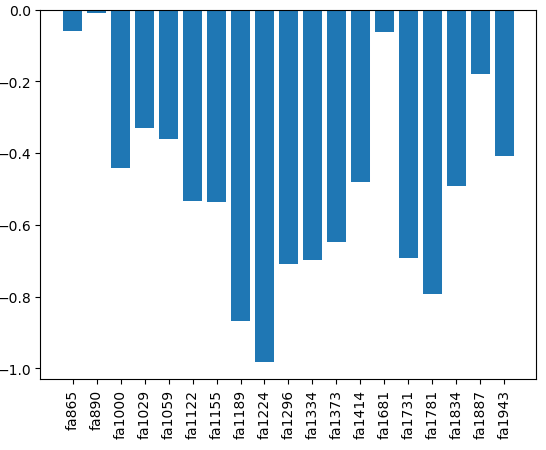


* Confusion matrix on the test set

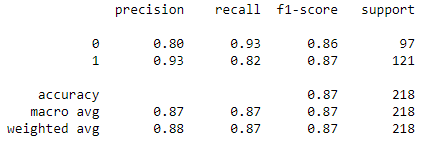


**C1.2 - Simple logistic regression before combining (Andrew)**

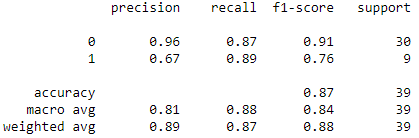
* Features used in the model



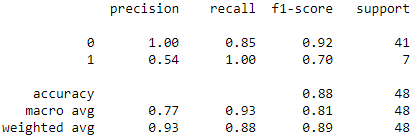
* Classification report on the training set



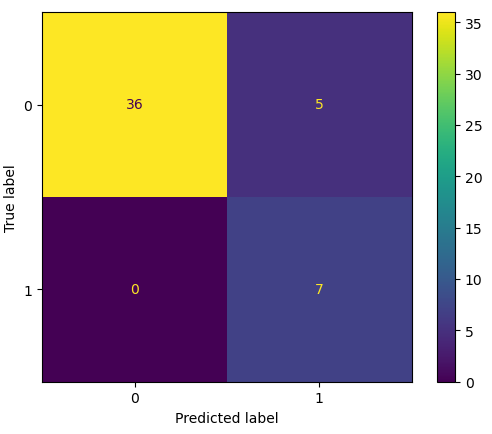
* Classification report on the validation set



* Classification report on the test set

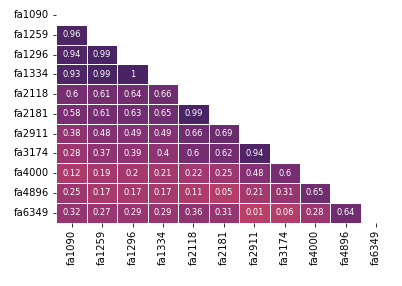


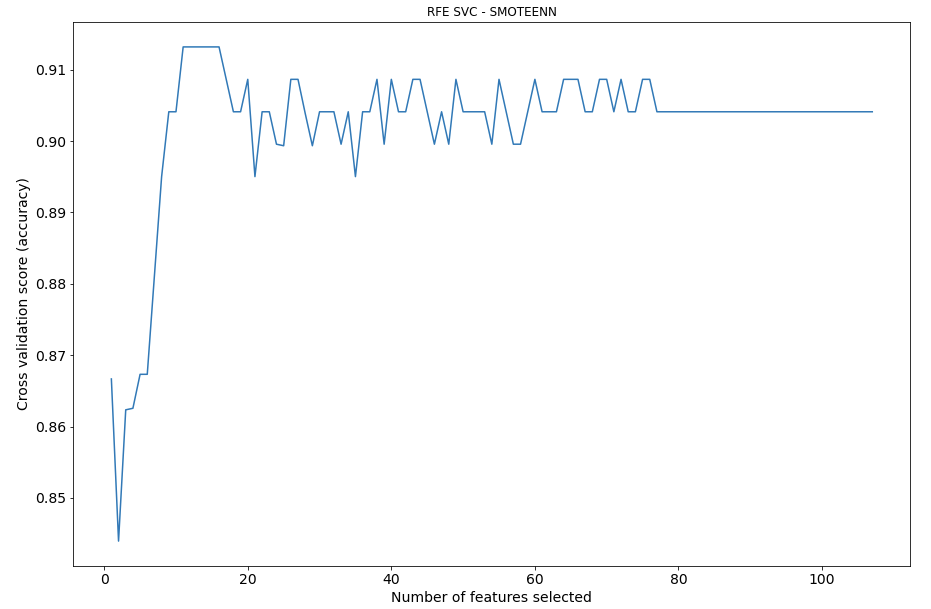
* Confusion matrix on the test set



## **C2 – SVM RESULTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sample | selector | n\_features | training\_recall | validation\_recall |
| Original | RFE SVC | 7 | 81 | 87.2 |
| NM2 | RFE SVC | 6 | 84.1 | 76.7 |
| SMOTE | RFE SVC | 16 | 89.6 | 76.7 |
| SMOTEENN | RFE SVC | 11 | 90.8 | 89.4 |
| SMOTETomek | RFE SVC | 28 | 92.9 | 86.1 |
| Original | RFECV RF | 12 | 81.8 | 87.2 |
| NM2 | RFECV RF | 9 | 84.1 | 67.8 |
| SMOTE | RFECV RF | 22 | 88.8 | 89.4 |
| SMOTEENN | RFECV RF | 68 | 93.5 | 73.3 |
| SMOTETomek | RFECV RF | 37 | 91.3 | 86.1 |
| Original | SFS KNN | 23 | 76.5 | 92.8 |
| NM2 | SFS KNN | 5 | 79.5 | 71.7 |
| SMOTE | SFS KNN | 12 | 73.5 | 60.6 |
| SMOTEENN | SFS KNN | 9 | 76.2 | 71.7 |
| SMOTETomek | SFS KNN | 39 | 91.7 | 89.4 |
| Original | SBS KNN | 5 | 79.2 | 94.4 |
| NM2 | SBS KNN | 5 | 79.5 | 77.8 |
| SMOTE | SBS KNN | 5 | 84.6 | 86.1 |
| SMOTEENN | SBS KNN | 7 | 85.8 | 71.7 |
| SMOTETomek | SBS KNN | 5 | 89.4 | 86.1 |

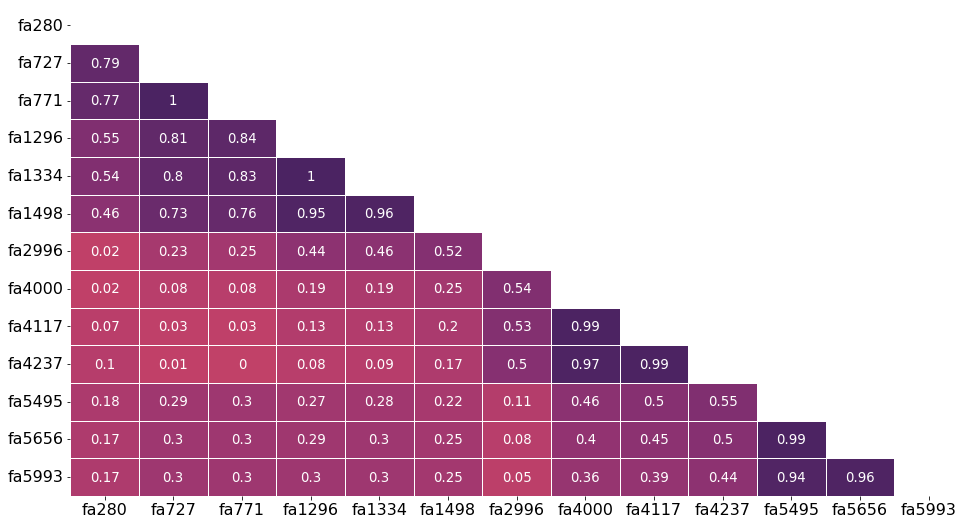




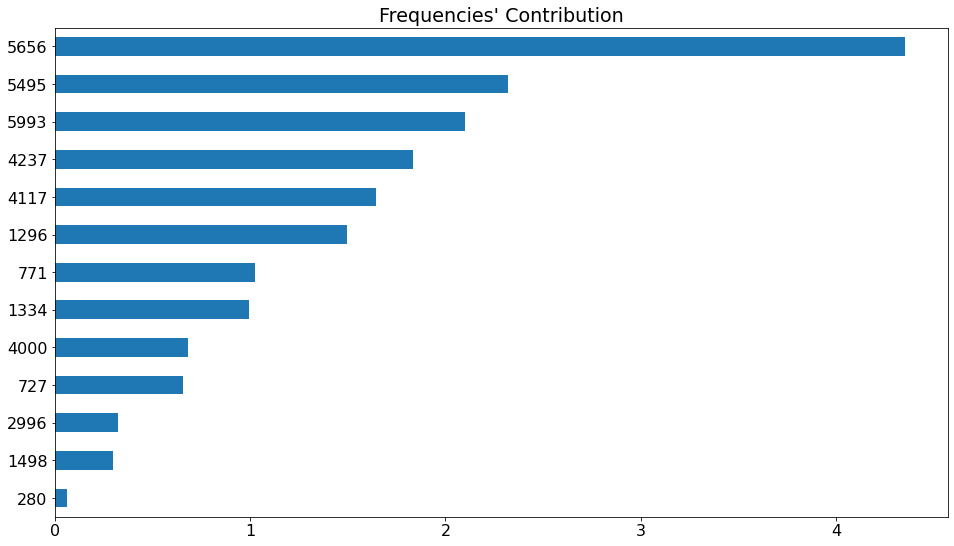
**C2.2 - SVM with Embedded Feature Selection**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | sample | regularisation | n\_features | training\_recall | validation\_recall |
| **0** | Original | l1 | 11 | 92.73 | 89.44 |
| **1** | Original | l2 | 13 | 91.19 | 73.33 |
| **2** | NM2 | l1 | 14 | 79.55 | 77.78 |
| **3** | NM2 | l2 | 7 | 86.36 | 74.44 |
| **4** | SMOTE | l1 | 21 | 92.31 | 87.78 |
| **5** | SMOTE | l2 | 25 | 95 | 76.67 |
| **6** | SMOTEENN | l1 | 14 | 96.29 | 82.22 |
| **7** | SMOTEENN | l2 | 22 | 96.6 | 66.11 |
| **8** | SMOTETomek | l1 | 12 | 95.67 | 80.56 |
| **9** | SMOTETomek | l2 | 13 | 89.37 | 91.67 |

**C2.3 - Best embedded SVM Frequency Correlation Plot**



## **C2.4**



## **C3 – TREE METHODS RESULTS**

**C3.1**

Decision tree (SMOTE over-sampling)

Timeline

Description automatically generated

**C3.2**

Random Forests

|  |
| --- |
| C3.2.1 Summary plot (SMOTEEN – SHAP values on test set) |
|  |
| C3.2.2 Force plot for a single instance |
| Graphical user interface, application  Description automatically generated |
| C3.3.3 Decision plot for a single instance |
| Chart, line chart  Description automatically generated |

# ARCHIVE

CHALLENGES(100 words)

There were 4 key challenges with this dataset that were all factors when choosing and refining our given models:

1. Imbalanced dataset: the *OverallPoF* field contained 201 passes and 38 fails, creating a very low representation of fails.
2. Small dataset: As mentioned, only 239 records were provided. This heavily limited the models we could utilise and was a factor toward choosing simpler models to work with. Neural networks and other deep learning methods were never feasible.
3. High dimensionality: The large number of features (107 frequencies) meant that we had to consider the likelihood of complex models overfitting data and the difficulty that would be faced by clustering approaches.
4. Multi-collinearity: the nature of the problem meant that there would possibly exist collinearity between some of the frequencies (input variables). The task of our project may be viewed as seeking the lowest average collinearity between a given frequency and all other frequencies.

Original sampling

To handle our imbalanced data, we utilised sampling methods that modifies the dataset and create a balanced distribution of the classes. This is commonly achieved by oversampling the minority class or undersampling the majority. According to He and Garcia [6], the resulting sample tend to improve the classifier accuracy. Hence, we tried out several different premiere sampling methods: NearMiss, SMOTE, SMOTEENN and SMOTETomek. NearMiss is an undersampling method that removes the instances of the majority class closest to the minority. He and Garcia [6] suggested NearMiss version 2 can provide competitive results in imbalanced learning. For oversampling, they discussed SMOTE as a powerful technique that synthetically creates artificial data based on the similarities between existing minority data. He and Garcia [6] further explained a combined method, which applies undersampling to clean out inherent overlapping created by SMOTE. As examples, they mentioned SMOTETomek and SMOTEENN, and showed that these methods will produce better defined class clusters compared to SMOTE, potentially leading to better classifier performance.

Andy’s FS

Machine learning models often suffer from a phenomenon called the ‘curse of dimensionality’. This occurs when there is a high dimensional data set containing irrelevant or redundant features that prevents fitting a model with an acceptable bias-variance trade off.

To overcome this problem, feature selection may be applied to reduce the dimensionality of the data prior to model fitting. There are three commonly known methods of feature selection: wrapper methods, embedded methods, and filter methods. These methods are further divided into supervised and unsupervised techniques which use appropriate statistical measures to determine the input variables that are most relevant to the model.

Depending on the data type of the response and input variables, an assortment of feature selection methods can be exploited to improve the performances of the classification models. Some of these methods that will be explored include, *Recursive Feature Elimination, Sequential Feature Selection, ANOVA testing, and shrinkage methods*.

#### Decision trees

Decision trees (DTs) are interpretable machine learning models that make predictions from decision-based rules and information learned from features in the dataset. In our model, these features are the range of frequencies that absorbance is measured at. We chose DTs primarily for their interpretability and presentable visualisation of the choices made by the model. The number of parameters in a DT also does not increase with the number of features, and this was beneficial for this project due to the relatively large number of frequencies used in the model (k=107). DTs also work well with non-linear datasets (which our project is addressing).

Our best-performing DT classifier used SMOTE over-sampling to balance the dataset, and GridSearchCV was used to optimise the parameters of the model. The top features, accuracy and loss are shown in the table below.

Davids model explainability

The explainability of our models was a large factor toward the initial choice of algorithms, and throughout the project. The goal of this project is to directly aid with decision-making in a clinical setting – an important objective for which a responsible and explainable methodology is paramount. If we cannot explain the logic and meaning behind our models, then we cannot expect their output to be applied in such a clinical setting.

The initial choices of our models were heavily influenced by how interpretable we thought they would be from Robyn’s perspective, or the perspective of other potential end users. Decision tree models, for example, create an appealing tree diagram that shows the decision made at each node, and the ranked importance of the features used in that tree.

Logistic regression produces coefficients that can be used to interpret the impact of its features on the outcome.

Where models were less interpretable, such as the Random Forest Regressor, we calculated and presented the SHAP values [figure ref].