Predicting Conductive Ear Conditions using Wideband Absorbance (WBA)

A. Jiang, A. Tan, D. Ika, D. Gunadi, E. Kurniady, J. Xie

INTRODUCTION

Objective

Wideband absorbance (WBA) tests detect hearing conditions by measuring the middle ear's absorbance across multiple frequencies, whereas traditional hearing tests previously measured only one frequency. This change produces effective results but adds complexity to the 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 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 spanned across 107 different frequencies (features). The output variable was whether the hearing test passed or failed, designated by the binary *OverallPoF* field (1 = fail; 0 = pass).

Contribution

The goal of this project is to provide audiologists with the best frequencies that are most effective for diagnosing hearing conditions. This will hopefully allow for a faster, directed approach to analysing the test output.

METHODS & RESULTS

SAMPLING

Considering our data is imbalanced, we applied sampling methods discussed in [1] to create a balanced distribution of the classes. The first 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 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.

FEATURE SELECTION

To fulfill the objective of choosing top frequencies, we experimented with various feature selection methods. According to [2], these methods are broadly categorised into filter, wrapper and embedded.

Filter methods rank feature importance using statistical criteria, such as variance or chi^2 and therefore are independent of any learning algorithms. In contrast, wrapper methods such as RFE and SFS rely on learning algorithms to iteratively evaluate and select the best feature subset. Embedded methods similarly use learning algorithms, specifically those that intrinsically perform feature selection, eliminating the need of evaluating one subset at a time. Popular examples of embedded methods include regularization models and decision trees.

MODELS

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 (our frequencies), as shown by Hasan et al. [3]. We chose DTs for their interpretability and visualisation of choices made by the model. Our best-performing DT classifier used *SMOTE* over-sampling to balance the dataset, and GridSearchCV was used to optimise the parameters of the model. As shown in Figure C3.1, the most influential frequencies were 1296Hz, 1155Hz, 4117Hz, 1943Hz, 2519Hz, 3363Hz, 280Hz, 1455Hz. The recall on the test set was 89%, and the loss was 6%.

Random Forests

Random Forests (RFs) are an ensemble of DTs that provide a more accurate and stable prediction than simple DTs. We used SHAP values to increase interpretability [4] – a method for explaining feature impact on predictions. Appendix C3.2.1 shows the top 3 frequencies on test set are 1189 Hz, 771Hz, and 1155Hz, and indicate that if the absorbance at 1189Hz is low, the chances of failing hearing tests increase, while a high absorbance of 1189Hz will decrease the probability of failing hearing tests.

Logistic Regression

Logistic regression is a method to predict the log probability of a binary event. Logistic regression's results are interpretable: Coefficients can be used to understand the importance of each feature towards the dependent variable [5]. Along with the regression, several regularization methods can be used to prevent over-fitting of the data [10].

Two approaches were used for Logistic regression, Multinomial LR using all frequencies for the model and Simple LR to create individual models per frequency. Regularization parameters L1, L2 and Elastic Net, were also used 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 (Appendix C1.1), using MinMax scaling and L1 (Lasso) regularization, while the Simple LR approach returns 87.51% and 92.78% recall on the training and test set (Appendix C1.2), using L2 (Ridge) 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.

SVM classifiers, in combination with feature selection (FS), were trained using different modified samples. In terms of FS methods, wrapper and embedded methods tend to result in better performing classifiers.

Specifically for wrapper selection methods, the RFE-SVM trained on *SMOTEENN* performed the highest with 91% and 89% recall on the train and validation set respectively (Appendix C2.1).

For embedded feature selection, ridge L2 penalty trained on *SMOTETomek* was able to effectively reduce the features to the 13 frequencies in Appendix C2.2. This approach is shown as model 10 in Appendix C2.3 and results in the best recall score of 92% on the validation set, with minimal difference from the training score.

DISCUSSION

COMPARING MODELS

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 Appendix B1.1. We find that most algorithms perform well on the *SMOTEENN* sampling method, which is consistent with Estabrooks, Jo and Japkowicz’s [8] finding that combining both oversampling and undersampling helps to deal with class-imbalance problem.

According to Appendix B1.2, 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 (Appendix A1.1). 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.

MODEL STRENGTHS AND LIMITATIONS

Tree-based models

DTs have the benefit of presenting a clear and visual thought process of the decisions made by the model, improving explainability and thus the confidence of end users (audiologists) in the model. Some negatives of DTs are that they tend to overfit datasets and can be heavily affected by small amounts of noise.

The advantage of random forest is that it may produce more reliable results than decision trees because it embeds multiple decision trees, and random forest reduces the degree of overfitting through voting. However, a random forest is like a black box with less interpretability.

Logistic Regression

Logistic regression models are interpretable, as they assigned 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 [5]. Regarding performance, logistic regression models are often outperformed by other methods on datasets with more complex relationships.

SVM

The inherent high dimensionality mechanism of SVM enables it to be effective in cases where number of features is high. SVM also has kernel tricks that can be applied when data is not linearly separable. These kernel tricks allow better separation of the data, albeit with the cost of explainability. Applying kernel functions are also computationally expensive, hence SVM is not suitable when data is large and noisy.

EXPLAINABILITY

In a recent survey on explainable AI [10], A. Barredo discusses the trade-off between interpretability and performance of models. He proposes that both Decision Trees (DTs) and Logistic Regression are high in model interpretability but have low accuracy. Conversely, he explains that SVMs have a higher accuracy but lower interpretability.

DTs are appealing because they show the decision made at each node and the corresponding importance of that node (feature). Similarly, logistic regression produces coefficients that can be used to determine feature importance. This contrasts with Linear SVMs, which produce coefficients on the boundary of separation (hyperplane) instead of the feature itself.

Where models were less interpretable (e.g. Random Forests), we calculated SHAP values. We used a force plot and decision plot to show how a single instance is predicted and adopted by the model. A summary plot was then shown for global interpretation (Appendix C3.2).

Regarding overall results, all models chose frequencies within the low-mid range of 1000-1400 Hz. This makes sense when observing F-Scores (ANOVA test) for each frequency (Appendix A1.2), which represent the greatest separation between the observed passes and fails at each frequency.

The frequencies that are best for diagnosing patients will possess absorbance ranges that have clear separation between passes and fails (Appendix A1.3). Where overlap is high between pass and fail absorbances, the models tend to reduce the importance of that frequency, this is because those frequencies would contribute less information towards developing a model that can distinguish between pass and fail.

CONCLUSION

Our approach has explored four different sampling methods to address class imbalances and four different classifiers (logistic regression, SVM, DTs and RFs) to model the binary classification problem. Across all models tested, sampling techniques that utilised both oversampling and undersampling (SMOTEENN or SMOTETomek) tended to perform better in training and testing.

The best model was selected based on two criteria, the relative interpretability of the model and the recall score on the test set. Despite other models such as SVM (with L2 regularisation) performing better, the simple decision tree classifier was selected due its inherent model interpretability and its competitive performance in training, validation and testing.

By comparing the results, strengths, limitations and explainability of the models we have arrive at the following conclusion. The top three selected frequencies across all methods were 1296 Hz, 1334 Hz and 1090 Hz. These frequencies present as the most influential in determining the presence of a hearing condition in the context of middle-ear testing.

FUTURE IMPROVEMENTS

The main caveat of this project is that the small number of observations (n=239) makes it difficult to rely on results. Next steps would see us continually seek more datapoints to re-run model testing and potentially investigate new models.

From there, the next steps would be to produce a tool for end-users based on the 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.

REFERENCES

1. H. He and E. A. Garcia, “Learning from Imbalanced Data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/tkde.2008.239.

2. J. Li *et al.*, “Feature Selection,” *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–45, Jan. 2018, doi: 10.1145/3136625.

3. Md. Rajib Hasan, Nur, F. Siraj, Mohd Shamrie Sainin, and S. Hasan, “Single decision tree classifiers’ accuracy on medical data,” 2015.

4. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” 2017, vol. 30. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

5. M. Pavlou, G. Ambler, S. Seaman, D. Iorio, and R. Z. Omar, “Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events,” *STATISTICS IN MEDICINE*, vol. 35, Art. no. 7, SI, 2016, doi: 10.1002/sim.6782.

6. H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY*, vol. 67, Art. no. 2, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

7. C. J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998, doi: 10.1023/a:1009715923555.

8. Estabrooks, T. Jo, and N. Japkowicz, “A Multiple Resampling Method for Learning from Imbalanced Data Sets,” *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, Feb. 2004, doi: 10.1111/j.0824-7935.2004.t01-1-00228.x.

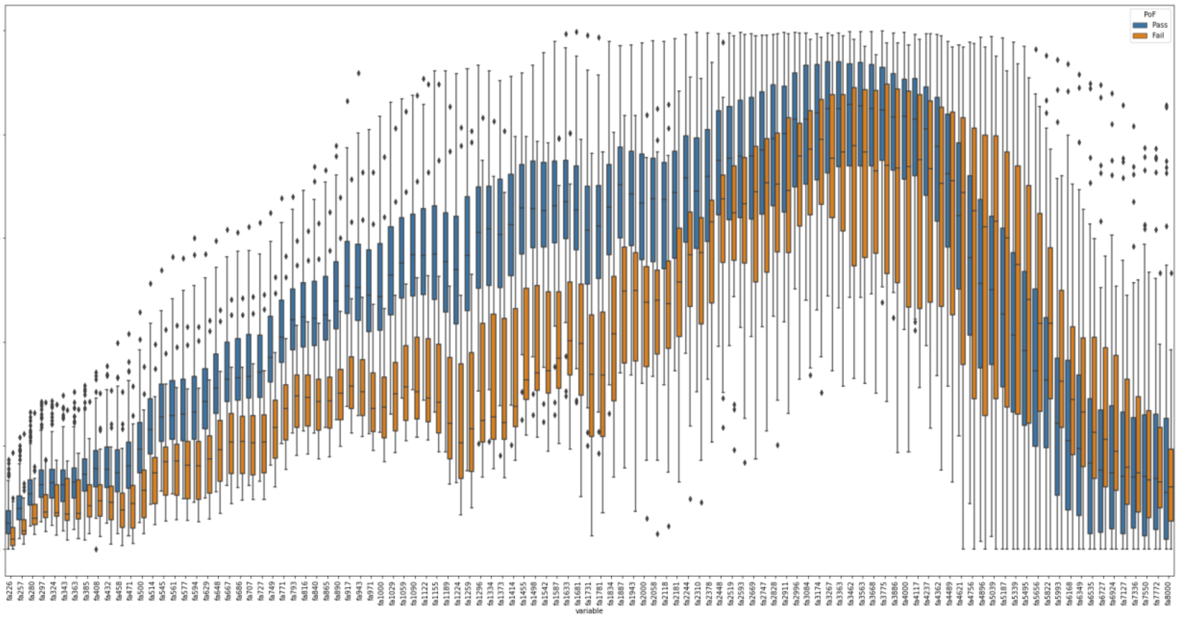
9. Arrieta *et al.*, “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *INFORMATION FUSION*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

10. P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

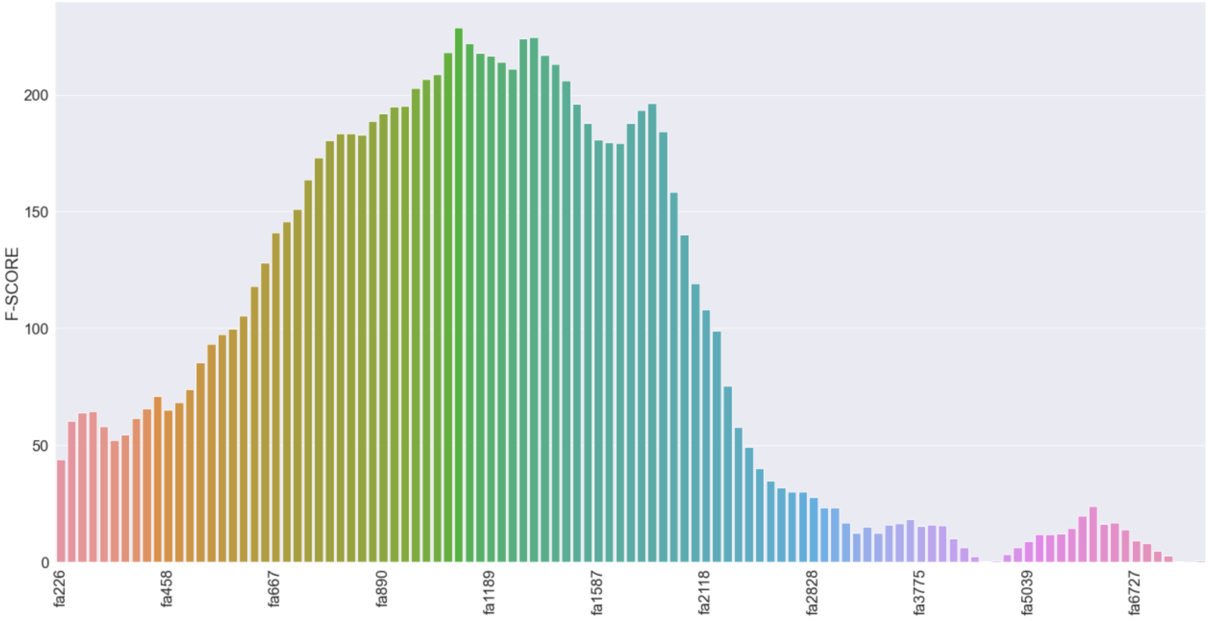
APPENDIX

A1 – Data Visualisations

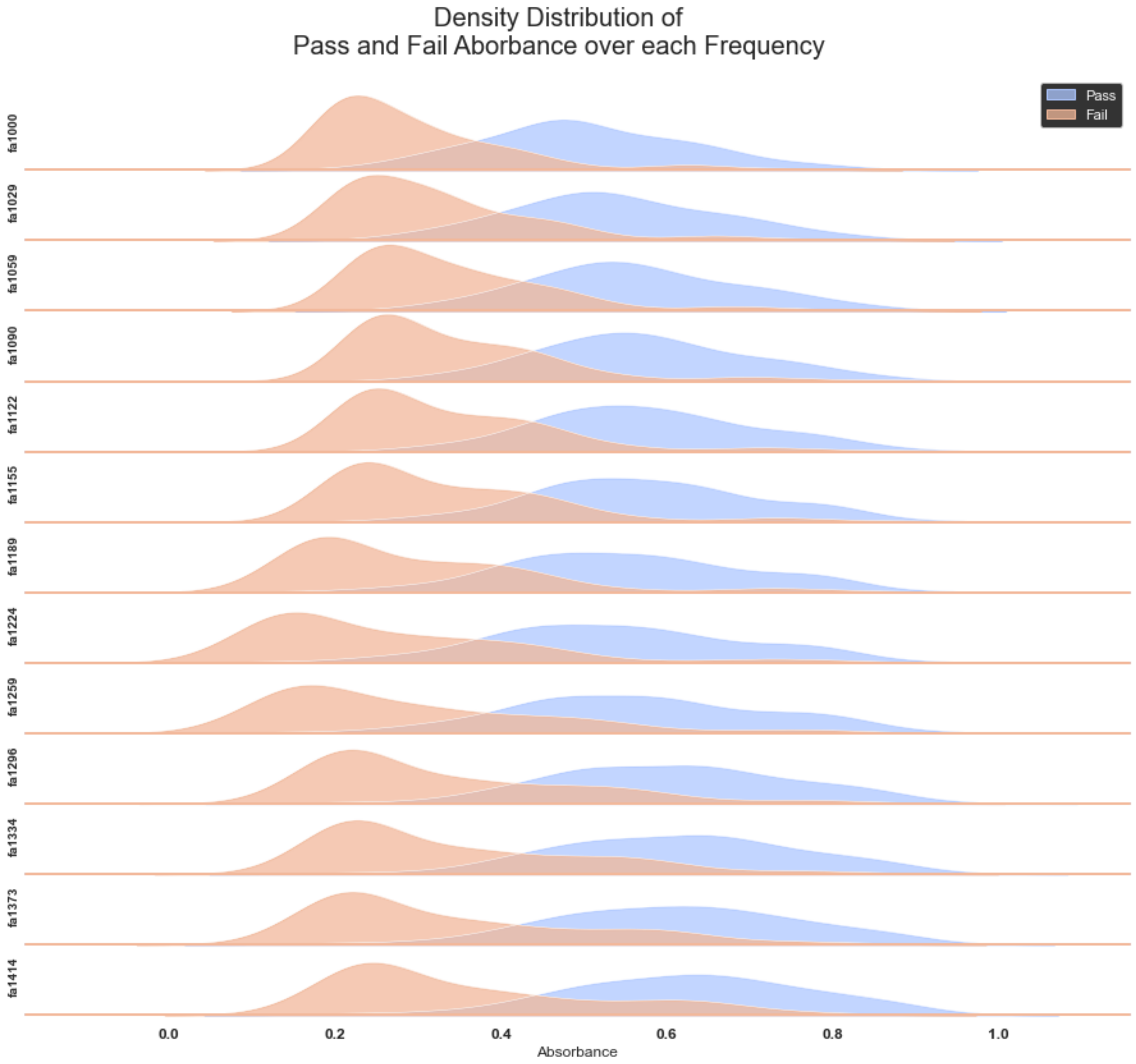
A1.1 – Plot of pass and fail absorbance (Blue=Pass, Orange =Fail)



A1.2 – Plot of F-Scores across all frequencies (univariate test)



A1.3 – Plot of density of absorbances for pass (blue) and fail (orange)



Example shown for 1000-1414 Hz

B1 – A SUMMARY OF EACH MACHINE LEARNING ALGORITHM

B1.1 - Summarise top frequencies for each machine learning algorithm

Model	Sample	Top Frequencies (Hz)	Recall (Test set)	Diff	Note
Logistic regression 1	SMOTEENN	1834, 1334, 1090, 4000, 3174, 5993, 2118, 4756, 1296, 6535'	0.93	0.05	Performed MinMax Scaling on the dataset and used L1 (Lasso) regularization to remove unnecessary features.
Logistic regression 2	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 1	SMOTEENN	1090, 1259, 1296, 1334, 2118, 2181, 2911, 3174, 4000, 4896, 6349	0.927	0.0137	Selection method used was a recursive feature elimination with a linear SVC estimator
SVM 2	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	1296, 1155, 4117, 1943, 2519, 3363, 0.058, 280, 1455	0.889	0.0573	Hyperparameters: {'criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 3, 'min_samples_split': 2, 'splitter': 'best'}
Random Forest	SMOTEENN	1189, 771, 1155, 971, 1090, 793, 1029, 1122, 1000, 943	0.903	0.0021	Hyperparameters: {'criterion': 'gini', 'max_depth': 3, 'n_estimators': 100}

Note: Top frequencies are rank based on the importance. Diff is calculated as the absolute recall difference between the training set and the test set.

B1.2 - Count the number of occurrences of top frequencies

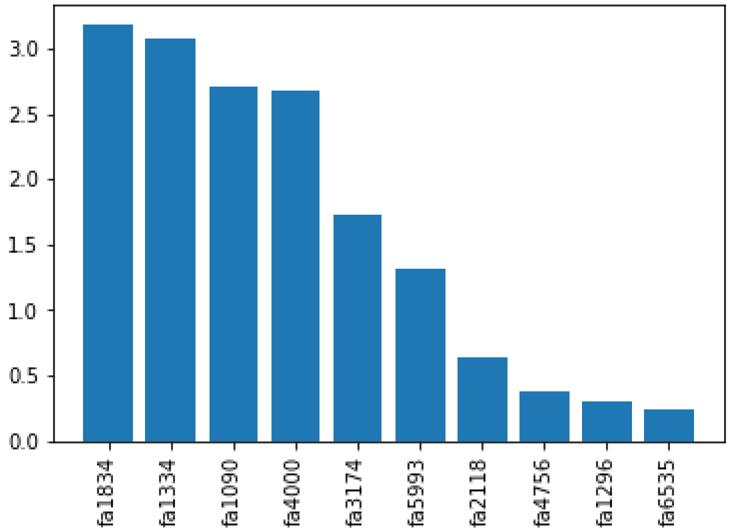
Frequency (Hz)	The number of occurrences	Model
1296	5	Logistic regression 1, Logistic regression 2, SVM 1, SVM 2, Decision tree
1334	4	Logistic regression 1, Logistic regression 2, SVM 1, SVM 2,
1090	3	Logistic regression 1, SVM 1, Random forest
1155	3	Logistic regression 2, Decision tree, Random forest
1000	2	Logistic regression 2, Random forest
1029	2	Logistic regression 2, Random forest
1122	2	Logistic regression 2, Random forest
1189	2	Logistic regression 2, Random forest
1834	2	Logistic regression 1, Logistic regression 2
1943	2	Logistic regression 2, Decision tree
280	2	SVM 2, Decision tree
4117	2	SVM 2, Decision tree

Note: This table shows top frequencies with more than 1 occurrence based on different ML algorithms.

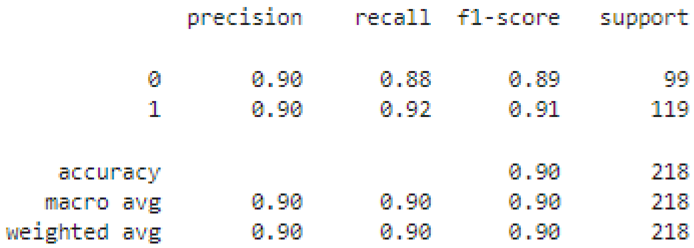
C1 – LOGISTIC REGRESSION RESULTS

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

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

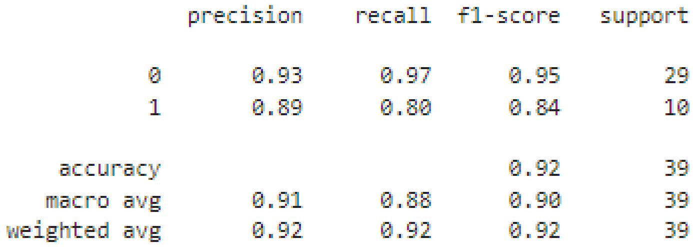


Classification report on the training set



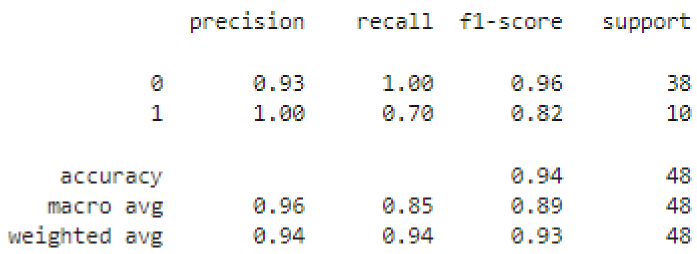
```
              precision    recall  f1-score   support

           0       0.90      0.88      0.89        99
           1       0.90      0.92      0.91       119

    accuracy                           0.90       218
   macro avg       0.90      0.90      0.90       218
weighted avg       0.90      0.90      0.90       218
```

Classification report on the validation set



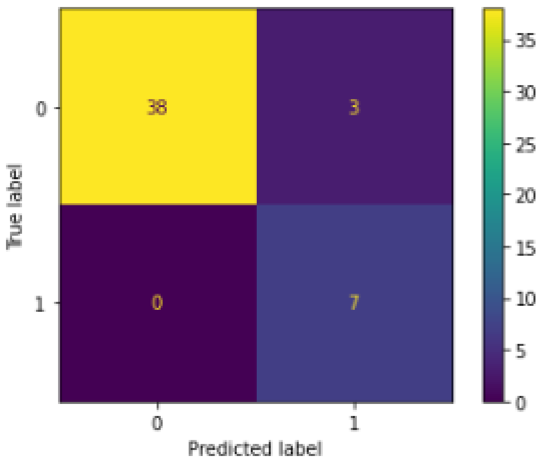
```
              precision    recall  f1-score   support

           0       0.93      0.97      0.95        29
           1       0.89      0.80      0.84        10

    accuracy                           0.92        39
   macro avg       0.91      0.88      0.90        39
weighted avg       0.92      0.92      0.92        39
```

Classification report on the test set



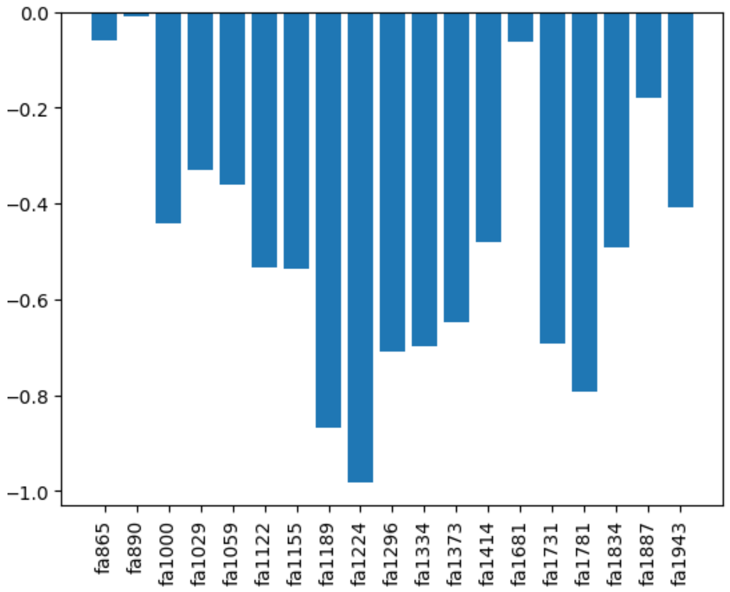
```
              precision    recall  f1-score   support

           0       0.93      1.00      0.96        38
           1       1.00      0.70      0.82        10

    accuracy                           0.94        48
   macro avg       0.96      0.85      0.89        48
weighted avg       0.94      0.94      0.93        48
```

Confusion matrix on the test set

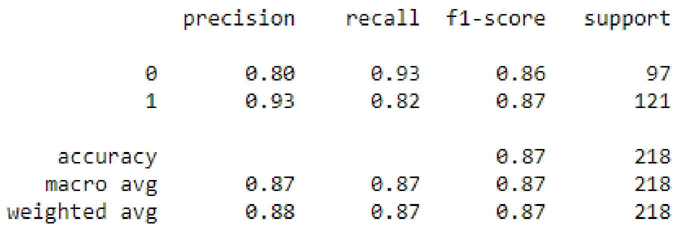


C1.2 - Simple logistic regression before combining

Feature importance

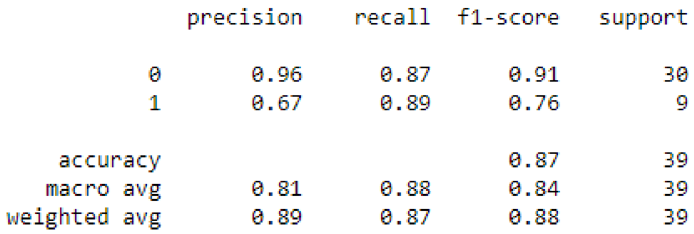


Classification report on the training set



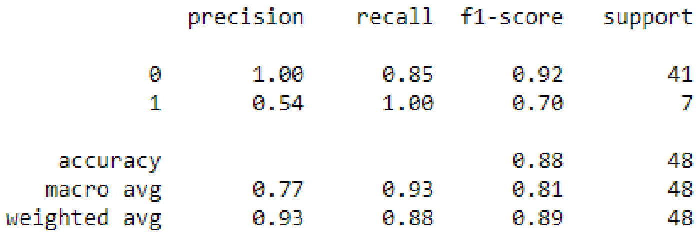
```
              precision    recall  f1-score   support

           0       0.80      0.93      0.86        97
           1       0.93      0.82      0.87       121

    accuracy                           0.87       218
   macro avg       0.87      0.87      0.87       218
weighted avg       0.88      0.87      0.87       218
```

Classification report on the validation set



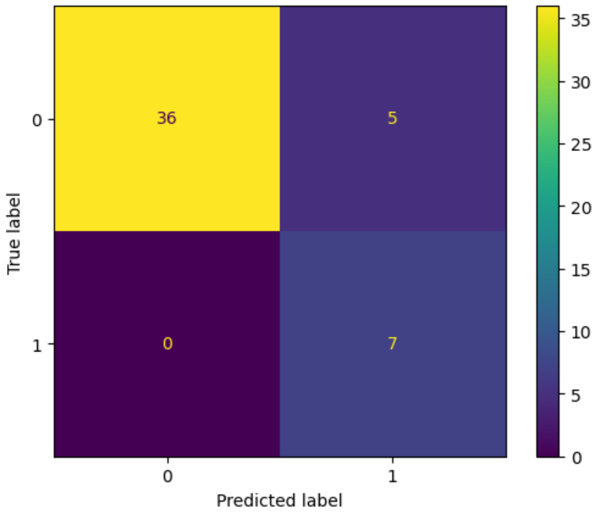
```
              precision    recall  f1-score   support

           0       0.96      0.87      0.91        30
           1       0.67      0.89      0.76         9

    accuracy                           0.87        39
   macro avg       0.81      0.88      0.84        39
weighted avg       0.89      0.87      0.88        39
```

Classification report on the test set



```
              precision    recall  f1-score   support

           0       1.00      0.85      0.92        41
           1       0.54      1.00      0.70         7

    accuracy                           0.88        48
   macro avg       0.77      0.93      0.81        48
weighted avg       0.93      0.88      0.89        48
```

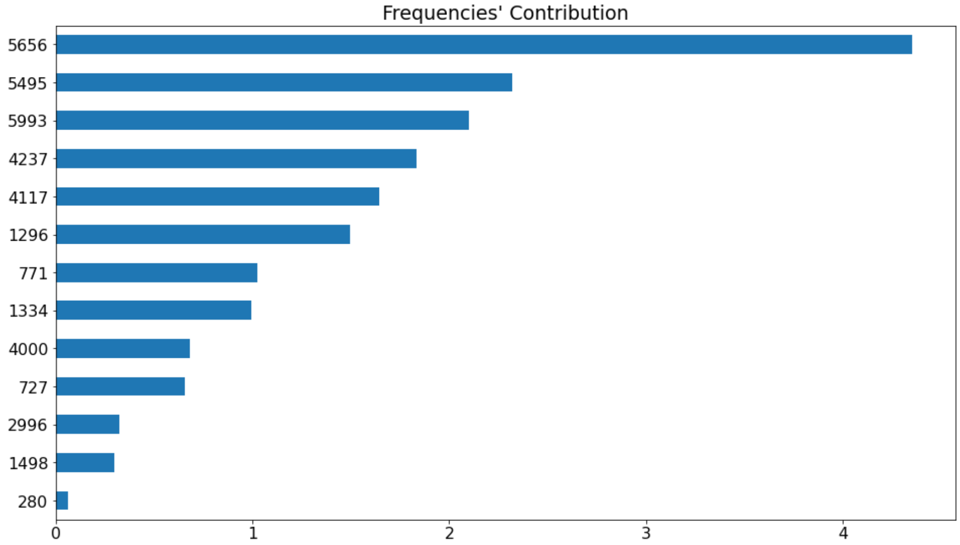
Confusion matrix on the test set



C2.1 – 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 - Features Selected by Ridge Embedded SVM with SMOTETomek

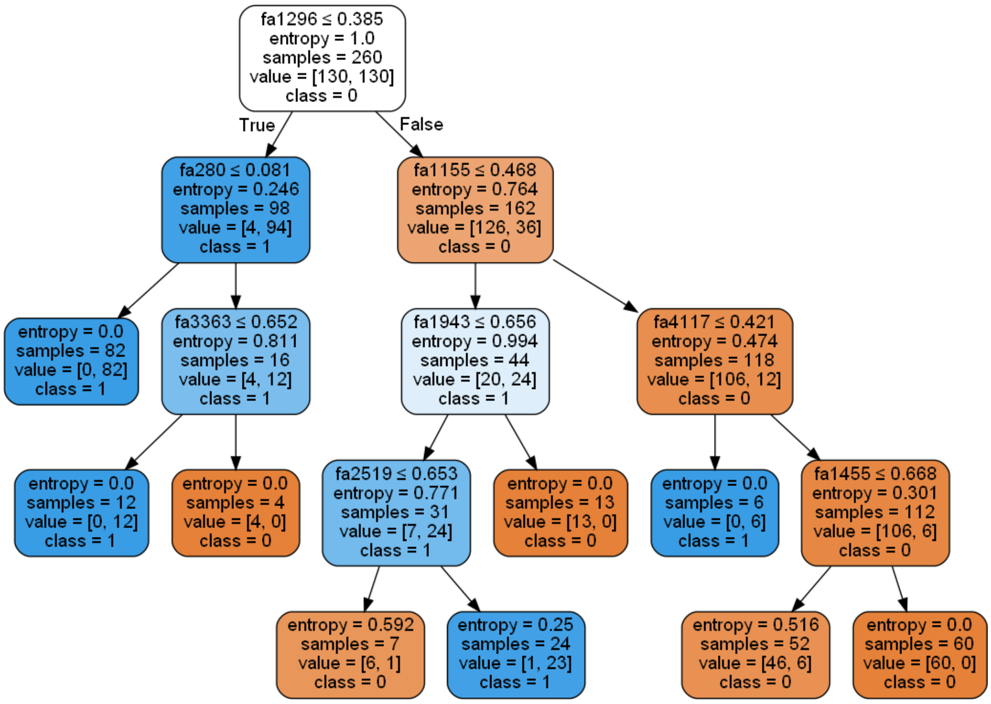


C2.3 - Comparison Table of SVM with Embedded Feature Selection

model	sample	regularisation	n_features	training_recall	validation_recall
1	Original	l1	11	92.73	89.44
2	Original	l2	13	91.19	73.33
3	NM2	l1	14	79.55	77.78
4	NM2	l2	7	86.36	74.44
5	SMOTE	l1	21	92.31	87.78
6	SMOTE	l2	25	95	76.67
7	SMOTEENN	l1	14	96.29	82.22
8	SMOTEENN	l2	22	96.6	66.11
9	SMOTETomek	l1	12	95.67	80.56
10	SMOTETomek	l2	13	89.37	91.67

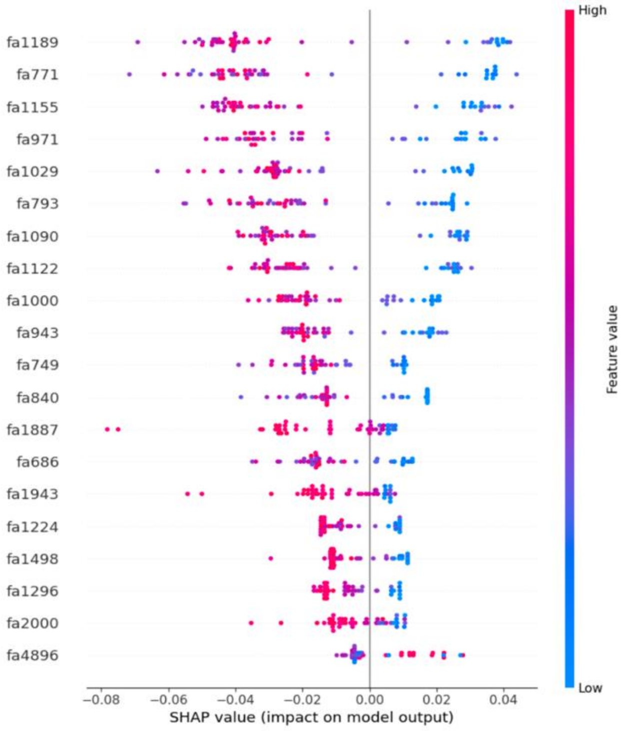
C3 – TREE METHODS RESULTS

C3.1 - Decision tree (SMOTE over-sampling)

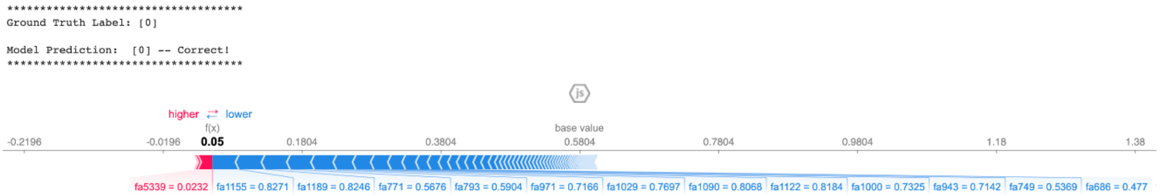


C3.2 - Random Forests

C3.2.1 - Summary plot (SMOTEEN –Test set)

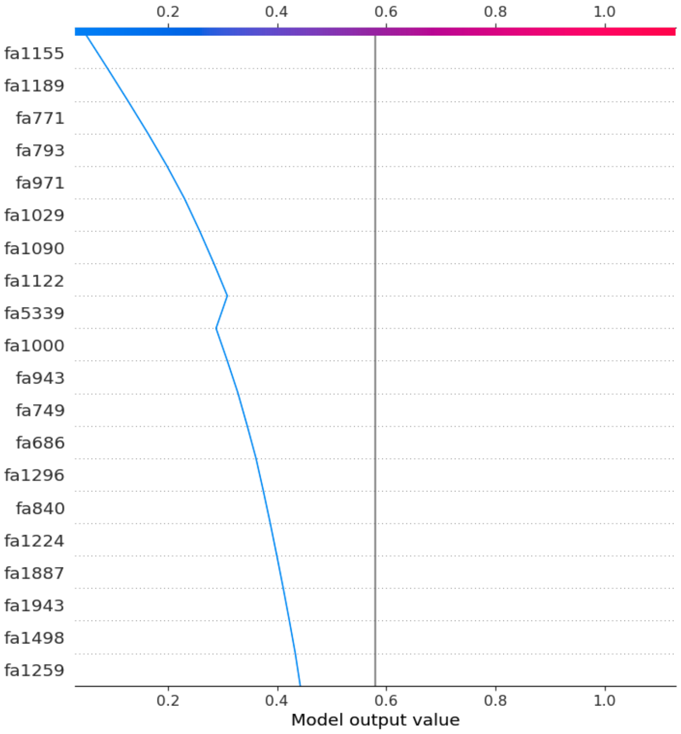


C3.2.2 - Force plot for a single instance



```
************************************
Ground Truth Label: [0]

Model Prediction:  [0] -- Correct!
************************************
```

C3.3.3 - Decision plot for a single instance



Group Contribution

Member Name	Student Number	Tasks for project	Skills contributed to the team & project
Andy Jiang	21124413	SVM Classifer with Feature Selection Methods	Project Management, Python, R and Machine Learning Statistical Analysis
Andrew Tan	22614179	Simple Logistic regression	Python – Machine Learning Statistics Knowledge
David Ika	21520699	- Decision Tree Classifier. - Helping review written work. - background research on problem statement (for proposal).	Python skills. Always facilitating discussion to increase collaboration and knowledge sharing.
Dennis Gunadi	22374535	Multinomial Logistic Regression, Variance Feature Selection,	Python and Machine Learning skills. Statistics Knowledge
Edward Kurniady	23220451	1. Exploratory analysis of the data 2. Data cleaning and sampling 3. Explored feature selection methods 4. Focused on embedded SVM	Biostatistical knowledge Machine learning using python Github
Jessie Xie	21918545	1.Audiology literature review 2.Exploratory data analysis 3.Data cleaning and transformation 4.Feature filter exploration 5.Tree-based models	Python Machine learning Statistics