Multimodal Pathological Voice Classification

Chun-Hsien Chen, Shu-Cheng Zheng, Yi-Cheng Hung, and Jia-Wei Liao

September 7, 2023

1 Algorithm and Model Architecture

In this session, we will introduce three models: Random Forest [1], LightGBM [2], and TabPFN [3]. Furthermore, data preprocessing and feature engineering techniques will be discussed in Section 3.

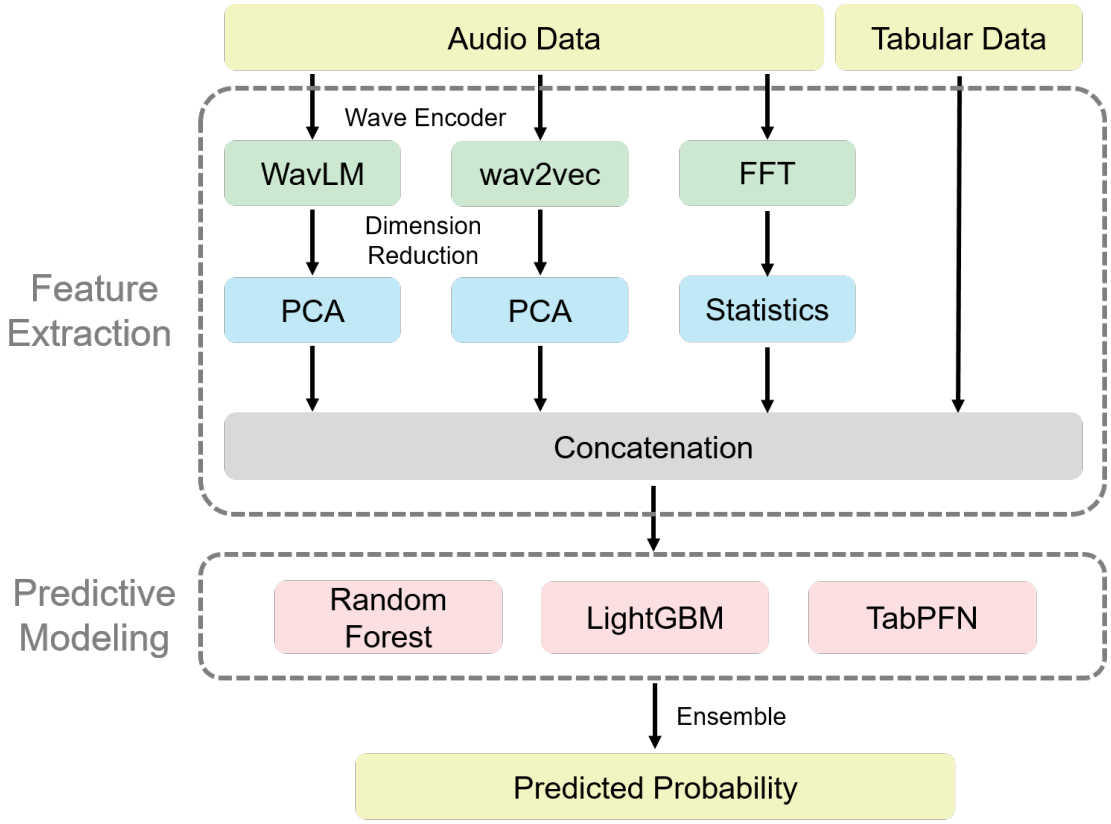


Figure 1: Model Architecture

The feature extraction process consists of two parts. In the first part, we employ the Fast Fourier Transform (FFT) to extract frequency features and calculate statistical indicators, constructing a global feature. The second part involves utilizing a deep learning-based pretraining model to extract local features, followed by dimension reduction using

Principal Component Analysis (PCA) to retain relevant feature combinations.

For the model training phase, we utilize machine learning-based tree models, namely Random Forest and LightGBM, and a deep learning-based transformer model called TabPFN for prediction purposes. An ensemble is performed on the predicted probabilities these models generate to obtain the final output.

1.1 Random Forest

Random Forest is an ensemble model that combines multiple decision trees. Each tree is trained on a random subset of the data and features. The GINI index measures the impurity of features and determines the splitting criteria at each node. The model employs a greedy algorithm where each tree is constructed by recursively partitioning the data based on the selected features. Bagging is applied to the ensemble by aggregating the predictions from individual trees, resulting in a more robust and accurate final prediction.

Parameter	Value
`n_estimators`	100
`criterion`	gini
`min_samples_split`	2
`min_samples_leaf`	1

1.2 LightGBM

LightGBM is a gradient-boosting framework that utilizes decision trees as weak learners. It employs a technique called residual learning, where subsequent trees are trained to learn and correct the residual errors of the previous trees. This iterative process gradually improves the model's performance by focusing on the remaining errors. LightGBM optimizes the construction of decision trees using a leaf-wise approach, which grows the tree by selecting the leaf node with the maximum reduction in the loss function. This strategy enables LightGBM to achieve faster training and better accuracy than traditional gradient-boosting methods.

Parameter	Value
`boosting_type`	gbdt
`num_leaves`	31
`learning_rate`	0.1
`n_estimators`	100
`class_weight`	balanced
`objective`	multiclass

1.3 TabPFN

TabPFN is a neural network architecture based on transformers, originally designed for tabular data. It incorporates causal inference to discover causal relationships among the different components or variables in the system. This enables TabPFN to capture complex interactions and dependencies between variables, allowing it to adapt well to unknown datasets and generalize to different contexts. Additionally, TabPFN is specifically designed to handle small data sets, where traditional deep learning models may struggle due to overfitting. It leverages techniques such as attention mechanisms and self-attention to extract meaningful features and make accurate predictions even with limited data.

Parameter	Value
`N_ensemble_configurations`	100

2 Innovative

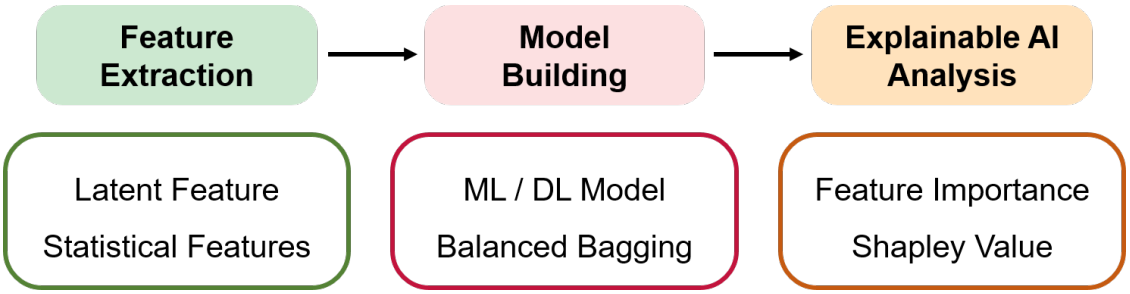


Figure 2: Training / Prediction Pipeline

In this competition, we present a novel pipeline for disease classification. Our approach combines traditional signal processing techniques with modern deep learning methods to effectively extract both global and local features, leveraging the unique strengths of each domain. Specifically, we employ specific signal processing algorithms, such as Fast Fourier Transform (FFT), to capture essential signal characteristics. Additionally, we utilize a powerful pretraining model to extract latent features using a zero-shot transfer method. Furthermore, to capture complex patterns and relationships within the data, we incorporate state-of-the-art models provided by external libraries [4, 5]. These models enhance our ability to handle intricate data structures and improve the overall performance of our classification system.

Our innovative methodology further exploits ensemble techniques, allowing us to capitalize on the complementary nature of traditional signal processing and deep learning. This integration of techniques has yielded superior performance compared to employing each approach individually.

To validate the significance of the extracted features, we compute feature importance indicators. These indicators provide quantitative measures of the impact of each feature on the model's performance, thereby confirming the effectiveness of our feature extraction process. Moreover, we utilize Shapley values [6] to gain a deeper understanding of the individual contributions of each feature. By quantifying the influence of each feature on the model's output, we can provide comprehensive explanations for the decision-making process of the model.

3 Data Preprocessing

This competition aims to use artificial intelligence technology to improve the detection and classification of voice disorders, with a focus on applications in the biomedical industry. The voice disorders are common in professions that rely heavily on the use of the audio, but they are difficult to detect and often require specialised medical professionals and equipment. In this competition, participants will use a non-contact approach that combines vocal signals and medical records, using dynamic sound and static text information to detect and classify diseases. The aim is to improve the quality of life for people with voice disorders by enabling early detection and treatment.

3.1 Exploratory Data Analysis

We have a dataset consisting of 2,000 annotated audio files, categorized based on the presence of the vowel sound " 阿" (pronounced as "a"). The dataset contains various classes of audio files, which have been corrected and classified. The corrected classifications of the audio files are as follows:

1. **Phonotrauma:** This category includes cases of phonotrauma, such as polyps, nodules, Reinke's edema, cysts, fibrous masses, and varices.

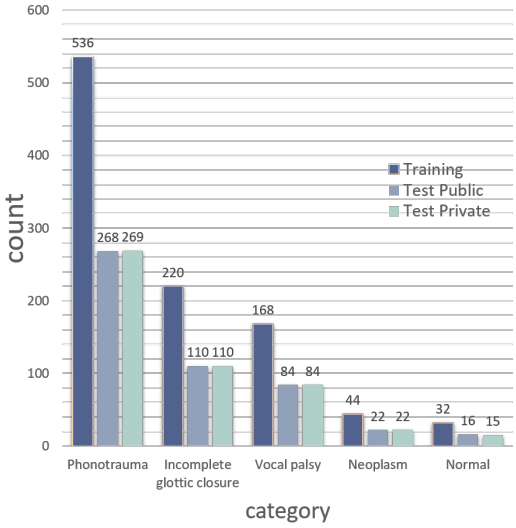
2. **Functional Dysphonia:** This category represents incomplete glottic closure and includes conditions like atrophy, sulcus, and presbyphonia.

3. **Vocal Palsy:** It encompasses cases of vocal cord paralysis or paresis.

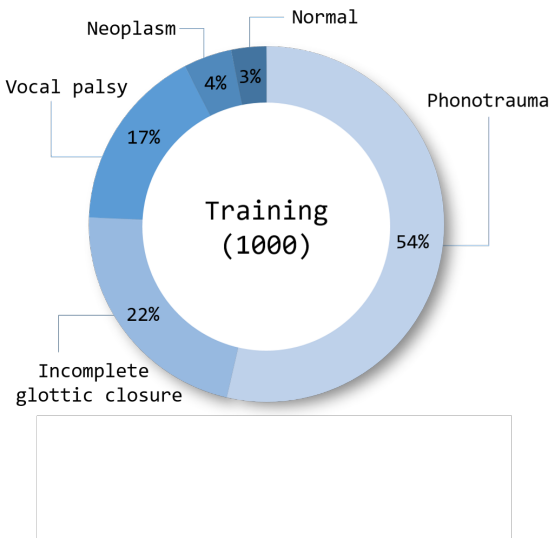
4. **Neoplasm:** This category includes vocal cord tumors, specifically neoplasms, and papillomas.

5. **Normal:** This category consists of audio files that were evaluated and determined to be normal, without any abnormal conditions. These files have a GRB value of 0.

In addition to the classification, the table provides the counts of audio files for each category, as well as the distribution of data for training, public testing, and private testing. The dataset comprises 1,073 audio files for phonotrauma, 440 files for incomplete glottic closure, 336 files for vocal palsy, 88 files for neoplasms, and 63 files classified as normal, totaling 2,000 audio files. Regarding data distribution, the phonotrauma category has 536 files for training, 268 files for public testing, and 269 files for private testing. The incomplete glottic closure category has 220 files for training, 110 files for public testing, and 110 files for private testing. The vocal palsy category consists of 168 training files, 84 public testing files, and 84 private testing files. The data statistics are presented in the following chart Figure 3a and Figure 3b.



(a) Overview each category counts



(b) Train set pie chart.

In the audio data, the duration of audio files ranges from 1 to 3 seconds, with 3-second audio files being the most common. Not all classes of audio files may have instances for each duration range. To mitigate the influence of audio length, we will introduce appropriate feature representation methods for audio data. The provided audio file the corresponding statistical graph is shown in Figure 5. Furthermore, in the tabular data

provided, we will impute missing values with zeros.

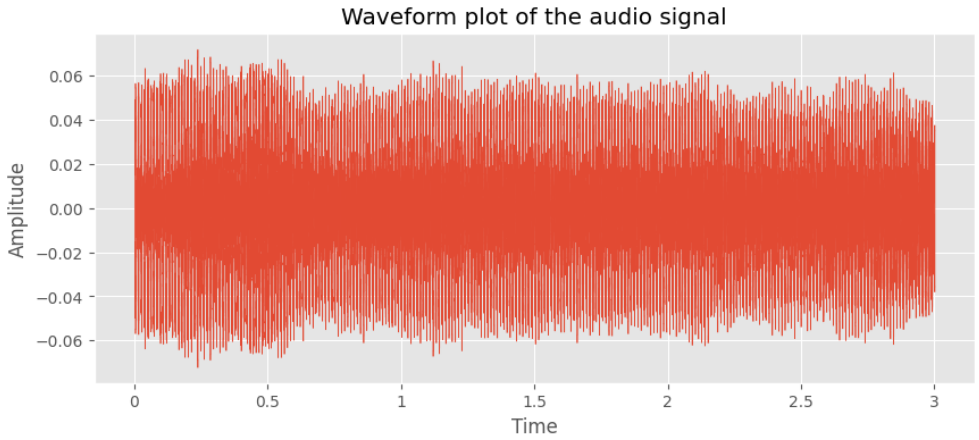


Figure 4: The visualization of the audio file

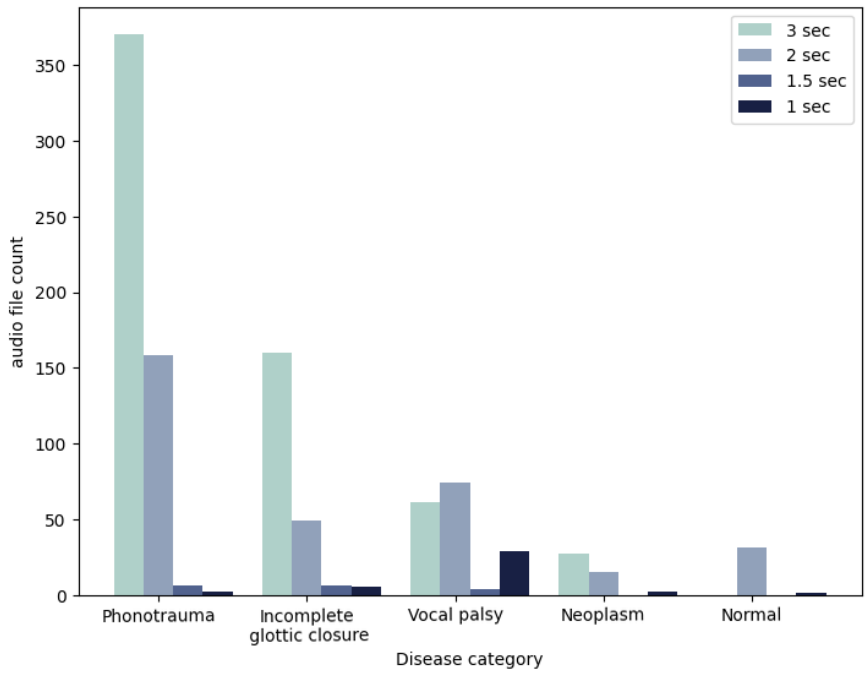


Figure 5: The audio length in each category.

Based on the reviewed data sources, we will proceed with further feature extraction.

3.2 Feature Extraction

For feature extraction, we divide it into two parts. In the first part, we adopt FFT to extract frequency features and compute statistical indicators to construct global features. In the second part, we employ a deep learning-based pretraining model to extract local features and perform dimension reduction using PCA, retaining useful feature combinations.

3.2.1 Latent Feature

We employ both wav2vec [7] and WavLM [8] to extract latent features from the audio data. These pretrained models offer complementary capabilities in capturing global and local audio characteristics. To address the challenge of high feature dimensionality, we utilize Principal Component Analysis (PCA) to reduce the dimensionality of the features to 9 dimensions. It prevents issues such as model learning difficulties and overfitting, allowing for a more efficient and effective representation of the audio data.

- **wav2vec**: This pretraining method utilizes contrastive learning. It divides the original speech waveform into fixed-length segments and rearranges them. The model aims to identify the correct order of the speech segments through self-contrastive learning. This pretraining process enables the model to acquire effective audio representations that capture crucial features within the speech signal.

- **WavLM**: It is a potent pretraining model that combines CNN and transformer architectures. It transforms speech waveforms into latent representations, effectively extracting local features. By leveraging both convolutional and transformer layers, WavLM can capture hierarchical and sequential information within the audio, leading to enhanced representation learning.

To select a suitable number of principal components, we consider the variance explained ratio displayed in Figure 6. The accumulated ratio gains little after 9 components. This observation suggests that reducing the dimensionality to 9 would be appropriate.

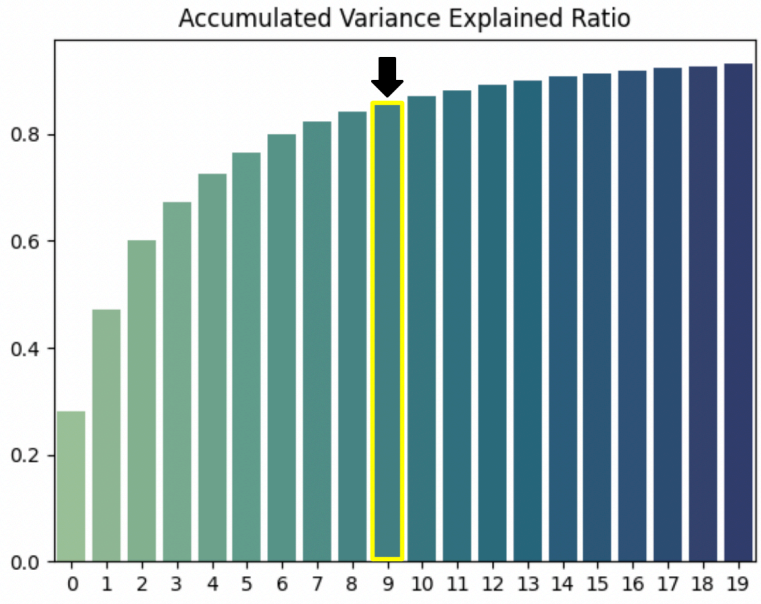


Figure 6: Accumulated variance explained ratio of principle components.

3.2.2 Statistical Feature

We employ FFT to extract frequency information from the audio and calculate statistical indicators such as the mean, median, and skewness of the signal. Additionally, we incorporate statistical indicators of the spectrogram into the feature representation. These methods enable the extraction of global features from the audio, capturing important characteristics of the overall signal.

4 Training Method

In the initial phase of our research, we considered three classification algorithms: RandomForest, LightGBM, and TabPFN, respectively. These algorithms were selected based on their suitability for addressing the specific task at hand.

To address the challenge of imbalanced data distribution, we employed the Balanced Bagging Classifier in conjunction with each algorithm. This technique helps to alleviate the bias introduced by class imbalance, thereby improving the overall performance of the models. By combining the predictions from multiple classifiers through a voting mechanism, we obtained an ensemble model that leverages the strengths of each individual classifier.

To enhance the generalization of each classifier, a comprehensive grid search was conducted using cross-validation. Cross-validation is a widely adopted technique for model evaluation, as it provides a robust estimate of the model's generalization performance. In our study, we employed stratified 5-fold cross-validation, ensuring that the data samples from different classes are represented proportionately in each fold. By maintaining class balance within each fold, we minimize the risk of biased performance evaluation due to the uneven distribution of classes. This approach enables us to obtain reliable performance metrics and make informed decisions regarding the selection and optimization of our classification models.

5 Analysis and Conclusion

In this section, we will describe the feature importance, model selection, and the obtained experimental results.

5.1 Experiment Result

The evaluation metrics in this competition is **UAR** (Unweighted Average Recall). It's a metric that measures the average recall across all classes in a classification model, regardless of class imbalances. It is calculated by dividing the number of correctly predicted instances by the total number of instances in each class and then averaging the recall values.

In Table 1, we found that the ensemble classifier, which was voted by LightGBM (LGB), RandomForest (RF), and TabPFN and fitted by Balanced Bagging Classifier (BB), consistently yielded the best results across different situations and datasets. Therefore, we selected this ensemble classifier as our final model for further analysis.

Model	**Data**		
	Audio only	Tabular only	Audio and Tabular
LGB	.381	.448	.526
RF	.309	.377	.356
TabPFN	.302	.372	.405
BB-LGB	.481	.514	.614
BB-RF	.501	**.554**	.608
BB-TabPFN	.489	.481	.586
BB-Voting	**.521**	.552	***.629**

* The highest cv-scores overall.

Table 1: The mean values of cross-validation scores in 5-stratified folds.

In addition, Figure 7, which displays the confusion matrix obtained by concatenating the prediction results of each fold in cross-validation, reveals that the voting model with balanced bagging classifier tends to predict a higher number of people for Neoplasm and Normal categories, even though the occurrence of these two types is relatively low. Furthermore, Table 3 provides more detailed information regarding the recall, precision, and accuracy metrics. The recall score is influenced by the denominator, which is the sum of false negatives and true positives. When the model predicts a large number of samples as false labels (false negatives), it has a slight impact on the recall score for categories that account for a majority of the data. Conversely, there is a significant increase in the recall score for the minor categories, thus improving the overall UAR score. However, it is important to note that precision and recall scores have a trade-off relationship. That is, when we strive to increase the recall score for the last two types, there is a notable

decrease in precision.

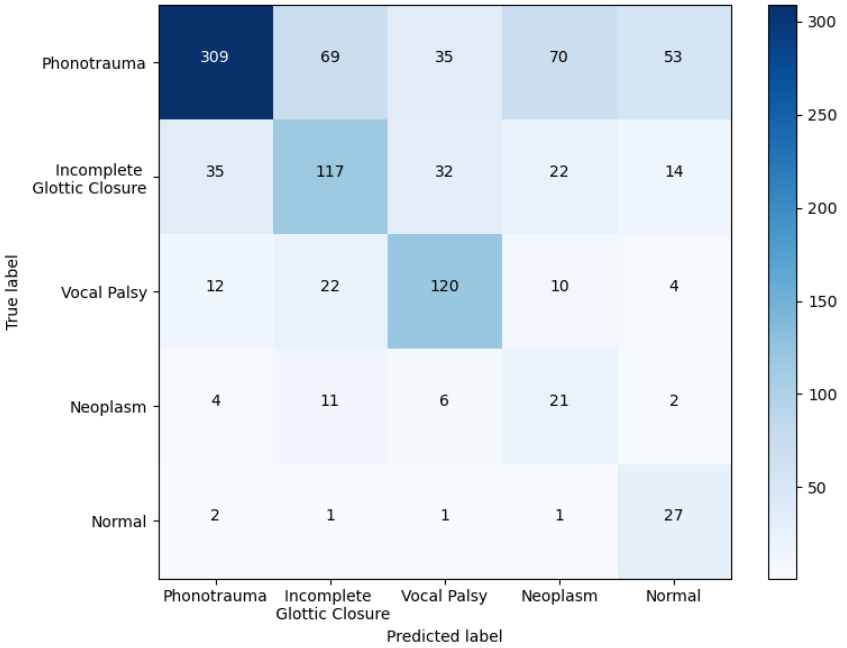


Figure 7: Confusion Matrix

	precision	recall	f1-score	support
Phonotrauma	0.854	0.576	0.688	536.000
Incomplete Glottic Closure	0.532	0.532	0.532	220.000
Vocal Palsy	0.619	0.714	0.663	168.000
Neoplasm	0.169	0.477	0.250	44.000
Normal	0.270	0.844	0.409	32.000
accuracy	0.594			
macro avg	0.489	0.629*	0.508	1000.000
weighted avg	0.695	0.594	0.621	1000.000

* The UAR score.

Table 2: The report for each types of diseases

Overall, these findings highlight the impact of the model's prediction tendencies on the recall, precision, and overall performance. By prioritizing the recall score for specific types, the model achieves a higher UAR score but at the cost of a decrease in precision, particularly for the last two types.

5.2 Feature Importance and SHAP Analysis

Impurity-based feature importances (Figure 8.) can be misleading for high cardinality features (many unique values), in our case, audio features (numerical) tend to have higher importance values than tabular features (categorical) if we use this kind of feature importance.

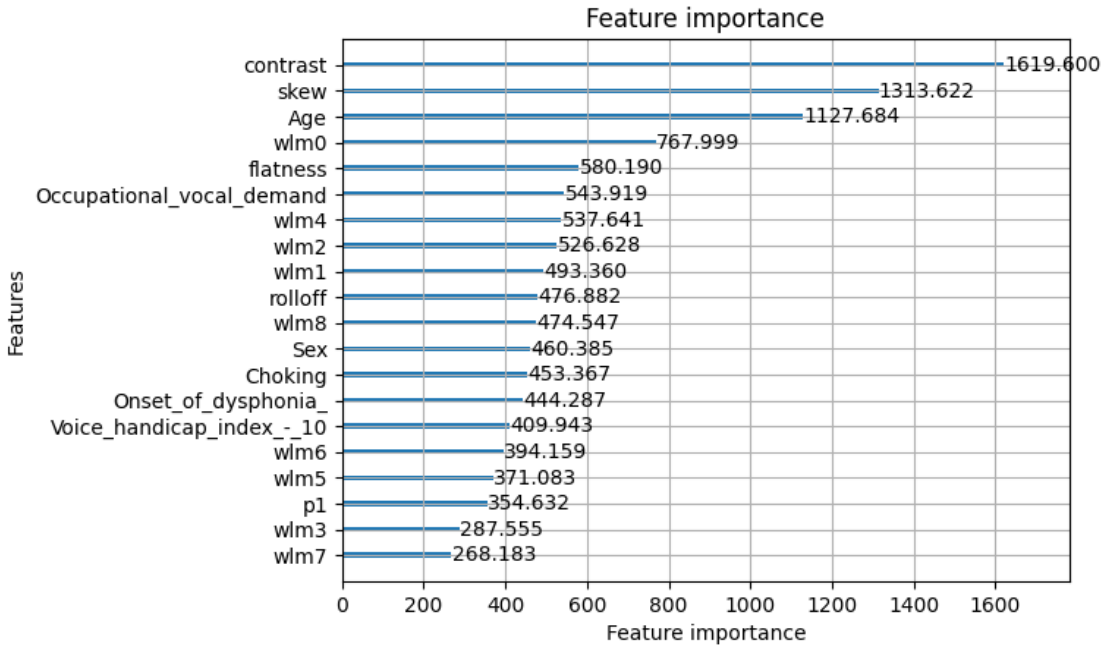


Figure 8: Impurity-based feature importance

In our approach, instead of relying on impurity-based feature importance, we utilize Shapley values (SHAP) as an alternative measure of feature importance. The objective of SHAP is to provide an explanation for the prediction of a particular instance by quantifying the contribution of each feature to the overall prediction. To compute the Shapley values, we employ the SHAP explanation method, which leverages concepts from coalitional game theory. In Figure 9-13, we present a visualization that illustrates the contribution of each feature to the final prediction probability. This visualization is achieved by representing the mean SHAP value for each class, enabling a better understanding of how different features impact the prediction for each specific class.

- **Phonotrauma**: In Figure 9, Age, mean, Occupational vocal demand, and median have the main contribution to pushing the model output from the base value (0.2) to the model output. For global interpretability, higher mean and median raise the predicted class probability. On the other hand, higher Occupational vocal demand and Age tend to lower the predicted class probability. Noted that for the force plot,

we take the average of the Shap values of the data in each class.

- **Incomplete Glottic Closure**: In Figure 10, Age, mean, contrast, wlm0, and median have the main contribution to pushing the model output from the base value to the model output. Additionally, higher Age and mean raise the predicted class probability. On the other hand, higher wlm0 tends to lower the predicted class probability.

- **Vocal Palsy**: In Figure 11, contrast, flatness, wlm0, choking, and Voice handicap index - 10 have the main contribution. Additionally, we see that lower contrast raises the probability, while higher flatness, wlm0, choking, and Voice handicap index - 10 also raise the probability.

- **Neoplasm**: In Figure 12, Smoking, Sex, wlm0, PPD, and wlm4 have the main contribution. The predicted class probability exhibits a positive relationship with smoking, wlm4, and PPD. The predicted class probability shows a negative relationship with Sex.

- **Normal**: In Figure 13, median, mean, Age, wlm4, and contrast have the main contribution. Age and wlm4 exhibit a negative relation with the predicted class probability, while median, mean, and contrast shows a positive relation.

Overall, we find that the features we extract from audio data provide valuable insights and contribute significantly to the predicted probability.

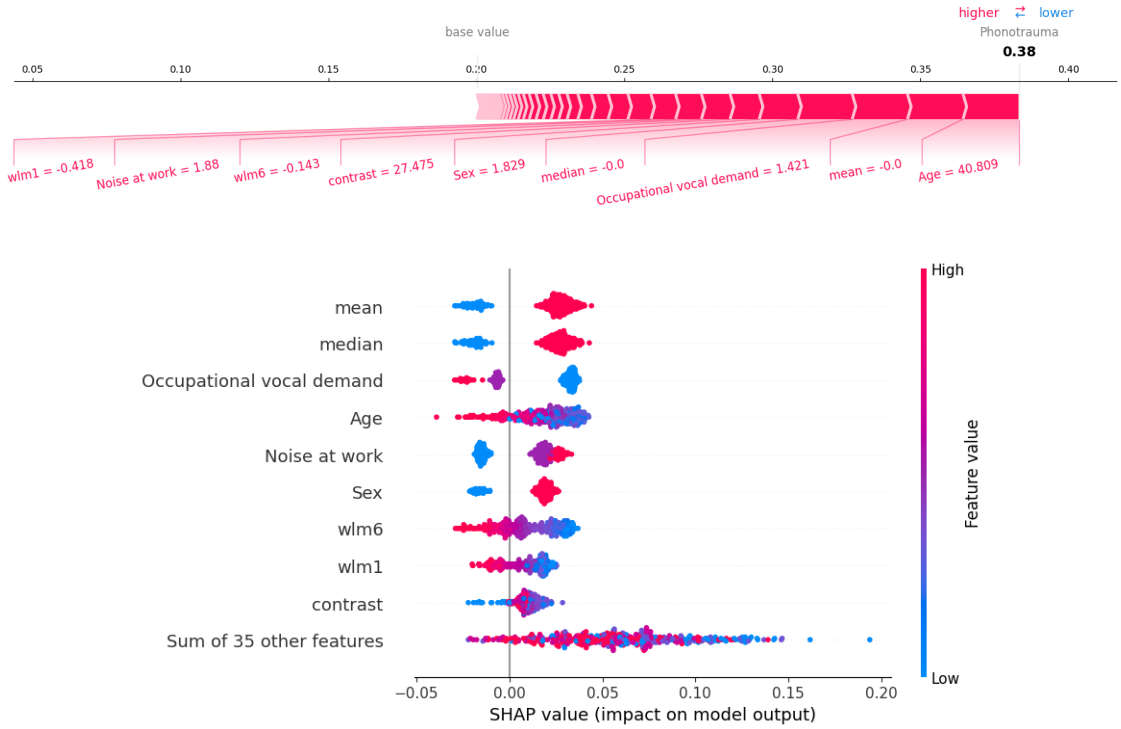


Figure 9: SHAP analysis of class *Phonotrauma*

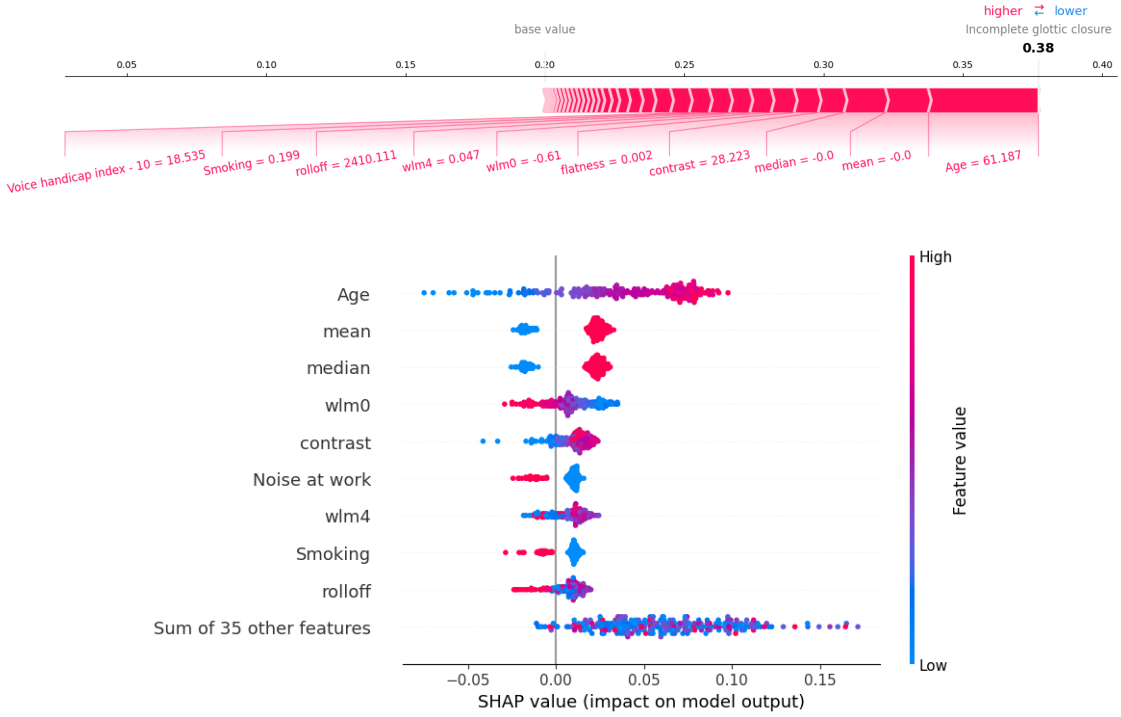


Figure 10: SHAP analysis of class *Incomplete Glottic Closure*

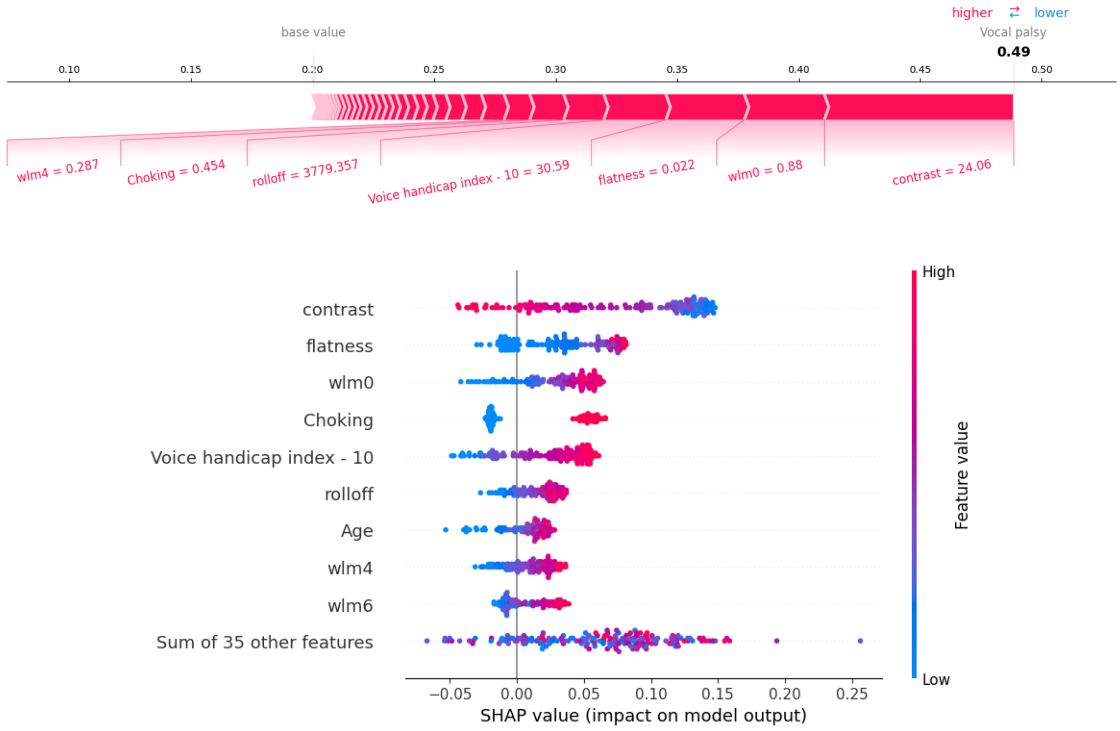


Figure 11: SHAP analysis of class *Vocal Palsy*

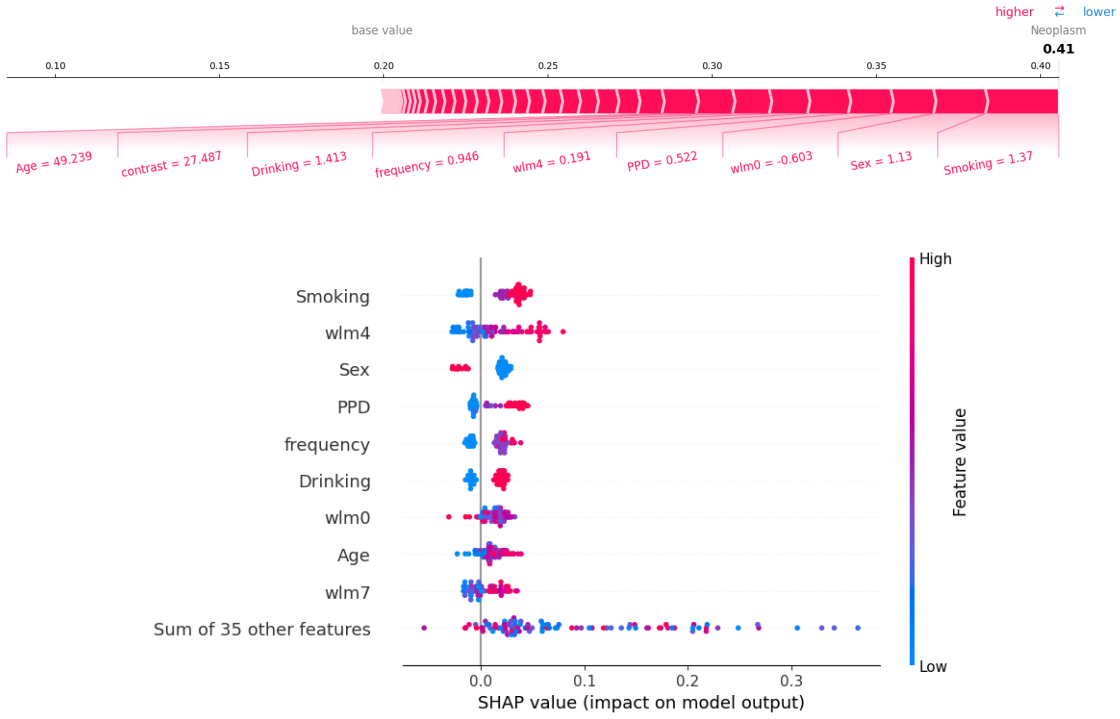


Figure 12: SHAP analysis of class *Neoplasm*

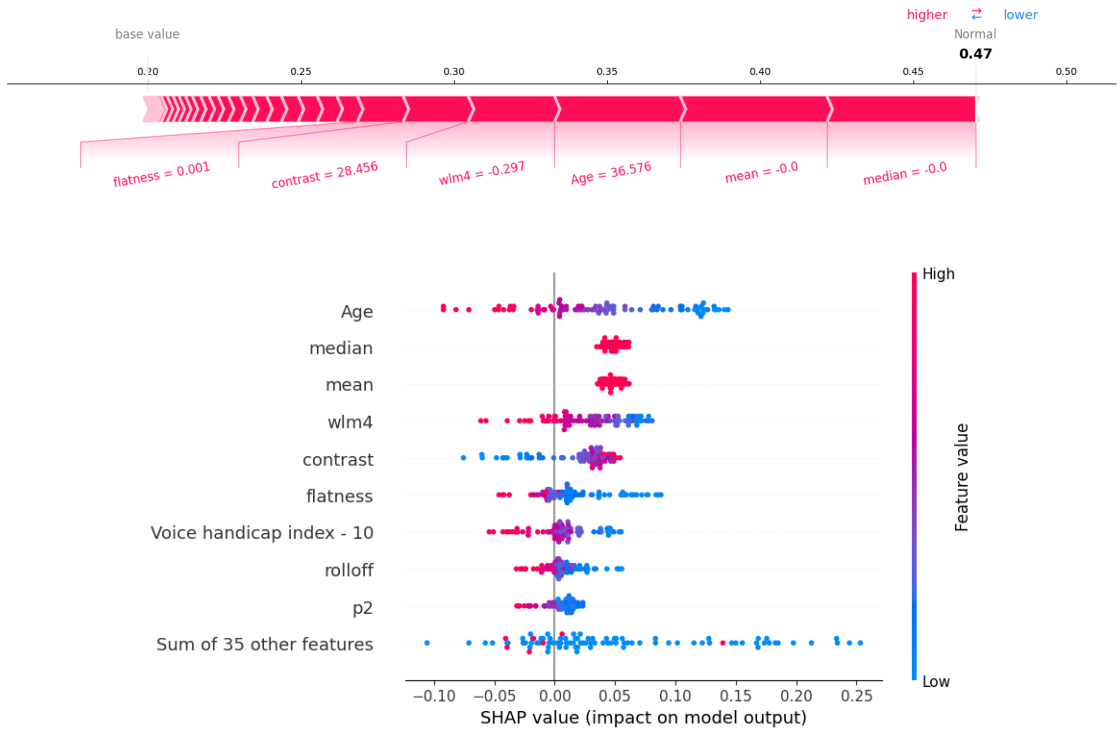


Figure 13: SHAP analysis of class *Normal*

5.3 Another Trial

We also attempted a purely computer vision approach, following the process of audio classification problems [9] [10]. In this competition, the participants were human beings, and the reference data [10] was used to differentiate sounds produced by different objects, with more noticeable differences in their corresponding audio frequencies. The concept involved transforming the audio into a spectrogram, which presents the time-frequency distribution of the sound's various frequency components as a heatmap.

We applied the Short-time Fourier Transform (STFT) to calculate the energy distribution of different frequencies in local segments of the audio. The choice of segment width affects the trade-off between time and frequency. To address the time-frequency trade-off issue of STFT, we employed three different widths for transformation and stacked the resulting outputs into a three-channel image. This image served as input for both CNN-based and Vision Transformer (ViT)-based models for prediction. To increase data diversity, we introduced methods such as adding noise and randomly shifting images. However, despite these efforts, we were unable to achieve a consistent improvement in the UAR metric, which remained oscillating between 0.45 and 0.55 without significant progress.

5.4 Conclusion

By prioritizing recall scores, we achieve a higher UAR score but observe a decrease in precision, especially for the last two types. Cross-validation ensures rigorous evaluation and selection of the most effective model. Our cross-validation scores show little difference between the public and private datasets, indicating the model's robustness. Additionally, we employ SHAP values to measure feature importance, considering their advantages over impurity-based feature importance, particularly for high cardinality features. Furthermore, we explore a computer vision approach as an alternative, but it did not lead to significant improvements in performance.

Model	Public Score	Private Score
BB-Voting	.657	.641

Table 3: The final UAR score of public and private dataset

6 Source Code

Our team has made our code accessible on GitHub, utilizing the MIT License. The code repository can be found at the following URL: https://github.com/SeanChenTaipei/Audio_Classification.

References

[1] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[2] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[3] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

[4] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[5] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[7] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[8] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[9] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*, 2020.

[10] KuanHaoHuang. T-brain 2021，tomofun 狗音辨識 ai 百萬挑戰賽：top 3% solution.