Prediction of dysphagia aspiration through machine learning-based analysis of patients' postprandial voices

Jung-Min Kim
Seoul National University
Min-Seop Kim
Dongguk University
Sun-Young Choi
Seoul National University Bundang Hospital
Ju Seok Ryu
jseok337@snu.ac.kr

Seoul National University Bundang Hospital

Research Article

Keywords: Dysphagia aspiration, Postprandial voice-based, Disease prediction model, Machine learning, Remote diagnosis and monitoring technology, Voice analysis

Posted Date: September 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3294017/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal of NeuroEngineering and Rehabilitation on March 30th, 2024. See the published version at https://doi.org/10.1186/s12984-024-01329-6.

Abstract

Background: Conventional diagnostic methods for dysphagia have limitations such as long wait times, radiation risks, and restricted evaluation. Therefore, voice-based diagnostic and monitoring technologies are required to overcome these limitations. Based on our hypothesis regarding the impact of weakened muscle strength and the presence of aspiration on vocal characteristics, this single-center, prospective study aimed to develop a machine-learning algorithm for predicting dysphagia status (normal, and aspiration) by analyzing postprandial voice limiting intake to 3cc.

Methods: This study was a single-center, prospective cohort study, conducted from September 2021 to February 2023, at the Seoul National University Bundang Hospital. A total of 204 participants were included, aged 40 or older, comprising 133 without suspected dysphagia and 71 with dysphagia-aspiration.Voice data from participants were collected and used to develop dysphagia prediction models using the Audio Spectrogram Transformer process with MobileNet V3. Male-only, female-only, and combined models were constructed using 10-fold cross-validation. Through the inference process, we established a model capable of probabilistically categorizing a new patient's voice as either normal or indicating the possibility of aspiration.

Results: The pre-trained models (mn40_as and mn30_as) exhibited superior performance compared to the non-pre-trained models (mn4.0 and mn3.0). The best-performing model, mn30_as, which is a pre-trained model, demonstrated an average AUC across 10 folds as follows: combined model 0.7879 (95% CI 0.7355-0.8403; max 0.9531), male model 0.7787 (95% CI 0.6768-0.8806; max 1.000), and female model 0.7586 (95% CI 0.6769-0.8402; max 0.9132). Additionally, the other models (pre-trained; mn40_as, non-pre-trained; mn4.0 and mn3.0) also achieved performance above 0.7 in most cases, and the highest fold-level performance for most models was approximately around 0.9.

Conclusions: This study suggests the potential of using simple voice analysis as a supplementary tool for screening, diagnosing, and monitoring dysphagia aspiration. By directly analyzing the voice itself, this method enables simpler and more remarkable analysis in contrast to conventional clinical evaluations. The postprandial voice-based prediction model holds implications for improving patient quality of life and advancing the development of non-invasive, safer, and more effective intervention methods.

Trial registration: This study was approved by the IRB (No. B-2109-707-303) and registered on clinicaltrials.gov (ID: NCT05149976).

Introduction

Dysphagia is a difficulty in swallowing food normally due to impaired movement in swallowing-related organs, which increases the risk of food passing into the airway. [1] The most common diagnostic method, the videofluoroscopic swallowing study (VFSS), requires specialized equipment typically found only in hospitals, resulting in long wait times and radiation risks. [2–4] Despite the availability of other diagnostic methods such as fiberoptic endoscopic evaluation of swallowing (FEES), manometry, and laryngeal electromyography, they also have limitations. [5–9] For example, FEES can only evaluate the pharyngeal stage and carries the risk of complications such as anterior or posterior epistaxis, and laryngospasm. [6] Meanwhile, manometry requires invasive procedures, and both manometry and laryngeal electromyography remain challenging to analyze. [7–9]

Thus, the current dysphagia diagnostic methods in clinical settings are limited in their ability to continuously monitor changes in a patient's condition over time. [10]

To overcome the limitations of conventional tests, researchers have focused on developing voice-based diagnostic and monitoring technologies for patients with dysphagia in clinical settings. [11–14] Dysphagia-induced food aspiration alters the airway vibrations, resulting in changes in voice quality and parameters. [14–16] Previous studies analyzing the voice of patients with dysphagia have reported significant changes in parameters such as RAP, SHIM, and NHR due to aspiration into the airway. [11–14] However, these studies often extracted specific vocal parameters rather than analyzing the patient's voice itself, which may limit their universal application in diagnosis and monitoring.

We hypothesized that patients with dysphagia may experience changes in their voice because weakened muscles and aspiration below the vocal cords. Additionally, it is assumed that a more precise assessment can be achieved through the application of Machine Learning to analyze patients' voices. Based on this hypothesis, the primary objective of this study was to explore the efficacy of machine learning into predicting dysphagia by analyzing the post-meal voices of patients. The ultimate goal was to establish the groundwork for the future development of an advanced dysphagia diagnosis and monitoring system.

Methods

Study Design

This single-center, prospective study was conducted from October 2021 to February 2023 at the Seoul National University Bundang Hospital. The study protocol was approved by the Seoul National University Bundang Hospital Institutional Review Board (IRB No.: B-2109-707-303) and registered at clinicaltrials.gov (ClnicalTrial.gov ID: NCT05149976). This study was conducted in accordance with the strengthening the reporting of observational studies in epidemiology (STROBE) guidelines.

Participants

The inclusion criteria for selecting study subjects are as follows: patients (1) who have signs and symptoms of dysphagia and are scheduled for VFSS, (2) can record ‘Ah ~ for 5 seconds’, and (3) normal subjects without dysphagia symptoms who can record voice as a normal. The exclusion criteria were as follows: (1) inability to speak, (2) inability to speak according to the researcher’s instructions, and (3) patients whose VFSS was reexamined.

Voice recordings were obtained with the consent of 285 participants, including 159 individuals without suspected dysphagia and 126 who underwent VFSS because of suspected dysphagia aspiration (PAS 5–7). Two participants (one from the normal group based on VFSS examination and one from the dysphagia aspiration group) with poor audio quality were excluded from the collected recordings. In the patient group, 1 participant aged < 40 years was included in the aspiration subgroup. To eliminate age-related bias in the patient's voice-based predictive model, 79 participants under the age of 40 years (comprising 75 participants without suspected dysphagia, 3 participants from the normal group by VFSS examination and 1 participant from the aspiration group) were excluded from the study population. The final study population consisted of 204 participants, categorized into the normal group (133 participants, including both individuals without suspected dysphagia and those who received a normal diagnosis based on VFSS), and the aspiration group (71 participants), based on VFSS interpretations by physicians. Figure 1 shows detailed flow chart of the recruitment of research subjects.

Voice Recording Procedures

After obtaining consent from the patient, a VFSS was performed using the modified Logemann protocol which is commonly used in domestic hospitals, to evaluate dysphagia. [17] During the test, the patient was instructed to repeat the sound 'Ah~' once or more for at least 5 seconds after consuming water (CUP), thin liquid (FT3), thickened liquid (LF), pureed food (SBD), soft and moist food (SF), and yogurt (YP), while their voice was recorded using a Sony ICD-TX660 recorder while limiting intake to 3cc. For the control group, which consisted of subjects without dysphagia, their voices were recorded once or more for at least 5 seconds before and after drinking water using a voice recording function on a mobile device.

In total, 411 voice files were collected, consisting of 217 files from the normal group (72 files for men, 145 files for women) and 194 files from the aspiration group. (148 files for men, 46 files for women). Among them, two data files from the normal group of men were excluded because they could not be analyzed. Therefore, 409 data files were utilized for the analysis.

Voice Data Preprocessing

Following the procedure outlined in Fig. 2, preprocessing was conducted on the voice data, and based on this, a machine learning model was constructed.

Step 1. Conversion of Voice Data Format

To make the acquired voice files suitable for machine learning, we performed two steps: (1) converting the stereo format to mono and (2) standardizing the data files, which were in various formats such as wav, m4a, and mp3, to the mp3 format. As a result, 671 data files (284 normal group files: men (99 files), women (185 files), 387 aspiration group files: men (295 files), women (92 files) were converted to mp3 format and utilized for model development.

Step 2. Creation of Train and Test Dataset for k-Fold Cross Validation

The mp3-formatted data were divided into training and testing sets in a ratio of approximately 9:1 for each group. For 10-fold cross-validation, the data has been divided into 10 sections based on individuals in each group. In other words, data from the same person is grouped together in the same fold. The range of these sections was varied to create 10-fold cross-validation datasets.

Step 3. Conversion of Voice Data to HDF5 Format for Model Training

To train MobileNet V3 with an efficient large-scale audio tagging model using voice data, we converted the data into a suitable format. This was achieved by modifying the create_h5pymp3_dataset.py code from PaSST (Patchout faSt spectrogram transformer, Apache-2.0 license) research and transforming the training/test data into HDF5 format files. [18, 19] The structure of the transformed HDF5 data consisted of the file name, audio data in MP3 format, and labeled information on normal, or aspiration in numeric form. The transformed data were saved as Dysphagia_post.train_mp3.hdf and Dysphagia_post.test_mp3.hdf.

Step 4. Preprocessing of Voice Data

Voice preprocessing was conducted using an efficient large-scale audio tagging model (Efficient AT Model, MIT license), which is widely utilized for audio classification tasks. [20, 21] This process involved defining the $AugmentMelSTFT$ class for audio augmentation and converting audio waveforms into Mel spectrogram format suitable for machine learning. It consists of several steps, including pre-emphasis filtering, short-time Fourier transform (STFT), power magnitude computation, and a Mel frequency filter bank. The hyperparameters such as the number of mels (128), sample rate (32,000), window length (800), hop size (320), number of fast Fourier transforms (FFT, 1024), etc. control the preprocessing process. The 'freqm (48)' and 'timem (192)' hyperparameters enable frequency and time masking for data augmentation, respectively. In summary, this code enables the augmentation of audio data and their transformation into a perceptually related Mel-spectrogram representation.

Development of Dysphagia Prediction Models.

The preprocessed voice data underwent an Audio Spectrogram Transformer process using multi-head attention pooling, which can improve learning performance in NLP and other speech analysis tasks. [20, 22] MobileNet V3 was utilized as the machine learning technique for voice training. Binary cross entropy with logits loss was used as the loss function to evaluate the predictive performance of the algorithm. [20] The two pre-trained models were named mn30_as, and mn40_as in accordance with the width_mult and hyperparameters in the Efficient AT Model. Similarly, two non-pre-trained models were designed with the same width_mult and hyperparameters as the pre-trained models, and were uniformly named mn3.0, and mn4.0, respectively. In situations where the dataset is limited, non-pre-trained models may encounter challenges in effectively extracting features. [22, 23] Therefore, this study conducted a comparison between the pre-trained and non-pre-trained models. [20, 21] The model constructed in this manner was validated for prediction accuracy using a k-fold cross-validation with k = 10. All the models were trained for 5e-5 learning rate, 12 number of workers, 150 epochs, and 64 batch sizes.

Inference design based on learned machine learning results.

The machine learning model trained using the before mentioned method was applied to actual patient voice data to determine the probability of normal, or aspiration using a technique divided into four stages: 'decode_mp3,' 'pad_or_truncate,' 'pydub_augment,' and 'audio_tagging'. In the 'decode_mp3' process, the input mp3 file was converted into an np.array waveform. The 'pad_or_truncate' process converted the audio waveform of the input mp3 file into a specific length of audio for discrimination. In our study, the patients' voices did not start at the beginning of the audio files. To minimize noise, we adjusted the $audio_{\leq}n > h$ by considering the actual length of the patient's voice, enabling us to effectively analyze the data by cutting it accordingly. The 'pydub_augment' process augmented the audio waveform data to improve prediction ability, while the 'audio_tagging' process transformed the augmented data into Mel spectrogram format. [20] Finally, these steps provide the probability of normal, or aspiration, based on the prediction of the model.

Outcome Variables

The primary outcome of this study is the area under the curve (AUC), considering the imbalanced distribution of data among groups in the medical field. Additionally, the degree of prediction for the model was analyzed from the perspectives of accuracy, mean average precision (mAP), recall, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score, and a final model was established.

Statistical Analysis

The baseline characteristics were analyzed, with mean ± SD used for continuous variables and number (%) for nominal variables. We used appropriate statistical tests to compare the baseline characteristics between the groups. We conducted a chi-square test for nominal variables and a Mann-Whitney U test for continuous variables. These tests were chosen because of violations of normality based on the Shapiro-Wilk test and sphericity assumptions based on Mauchly's test of sphericity. The significance level was set at $p < 0.05$. The performance of each model was evaluated using metrics such as the AUC, accuracy, mAP, recall, specificity, PPV, NPV, and F1-score. Additionally, the performance of each model was calculated for each fold, and the average values across the ten folds were selected as the final predictions for the model. All the analyses were conducted using Python and Google Colaboratory Pro + GPU A100. Statistical analysis and machine learning modeling were conducted between January and July 2023.

Results

Table 1 shows the demographic characteristics of all the study subjects.

Table 1
Demographic Characteristics

	Normal	Aspiration	p-value
Sex (N (%))			
Men	45 (33.83%)	53 (74.65%)	< 0.001*
Women	88 (66.17%)	18 (25.35%)	(x^2: 29.28, df: 1)
Age (Mean ± SD)			
Total	61.34 ± 12.92	72.45 ± 12.01	< 0.001**
Men	63.18 ± 13.13	72.45 ± 11.66	< 0.001**
Women	60.40 ± 12.79	72.44 ± 13.34	0.001**
Diagnosis (N (%))			
Central Nervous System Disorders	18 (13.53%)	19 (26.76%)	< 0.001*
Digestive System and Dental Disorders	3 (2.26%)	10 (14.08%)	(x^2: 35.19, df: 6)
Pulmonary Disorders	3 (2.26%)	9 (12.68%)	
Other Cancers	6 (4.51%)	3 (4.23%)	
Vocal Fold Disorders	2 (1.50%)	2 (2.82%)	
Aging-Related Disorders	13 (9.77%)	7 (9.86%)	
None	88 (66.17%)	21 (29.58%)	

* The Chi-square test results show a significant difference. To address gender bias, separate models were constructed for each gender (male and female). The data was then divided into 10 folds for each gender. After that, the results were combined in the gender-neutral model, effectively removing any gender-related biases.

** The Mann-Whitney U test results indicate a significant difference between the two groups. However, to eliminate bias, participants under the age of 40 were excluded from the analysis.

For the 10-fold cross-validation, male-only, female-only, and combined (men + women) models were constructed. Table 2 shows the average predictive performance of the combined (men + women) model across 10 folds. Regarding the primary outcome, the average AUC values were mn40_as = 0.7798 (95% CI 0.7130–0.8465; max in 10 folds 0.9777) and mn30_as = 0.7879 (95% CI 0.7355–0.8403; max in 10 folds 0.9531) for the pre-trained models and mn4.0 = 0.7658 (95% CI 0.7069–0.8246; max in 10 folds 0.9324), mn3.0 = 0.7603 (95% CI 0.6950–0.8256; max in 10 folds 0.8973) for the non-pre-trained models. Owing to the smaller amount of available data, the pre-trained models (mn40_as and mn30_as) demonstrated higher performance than the non-pre-trained models (mn4.0 and mn3.0). In addition, all models consistently showed high prediction accuracy in analyzing a person's voice, with metrics such as accuracy, mAP, recall, specificity, PPV, NPV, and F1-score exceeding approximately 70%.

Table 2. The levels of prediction for combined (men + women) model

Model	Pre-trained models		Non-pre-trained models	
	mn40_as	mn30_as	mn40	mn30
AUC (Area Under the Curve)				
AUC average (95% CI)	0.7798 (0.7130, 0.8465)	0.7879 (0.7355, 0.8403)	0.7658 (0.7069, 0.8246)	0.7603 (0.6950, 0.8256)
AUC max in 10 folds	0.9777	0.9531	0.9324	0.8973
Accuracy (%)				
Accuracy average (95% CI)	71.81 (65.81, 77.81)	69.51 (64.90, 74.12)	73.37 (66.31, 80.44)	73.39 (67.20, 79.58)
Accuracy max in 10 folds	92.31	82.69	95.24	89.29
mAP (Mean Average Precision, %)				
mAP average (95% CI)	79.15 (72.89, 85.41)	80.04 (75.03, 85.06)	78.26 (73.72, 82.79)	77.12 (71.62, 82.63)
mAP max in 10 folds	97.97	95.77	93.17	89.58
Recall (%)				
Recall average (95% CI)	71.85 (65.54, 78.16)	69.45 (65.04, 73.86)	73.37 (66.31, 80.44)	73.22 (66.96, 79.48)
Recall max in 10 folds	92.31	82.69	95.24	89.29
Specificity (%)				
Specificity average (95% CI)	71.97 (65.70, 78.25)	70.08 (65.78, 74.39)	72.82 (65.76, 79.88)	72.78 (66.57, 78.98)
Specificity max in 10 folds	92.26	83.04	93.75	88.39
PPV (Positive Predictive Value, %)				
PPV average (95% CI)	71.62 (65.50, 77.73)	70.25 (65.67, 74.83)	73.46 (66.20, 80.72)	73.07 (66.79, 79.35)
PPV max in 10 folds	92.26	82.89	95.49	87.74
NPV (Negative Predictive Value, %)				
NPV average (95% CI)	71.67 (65.51, 77.84)	70.20 (65.69, 74.72)	73.46 (66.20, 80.72)	73.07 (66.79, 79.35)
NPV max in 10 folds	92.26	82.89	95.49	87.74

Model	Pre-trained models		Non-pre-trained models	
	mn40_as	mn30_as	mn40	mn30
AUC (Area Under the Curve)				
F1 Score				
F1 Score average (95% CI)	0.7189 (0.6564, 0.7814)	0.6951 (0.6520, 0.7382)	0.7321 (0.6609, 0.8033)	0.7313 (0.6683, 0.7942)
F1 Score max in 10 folds	0.9231	0.8271	0.9519	0.8933

* All metrics represent the predictive performance on the Test Data. The results presented in this table are the average predictive performance (95% CI) across all folds of each model after performing 10-fold cross-validation.

Table 3 presents the average predictive performance for each sex (men and women) across the 10 folds. The average AUC values for the pre-trained model, using mn40_as, were 0.7673 (95% CI 0.6516-0.8829; max in 10 folds 1.000) and 0.7692 (95% CI 0.6828-0.8556; max in 10 folds 0.9375) for the male and female model, respectively. Additionally, for the pre-trained model using mn30_as, the AUC values were 0.7787 (95% CI 0.6768-0.8806; max in 10 folds 1.000) and 0.7586 (95% CI 0.6769-0.8402; max in 10 folds 0.9132) for the male and female models, respectively. For the non-pre-trained model, using mn4.0, the AUC values were 0.7239 (95% CI 0.6230-0.8247; max in 10 folds 0.9000) and 0.7575 (95% CI 0.6476-0.8674; max in 10 folds 0.9412) for the male and female models, respectively. For the non-pre-trained model using mn3.0, the AUC values were 0.6784 (95% CI 0.5713-0.7856; max in 10 folds 0.9603) and 0.7007 (95% CI 0.5494-0.8519; max in 10 folds 1.000) for the male and female models, respectively. Figure 3 presents the macro-average ROC across 10 folds for each model.

Table 3
The levels of prediction for gender-specific model

Model	Male Models				Female Models			
	Pre-trained models		Non-pre-trained models		Pre-trained models		Non-pre-trained models	
	mn40_as	mn30_as**	mn40	mn30	mn40_as	mn40_as	mn40	mn30
AUC (Area Under the Curve)								
AUC average (95% CI)	0.7673 (0.6516, 0.8829)	0.7787 (0.6768, 0.8806)	0.7239 (0.6230, 0.8247)	0.6784 (0.5713, 0.7856)	0.7692 (0.6828, 0.8556)	0.7586 (0.6769, 0.8402)	0.7575 (0.6476, 0.8674)	0.7007 (0.5494, 0.8519)
AUC max in 10 folds	1.0000	1.0000	0.9000	0.9603	0.9375	0.9132	0.9412	1.0000
Accuracy								
Accuracy average (95% CI)	78.41 (72.54, 84.28)	80.60 (74.16, 87.03)	81.98 (77.88, 86.08)	80.03 (76.29, 83.77)	65.18 (58.77, 71.60)	65.77 (60.49, 71.04)	64.09 (50.40, 77.79)	50.79 (40.00, 61.57)
Accuracy max in 10 folds	96.08	96.15	90.20	94.12	80.00	80.77	93.94	69.23
mAP (Mean Average Precision)								
mAP average (95% CI)	79.60 (69.58, 89.61)	80.15 (71.25, 89.05)	75.52 (67.43, 83.60)	72.21 (63.66, 80.77)	66.27 (49.73, 82.81)	74.10 (66.44, 81.76)	75.51 (65.59, 85.42)	72.45 (60.36, 84.55)
mAP max in 10 folds	100.00	100.00	90.06	94.22	93.43	91.43	94.88	100.00
Recall								
Recall average (95% CI)	78.23 (72.38, 84.08)	80.60 (74.16, 87.03)	81.98 (77.88, 86.08)	80.32 (76.66, 83.99)	65.93 (59.40, 72.46)	65.77 (60.49, 71.04)	64.88 (51.41, 78.36)	50.79 (40.00, 61.57)
Recall max in 10 folds	96.08	96.15	90.20	94.12	80.00	80.77	93.94	69.23
Specificity								

Model	Male Models				Female Models			
	Pre-trained models		Non-pre-trained models		Pre-trained models		Non-pre-trained models	
	mn40_as	mn30_as**	mn40	mn30	mn40_as	mn40_as	mn40	mn30
Specificity average (95% CI)	70.66 (61.63, 79.70)	73.88 (64.17, 83.59)	69.70 (62.53, 76.87)	66.68 (59.85, 73.52)	62.99 (54.03, 71.96)	58.34 (50.27, 66.41)	58.02 (48.14, 67.91)	52.81 (49.11, 56.52)
Specificity max in 10 folds	97.62	96.88	85.32	87.70	83.33	80.95	93.33	65.62
PPV (Positive Predictive Value)								
PPV average (95% CI)	72.74 (64.66, 80.83)	78.87 (70.42, 87.33)	78.90 (68.58, 89.22)	76.23 (66.27, 86.18)	62.17 (53.44, 70.91)	59.66 (49.52, 69.80)	46.15 (30.08, 62.22)	33.79 (22.66, 44.92)
PPV max in 10 folds	91.67	95.45	94.23	91.42	83.33	86.96	95.00	64.29
NPV (Negative Predictive Value)								
NPV average (95% CI)	72.74 (64.66, 80.83)	78.96 (70.58, 87.34)	78.90 (68.58, 89.22)	76.23 (66.27, 86.18)	62.28 (53.55, 71.01)	59.66 (49.52, 69.80)	46.20 (30.11, 62.29)	33.79 (22.66, 44.92)
NPV max in 10 folds	91.67	95.45	94.23	91.42	83.33	86.96	95.00	64.29
F1 Score								
F1 Score average (95% CI)	0.7721 (0.7088, 0.8354)	0.7938 (0.7214, 0.8662)	0.7923 (0.7334, 0.8513)	0.7689 (0.7147, 0.8231)	0.6476 (0.5850, 0.7102)	0.6401 (0.5838, 0.6965)	0.5558 (0.4022, 0.7094)	0.3940 (0.2853, 0.5028)
F1 Score max in 10 folds	0.9623	0.9618	0.9040	0.9398	0.8000	0.8039	0.9388	0.5664

* The table shows average predictive performance across all folds of each model after 10-fold cross-validation.

** In the case of the male model's performance on mn30_as, the highest results were achieved across all metrics in one of the folds, with all indicators reaching 100%. However, considering that this could be indicative of overfitting to the specific data configuration of the train and test datasets, the results from the second-highest performing fold were reported instead.

The program that we aim to develop for the constructed model is shown in Fig. 4 determines the probabilities, in percentage, of classifying a patient's voice input into normal or aspiration during the inference stage. This is the final output of the study.

Discussion

This study applied an efficient large-audio tagging model [20], which is known for its outstanding performance in sound analysis, to predict the presence of postprandial dysphagia at two levels (normal and aspiration). It demonstrates a high predictive performance, with the majority of the models achieving an AUC value of over 0.75, considering the diversity of people's voices. In particular, the mn30_as model, which had the highest number of hyperparameters among the trained models, demonstrated an AUC of approximately 0.7879 in the combined model and 0.7787 in the male model, indicating good performance in predicting dysphagia aspiration. Additionally, all other predictive performance measures for the combined and male models yielded high results, exceeding 70%.

Various studies on dysphagia aspiration have been conducted using non-invasive methods. The 3-ounce water swallow test showed a sensitivity of 59–96.5% and specificity of 15–59% when compared with FEES and VFSS. [24–26] The Gugging swallowing screen test had a sensitivity of 100% and a specificity of 50–69% in acute stroke patients. [27] Sensitivity and specificity for dysphagia based on language and speech-related dysfunctions were reported as follows: aphasia (36% and 83%, respectively), dysarthria (56% and 100%, respectively), and a combination of variables (64% and 83%, respectively). [28] Dysphonia, dysarthria, gag reflex, cough, and voice changes were used as diagnostic performance measures. [29] Other screening tools, such as the food intake level scale (FOIS), modified Mann assessment of swallowing ability test, and volume-viscosity swallow test (V-VST), etc., were also developed and subjected to performance validation. [16, 26, 30–37] While predictive performance varies depending on the research techniques, all of them require expert intervention for accurate diagnosis and monitoring, posing limitations on their applicability for everyday life monitoring. Efforts to observe voice changes during dysphagia monitoring are ongoing. [11–14, 38, 39]

Most previous studies on voice analysis in patients with dysphagia have focused on analyzing frequency perturbation measures (RAP, Jitter, PPQ, etc.), amplitude perturbation measures (Shimmer, APQ, etc.), and noise analysis (NHR) to differentiate between high- and low-risk groups. [11–14, 38, 39] Additionally, vocal intensity (MVI) and vocal duration measures (MPT) were used as voice analysis indicators. [38] Moreover, some studies have analyzed the correlations between these measures and established clinical diagnostic indicators for dysphagia, such as the penetration-aspiration scale (PAS), videofluoroscopic dysphagia scale (VDS), and American speech-language-hearing association national outcome measurement system swallowing scale (ASHA-NOMS). [38] Some studies have employed the Praat program to extract these sound parameters and analyze each indicator, either using voice-only or combining voice with clinical data indicators, trained with algorithms such as Logistic Regression, Decision Tree, Random Forest, SVM, GMM, and XGBoost. [12] Another study reported the results of dysphagia prediction using specific phonation or articulation features trained using support vector machine (SVM), random forest, and other methods. [39] However, these studies have limitations in that they only analyzed specific numerical indicators of voice and failed to analyze the overall voice itself.

Therefore, in this study, we trained a dysphagia prediction model using the entire voices of patients, represented as mel-spectrograms. Our model design focused on noise removal, prediction performance, and light-weighting for mobile integration. To reduce the noise from audio files, we implemented preprocessing steps from an efficient large-scale audio tagging model, resulting in improved prediction performance. [20, 21] Regarding the second consideration, we experimented with different models including the ResNet model, which is known for its excellent performance in CNN image recognition. [40, 41] However, its accuracy was relatively low. We also found that training the model solely on the Jitter, RAP, and Shimmer parameters did not yield stable results. Considering the recent advancements in machine learning for sound analysis, we ultimately chose the current learning model. Moving on to the third consideration, we focused on model light-weighting, to achieve real-time dysphagia diagnosis, monitoring, and intervention in mobile or resource-constrained environments. We converted the audio data from stereo to mono format, improving efficiency by eliminating the need for simultaneous processing of the two channels and enhancing voice recognition accuracy. [42] Additionally, we unified and compressed the files into mp3 format for real-time processing on mobile devices. [43, 44] Utilizing the HDF5 data format provides faster loading, increased storage efficiency, and compatibility with various programming languages. [45, 46] Throughout the study, we prioritized a compact model that occupied less storage space and enabled fast prediction of speech impairments. Employing MobileNetV3, a light-weighting and high-performance model, ensures the efficient execution of mobile devices. [47] We adapted the efficient large-scale audio tagging model [20, 21] as a reference, tailored to our specific data environment.

This study developed a model to predict dysphagia - aspiration based on the postprandial voice. The expected benefits of this study are as follows. First, by determining the occurrence of aspiration and providing clinicians with more parameters through voice, it enhances the clinical utility compared to previous studies. Second, it is anticipated that the diagnosis time for both outpatient and inpatient cases will be significantly reduced, providing additional diagnostic parameters for a more accurate assessment of dysphagia. Third, this study is expected to lay the groundwork for designing diagnostic, treatment, and management systems by integrating them with future developments, such as a mobile application-based dysphagia meal guide monitoring system.

Limitations

This study has several limitations. First, owing to the limited availability of voice data for individuals with dysphagia, we did not create a validation set, instead, we used a 9:1 training-to-testing data split (10-fold cross-validation). Second, due to the limited number of recruited female aspiration subjects, the female model showed lower performance compared with the combined model and male model. Third, voice data collection for healthy individuals and patients with dysphagia occurred in different environments and with varying numbers of participant, whereas the diet types were not standardized. Fourth, as a mel-spectrogram-based machine learning model, we lacked characteristic parameter extraction, which is similar to conventional voice indicators. In future studies, we aim to develop a more predictive model with better performance by recording a more diverse range of voices and diet types in patients with dysphagia, and comparing voice changes before and after meals.

Conclusions

This study suggests the potential of simple voice analysis as a supplementary tool for screening, diagnosing, and monitoring dysphagia. Our high-performance postprandial voice-based prediction model highlights the possibility of using voice-based technology for the diagnosis and management of dysphagia. By analyzing the voice itself, this method allows for easier and outstanding analysis compared to traditional clinical evaluations such as VFFS or FEES. Moreover, it empowers patients to record their voices at home, enabling self-monitoring of aspirations in daily life, while providing clinical practitioners with valuable everyday data to track changes. By identifying aspiration in patients' daily lives, this approach has the potential to improve patients' quality of life and enable the development of non-invasive, safer, and more effective intervention methods.

Abbreviations

- VFSS: Videofluorocopic swallowing study
- FESS: Fiberoptic endoscopic evaluation of swallowing
- STROBE: The strengthening the reporting of observational studies in epidemiology
- CUP: Consuming water
- FT3: Thin liquid
- LF: Thickened liquid
- SBD: Pureed food
- SF: Soft and moist food
- YP: Yogurt
- HDF5: Hierarchical data format version 5
- AUC: Area under the curve
- mAP: Mean average precision
- PPV: Positive predictive value
- NPV: Negative predictive value
- RAP: Relative average perturbation
- PPQ: Pitch period quotient
- APQ: Amplitude perturbation quotient
- NHR: Noise-to-harmonic ratio
- SHIM: Shimmer percent
- MVI: Maximal voice intensity
- MPT: Maximum phonation time
- SVM: Support vector machine
- GMM: Gaussian mixture model

Declarations

Ethics approval and consent to participate: This study was approved by the Seoul National University Bundang Hospital Institutional Review Board (IRB No.: B-2109-707-303). The study was conducted on patients scheduled for VFSS who consented after receiving an explanation about the research before participating. The normal control group was comprised of only from individuals who agreed to participate after seeing the recruitment notice for this study.

Consent for publication: The patients' voice recordings were anonymized using de-identification numbers. Following a verbal explanation by the researcher, written consent was obtained from the participants for the publication of this paper.

Availability of data and materials

- **Data availability:** All data in this study is available after de-identification upon request. The data that support the findings of this study are available from the first author, Jung-Min Kim (owljm@snu.ac.kr), upon reasonable request.
- **Code availability:** The code will be publicly available on GitHub before publication. If you need information about the code, you can request access from the first authors, Jung-Min Kim (owljm@snu.ac.kr) or Min-Seop Kim (tjqtjq0516@gmail.com).

Competing interests: Dr Ryu, Jung-Min Kim, and Min-Seop Kim reported owing patent No. 10-2023-0095566. This patent is owned by RS Rehab and Bundang Seoul National University Hospital. No other disclosures were reported.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2C1007780) This work was Supported by grant no 14-2022-0017 from the SNUBH Research Fund.

Role of the Funder/Sponsor: The funder was not involved in the study design and conduct of any protocol for this study, such as data collection, management, analysis, interpretation, and submit or publication of this manuscript.

Authors' Contributions: Jung-Min Kim and Min-Seop Kim had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Also, both are contributed to this paper and should therefore be regarded as first authors. (Jung-Min Kim: owljm@snu.ac.kr, Min-Seop Kim: tjqtjq0516@gmail.com)

Concept and design: Ryu.

Acquisition data: Choi.

Analysis, or interpretation of data: Jung-Min Kim, Min-Seop Kim.

Drafting of the manuscript: Jung-Min Kim.

Critical revision of the manuscript for important intellectual content: Jung-Min Kim.

Statistical analysis: Jung-Min Kim, Min-Seop Kim.

Obtained funding: Ryu.

Administrative, technical, or material support: Jung-Min Kim, Min-Seop Kim.

Supervision: Ryu.

References

1. Matsuo K, Palmer JB. Anatomy and physiology of feeding and swallowing: normal and abnormal. Phys Med Rehabil Clin North Am. 2008;19(4):691–707.
2. Re GL, et al. Swallowing evaluation with videofluoroscopy in the paediatric population. Acta Otorhinolaryngol Ital. 2019;39(5):279.
3. Costa MMB. *Videofluoroscopy: the gold standard exam for studying swallowing and its dysfunction*. 2010, SciELO Brasil. p. 327–8.
4. Na YJ et al. *Thyroid cartilage loci and hyoid bone analysis using a video fluoroscopic swallowing study (VFSS)*. Medicine, 2019. 98(30).
5. Lind CD. Dysphagia: evaluation and treatment. Gastroenterol Clin. 2003;32(2):553–75.
6. Nacci A, et al. Fiberoptic endoscopic evaluation of swallowing (FEES): proposal for informed consent. Acta Otorhinolaryngol Ital. 2008;28(4):206.
7. Ryu JS, Park D, Kang JY. Application and interpretation of high-resolution manometry for pharyngeal dysphagia. J Neurogastroenterol Motil. 2015;21(2):283.
8. Kunieda K, et al. Relationship between tongue pressure and pharyngeal function assessed using high-resolution manometry in older dysphagia patients with sarcopenia: a pilot study. Dysphagia. 2021;36:33–40.
9. Vaiman M, Eviatar E. Surface electromyography as a screening method for evaluation of dysphagia and odynophagia. Head Face Med. 2009;5(1):1–11.
10. Jayatilake D, et al. Smartphone-based real-time assessment of swallowing ability from the swallowing sound. IEEE J translational Eng health Med. 2015;3:1–10.
11. Ryu JS, Park SR, Choi KH. Prediction of laryngeal aspiration using voice analysis. Am J Phys Med Rehabil. 2004;83(10):753–7.
12. Park H-Y, et al. Post-stroke respiratory complications using machine learning with voice features from mobile devices. Sci Rep. 2022;12(1):16682.
13. Waito A, et al. Voice-quality abnormalities as a sign of dysphagia: validation against acoustic and videofluoroscopic data. Dysphagia. 2011;26:125–34.
14. Kang YA, et al. Detection of voice changes due to aspiration via acoustic voice analysis. Auris Nasus Larynx. 2018;45(4):801–6.
15. Salghetti A, Martinuzzi A. Dysphagia in cerebral palsy. Eastern J Med. 2012;17(4):188.

16. Brodsky MB, et al. Screening accuracy for aspiration using bedside water swallow tests: a systematic review and meta-analysis. Chest. 2016;150(1):148–63.
17. Logemann JA. *Manual for the videofluoroscopic study of swallowing*. Pro-Ed ed. Vol. 2. 1993, Texas: Ausin.
18. Koutini K et al. *Efficient training of audio transformers with patchout.* arXiv preprint arXiv:2110.05069, 2021.
19. kkoutini F-RS. *PaSST-Efficient Training of Audio Transformers with Patchout*. 2023; Available from: https://github.com/kkoutini/PaSST.
20. Schmid F, Koutini K, Widmer G. *Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation*. in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023. IEEE.
21. fschmid56, t., Joemgu7. EfficientAT. 2023; Available from: https://github.com/fschmid56/EfficientAT.
22. Gong Y, Chung Y-A, Glass J. *Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021. 29: p. 3292–3306.
23. Lou S et al. *Audio-text retrieval in context*. in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022. IEEE.
24. Suiter DM, Leder SB. Clinical utility of the 3-ounce water swallow test. Dysphagia. 2008;23:244–50.
25. Garon BR, Engle M, Ormiston C. Reliability of the 3-oz water swallow test utilizing cough reflex as sole indicator of aspiration. J Neurologic Rehabilitation. 1995;9(3):139–43.
26. Edmiaston J, et al. Validation of a dysphagia screening tool in acute stroke patients. Am J Crit Care. 2010;19(4):357–64.
27. Trapl M, et al. Dysphagia bedside screening for acute-stroke patients: the Gugging Swallowing Screen. Stroke. 2007;38(11):2948–52.
28. Bahia MM, Mourao LF, Chun RYS. Dysarthria as a predictor of dysphagia following stroke NeuroRehabilitation. 2016;38(2):155–62.
29. Daniels SK, et al. Aspiration in patients with acute stroke. Arch Phys Med Rehabil. 1998;79(1):14–9.
30. Nishiwaki K, et al. Identification of a simple screening tool for dysphagia in patients with stroke using factor analysis of multiple dysphagia variables. J Rehabil Med. 2005;37(4):247–51.
31. Kunieda K, et al. Reliability and validity of a tool to measure the severity of dysphagia: the Food Intake LEVEL Scale. J Pain Symptom Manag. 2013;46(2):201–6.
32. Crary MA, Mann GDC, Groher ME. Initial psychometric assessment of a functional oral intake scale for dysphagia in stroke patients. Arch Phys Med Rehabil. 2005;86(8):1516–20.
33. Antonios N, et al. Analysis of a physician tool for evaluating dysphagia on an inpatient stroke unit: the modified Mann Assessment of Swallowing Ability. J Stroke Cerebrovasc Dis. 2010;19(1):49–57.
34. Clavé P, et al. Accuracy of the volume-viscosity swallow test for clinical screening of oropharyngeal dysphagia and aspiration. Clin Nutr. 2008;27(6):806–15.
35. Audag N, et al. Screening and evaluation tools of dysphagia in adults with neuromuscular diseases: a systematic review. Therapeutic Adv chronic disease. 2019;10:2040622318821622.

36. Zhang P-p, et al. Diagnostic accuracy of the eating assessment tool-10 (EAT-10) in screening dysphagia: a systematic review and meta-analysis. Dysphagia. 2023;38(1):145–58.
37. Rofes L, et al. Sensitivity and specificity of the Eating Assessment Tool and the Volume-Viscosity Swallow Test for clinical evaluation of oropharyngeal dysphagia. Neurogastroenterology & Motility. 2014;26(9):1256–65.
38. Song Y-J, et al. Predicting Aspiration Using the Functions of Production and Quality of Voice in Dysphagic Patients. J Korean Dysphagia Soc. 2022;12(1):50–8.
39. Roldan-Vasco S, et al. Machine learning based analysis of speech dimensions in functional oropharyngeal dysphagia. Comput Methods Programs Biomed. 2021;208:106248.
40. Hershey S et al. CNN architectures for large-scale audio classification. in 2017 *ieee international conference on acoustics, speech and signal processing (icassp)*. 2017. IEEE.
41. He K et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
42. Sun S. *Digital audio scene recognition method based on machine learning technology.* Scientific Programming, 2021. 2021: p. 1–9.
43. Pollak P, Behunek M. Accuracy of MP3 Speech Recognition Under Real-World Conditions. Electrical Engineering, Czech Technical University in Prague; 2011.
44. Fuchs R, Maxwell O. *The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range*. in *Speech Prosody*. 2016.
45. Group H. *The board of trustees of the University of Illinois: "introduction to HDF5"*. 2006; Available from: http://web.mit.edu/fwtools_v3.1.0/www/H5.intro.html.
46. Ji Y et al. *HDF5-based I/O optimization for extragalactic HI data pipeline of FAST*. in *Algorithms and Architectures for Parallel Processing: 19th International Conference, ICA3PP* 2019, Melbourne, VIC, Australia, December 9–11, 2019, Proceedings, Part II 19. 2020. Springer.
47. Howard A et al. *Searching for mobilenetv3*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

Figures

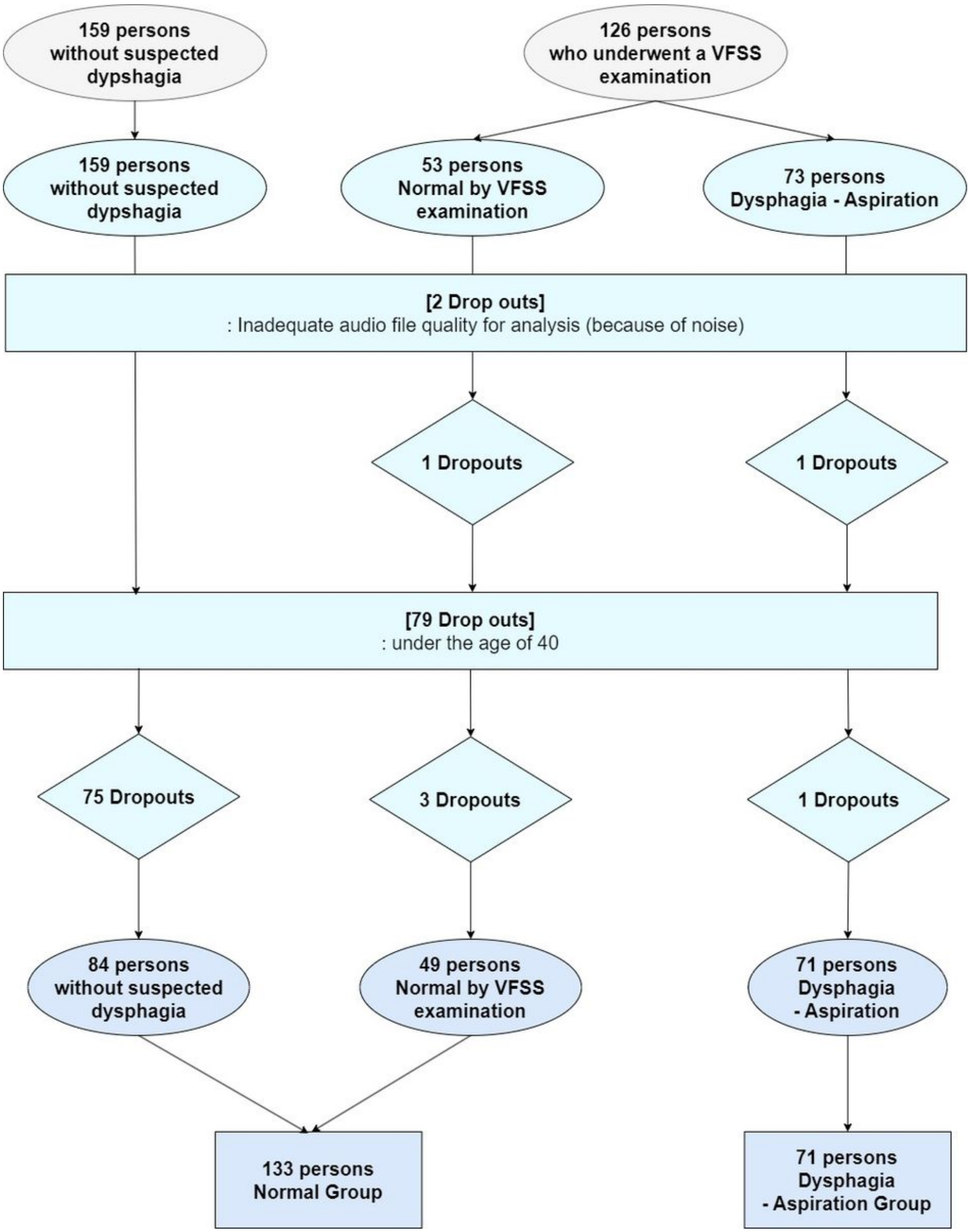
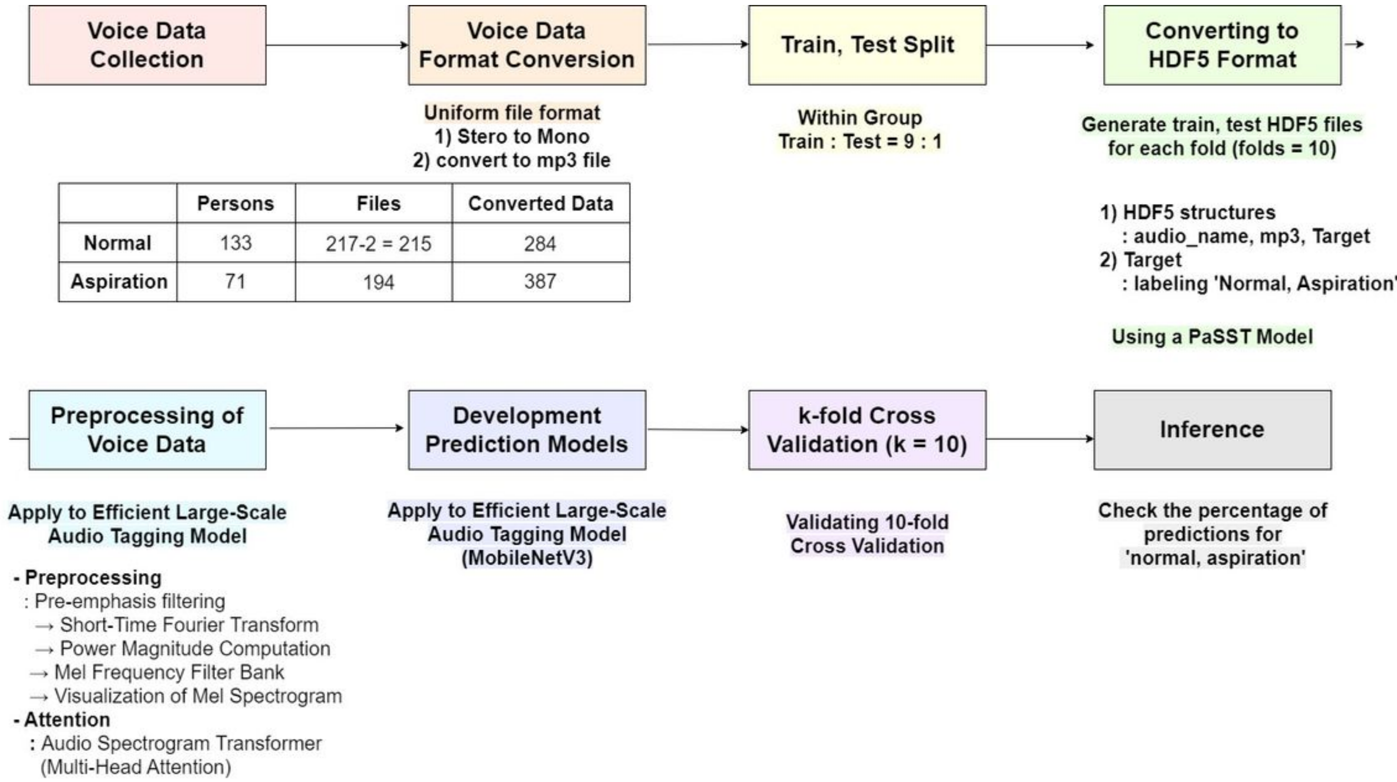


Figure 1

Flowchart of the Dysphagia Voice Cohort



	Persons	Files	Converted Data
Normal	133	217-2 = 215	284
Aspiration	71	194	387

Figure 2

Overview of Voice Data Preprocessing and Modeling

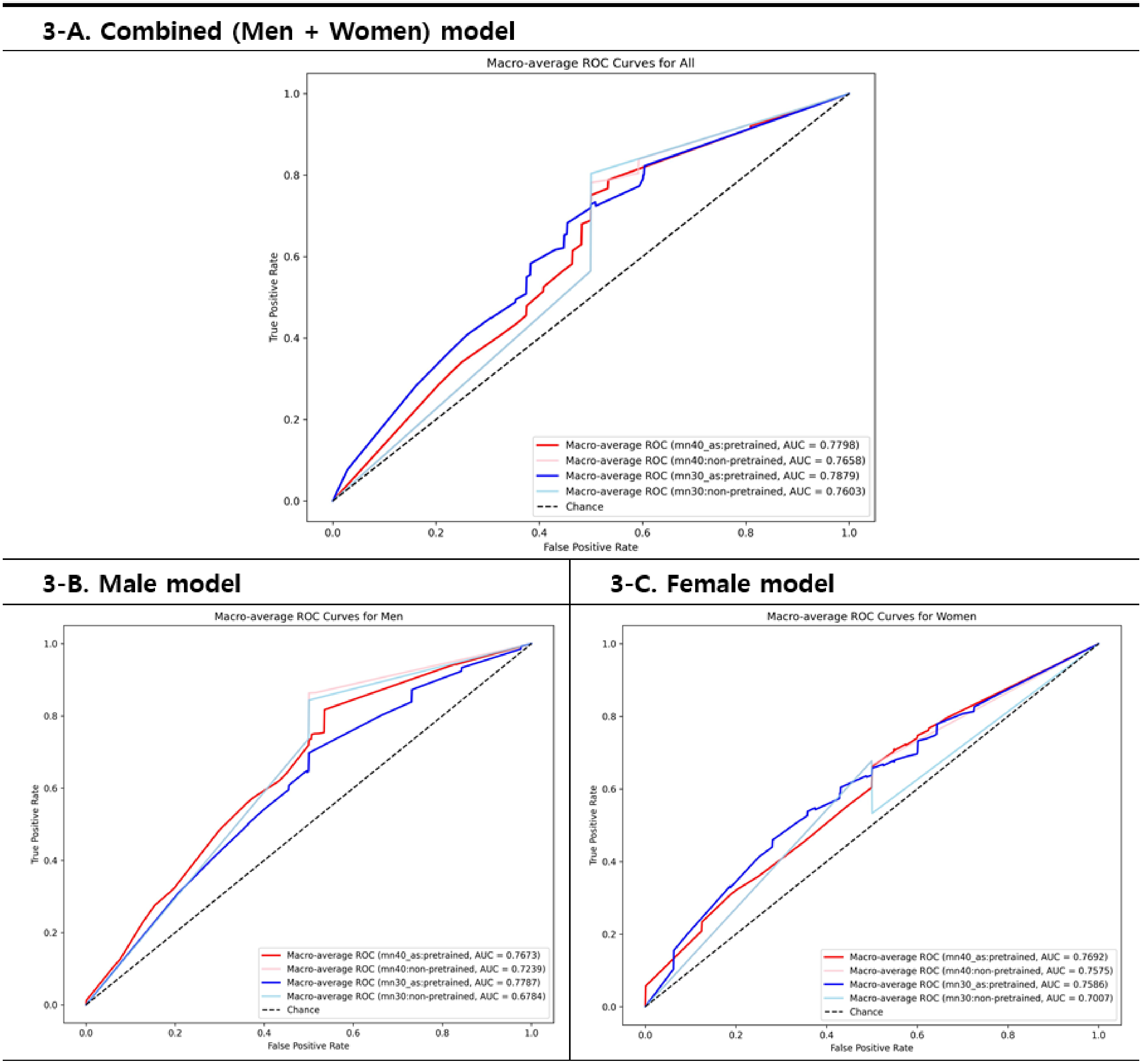


Figure 3

ROC Curve for each prediction model

The pre-trained models demonstrated higher performance compared to the non-pre-trained models. Among the four models, the mn30_as (pre-trained model) performed the best on average. The ROC curve was plotted, and the AUC (Area Under the Curve) was calculated.

```
************* Acoustic Event Detected: *****************
aspiration: 0.927
normal: 0.072
********************************************************
```

Figure 4

Inference

After evaluating one example of postprandial voice data that was not used during model training, it was observed that when classifying it as aspiration, the model assigned a probability of 92.7%. The output window displayed the results as mentioned earlier.