scientific reports

OPEN

Deep learning approach for dysphagia detection by syllable-based speech analysis with daily conversations

Seokhyeon Heo[1,4], Kyeong Eun Uhm[1,4], Doyoung Yuk[1], Bo Mi Kwon[1], Byounghyun Yoo[2], Jisoo Kim[3,5]✉ & Jongmin Lee[1,5]✉

Dysphagia, a disorder affecting the ability to swallow, has a high prevalence among the older adults and can lead to serious health complications. Therefore, early detection of dysphagia is important. This study evaluated the effectiveness of a newly developed deep learning model that analyzes syllable-segmented data for diagnosing dysphagia, an aspect not addressed in prior studies. The audio data of daily conversations were collected from 16 patients with dysphagia and 24 controls. The presence of dysphagia was determined by videofluoroscopic swallowing study. The data were segmented into syllables using a speech-to-text model and analyzed with a convolutional neural network to perform binary classification between the dysphagia patients and control group. The proposed model in this study was assessed in two different aspects. Firstly, with syllable-segmented analysis, it demonstrated a diagnostic accuracy of 0.794 for dysphagia, a sensitivity of 0.901, a specificity of 0.687, a positive predictive value of 0.742, and a negative predictive value of 0.874. Secondly, at the individual level, it achieved an overall accuracy of 0.900 and area under the curve of 0.953. This research highlights the potential of deep learning modal as an early, non-invasive, and simple method for detecting dysphagia in everyday environments.

Keywords Dysphagia, Deep learning, Conversations, Syllable-based speech analysis, Speech-to-text model, Artificial intelligence

The act of swallowing involves complex interactions of neurophysiological and anatomical processes. Dysphagia, which is defined as difficulty in swallowing, occurs when these functions are impaired. This condition is predominantly seen in older adults and can lead to serious health complications, including malnutrition, dehydration, choking, and aspiration pneumonia. Therefore, early detection of dysphagia is crucial for preventing health complications and maintaining quality of life in older individuals.

The current gold standard for diagnosing dysphagia is the videofluoroscopic swallowing study (VFSS), which involves radiologically recording the swallowing process using a mixture of substances of various viscosities mixed with a contrast agent. This procedure allows the visual assessment of the structures and movement of the oral cavity, pharynx, and esophagus. However, the procedure involves radiation exposure and carries risks of aspiration and choking. Additionally, the requirement for specialized equipment and trained personnel makes it challenging to detect dysphagia in everyday environments.

Considering these disadvantages of VFSS, researches have explored non-invasive methods for detecting dysphagia[1]. Among these, one of the most direct approaches is analyzing swallowing signals using sensors. This method involves cervical auscultation or employing devices such as accelerometers to collect and analyze swallowing signals to detect dysphagia[2–5]. However, this approach has limitations, including discomfort from attaching the device to the neck to collect swallowing signals.

[1]Department of Rehabilitation Medicine, Konkuk University Medical Center, 120-1 Neungdong-ro, Gwangjin-gu, Seoul 05030, Republic of Korea. [2]Center for Artificial Intelligence, Korea Institute of Science and Technology, 5 Hwarangro14-gil, Seongbuk-gu, Seoul 02792, Republic of Korea. [3]Faculty of Software Major in Artificial Intelligence, Jeju National University, 102 Jejudaehak-ro, Jeju-si 63243, Republic of Korea. [4]These authors contributed equally: Seokhyeon Heo and Kyeong Eun Uhm. [5]These authors jointly supervised this work: Jisoo Kim and Jongmin Lee. ✉email: clionelove@jejunu.ac.kr; leej@kuh.ac.kr

Accordingly, voice analysis of patients with dysphagia has been introduced. Individuals suffering from dysphagia experience alterations in their voice quality, such as a 'wet voice'[6], due to the nature of the anatomical structures responsible for vocalization and swallowing. The majority of previous studies on voice analysis for dysphagia have focused on phonetic analysis, examining differences between the voices of individuals with dysphagia and healthy controls. During this process, phonetic parameters such as fundamental frequency, jitter, and relative average perturbation (RAP) were used[7,8]. The unique characteristics of the voices of patients with dysphagia have been identified in previous studies, and these findings need to be considered when developing a new non-invasive diagnostic tool for dysphagia.

The development of artificial intelligence (AI) has achieved significant success in various areas and has expanded its application scope to healthcare and medical technology, proving its diverse possibilities[9–12]. Moreover, AI is being applied to classify various sounds, including vocal signals in the healthcare field[13,14,15]. In particular, machine learning has been utilized to analyze the vocalizations of patients with dysphagia[16,17]. These studies aimed to identify specific patterns in the voices of patients with dysphagia by analyzing various sound features, including vowel phonation sounds, conversational speech, and coughing sounds[17,18,19]. Furthermore, recent attempts have also been made to detect dysphagia using convolutional neural network (CNN) models based on voice signals, such as specific vowel sounds, without relying on complex audio features[18].

In this research, a novel deep learning model was developed to detect dysphagia by utilizing the characteristics of Korean language. Korean is a phonogrammatic language, where the letters correspond to sounds. This allows the segmentation and analysis of conversational data at the syllable-level. The approach for segmenting conversational voice data into syllables is simpler compared to obtaining voice quality data, such as phonetic parameters. In addition, syllable-based speech analysis has not been previously attempted in the detection of dysphagia in prior research. Based on this, the deep learning model was developed to effectively identify dysphagia through syllable-based speech analysis by segmenting daily conversations into syllables using a speech-to-text (STT) model. The efficacy of the proposed model was validated through two approaches: a syllable-segmented analysis of the entire data and an individual-level evaluation.

Method

This study aimed to develop a deep learning model for the early detection of dysphagia using daily conversational audio data. The conversational audio data were segmented into syllables using a STT model. These syllable-segmented audio data were then converted into a 2D form through the log-Mel spectrogram. The transformed data were used for training and testing by a CNN-based model. The overall process of this work is illustrated in Fig. 1.

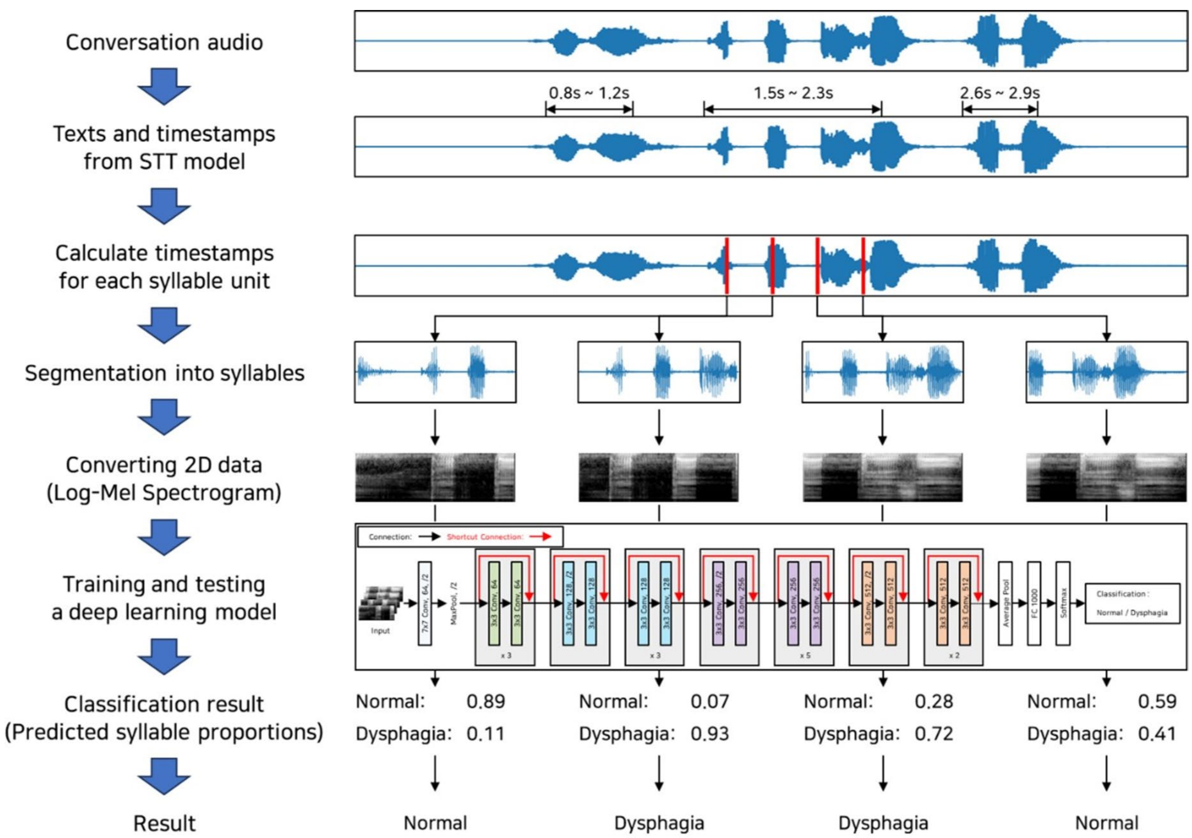

Fig. 1. The overview of the proposed classification process.

Data source and collection

This study was conducted with participants suspected of having dysphagia at a single tertiary university hospital from July to November 2023. The inclusion criteria were participants over the age of 19 who were suspected of having dysphagia, underwent VFSS, and provided informed consent. Patients unable to give voluntary informed consent due to cognitive impairment or unable to record their voice were excluded from the study. The institutional review board of Konkuk University Medical Center reviewed and approved the study protocol (approval number KUMC 2023-08-068). Written informed consent was obtained from all participants. All research procedures were performed in accordance with relevant guidelines and regulations.

All conversations consisted of interactions between the patient and a single researcher and were recorded in an isolated space to eliminate background noise. The embedded microphone of the Galaxy S22 smartphone (Samsung, Suwon, Republic of Korea) was used for recording, placed within the 30 cm of the participant's face. The recording of conversations, which had no fixed topic, lasted approximately 10 min for each participant. No additional equipment was used.

The recorded audio data were saved as WAV files with a quality of 44.1 kHz and 16 bits. All voices were separated to ensure that only the respective patient's voice was included, using Audacity software.

All data from the participants were classified into either the dysphagia group or the control group based on the results of their VFSS. The VFSS was performed according to the standard protocol of the institution by a physiatrist specializing in swallowing disorders. Four types of dysphagia diets with different consistencies, including liquids and semisolids, were tested by mixing them with contrast media. For each type of dysphagia diet, volumes of 3 cc and 7 cc were used as small and large amounts, respectively. The determination of dysphagia was based on the presence of any penetration or aspiration in the pharyngeal phase as identified through VFSS. Penetration was defined as when the test diet enters the airway but does not pass below the vocal folds, while aspiration is defined as the test diet passing through the vocal folds.

Additional conversational data were collected from YouTube to supplement the normal control group. This data consists of voices from age-matched individuals who have been confirmed to have no medical problems.

Data structure and generation

The audio data used in the study were divided into two groups: 24 control individuals and 16 dysphagia patients. These data were collected from 16 enrolled control participants, the YouTube data of 8 individuals for normal control, and 16 patients with dysphagia. The entire dataset comprised of a total duration of 5109 s for the normal group and 3461 s for the dysphagia group. The collected data were segmented into syllables using a STT model. In this segmentation process, the audio data for the normal group was divided into 15,980 syllables, while those for the dysphagia group were divided into 7830 syllables. The segmented audio data were then split into training and testing sets, with 18,951 syllables allocated for training and 4859 syllables for testing. Generally, datasets are divided into training, validation, and test sets. However, in this study, the validation set was excluded in order to maximize the amount of data available for training. The detailed structure of the data is shown in Fig. 2.

In this study, Clova Speech[20], a STT model, was used for syllable segmentation of conversational voice data. Clova Speech is specialized in Korean, demonstrating excellent voice recognition performance for the Korean language. This model converts audio data into text and provides start and end timestamps for each word. The text of the audio and the timestamp information for each word were used to segment the audio data into syllables. Midpoints for each syllable were calculated by dividing the duration of each word by the number of syllables comprising it, using the timestamps. The midpoint represents the central point of a given syllable, and this is

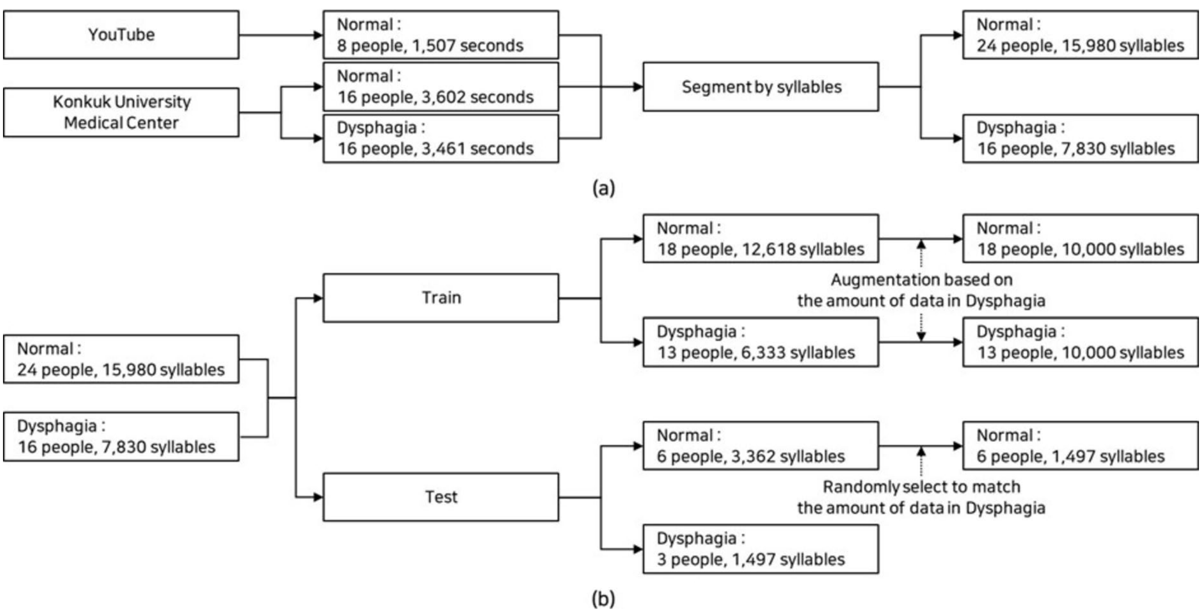

Fig. 2. Construction of dataset (**a**) dataset of syllable-based segmentation (**b**) dataset for train and test.

used as a basis for segmenting everyday conversational voice data into syllables. The segmentation time for each syllable was set to 0.6 s in order to fully include the sound of each syllable and ensure accurate recognition. The segmentation time was determined by listening to the directly segmented data to consider the time necessary to include the characteristics of the syllable. In the training set, the segmentation time was set to 0.8 s to enable data augmentation by subsequently selecting 0.6 s. The details of the data generation process are shown in Fig. 3.

To increase the number and diversity of the training dataset, a method was employed where a 0.6-s segment is selected based on a random starting point between 0 and 0.2 s, and this was combined with noise sounds. As a result, 10,000 audio files were augmented for each group, from the 12,618 syllables in the normal group and the 6333 syllables in the dysphagia group. The number of audio files for augmentation was determined based on the count of data from the dysphagia group.

Each syllable audio data was converted into a log-Mel spectrogram. The log-Mel spectrogram represents the frequency spectrum of a signal that changes over time, transforming a 1D audio signal into a 2D format. This method is achieved through the short-time Fourier transform (STFT) of the audio signal. The STFT divides the audio signal into short time segments and performs the fast Fourier transform (FFT) for each segment, extracting the frequency changes over time. After applying the Mel scale and logarithm, the frequency bands are adjusted to match human auditory characteristics, enabling the capture of important features in the voice signal.

The processing of syllable-level audio data used in this study was as follows. Firstly, the Librosa Python library was utilized to convert the syllable-level audio signals into log-Mel spectrograms. Librosa is a library that provides various functions for audio analysis, facilitating the implementation of complex signal processing. In the conversion process to log-Mel spectrogram, the length of the FFT was set to 512 samples, the Hann window length was 512 samples, and the hop length was 128 samples to define the number of samples between successive frames. Furthermore, the number of Mel bands was set to 64, dividing the frequency domain into Mel scale. Subsequently, the transformed Mel spectrogram was converted into a logarithmic scale. Afterward, the log-Mel spectrogram was normalized to transform it into 2D data in grayscale, with values ranging between 0 and 255. Finally, the generated data was resized to dimensions of (64,208) through zero-padding. This processed 2D data was used for neural network training and testing.

Develop a detection model for dysphagia

In this study, a CNN-based model was used for the training and testing of syllable-level audio data. CNN is a type of deep learning model widely used in the fields of computer vision, utilizing convolutional operations as its core function. A key feature of CNN is its ability to effectively recognize and extract local patterns, which is particularly useful in recognizing visual patterns in 2D data and has recently shown high performance in audio classification as well. The architecture of a CNN comprises multiple convolutional layers and pooling layers, which extract important features from the input data. Convolutional layers use filters to scan the input data, creating feature maps that emphasize important characteristics of the data. Pooling layers reduce the size of these feature maps, decrease computational complexity, and help prevent overfitting. The combination of these layers enables the model to effectively learn complex patterns and correlations.

ResNet is a CNN-based model that enhances neural network performance by applying the concept of residual learning[21]. The fundamental component of ResNet is the residual block, which includes a shortcut connection between the input and output of the network. This connection adds the input data directly to the block's output, introducing the concept that what the model needs to learn is the 'residual' between the input and output. This contrasts with traditional neural networks, which directly learn the output values themselves. Through this residual learning, issues such as gradient vanishing or exploding are mitigated. This enables effective learning in deep networks and has significantly improved performance in various tasks in the field of computer vision.

In this study, the ResNet-34 model was used. ResNet-34 has advantages such as being easy to train, delivering good performance resulting in improved accuracy, and having the capability to learn complex features.

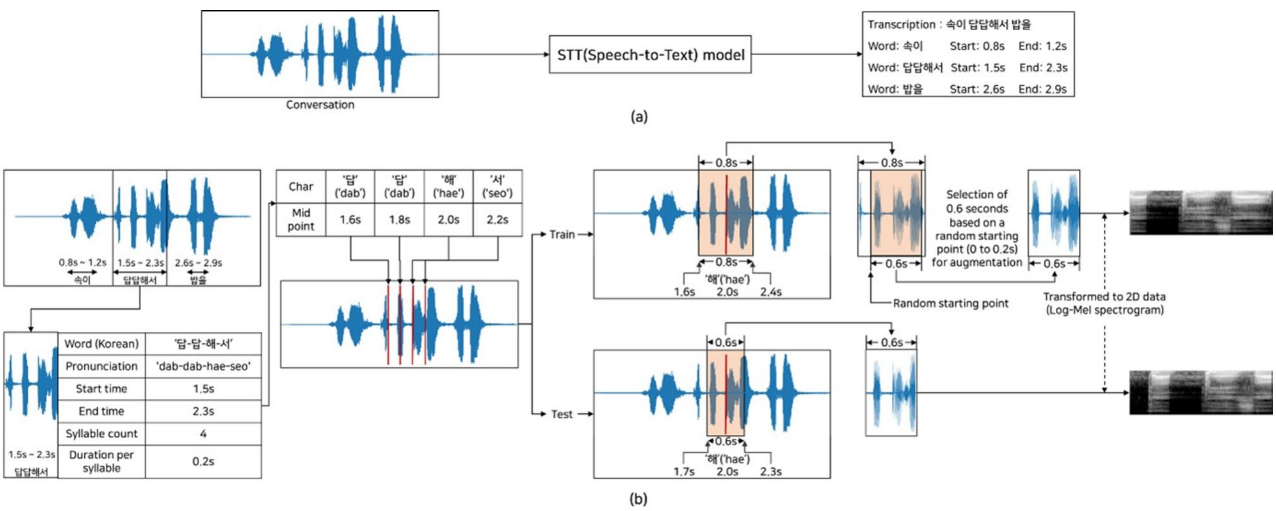

Fig. 3. Method of segmenting conversation into syllable units.

To compare performance improvements based on the complexity of neural network architectures, ResNet-18, ResNet-34, and ResNet-50 were evaluated. The ResNet-34 was selected, as no significant differences in performance were noted between the models. The training was conducted over a total of 5 epochs, and the size of each minibatch was set to 16. The cross-entropy loss function and Adam optimizer were applied. The learning rate was set to 0.001. The input data and the architecture of ResNet-34 are illustrated in Fig. 4.

Development environment

The classification system was developed using Python 3.11.4 and the PyTorch 2.0.1 deep learning framework. The computer used for training was equipped with 13th Gen Intel® Core™ i9-13900K processor, a GeForce RTX 4090 (24 GB) GPU and 64 GB RAM. The operating system was Windows 11 Pro. Clova Speech API version 1.9.0 was used as an STT model in this study.

Statistical analysis

The primary objective of this study was to evaluate the effectiveness of the proposed model in determining the presence of dysphagia using daily conversational audio data. To achieve this goal, two evaluation methods were applied. The first analysis involved binary classification, which aimed to classify each piece of syllable data as either 'dysphagia' or 'normal'. Based on this approach, data from the study were analyzed to estimate the overall sensitivity and specificity for diagnosing dysphagia according to the results of VFSS. Additionally, positive predictive value (PPV) and negative predictive value (NPV) for dysphagia were calculated.

Additionally, an individual-based evaluation was performed. This second approach assessed the presence of dysphagia for each individual by categorizing the test set into individual units for analysis. The determination of dysphagia in each individual's conversation was based on the proportion of syllable data classified as dysphagia within their dataset. Ultimately, if the proportion of syllables classified as dysphagia exceeded the threshold of 0.5, that individual was classified as having dysphagia; otherwise, they were classified as normal. These two evaluation methods are illustrated in Fig. 5.

Results

Demographic characteristics of the enrolled participants and information from the YouTube data are presented in Table 1. The majority of the participants were stroke patients.

The model proposed in this study demonstrated a total accuracy of 0.794 in diagnosing dysphagia. Specifically, it revealed a sensitivity of 0.901, a specificity of 0.687, a PPV of 0.742, and a NPV of 0.874 (Fig. 6).

Secondly, in the individual-level evaluation of all 40 subjects, the results showed an accuracy of 0.833 in the control group and 1.0000 in the dysphagia group, leading to an overall accuracy of 0.900. The area under the curve (AUC) was 0.953 (Fig. 7).

Discussion

Dysphagia has a high prevalence among older adults and can lead to serious complications, such as aspiration pneumonia. This significantly deteriorates patients' quality of life and has a substantial impact on their healthy living. Therefore, early detection of dysphagia is crucial for preventing such complications and improving the health status of older adults.

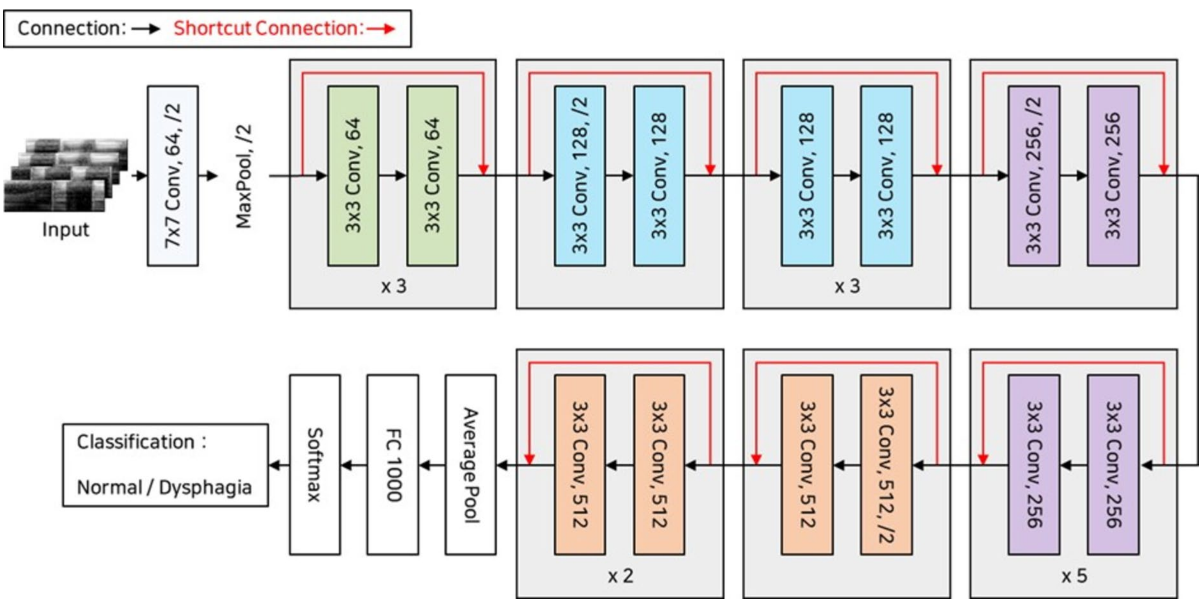

Fig. 4. Architecture of ResNet-34.

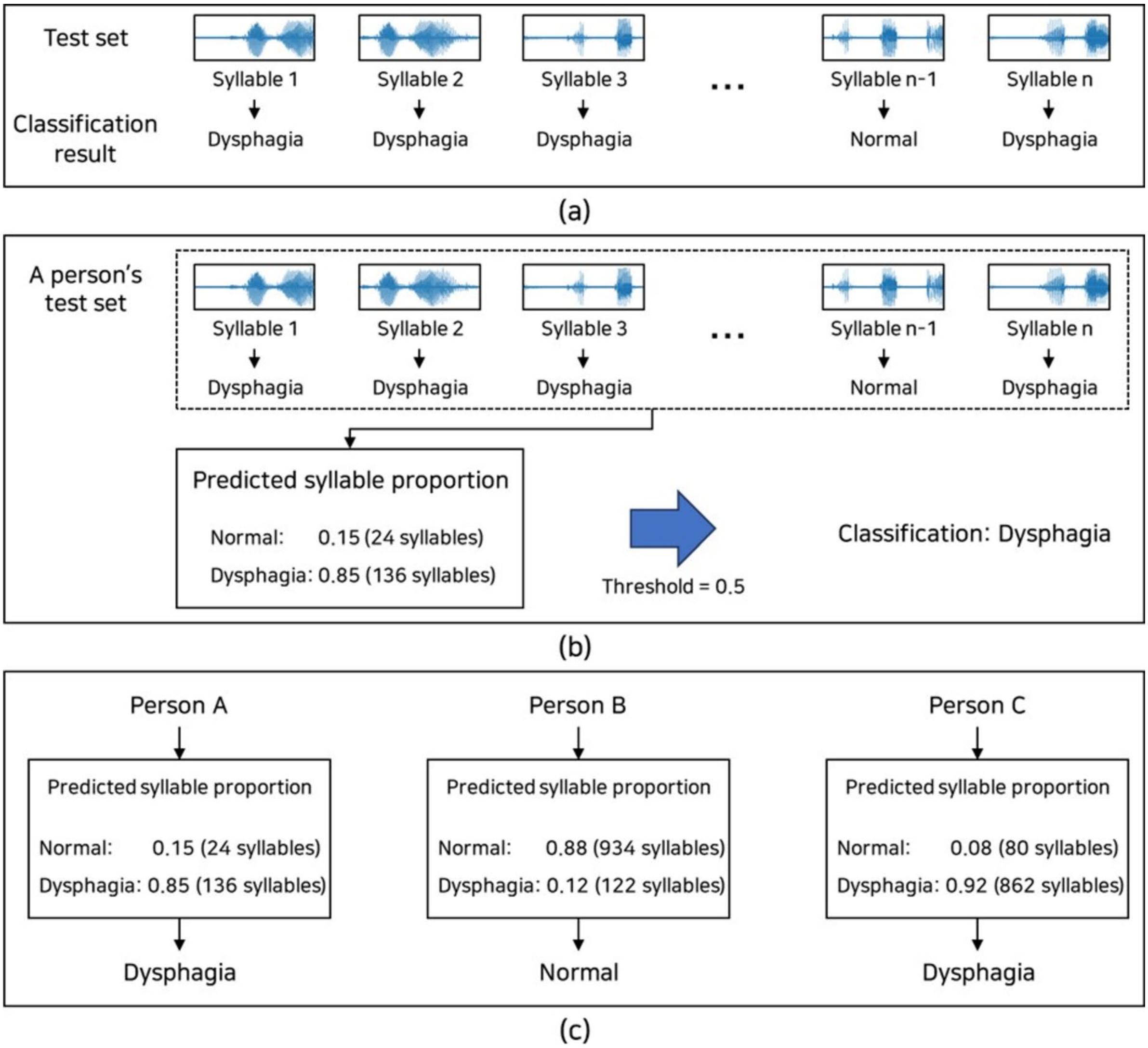

Fig. 5. Evaluation methods (**a**) syllable-segmented analysis (**b**) individual-level evaluation (**c**) individual-level classification process.

	Enrolled control	Collected control (YouTube)	Dysphagia
N	16	8	16
Age (years), mean ± SD	66.3 ± 12.6	61.4 ± 9.1	67.7 ± 16.1
Male: famale (N)	9:7	8:0	11:5
Height (cm), mean ± SD	161.0 ± 10.7		164.8 ± 10.6
Weight (kg), mean ± SD	61.9 ± 13.6		61.5 ± 16.0
BMI (kg/m^2), mean ± SD	23.9 ± 4.4		22.5 ± 4.7

Table 1. Demographic features of participants and collected data.

This study introduced a novel methodology for the diagnosis of dysphagia. The STT model segmented daily conversational audio data into syllables, and a newly developed deep learning model analyzed this data to distinguish between dysphagia and normal conditions. Using conversational audio data suggests that this method can closely approximate the real-life environment of patients.

Through the syllable-based speech analysis, the study achieved a total accuracy of 0.794, with a sensitivity of 0.901, specificity of 0.687, PPV of 0.742, and NPV of 0.874. Moreover, individual-level evaluation to detect the presence of dysphagia in specific individuals showed an overall accuracy of 0.900 with an AUC of 0.953. Previous studies that examined the efficacy of using a machine learning model with voice for dysphagia classification showed a sensitivity of 0.887–0.947 and a specificity of 0.779–0.845[16,17,18]. The current machine learning

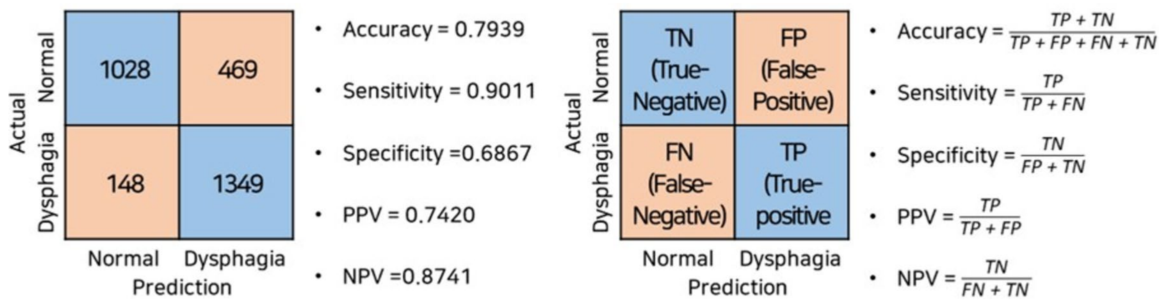

Fig. 6. Result of syllable-segmented classification for dysphagia.

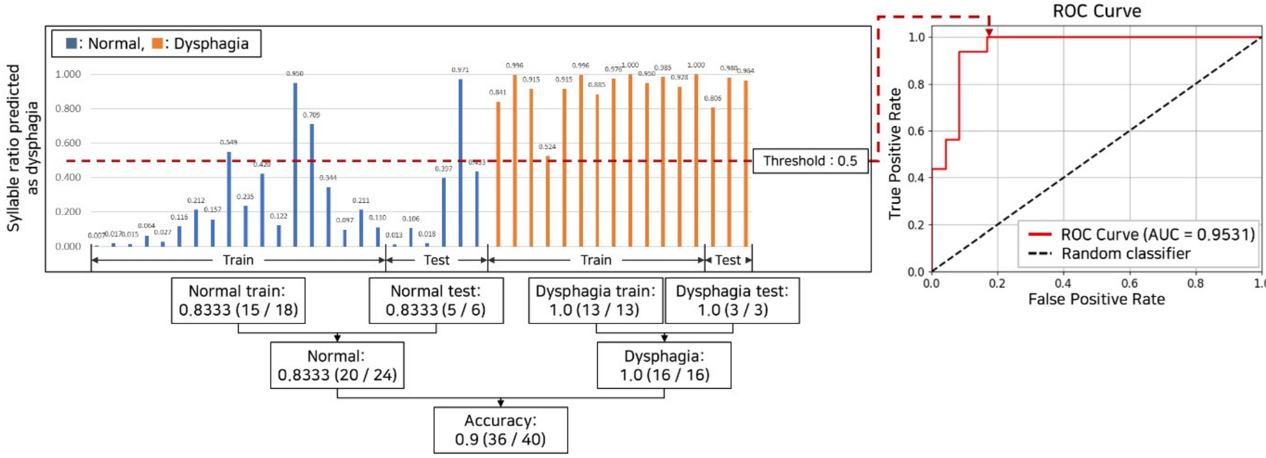

Fig. 7. Result of individual-level classification for dysphagia and ROC curve.

model could accurately diagnose dysphagia subjects, demonstrating comparable outcomes to previous research. However, four normal subjects without dysphagia were classified as dysphagia, and three of them had suffered from dysarthria. Future research might be needed to differentiate between dysarthria and dysphagia using conversational speech sounds.

In previous studies, specific sounds, such as the vocalization of specific vowels and sentences, and coughing sounds were used to detect dysphagia[16,17,18,19]. One study employing machine learning algorithm examined the risk of respiratory complications among stroke patients using a single vowel phonation of /e/[16]. Another study utilized machine learning-based speech analysis to discriminate between dysphagia patients and healthy controls[17]. That study recorded three speech tasks: (1) sustained vowels phonation including \ a \, \ e \, \ i \, \ o \ and \ u \; (2) rapid repetitions of the syllables pa-ta- ka; (3) a spontaneous monologue with approximately 90 s. Additionally, voluntary coughing, high-pitch gliding, and counting from 1 to 5 were used for dysphagia classification[18]. The unique characteristic of the current research is that the developed deep learning model was trained on daily conversational audio data, which has been segmented into syllables. This approach not only enhances the availability of data but also improves the accessibility of diagnosing dysphagia in everyday environments.

There are several limitations in this study. First, the study featured a relatively small sample size. Further study with larger sample size will improve the generalizability and robustness of the findings. Second, the severity of dysphagia was not accounted for in the analysis using this deep learning approach. Therefore, future studies are necessary to incorporate severity levels. Third, the data obtained from YouTube were not guaranteed to be of similar quality as the recordings from the enrolled participants. Lastly, the study did not include a validation data set to maximize the size of the training set. However, utilizing a separate validation set could have reduced risk of overfitting and provided more robust evaluation of the model,

The purpose of this study was to develop a deep learning model for the detection of dysphagia using everyday conversational voice data, analyzed through a CNN model. The results suggest that the proposed deep learning model demonstrates promising performance in dysphagia classification. This research highlights the potential for early, non-invasive, and simple detection of dysphagia in everyday environments.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to protection of personal information but are available from the corresponding author on reasonable request.

Received: 22 March 2024; Accepted: 21 August 2024
Published online: 31 August 2024

References

1. Lai, D. K. H. *et al.* Computer-aided Screening of aspiration risks in dysphagia with wearable technology: A systematic review and meta-analysis on test accuracy. *Front. Bioeng. Biotechnol.* **11**, 1205009 (2023).
2. Subramani, S., Rao, A., Roy, A., Hegde, P. S., & Ghosh, P. K. SegNet-based deep representation learning for dysphagia classification. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* https://sigport.org/documents/segnet-based-deep-representation-learning-dysphagia-classification (2022).
3. Roldan-Vasco, S., Restrepo-Uribe, J. P., Orozco-Duque, A., Suarez-Escudero, J. C. & Orozco-Arroyave, J. R. Analysis of electrophysiological and mechanical dimensions of swallowing by non-invasive biosignals. *Biomed. Signal Process. Control* **82**, 104533 (2023).
4. Dudik, J. M. *et al.* Deep learning for classification of normal swallows in adults. *Neurocomputing* **285**, 1–9 (2018).
5. Shu, K., Mao, S., Coyle, J. L. & Sejdić, E. Improving non-invasive aspiration detection with auxiliary classifier wasserstein generative adversarial networks. *IEEE J. Biomed. Health Inf.* **26**, 1263–1272 (2021).
6. Warms, T. & Richards, J. ``Wet voice" as a predictor of penetration and aspiration in oropharyngeal dysphagia. *Dysphagia* **15**, 84–88 (2000).
7. Kang, Y. A., Kim, J., Jee, S. J., Jo, C. W. & Koo, B. S. Detection of voice changes due to aspiration via acoustic voice analysis. *Auris Nasus Larynx* **45**, 801–806 (2018).
8. Ryu, J. S., Park, S. R. & Choi, K. H. Prediction of laryngeal aspiration using voice analysis. *Am. J. Phys. Med. Rehabil.* **83**, 753–757 (2004).
9. Kim, J., Park, C. M., Kim, S. Y. & Cho, A. Convolutional neural network-based classification of cervical intraepithelial neoplasias using colposcopic image segmentation for acetowhite epithelium. *Sci. Rep.* **12**, 17228 (2022).
10. Kim, H., Lee, H., Ahn, S., Jung, W. K. & Ahn, S. H. Broken stitch detection system for industrial sewing machines using HSV color space and image processing techniques. *J. Comput. Des. Eng.* **10**, 1602–1614 (2023).
11. Ko, J. U., Lee, J., Kim, T., Kim, Y. C. & Youn, B. D. Deep-learning-based fault detection and recipe optimization for a plastic injection molding process under the class-imbalance problem. *J. Comput. Des. Eng.* **10**, 694–710 (2023).
12. Lee, Y. R., Jung, S. H., Kang, K. S., Ryu, H. C. & Ryu, H. G. Deep learning-based framework for monitoring wearing personal protective equipment on construction sites. *J. Comput. Des. Eng.* **10**, 905–917 (2023).
13. Kim, J., Lee, H., Jeong, S. & Ahn, S. H. Sound-based remote real-time multi-device operational monitoring system using a convolutional neural network (CNN). *J. Manuf. Syst.* **58**, 431–441 (2021).
14. Atif, M., Muralidharan, S., Ko, H. & Yoo, B. COVID-beat: a low-cost breath monitoring approach for people in quarantine during the pandemic. *J. Comput. Des. Eng.* **9**, 992–1006 (2022).
15. Matsushita, Y., Tran, D. T., Yamazoe, H. & Lee, J. H. Recent use of deep learning techniques in clinical applications based on gait: A survey. *J. Comput. Des. Eng.* **8**, 1499–1532 (2021).
16. Park, H. Y. *et al.* Post-stroke respiratory complications using machine learning with voice features from mobile devices. *Sci. Rep.* **12**, 16682 (2022).
17. Roldan-Vasco, S., Orozco-Duque, A., Suarez-Escudero, J. C. & Orozco-Arroyave, J. R. Machine learning based analysis of speech dimensions in functional oropharyngeal dysphagia. *Comput. Methods Programs Biomed.* **208**, 106248 (2021).
18. Kim, H., Park, H. Y., Park, D., Im, S. & Lee, S. Non-invasive way to diagnose dysphagia by training deep learning model with voice spectrograms. *Biomed. Signal Process. Control* **86**, 105259 (2023).
19. Cesarini, V. *et al.* A machine learning-based voice analysis for the detection of dysphagia biomarkers. In *IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT)*, 407–411 (2021).
20. Naver. *CLOVA Speech*. https://clova.ai/speech (2024) **(in Korean)**.
21. He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 770–778 (2016).

Author contributions

Conceptualization: K.E.U., B.Y., J.K, and J.L.; Methodology: S.H., D.Y., B.M.K., K.E.U., and J.K.; Investigation: K.E.U., D.Y., B.M.K., and J.L.; Formal analysis: S.H. and J.K.; Writing—original draft: S.H. and K.E.U.; Writing—review and editing: J.K. and J.L. Funding acquisition: B.Y.; Supervision: J.K. and J.L. All authors reviewed the manuscript.

Funding

This work was supported by the Industrial Technology Innovation Program (20012462) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.K. or J.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher'S note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.