Multimodal Classification of Voice Pathology

*Machine Learning Final Project, The 2023 AI CUP Spring Competition

Liang-Jen Huang
Institute of Data Science, NCKU
January 9, 2025

***Abstract*—This is the topic of the 2023 Spring AI Cup competition, aiming to use a multimodal approach by combining voice signals and medical history records to achieve automatic detection and classification of pathological voices. Competition rules can be found at: https://tbrain.trendmicro.com.tw/Competitions/Details/27. My report and detailed code are available at: https://github.com/edogawa-liang/multimodal-pathological-voice-classification**

I. Introduction

In recent years, with the increasing pressures of modern life, various lifestyle-related diseases have drawn significant attention from the medical community. Among them, voice disorders are commonly observed in professions requiring extensive voice use, such as teachers, salespeople, lecturers, and retail vendors. Due to the vocal cords being located deep within the throat, diagnosing voice-related disorders is particularly challenging, requiring specialized medical equipment operated by trained professionals for proper diagnosis and treatment. Additionally, modern busy lifestyles often lead to delays in seeking medical attention.

During the recent COVID-19 pandemic, conducting oral endoscopic examinations posed a risk of droplet transmission. If artificial intelligence could be applied in a non-contact manner by combining dynamic voice signals with static medical history records to detect and classify throat-related conditions, it would enable early detection and treatment. This advancement could be a major breakthrough, bringing significant benefits to all individuals suffering from voice disorders.

II. Data Description

The data for this study were sourced from the Far Eastern Memorial Hospital's ENT Department Voice Database and the Ministry of Education's AI Competition and Data Annotation Collection Project. The dataset includes medical history records in CSV format and voice signals in WAV format. It consists of 1,000 patients in the training set and 500 patients in the test set. The medical history records contain 26 variables, while the voice signals are 1 to 3 seconds of sustained "Ah" sounds. The variables in the dataset are introduced in Table I.

The prediction target is the "Disease category", classified into five categories: 1. Phonotrauma 2. Incomplete glottic closure 3. Vocal palsy 4. Neoplasm 5. Normal

This classification aims to identify pathological voice conditions based on the multimodal data provided.

In addition, these five categories suffer from severe class imbalance, as shown in Figure 1.

TABLE I
Variable Description

Variable	Scale	Description
Sex	1/2	Male/Female
Age	Numbers	
Narrow pitch range	0/1	No/Yes
Decreased volume	0/1	No/Yes
Fatigue	0/1	No/Yes
Dryness	0/1	No/Yes
Lumping	0/1	No/Yes
Heartburn	0/1	No/Yes
Choking	0/1	No/Yes
Eye dryness	0/1	No/Yes
PND	0/1	No/Yes
Diabetes	0/1	No/Yes
Hypertension	0/1	No/Yes
CAD	0/1	No/Yes
Head and Neck Cancer	0/1	No/Yes
Head injury	0/1	No/Yes
CVA	0/1	No/Yes
Smoking	0/1/2/3	Never/past/active/e-cigarette
PPD	Numbers	Pack (of cigarettes) per day
Drinking	0/1/2	Never/past/active
Frequency	0/1/2/3	Not/occasionally/weekly/daily
Onset of dysphonia	1/2/3/4/5	Sudden/Gradually/On and off/Since childhood
Noise at work	1/2/3	Not/a little/noisy
Diurnal pattern	1/2/3/4	Worse in the morning/Worse in the afternoon/
Occupational vocal demand	1/2/3/4	Always/Frequent/Occasional/Minimal
Voice handicap index-10	0 to 40	

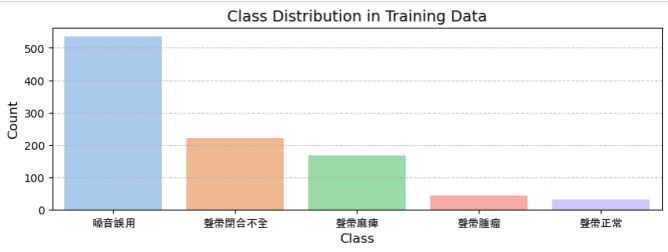

Fig. 1. Number of Samples in the Five Categories

III. Method

In this analysis, various data processing techniques and modeling strategies were progressively employed to address the prediction problem effectively. The steps undertaken are detailed as follows:

A. Handling Missing Values in Medical History

Variables with missing values included PPD (packs per day) and VHI-10 (Voice Handicap Index). Different imputation

methods were applied. For PPD, missing values were filled with zero for samples where the patients never smoked or had quit smoking, while the median of the feature was used for the rest. For VHI-10, missing values were imputed using the KNN Imputer to ensure data integrity.

B. Medical History Data Preprocessing

For the medical history data, nominal variables were encoded using both One Hot Encoding and Multiple Correspondence Analysis (MCA), and the impact of these methods on model performance was compared. Ordinal variables, due to their inherent ranking structure, were left unprocessed. Additionally, two models were used to select key features: LightGBM based on feature importance and Multinomial Logistic Regression based on coefficients. The top 15 most influential variables were selected for modeling.

C. Voice Signal Processing and Feature Extraction

For voice signals, all audio files were standardized to a duration of two seconds. For audio files shorter than two seconds, the middle segment was duplicated to pad the length. For audio files longer than two seconds, the middle portion was removed to reduce the length. In terms of feature extraction, a pre-trained Wav2Vec2 model was fine-tuned using pathological voice labels, and 784-dimensional feature vectors were extracted from the audio data.

To further refine the features and reduce dimensionality, PCA (Principal Component Analysis) was applied. However, unlike traditional PCA, which retains only the top principal components, it was hypothesized that the subtle features of pathological voices might be hidden in later principal components rather than being captured by the top components dominated by global structures (e.g., volume, pitch variability, or frequency). For instance, the top components might represent general patterns such as overall loudness or pitch, while the subtle features of pathological voices, such as fine tremors, minor pitch modulations, or breathiness, might appear in later components. Based on this hypothesis, the PCA results were further refined to select 36 principal components by feature selection using models, aiming to retain the most critical information related to pathology while effectively reducing noise.

IV. Experiment

Due to the severe class imbalance in the dataset, weighted sampling was applied during model training. Higher weights were assigned to underrepresented classes to mitigate bias in the model. The following five versions were experimented with:

- **Ver.1:** Using medical history data (Feature Selection).
- **Ver.2:** Combining medical history data (Feature Selection) with voice data PCA (Feature Selection).
- **Ver.3:** Combining medical history data (Feature Selection + MCA) with voice data PCA (Feature Selection).
- **Ver.4:** Combining medical history data (Feature Selection + MCA) with voice data PCA (Top Principal Components).
- **Ver.5:** Combining medical history data (Feature Selection) with voice data PCA (Top Principal Components).

TABLE II
Model Performance Comparison

Version	Model	Accuracy	Recall	Precision	F1-Score
Ver.1	Logistic	0.504	0.433	0.406	0.379
	SVM	0.566	0.485	0.436	0.435
	XGB	0.632	0.376	0.384	0.377
	LGBM	0.670	0.406	0.425	0.406
Ver.2	Logistic	0.558	0.503	0.459	0.438
	SVM	0.612	0.538	0.476	0.477
	XGB	0.710	0.422	0.413	0.416
	LGBM	0.714	0.419	0.414	0.414
Ver.3	Logistic	0.560	0.476	0.453	0.438
	SVM	0.600	0.522	0.479	0.470
	XGB	0.710	0.420	0.478	0.424
	LGBM	0.716	0.415	0.416	0.412
Ver.4	Logistic	0.564	0.542	0.476	0.460
	SVM	0.622	0.504	0.461	0.469
	XGB	0.714	0.425	0.487	0.430
	LGBM	0.698	0.402	0.411	0.402
Ver.5	Logistic	0.598	0.603	0.517	0.497
	SVM	0.642	0.578	0.503	0.509
	XGB	0.708	0.444	0.511	0.454
	LGBM	0.700	0.415	0.412	0.411

Experimental results demonstrated that combining medical history data with voice signals significantly improved model performance. When MCA was applied to process medical history data, the performance of Logistic Regression and SVM declined, while tree-based models such as LightGBM and XGBoost showed notable improvements. On the other hand, when only the top principal components extracted through PCA were used, Logistic Regression, SVM, and XGBoost showed slight improvements, whereas LightGBM's performance decreased. This indicates that subtle variations in voice signals can be captured and utilized more effectively by LightGBM. Overall, the analysis suggests that using only the top principal components that explain the most variance is often sufficient to build effective models.

V. Conclusion

This study successfully combined voice features with medical history data to significantly enhance the accuracy of pathological voice classification. The results demonstrated that integrating multimodal data could offer a more comprehensive understanding of voice disorders. Notably, the testing performance ranked 20th out of 371 on the public leaderboard, underscoring the effectiveness of the proposed methodology compared to other approaches.

The experiments further revealed that while medical history data effectively captured key categorical and ordinal features, voice signals contributed subtle yet critical variations, which were better leveraged by models such as Logistic Regression and LightGBM. Moreover, the application of wav2vec with PCA for feature engineering highlighted the importance of retaining both global and fine-grained details to optimize model performance.

VI. Future Work

Future research can enhance the proposed framework by transitioning from the current two-stage design to an end-to-end approach. The current method generates embeddings for medical history and voice data independently and then concatenates them for modeling. However, this design lacks direct interaction between the two modalities during feature extraction, and the final classification model is not optimized for embedding generation.

An end-to-end deep learning architecture can address these limitations by directly integrating the embedding generation process with the final classification objective. This approach would enable seamless interaction between medical history and voice data. Additionally, incorporating a Cross Attention mechanism could facilitate learning the relationships between the two modalities, mapping them into a shared latent space for better interaction and more effective feature utilization.