npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01417-w

Multimodal deep ensemble classification system with wearable vibration sensor for detecting throat-related events

Yonghun Song [1,6], Inyeol Yun [2,6], Sandra Giovanoli[3], Chris Awai Easthope[3] & Yoonyoung Chung[1,4,5]

Dysphagia, a swallowing disorder, requires continuous monitoring of throat-related events to obtain comprehensive insights into the patient's pharyngeal and laryngeal functions. However, conventional assessments were performed by medical professionals in clinical settings, limiting persistent monitoring. We demonstrate feasibility of a ubiquitous monitoring system for autonomously detecting throat-related events utilizing a soft skin-attachable throat vibration sensor (STVS). The STVS accurately records throat vibrations without interference from surrounding noise, enabling measurement of subtle sounds such as swallowing. Out of the continuous data stream, we automatically classify events of interest using an ensemble-based deep learning model. The proposed model integrates multiple deep neural networks based on multi-modal acoustic features of throat-related events to enhance robustness and accuracy of classification. The performance of our model outperforms previous studies with a classification accuracy of 95.96%. These results show the potential of wearable solutions for improving dysphagia management and patient outcomes outside of clinical environments.

Dysphagia, characterized by difficulty in swallowing, arises from various etiologies, including neurological disorders, surgical complications, and muscular impairments[1,2]. This disorder poses substantial health risks such as malnutrition, dehydration, and aspiration pneumonia[3]. Consequently, patients suffer from severe deterioration in quality of life, which leads to a notable increase in mortality[4]. The severity of dysphagia necessitates prompt and effective treatment to prevent life-threatening complications and improve patient outcomes. Effective management of dysphagia relies on the monitoring of multiple throat-related events[5]. For example, coughing and throat clearing during meals can indicate abnormalities in the neuromuscular functions of the pharynx and larynx[6]. This information enables healthcare professionals to devise appropriate therapeutic interventions tailored to each patient's condition. However, current treatment protocols for dysphagia are predominantly confined to clinical settings, which inherently limits the frequency and scope of clinical assessments[7]. Moreover, significant discrepancies exist between patient's capacity observed in the clinic and actual day-to-day performance[8]. These differences can be attributed to the dynamic nature of symptoms, which fluctuate in response to emotional states, levels of physical fatigue, and other daily variables[9]. Consequently, there is an urgent demand for the development and integration of ubiquitous monitoring devices for continuous and real-time tracking of throat-related events outside of the clinic[10].

Numerous studies have developed wearable sensor systems for continuously monitoring throat-related events[11–13]. For example, piezoresistive and surface electromyography (sEMG) sensors were utilized to detect swallowing events[14,15]. Combining these sensors with machine learning algorithms can evaluate various swallowing-related muscle activities, including frequency, duration, and force. However, to comprehensively diagnose the functions of the larynx and pharynx, it is necessary to observe a variety of throat-related events, not just swallowing[3]. Microphone-based wearable devices can acquire a wide range of sounds, such as coughing and speaking[16,17]. The measured sounds were classified into several throat-related events using artificial intelligence algorithms[18]. For instance, deep neural networks (DNNs) were applied to detect coughing sounds, but they often struggle to distinguish coughing from comparable events, such as throat clearing[19]. Some studies employed convolutional neural networks (CNNs) to classify sounds with similar acoustic characteristics by analyzing temporal-frequency features[20,21]. Additionally, other researchers have

[1]Department of Electrical Engineering, Pohang University of Science and Technology, Pohang, Korea. [2]Future IT Innovation Laboratory, Pohang University of Science and Technology, Pohang, Korea. [3]Data Analytics & Rehabilitation Technology (DART), Lake Lucerne Institute (LLUI) & cereneo Center for Interdisciplinary Research (CEFIR), Vitznau, Switzerland. [4]Department of Semiconductor Engineering, Pohang University of Science and Technology, Pohang, Korea. [5]Center for Semiconductor Technology Convergence, Pohang University of Science and Technology, Pohang, Korea. [6]These authors contributed equally: Yonghun Song, Inyeol Yun. e-mail: ychung@postech.ac.kr

combined microphone and sEMG sensors with CNN and long short-term memory algorithms to perform multi-event classification, enabling accurate recognition of swallowing, coughing, and chewing sounds[22]. Although various models were developed to classify sounds captured by microphones, these systems remain highly vulnerable to background noise, hindering the accurate detection of throat-related events in daily life[16–22].

Recent studies have utilized skin-attachable sensors to acquire acoustic medical data without ambient interference[13,23]. These accelerometer-based sensors were attached at the suprasternal notch to detect vibration signals such as coughing, speaking, and throat clearing[23–26]. However, the form factor restricts mounting the sensing part onto the curved and narrow neck, an optimal location for capturing subtle acoustic features with high quality[27]. In processing the measured data, DNNs were applied to analyze and classify diverse vibrations, facilitating multi-event monitoring[28]. For example, a random forest classifier assessed the severity of dysphagia, and a support vector machine quantified the swallowing variability[24]. However, these models were trained on small datasets, resulting in overfitting that degrades their reliability and generalizability[29]. In comparative studies on limited datasets, CNN models outperformed recurrent neural networks in classifying throat-related events[28]. Based on these results, several studies have employed scalograms and mel spectrograms as inputs in residual neural networks (ResNets) for image-based event classification[23,28]. However, the single-network models rely on limited features from acoustic data, leading to classification accuracies below 89%[23,25,26,28]. Therefore, further advancements in both sensor hardware and classification models are necessary to enable accurate and continuous monitoring of throat-related events.

In this study, we propose an advanced healthcare system devised for real-time monitoring of throat-related events in daily life. A soft skin-attachable throat vibration sensor (STVS) was designed by isolating a small and flexible sensing part, ensuring precise adherence to the curved neck skin. This facilitates uninterrupted measurements of vibrational signals emanating from the larynx and pharynx, capturing various throat activities, including coughing, speaking, swallowing, and throat clearing. Additionally, we developed a deep ensemble model to classify these activities by integrating diverse neural networks trained on multi-modal data, encompassing both time-series features and acoustic spectral image patterns. The model architecture was optimized through performance evaluations based on various combinations of neural networks and ensemble algorithms. As a result, we achieved an outstanding performance with a classification accuracy of 95.96% and an area under the receiver operating characteristic curve of 0.99 on a test dataset, which comprised a range of throat-related events in multiple languages. This innovative system holds significant potential for substantially enhancing the management of dysphagia by providing continuous and accurate monitoring of throat-related events.

Results

Overall process

In this study, we developed a highly accurate classification system for detecting major throat-related events, such as coughing, speaking, swallowing, and throat clearing. We engineered a soft skin-attachable throat vibration sensor (STVS) that conforms to the neck skin contours to acquire high-quality acoustic medical data. Conventional acoustic microphones can measure a broad spectrum of high-quality sounds; however, they struggle to detect subtle signals, such as swallowing, and are susceptible to interference from ambient noise[16,30]. On the other hand, our STVS can precisely measure the signals originating from the user's throat vibration, completely unaffected by external noise. As shown in Fig. 1a, the STVS was positioned on the neck skin to measure throat vibrations. The STVS hardware comprises sensing and controller parts that are interfaced by serpentine interconnects, as depicted in Fig. 1b. The sensing part was placed at the optimal location above the laryngeal prominence to capture the highest quality of vocal signal[27]. The controller part was responsible for digital filtering, data storage and transmission, and battery management. Since the controller part exhibited less flexibility due to its multi-chip composition, it was positioned on the relatively flat side of the neck. The connection between the sensing and controller parts was facilitated through a serpentine structure to ensure stable signal transmission during neck movements. The controller part was folded in half (see Supplementary Fig. 1) and connected to a battery. Then, the entire STVS was encapsulated with a soft polymer for enhanced wearability, as shown in Fig. 1c. The dimension of the encapsulated sensing part was 0.7 cm in length, 0.9 cm in width, and approximately 0.3 cm in thickness.

An overall process for developing and applying the classification model to detect throat-related events pertaining to dysphagia is illustrated in Fig. 1d. We gathered extensive acoustic medical data through the STVS, encompassing four key events: coughing, speaking, swallowing, and throat clearing. This dataset comprises 9000 data segments, each with a duration of 625 ms, obtained from 32 subjects at a sampling rate of 6400 Hz. The subjects, representing a diverse range of ages and genders, were distinctly categorized for the training and testing phases (see Supplementary Table 1). During the training, we utilized an acoustic dataset consisting of utterances in the English language. The acoustic data was transformed into the frequency domain, facilitating the training of various acoustic features in both time and frequency domains. This extensive feature extraction strategy mitigated the risk of overfitting and enhanced generalization performance for unseen data[31]. Each network was integrated through an ensemble algorithm to increase the classification accuracy and robustness against outlier data. The developed classification model was evaluated using a separate test dataset composed of English and other diverse languages, including French, German, Spanish, and Korean. Training in English and testing in multiple languages validate the model's generalization ability across various linguistic boundaries, demonstrating its applicability in global contexts[32,33]. The proposed model exhibited remarkable performance compared to previous studies that classified acoustic medical data using microphones, as shown in Table 1 and Fig. 1e[19–21,23,25,26,28]. Using a throat microphone enabled the detection of subtle sounds that conventional acoustic microphones typically overlook, thereby expanding the range of classifiable events. Furthermore, implementing the multi-domain ensemble method facilitated an exceptionally high classification accuracy of 95.96%, significantly exceeding the benchmarks accomplished by previous studies[23,25,26,28].

Stretchability of the soft skin-attachable throat vibration sensor

The sensing part of the STVS is strategically placed above the center of the laryngeal prominence for optimal detection of vocal cord vibrations, thereby achieving a high signal-to-noise ratio (SNR)[27]. Due to the spatial constraints of this region, we segregated the wearable hardware into two distinct parts: the sensing part above the laryngeal prominence and the controller part on the flatter side of the neck. These two parts were connected by a serpentine interconnect, designed to be highly stretchable and robust against external strains (see Supplementary Fig. 2)[34,35]. As shown in Fig. 2a, we measured the variation in resistance when the serpentine interconnect was stretched to 100%. It is generally known that the serpentine structure causes minimal resistance change in the metal film even at this maximum stretching level[36,37]. We demonstrated the physical durability of this structure by measuring the resistance between two points across the serpentine interconnect. Figure 2b shows the minimal change in relative resistance due to the strain, revealing a mere 0.19% increase at 100% strain. These results underscore that the serpentine interconnect stably transmits signals even when the wearer's neck is moving. Although a marginal rise in resistance was noted with the strain, there was negligible resistance change after 5000 cyclic loadings, as shown in Fig. 2c.

Measurement protocol and throat-related events

Subjects wearing the STVS participated in a structured experimental protocol, as shown in Fig. 3a. All subjects conducted five repetitions of coughing, speaking, swallowing, and throat clearing in a controlled environment, maintaining a rest interval of 2 s between each event (see Supplementary Fig. 3). The data was stored in the sensor memory and instantaneously transmitted to a database via Bluetooth low-energy (BLE).

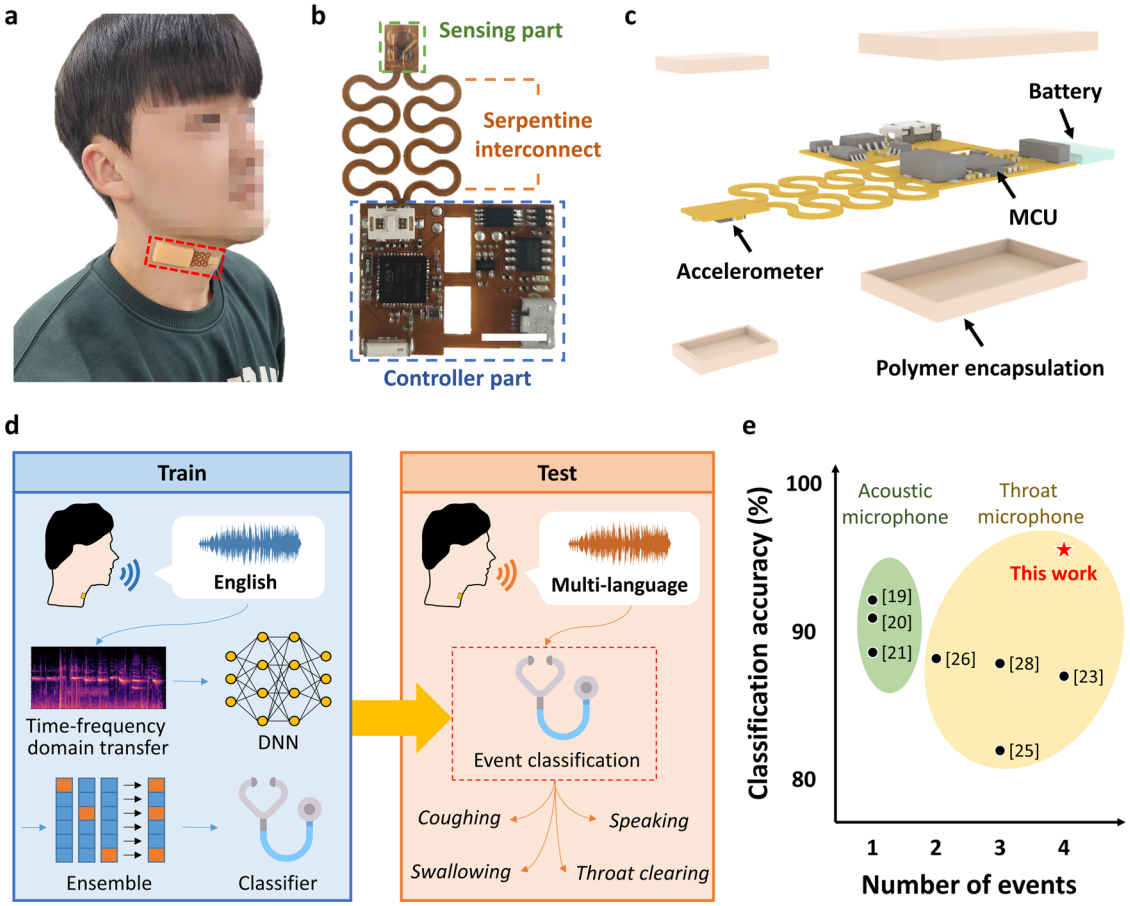

Fig. 1 | Overview of soft skin-attachable throat vibration sensor (STVS) system for classifying throat-related events: coughing, speaking, swallowing, and throat clearing. a Photo of a subject wearing the STVS above the laryngeal prominence position. **b** Magnified image of the STVS without polymer encapsulation. A sensing part of the STVS forms a conformal contact with the neck skin. Scale bar: 1 cm. **c** Image of the STVS with integrated components. **d** Experimental process for designing and applying an ensemble-based deep learning model to classify throat-related events. The training dataset consists of events obtained from subjects using English. In contrast, the test dataset comprises events from subjects with diverse linguistic backgrounds, such as English, French, German, Spanish, and Korean. **e** Comparison of the classification accuracy and the number of classifiable events with previous studies that detect major throat-related events using microphone devices.

Table 1 | Comparison of the proposed wearable sensor system with previous studies for detecting throat-related events

Reference	Hardware	Dataset	Model	Classification events	Accuracy
19	AM[a]	5640 segments, 9 subjects	DNN trained with SGD and momentum	Coughing	92.3%
20	AM	15,591 segments, 43 subjects	CNN with bagging	Coughing	90.9%
21	AM	25,537 segments 43 subjects	Scalable compact CNN	Coughing	88.8%
23	TM[b] (on the suprasternal notch)	8446 segments, 10 subjects	ResNet50	Coughing, Laughing, Speaking, Throat clearing	87.2%
25	TM (on the suprasternal notch)	120 segments, 4 subjects	Support Vector Machine	Coughing, Laughing, Throat clearing	81.9%
26	TM (on the suprasternal notch)	36 subjects	CNN followed by ResNet50	Coughing, Talking	88.3%
28	TM (side of the neck)	4639 segments, 15 subjects	ResNet18	Coughing, Swallowing, Throat clearing	87.9%
This work	TM (above the laryngeal prominence)	9000 segments, 32 subjects	LightGBM ensemble, WaveNet, EfficientNet	Coughing, Speaking, Swallowing, Throat clearing	96.0%

[a]AM: acoustic microphone.
[b]TM: throat microphone.

The acquired data was refined in a series of signal-processing steps to build a training dataset. First, max absolute normalization was applied to reduce variability in signal amplitude depending on individual characteristics. This method preserves the original distribution of positive and negative values, retaining the integrity of the data. Second, an adaptive filter was utilized to improve the SNR of the data[38]. The filter dynamically adjusts its coefficients in response to the input signal's magnitude, allowing it to accurately track and filter out unwanted noise. This real-time adjustment helps to effectively

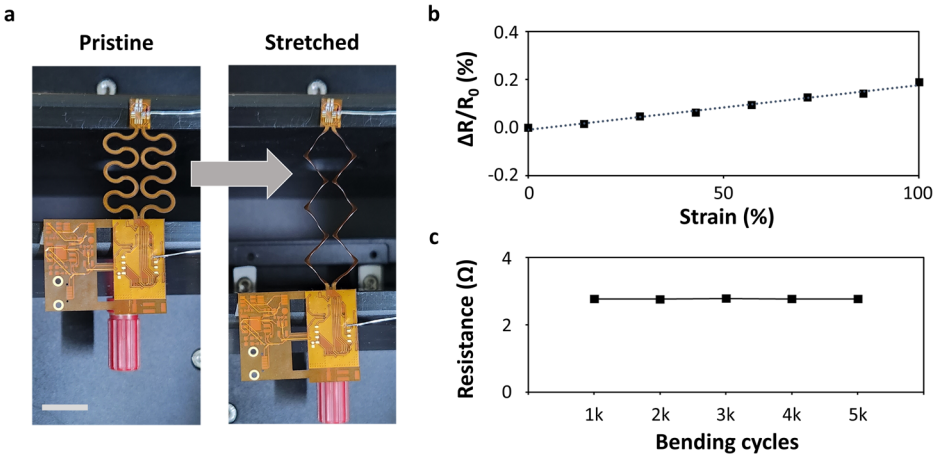

Fig. 2 | Characteristics of the serpentine interconnect in soft skin-attachable throat vibration sensor. a Image of experimental setup for the stretchability test. The serpentine interconnect was in a pristine state (left) and stretched to 100% (right). Scale bar: 1 cm. **b** Resistance variations in response to the applied strain. The relative resistance was increased by 0.19% after a 100% stretch. **c** Resistance across two further points connected by the serpentine interconnect. The resistance value did not change even after 5000 cyclic loadings.

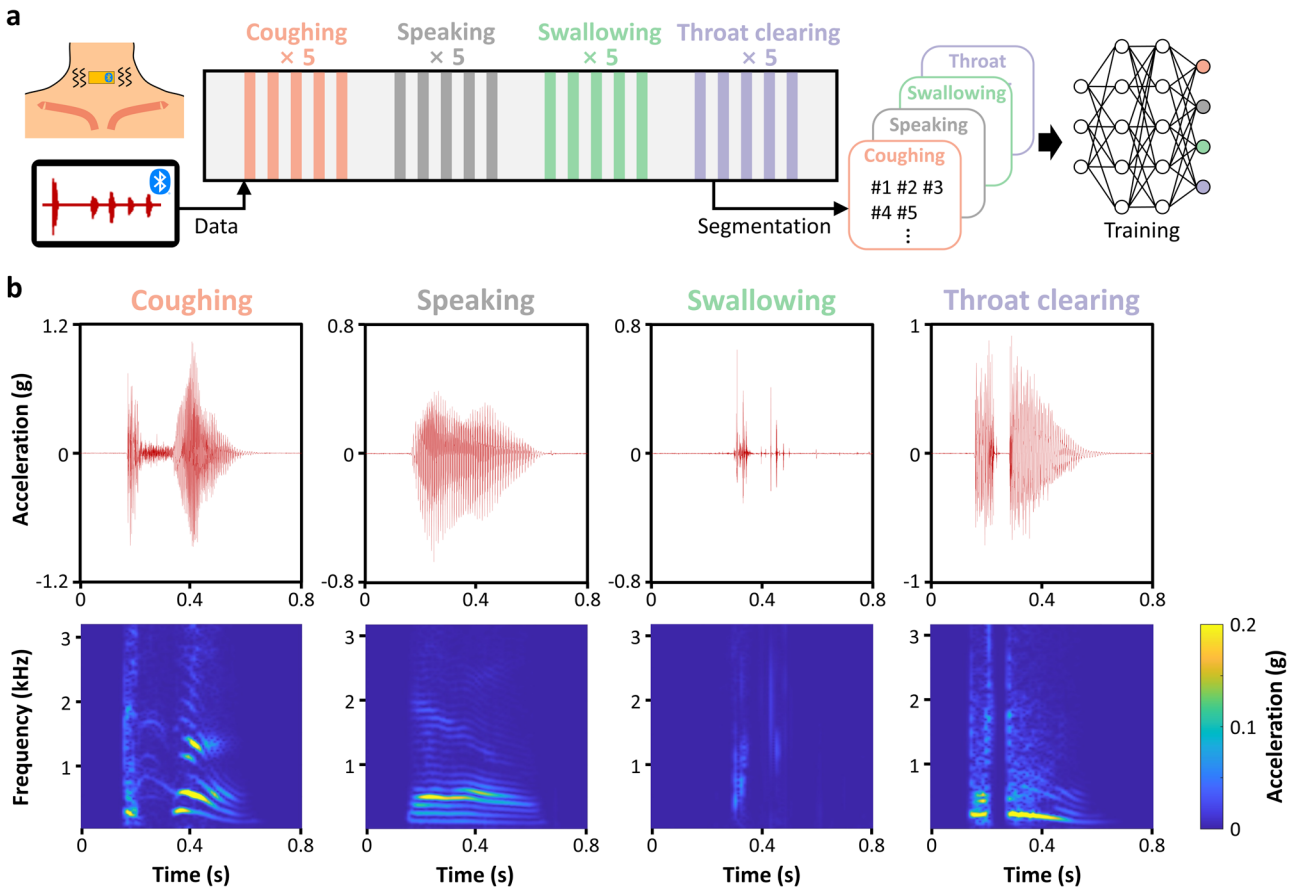

Fig. 3 | Experimental protocol and acquired signals from the soft skin-attachable throat vibration sensor (STVS). a Schematic of measurement protocol and data preprocessing steps for acquiring throat-related events using the STVS. Subjects repeated four distinct events—coughing, speaking, swallowing, and throat clearing—five times each. All subjects followed the same controlled protocol. The data was segmented based on the peak amplitude and utilized as inputs for network training. **b** Example waveforms and spectrograms for each event. Signals related to the vibration of the pharynx and larynx were captured around the neck. The STVS was configured with a sampling rate of 6400 Hz and a dynamic range of ±4 g. The spectrogram data was obtained using a short-time Fourier transform with a Hanning window frame width of 40 ms and an overlap of 75%.

smooth the signal and remove thermal noise, which is inherent in the hardware, resulting in a high-quality signal. Third, the processed data was segmented into 625 ms windows centered on peaks surpassing a predefined threshold. This threshold was set at six times the mean absolute value of signals measured in the absence of throat-related events. The 625 ms segment duration was optimized by systematically evaluating captured events across various segment lengths. A length of 4000 samples (equivalent to 625 ms at our sampling rate) was consistently effective in detecting individual throat-related events without overlapping multiple occurrences. This process yielded labeled data consisting of 2134 coughing, 2440 speaking, 2189 swallowing, and 2237 throat-clearing events (see Supplementary Table 2).

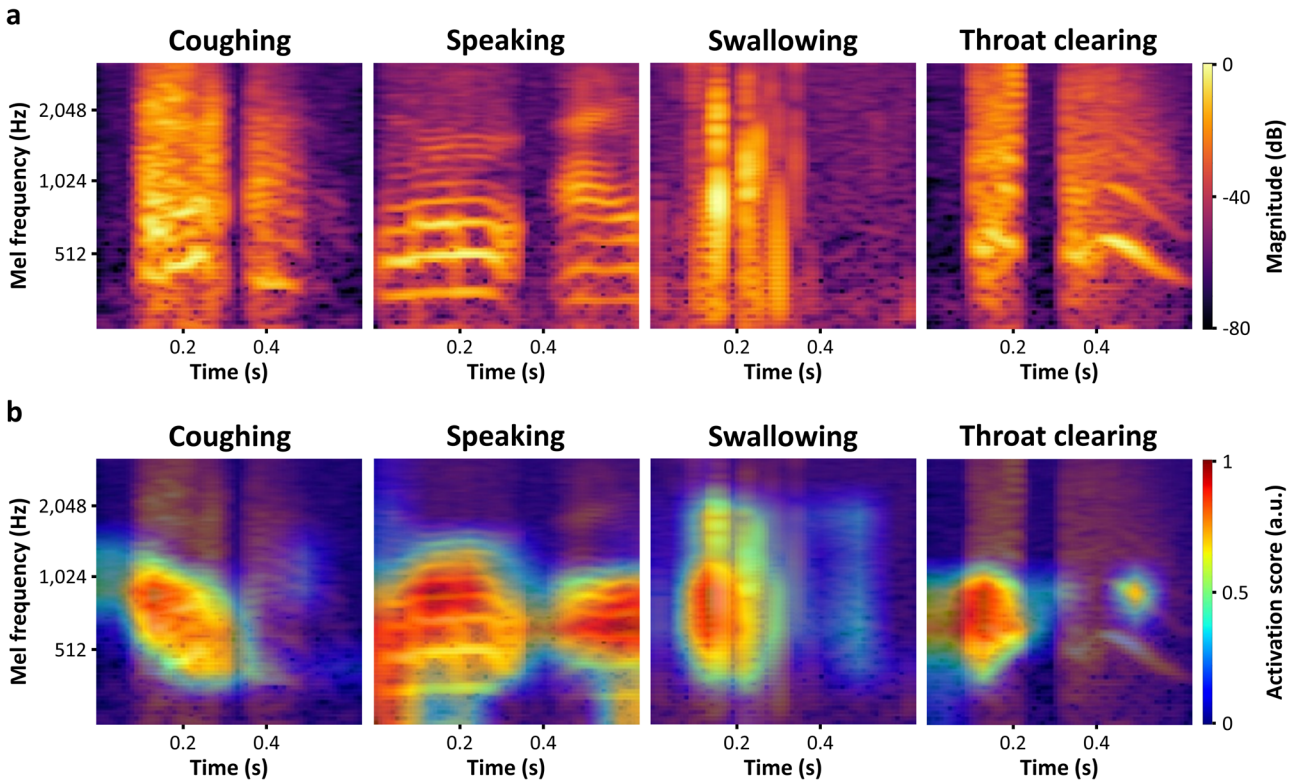

Fig. 4 | Visualization of activation regions in the image-based classification network. a Mel spectrograms and **b** its gradient-weighted class activation mapping (Grad-CAM) results from the EfficientNet for various events: coughing, speaking, swallowing, and throat clearing. The Grad-CAM provided a visual representation of crucial pixels in the input image as a heatmap. The heatmap of the speaking highlights the harmonic components, while the swallowing event generates the prominent heatmap over several spike-shaped signals. The coughing and throat-clearing events show significant heatmaps around the 0.1 s.

Examples of the waveform and spectrogram of the four events are shown in Fig. 3b. The spectrograms are generated through a short-time Fourier transform (STFT) and a Hanning window with a frame width of 40 ms and an overlap of 75%. The main features of coughing are the broad-bandwidth response, followed by the rapid decrease in signal intensity. On the other hand, speaking exhibits a harmonic structure in the spectrogram. This distinct harmonic pattern consists of integer multiples of the fundamental frequency (f_0). Swallowing is characterized by multiple spike-shaped peaks and completed within a relatively short duration of 0.2 s, compared to other events. Throat clearing exhibits the spectrogram similar to coughing, but a sustained low-frequency component is distinctly observed. The distinct spectral characteristics among these event signals, such as various frequency components and amplitude patterns, provide foundational insights for developing the classification model.

Grad-CAM analysis of conventional image-based events classification methods

We evaluated representative state-of-the-art neural networks, ResNet50 and EfficientNet, for classifying coughing, speaking, swallowing, and throat-clearing events based on image patterns extracted from the time-series acoustic data[39,40]. Most previous studies on throat-related event classification have predominantly utilized these image-based networks instead of one-dimensional input networks to extract complex features from signals[23,28]. We first transformed throat vibration signals into spectrogram and mel spectrogram images to train each network. Augmentation methods were then applied to increase the data variability[41]. The amount of training data was increased thirteenfold, thereby reducing the overfitting and enhancing the robustness of the trained networks. The classification accuracy after augmentation exhibited an average improvement of 5.03% for ResNet50 and 1.28% for EfficientNet (see Supplementary Fig. 4). However, all image-based neural networks exhibited relatively low classification accuracy in distinguishing coughing and throat-clearing events in the confusion matrices (see Supplementary Fig. 5).

We analyzed the reasons for the low accuracy of certain events using the gradient-weighted class activation mapping (Grad-CAM) method. The Grad-CAM represents the activated regions in the image as a heatmap when a network performs classification tasks[42]. We extracted heatmaps from the EfficientNet, which performed the highest accuracy among individual classification networks. The mel spectrograms and heatmaps corresponding to throat-related events are shown in Fig. 4a and b, respectively. The activation regions for speaking are primarily concentrated in harmonic components below 1 kHz. On the other hand, the heatmap of the swallowing event appears in multiple spike-shaped signals across a broadband frequency range. These distinctive Grad-CAM results demonstrate that each event can be accurately distinguished in the current image domain. However, the mel spectrogram of coughing exhibits similar patterns to throat clearing, with both events showing identical heatmaps in the frequency range of 500 to 1000 Hz around 0.1 s. This result emphasizes the necessity of classification models that utilize multi-modal data and ensemble techniques to enhance accuracy by recognizing diverse acoustic features of throat-related events rather than solely relying on image-based single networks[43,44].

Ensemble-based deep learning model development for event classification

We addressed the limitations of image-based classification networks by utilizing various types of input data and multiple networks. Figure 5 illustrates the architecture of the multi-modal ensemble model, which classifies throat-related events with high accuracy. We augmented the training data thirteenfold and transformed it into two domains: (1) time domain as waveform and f_0 and (2) time-frequency domain as spectrogram and mel spectrogram (see Supplementary Fig. 6). Time domain signals were used to train the WaveNet, optimized for audio signal processing[45]. We also utilized

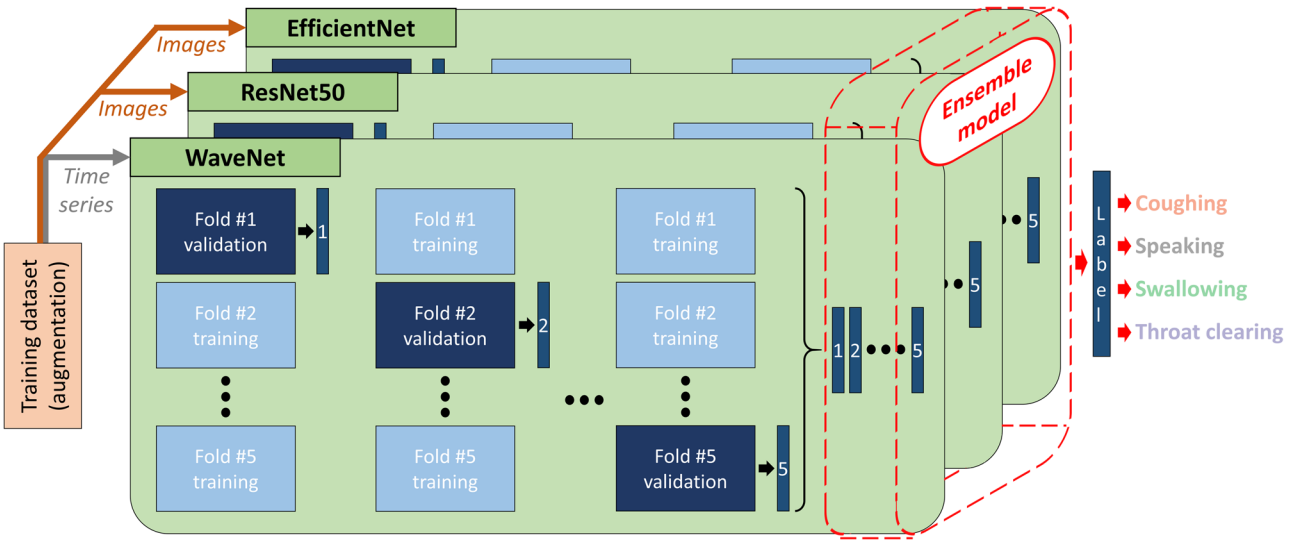

Fig. 5 | Ensemble-based deep learning model for event classification. Our deep learning architecture combines various deep neural networks into an ensemble method. After augmentation, a training dataset was pre-processed into time series (waveform and fundamental frequency) and images (spectrogram and mel spectrogram). WaveNet was trained on time-series data, capturing sequential patterns, while ResNet50 and EfficientNet were trained on image data, focusing on spatial features. The networks were trained using a fivefold cross-validation method, and each prediction from the validation dataset served as the input data for the ensemble model. The ensemble-based deep learning model, trained with various features, accurately classifies throat-related events.

the time-frequency domain signals, which resemble images, to train ResNet50 and EfficientNet, both renowned for their outstanding performance in image classification tasks[39,40]. This approach enabled each network to learn various features and patterns from the acoustic medical data. Detailed information regarding data preprocessing and the structure of each network is provided in the "Methods" section.

The individual networks were trained using a fivefold cross-validation approach. The validation results, consisting of class probabilities generated by each neural network, were concatenated to train ensemble algorithms using a stacking method[46]. We utilized several ensemble algorithms, including gradient boosting machine (GBM), random forest (RF), extreme gradient boosting (XGBoost), LightGBM, extra trees, adaptive boosting (AdaBoost), and support vector machine (SVM)[47–53]. Each of these ensemble models has unique strengths: GBM and XGBoost excel in handling structured data and complex interactions, RF and extra trees offer robustness through random feature selection, AdaBoost corrects misclassifications iteratively, SVM performs well in high-dimensional spaces, and LightGBM effectively handles large datasets with high speed and performance. Therefore, we compared the classification accuracy of each ensemble algorithm combining different neural networks, and heuristically optimized the model architecture (see Supplementary Table 3). The highest performance was achieved by stacking WaveNet, trained on waveforms, with EfficientNet, trained on mel spectrograms, and further integrating them using the LightGBM ensemble algorithm. This outcome highlights the synergy achieved by leveraging the complementary strengths of each network: WaveNet for capturing intricate acoustic features from waveforms and EfficientNet for discerning patterns within mel spectrograms[40,45]. The stacking ensemble technique, such as LightGBM, integrates these networks to comprehensively analyze diverse aspects of acoustic data, leading to a substantial performance improvement.

Performance evaluation of the proposed model

The proposed model was evaluated on the test dataset, which includes coughing, speaking, swallowing, and throat-clearing events. Figure 6a presents a normalized confusion matrix of our deep ensemble model. The ensemble model exhibited outstanding classification accuracies for each event: coughing (93%), speaking (98%), swallowing (100%), and throat clearing (90%). A comparison between the proposed model and previous studies regarding the classification of throat-related events is summarized in Table 1[19–21,23,25,26,28]. Multi-label classification models reported in previous studies show low accuracy due to their inability to reflect diverse characteristics to discern each class[23,25,26,28]. However, due to hardware configurations and measurement data differences across studies, directly comparing prior classification models using our dataset is challenging. To address this, we employed well-known models to quantitatively demonstrate that the lower accuracy, typically observed with single networks in previous studies, can be significantly improved through ensemble methods[39,40,45]. As shown in Fig. 6b, the proposed model outperformed the accuracy of individual networks. Our model showed higher accuracy across all events than the mel spectrogram-based EfficientNet, which exhibited the second-highest performance, notably achieving a 5% increase in coughing classification (see Supplementary Fig. 5). Our approach using multi-modal data and the ensemble method enabled performance enhancement by facilitating the identification of distinguishing features for throat-related events. Additionally, our model exhibited consistently high accuracy for various language groups (see Supplementary Fig. 7). The speaking events maintained high classification accuracy across all groups, indicating the minimal impact of linguistic differences on the model's performance. This result suggests that throat-related events are classified based on generic acoustic features, such as harmonic patterns, rather than linguistic characteristics[54]. However, slight variations in classification accuracy were observed across language groups, particularly in the German and Spanish groups. These variations were due to lower accuracy in classifying throat-clearing events caused by differences in throat-clearing signal patterns across individuals[55]. The macro-averaged receiver operation characteristic (ROC) curves are illustrated in Fig. 6c. The proposed model achieved an area under the ROC curve (AUC) of greater than 0.99, indicating high sensitivity and specificity with excellent discrimination between positive and negative classes.

Practical demonstration of the proposed algorithm

We demonstrated the proposed system in daily-life environments, confirming its outstanding accuracy for classifying throat-related events and highlighting its superior noise immunity, sound sensitivity, and hardware durability. Figure 7a illustrates the collection of acoustic signals utilizing the STVS while a subject walked at a consistent speed of 4 km/h. The STVS remained securely attached to the subject's neck after the experiment, and acoustic signals emanating from throat vibrations were recorded precisely,

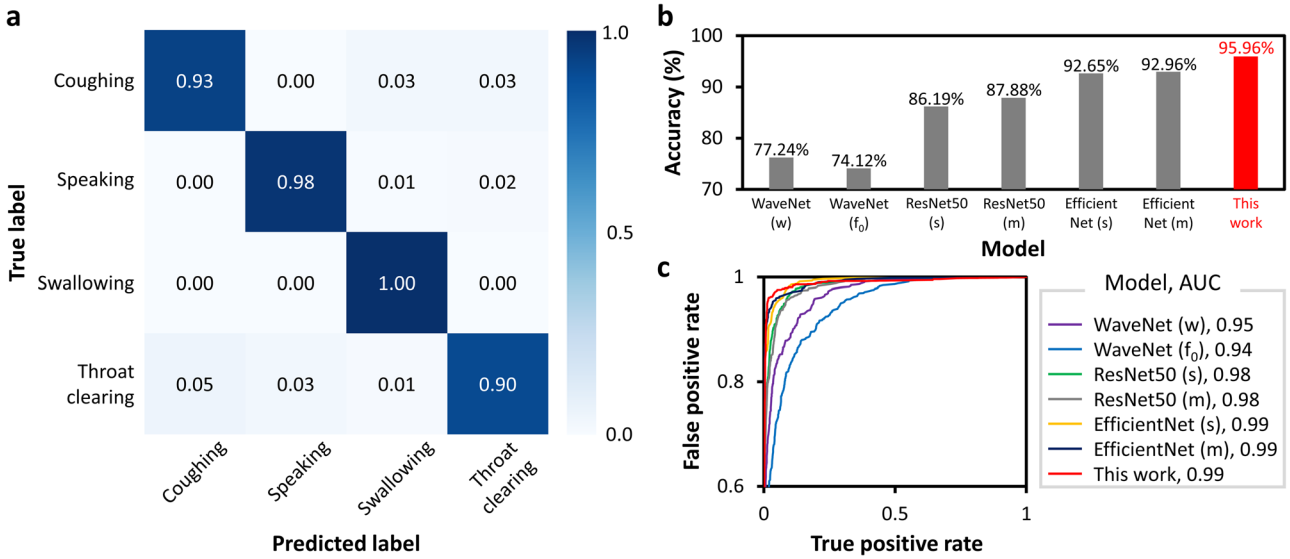

Fig. 6 | Performance metric of our ensemble-based deep learning model. **a** Normalized confusion matrix of the ensemble model on the test dataset. All event classification accuracies exceed 90%. The evaluation of **b** accuracy and **c** macro-averaged receiver operating characteristic (ROC) curves across single neural networks and proposed ensemble model on the test dataset. The abbreviations denoted distinct preprocessing techniques: 'w' for waveform, 'f_0' for fundamental frequency, 's' for spectrogram, and 'm' for mel spectrogram. The ensemble model achieves the highest accuracy of 95.96% and an area under the ROC curve (AUC) value of 0.99 for classifying the four events.

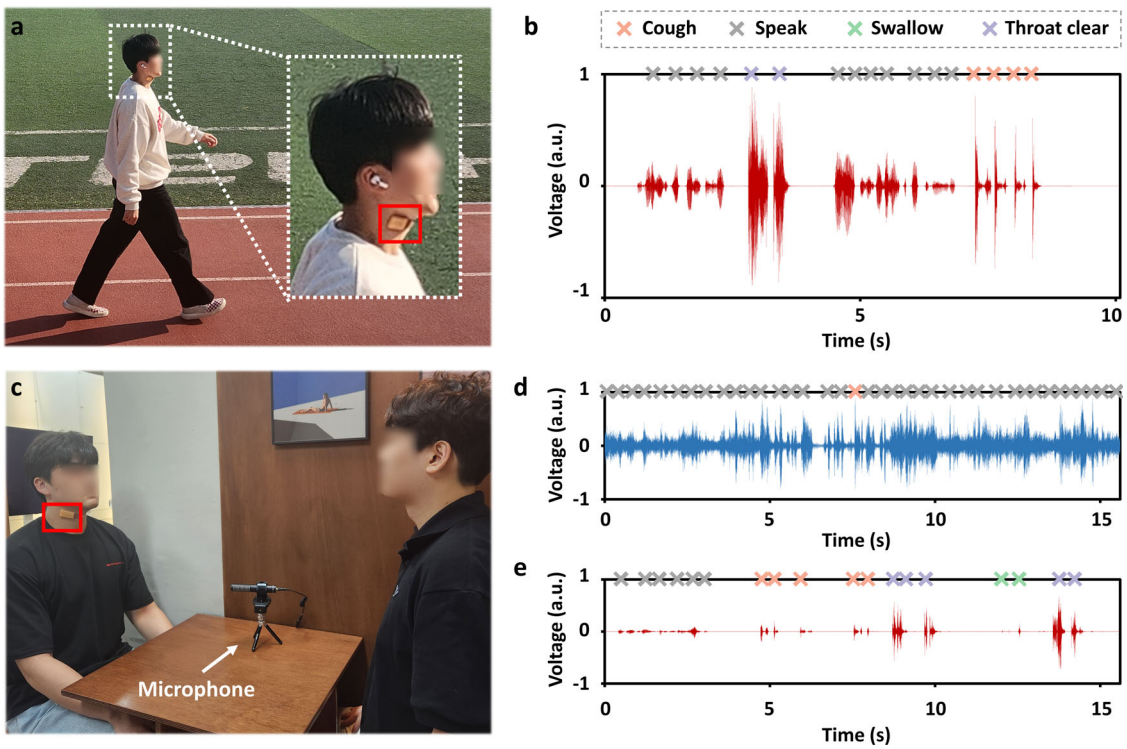

Fig. 7 | Throat vibration monitoring in daily life with the soft skin-attachable throat vibration sensor (STVS). **a** Experimental setup of vibration signal measurement during walking on a running track with a 4 km/h speed. **b** Vibration signal measured by the STVS. The outstanding stretchability of the STVS mitigated motion artifacts caused by walking. The throat-related events, such as coughing, speaking, swallowing, and throat clearing, were precisely detected through the proposed classification model. **c** An experimental setup was used to measure throat vibration signals while a subject was conversing with another person. **d** Vocal signal measured by a conventional acoustic microphone. The sounds caused by throat-related events in the subject interfered with conversations of another subject and ambient noise. **e** Throat vibration signal measured by the STVS. The measured signal is clear and accurately classified.

as shown in Fig. 7b. Our classification model accurately identified coughing and throat-clearing events during free speech. Additionally, we conducted a subsequent test at an increased speed of 8 km/h to evaluate the impact of motion artifacts (see Supplementary Fig. 8). Motion artifacts and baseline wandering caused by running were removed through a high-pass filter embedded in the STVS. Our classification system achieved an accuracy of 95.16% under this vigorous condition, comparable to the 95.96% attained with the test dataset, which was measured in a static environment. This

result demonstrates the system's robustness and reliability in real-world scenarios. Furthermore, we assessed the impact of environmental factors such as head movement and sweat (see Supplementary Fig. 9). The subject intentionally moved their head while performing throat-related events, and 1 mL of deionized water was applied between the sensor and the neck skin to simulate sweat. These factors did not affect the STVS's accuracy in detecting the four throat-related events. As depicted in Fig. 7c, we also demonstrated the detection of throat-related events during conversations. The subject performed throat-related events during free speech without additional instructions. To highlight the noise immunity of the STVS, voice signals were simultaneously measured using both the STVS and a conventional acoustic microphone (BOYA, BY-PVM50). The acoustic microphone captures the user's voice and extraneous sounds, such as other people's conversations and environmental noise. Therefore, the sounds produced when swallowing or throat clearing can be masked by other people's voices, limiting accurate classification of throat-related events, as shown in Fig. 7d. In contrast, the STVS specifically measured the signals resulting from the subject's throat vibrations, regardless of external noise. As shown in Fig. 7e, the proposed classification system, encompassing the STVS and the multi-modal ensemble model, successfully detected all throat-related events precisely. Additionally, free speech signals, such as loud talking and laughing, were accurately classified as speaking by our model (see Supplementary Fig. 10).

Discussion

In this study, we developed an advanced healthcare system for real-time monitoring of throat-related events, devising a soft skin-attachable throat vibration sensor (STVS) and an ensemble classification model. The STVS was designed with a small and stretchable form factor, ensuring continuous signal monitoring even during movement. Moreover, our model achieved high accuracy in classifying throat-related events such as coughing, swallowing, speaking, and throat clearing by integrating multi-modal data and optimizing neural network architectures. This advancement offers a significant improvement in the management of dysphagia compared to previous approaches.

Most previous studies employed conventional acoustic microphones, which struggle to detect subtle sounds such as swallowing events and are susceptible to ambient noise, thereby complicating the precise observation of various throat-related events[16–22,56]. Several studies have proposed skin-attachable sensors to detect multiple events from throat vibrations[23–26,57]. However, the limited form factor of these sensors presents challenges in attaching them to the laryngeal prominence, which is optimal for precisely capturing vocal cord vibrations[23–27]. In this study, we successfully affixed sensors to the narrow and contoured laryngeal prominence through a strategy of hardware dualization, facilitating the acquisition of high SNR acoustic signals. We experimentally measured the vibration signals above the laryngeal prominence, and this location exhibited significantly higher sensitivity and SNR compared to the suprasternal notch and the side of the neck in previous methods (see Supplementary Fig. 11)[23–26,28]. Furthermore, the stretchable serpentine structure of the wearable hardware was meticulously adapted to the anatomical contours of the neck, ensuring consistent and stable signal acquisition even when the wearer is moving. Our sensor's adaptability to the dynamic movements of the neck marks a considerable improvement in patient comfort and data reliability, which are often compromised in traditional monitoring methods.

Most of the previous studies on the classification of throat-related events utilized single neural networks such as ResNet and SVM[23,25,26,28]. These networks often rely on a single type of input data, which limits the diversity of features and increases the risk of overfitting. The limitation was evident in the Grad-CAM results from our analysis of EfficientNet. This image-based neural network struggles to accurately classify similar signals like coughing and throat clearing, resulting in lower classification accuracy and limited generalizability. Our approach leverages signal domain diversification and an optimized ensemble technique for multi-feature extraction to address these issues. The proposed ensemble technique combined the strengths of various neural networks, allowing for a more nuanced understanding of complex acoustic signals associated with different throat-related events. This method achieved the highest accuracy for throat-related events classification compared to previous studies. Even though we utilized a few representative networks in the ensemble process for model development, this approach can be integrated with a broader range of networks, potentially contributing to enhanced performance in future work.

In conclusion, our innovative healthcare system ensures precise detection of throat-related events through significant advancements in both hardware and software. This breakthrough highlights the effectiveness of our sensor system in a critical healthcare application and establishes a new standard for the development of diagnostic tools aimed at enhancing the prognosis and management of dysphagia. By providing a reliable, real-time monitoring solution, our system holds the potential to significantly improve patient outcomes, enabling timely interventions and more personalized treatment strategies. As we refine and expand upon this technology, we envision a future where advanced diagnostic tools like ours become integral to managing dysphagia and other related conditions, ultimately contributing to better health and quality of life for patients.

Methods

Fabrication of hardware and soft encapsulation

The flexible printed circuit board (fPCB) was fabricated by a PCB manufacturer (Hansaem Digitec, Korea). Multiple integrated circuit (IC) chips were electrically attached to the fPCB through an infrared IC heater (Puhui Electric Technology, T-962) with solder paste (ChipQuik, SMD4300AX10T5).

The STVS included the following elements: (1) an inertial measurement unit (IMU, TDK InvenSense, IIM-42652) to measure the z-axis acceleration on the throat, featuring a sampling rate of 6400 Hz and a dynamic range of ±4 g; (2) a microcontroller unit (MCU, STMicroelectronics, STM32WB55CGU6) to control the system and enable wireless communication; (3) a 2 Gb NAND flash memory (Micron, MT29F2G01) to store the acquired data; (4) a battery management IC (Shenzhen Fuman Elec, TC4056A) and a 50 mAh Li-Po battery to supply power; and (5) associated passive components such as resistors, capacitors, and inductors. The MCU controlled the IMU and NAND flash memory via a serial peripheral interface. A high-pass filter (cut-off frequency: 20 Hz) was embedded in the MCU to remove baseline wandering from the acceleration signals. The processed signals were then transmitted to a mobile device through BLE (see Supplementary Fig. 12). The quantitatively calculated battery duration of the STVS was 5.2 h with a 50 mAh Li-Po battery.

A soft encapsulation was made of polymer (XINUS SILICONE, SH3264) to protect the hardware. The bottom and top layers of the encapsulation structure were formed using female and male polylactic acid molds produced with a 3D printer. The uncured polymer was poured into the female mold and covered with the male mold. Subsequently, the assembly was cured at 50 °C for 6 h in an oven. Two layers of the encapsulation were fabricated using the same procedure. After all layers were manufactured, the fPCB was mounted between the bottom and top layers. Then, the uncured polymer was poured onto the contact surface of the layers and cured at 50 °C for 6 h to establish a robust mechanical bond. The underside of the completed encapsulation was bonded with double-sided silicone acrylate adhesive (3 M, 2477p) for attaching the device to the skin.

Subjects and dataset

We conducted the measurement with subjects covering diverse ages, genders, and linguistic backgrounds (see Supplementary Table 1). The study was approved by the Institutional Review Board of Pohang University of Science and Technology (PIRB-2023-E010) and by the Ethics Committee of Northwestern and Central Switzerland (BASEC-No: Req-2022-0082). All the subjects provided informed consent following a comprehensive explanation of the study procedures. Additionally, all individuals depicted in the images provided written consent for publication.

The data acquired with the STVS were categorized into training and test datasets based on spoken language during the study (see Supplementary

Table 2). The training dataset included English-speaking subjects' throat-related events, such as coughing, speaking, swallowing, and throat clearing. Conversely, the test dataset comprised these same events, but they were derived from subjects with various linguistic backgrounds, including English, French, German, Spanish, and Korean. Our experimental protocol sampled spoken words from the TensorFlow speech commands dataset in each language's native form[58].

Computing infrastructure

The hardware setup for data processing included an Intel i7-6800K Processor, 128GB RAM, and 8TB SSD storage. On the software front, we utilized the Windows 10 operating system. We employed Python 3.11.5 to develop our application, leveraging its extensive libraries for data analysis and machine learning. This integrated computing infrastructure provided a reliable and scalable platform for conducting our extensive data analyses.

Signal preprocessing for deep learning model

Data augmentation techniques were implemented to enhance the training dataset's diversity and prevent overfitting[59]. The data was expanded thirteenfold through flipping, moving, scaling, and noise injection operations[60]. This augmentation process generated 25,792 coughing, 27,391 speaking, 26,364 swallowing, and 27,157 throat-clearing events.

The training dataset was then transformed into the waveform, f_0, spectrogram, and mel spectrogram (see Supplementary Fig. 6). f_0 values were estimated using the robust algorithm for pitch tracking (RAPT), which leverages normalized cross-correlation function to track the pitch in speech data[61]. The waveform and f_0, representing time-series data, were utilized as inputs for the WaveNet. Spectrogram and mel spectrogram images were generated using the STFT. The frame and overlap of the Hanning window for STFT were set to 80 ms and 87.5%, respectively, and the coefficient of the mel filter bank was 128. These images were resized to 224×224 and normalized before serving as inputs for the ResNet50 and the EfficientNet.

Deep neural network

We utilized three state-of-the-art neural networks (WaveNet, ResNet50, and EfficientNet) to distinguish throat-related events. The structures of these three networks were as follows:

1. In the WaveNet, each layer comprised 32 input and 32 output channels[45]. Specific layers adjusted dilation rates to effectively capture information from the distant layers. In addition, the residual and skip modules addressed gradient vanishing issues. The final output was predicted through a fully connected layer with 256 features, mapping to four classes.
2. The ResNet50 processed input images with a dimension of 224×224, utilizing a 7×7 convolutional layer to detect 64 features[39]. Each layer incorporated residual blocks, employing 1×1 convolution to match input and output dimensions. After the last convolutional layer, the global average pooling layer reduced the spatial dimension by calculating the average values. Finally, a 2048-dimensional fully connected layer predicted the output class.
3. The EfficientNet processed input images sized 224×224[40]. Initially, a padding module with a 3×3 kernel was employed, featuring 32 input and 32 output channels. Then, the network's MBConv block incorporated depthwise separable convolution with various kernel sizes and strides. Additionally, the squeeze-and-excitation mechanism emphasized inter-channel relationships, enhancing the network's expressive power. Following the last block, the output class was predicted through global average pooling and a fully connected layer with 1280 features.

The training process for all deep neural networks followed a fivefold cross-validation procedure, employing the Adam optimizer and utilizing cross-entropy as the loss function. During training, the increase in accuracy and reduction in loss across all networks indicated the effective capture of underlying patterns within the data (see Supplementary Fig. 13).

Gradient-weighted class activation mapping

The gradient-weighted class activation mapping (Grad-CAM) technique was employed to highlight the crucial regions in the image relevant to the network's decisions[42]. This process entails computing the gradients of the predicted class in the last convolutional layer to generate a heatmap. The redder regions within the heatmap indicate the more significant features in the neural network prediction.

Ensemble model

The architectures of the seven algorithms used to ensemble individual networks were as follows:

1. The GBM employed the deviance loss function and constructed an ensemble of decision trees[47]. It sequentially fitted new trees to correct the errors of the previous ones, with a learning rate of 0.1. The individual trees had a maximum depth of 3, and the ensemble comprised 100 such trees. The model utilized the entire training dataset for growing trees, setting the minimum samples for node splitting to 2 and leaf nodes to 1.
2. The RF classifier created an ensemble of decision trees with 100 trees[48]. Each tree was constructed using the Gini criterion for impurity, with no restriction on the maximum depth. The model established a minimum of 2 samples for node splitting and 1 sample for leaf nodes. The number of features to consider for the best split was automatically determined.
3. The XGBoost classifier was a boosted tree algorithm with 100 trees and a learning rate of 0.148[49]. Each tree had a maximum depth of 3 and utilized the entire training dataset. The feature sampling rate was set to 1 when constructing each tree, considering all features. Additionally, regularization terms were set to zero.
4. The LightGBM enhances the speed of GBM, using a leaf-wise splitting strategy[50]. It had 31 maximum leaves per tree, unlimited tree depth, and a learning rate of 0.1 across 100 boosting iterations. These parameters were selected to enhance performance and efficiency in capturing intricate patterns within the data.
5. The extra trees classifier formed an ensemble of 100 decision trees, using the Gini criterion for impurity and imposing no restrictions on the maximum tree depth[51]. Like the RF classifier, the number of features considered for the best split was determined automatically.
6. The AdaBoost classifier built an ensemble of 50 weak learners with a learning rate of 1.0, using the SAMME.R algorithm[52]. The ensemble adapted to misclassified samples, emphasizing difficult-to-classify instances in subsequent iterations.
7. The SVM classifier operated as a nonlinear classification model using the radial basis function (RBF) kernel[53]. The RBF kernel was characterized by parameters such as gamma (set to scale) and a default degree of 3 for the polynomial kernel. Using the regularization parameter, it sought the hyperplane that best separated classes in the feature space. Finally, the decision function was determined by the one-vs-rest scheme for multi-class classification.

Data availability

The data supporting this study's findings are publicly available on Zenodo at https://doi.org/10.5281/zenodo.12630148.

Code availability

The ensemble-based deep learning model codes are available on GitHub at https://github.com/yonghunsong/Throat-related-events-classification.git.

Received: 13 July 2024; Accepted: 21 December 2024;
Published online: 07 January 2025

References

1. Clavé, P. & Shaker, R. Dysphagia: current reality and scope of the problem. *Nat. Rev. Gastroenterol. Hepatol.* **12**, 259–270 (2015).
2. Sungsinchai, S., Niamnuy, C., Wattanapan, P., Charoenchaitrakool, M. & Devahastin, S. Texture modification technologies and their

opportunities for the production of dysphagia foods: a review. *Compr. Rev. Food Sci. Food Saf.* **18**, 1898–1912 (2019).
3. Labeit, B. et al. The assessment of dysphagia after stroke: state of the art and future directions. *Lancet Neurol.* **22**, 858–870 (2023).
4. Rommel, N. & Hamdy, S. Oropharyngeal dysphagia: manifestations and diagnosis. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 49–59 (2016).
5. Murry, T., Carrau, R. L. & Chan, K. *Clinical Management of Swallowing Disorders* (Plural Publishing, 2020).
6. Hammond, C. A. S. & Goldstein, L. B. Cough and aspiration of food and liquids due to oral-pharyngeal dysphagia: ACCP evidence-based clinical practice guidelines. *Chest* **129**, 154S–168S (2006).
7. Cook, I. J. Diagnostic evaluation of dysphagia. *Nat. Rev. Gastroenterol. Hepatol.* **5**, 393–403 (2008).
8. Feng, W. Diagnosis of post-stroke dysphagia: towards better treatment. *Lancet Neurol.* **22**, 778–779 (2023).
9. Guo, W. J. et al. Effects of anxiety and depression and early detection and management of emotional distress on length of stay in hospital in non-psychiatric inpatients in China: a hospital-based cohort study. *Lancet* **394**, S83 (2019).
10. Rafeedi, T. et al. Wearable, epidermal devices for assessment of swallowing function. *NPJ Flex. Electron.* **7**, 52 (2023).
11. Kang, Y. J. et al. Soft skin-interfaced mechano-acoustic sensors for real-time monitoring and patient feedback on respiratory and swallowing biomechanics. *NPJ Digit. Med.* **5**, 147 (2022).
12. Xu, H. et al. A fully integrated, standalone stretchable device platform with in-sensor adaptive machine learning for rehabilitation. *Nat. Commun.* **14**, 7769 (2023).
13. Lee, K. et al. Mechano-acoustic sensing of physiological processes and body motions via a soft wireless device placed at the suprasternal notch. *Nat. Biomed. Eng.* **4**, 148–158 (2020).
14. Ramírez, J. et al. Metallic nanoislands on graphene for monitoring swallowing activity in head and neck cancer patients. *ACS Nano* **12**, 5913–5922 (2018).
15. Zhang, D. et al. Stretchable and durable HD-sEMG electrodes for accurate recognition of swallowing activities on complex epidermal surfaces. *Microsyst. Nanoeng.* **9**, 115 (2023).
16. Liaqat, D. et al. Coughwatch: real-world cough detection using smartwatches. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*., 8333–8337 (2021).
17. Hoyos-Barceló, C., Monge-Álvarez, J., Shakir, M. Z., Alcaraz-Calero, J. M. & Casaseca-de-La-Higuera, P. Efficient k-NN implementation for real-time detection of cough events in smartphones. *IEEE J. Biomed. Health Inform.* **22**, 1662–1671 (2017).
18. Alqudaihi, K. et al. Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. *IEEE Access* **9**, 102327–102344 (2021).
19. Kadambi, P. et al. Towards a wearable cough detector based on neural networks. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*. 2161–2165 (2018).
20. Barata, F. et al. Towards device-agnostic mobile cough detection with convolutional neural networks. In *Proc. IEEE Int. Conf. Healthc. Inform*. 1–11 (2019).
21. Peng, P. et al. Design of an efficient CNN-based cough detection system on lightweight FPGA. *IEEE Trans. Biomed. Circuits Syst.* **17**, 116–128 (2023).
22. Shin, B. et al. Automatic clinical assessment of swallowing behavior and diagnosis of silent aspiration using wireless multimodal wearable electronics. *Adv. Sci.* **11**, 2404211 (2024).
23. Ni, X. et al. Automated, multiparametric monitoring of respiratory biomarkers and vital signs in clinical and home settings for COVID-19 patients. *Proc. Natl Acad. Sci. USA* **118**, e2026610118 (2021).
24. O'Brien, M. K. et al. Advanced machine learning tools to monitor biomarkers of dysphagia: a wearable sensor proof-of-concept study. *Digit. Biomark.* **5**, 167–175 (2021).
25. Jeong, H. et al. Differential cardiopulmonary monitoring system for artifact-canceled physiological tracking of athletes, workers, and COVID-19 patients. *Sci. Adv.* **7**, eabg3092 (2021).
26. Tzavelis, A. et al. Development of a miniaturized mechanoacoustic sensor for continuous, objective cough detection, characterization and physiologic monitoring in children with cystic fibrosis. *IEEE J. Biomed. Health Inform.* **28**, 5941–5952 (2024).
27. Song, Y. et al. Study on optimal position and covering pressure of wearable neck microphone for continuous voice monitoring. In *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc*. 7340–7343 (2021).
28. Groh, R., Lei, Z., Martignetti, L., Li-Jessen, N. Y. & Kist, A. M. Efficient and explainable deep neural networks for airway symptom detection in support of wearable health technology. *Adv. Intell. Syst.* **4**, 2100284 (2022).
29. Lever, J., Krzywinski, M. & Altman, N. Points of significance: model selection and overfitting. *Nat. Methods* **13**, 703–704 (2016).
30. Sazonov, E. et al. Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior. *Physiol. Meas.* **29**, 525 (2008).
31. Ganaie, M. A., Hu, M., Malik, A., Tanveer, M. & Suganthan, P. N. Ensemble deep learning: a review. *Eng. Appl. Artif. Intell.* **115**, 105151 (2022).
32. Zuluaga-Gomez, J., Ahmed, S., Visockas, D. & Subakan, C. CommonAccent: exploring large acoustic pretrained models for accent classification based on common voice. In *Proc. Interspeech*, 5291–5295 (2023).
33. Lakhotia, K. et al. On generative spoken language modeling from raw audio. *Trans. Assoc. Comput. Linguist.* **9**, 1336–1354 (2021).
34. Matsuhisa, N., Chen, X., Bao, Z. & Someya, T. Materials and structural designs of stretchable conductors. *Chem. Soc. Rev.* **48**, 2946–2966 (2019).
35. Widlund, T., Yang, S., Hsu, Y. Y. & Lu, N. Stretchability and compliance of freestanding serpentine-shaped ribbons. *Int. J. Solids Struct.* **51**, 4026–4037 (2014).
36. Yin, L. et al. From all-printed 2D patterns to free-standing 3D structures: controlled buckling and selective bonding. *Adv. Mater. Technol.* **3**, 1800013 (2018).
37. Xu, S. et al. Stretchable batteries with self-similar serpentine interconnects and integrated wireless recharging systems. *Nat. Commun.* **4**, 1543 (2013).
38. Song, Y., Kim, Y., Jeung, J., Yun, I. & Chung, Y. Voice monitoring system for vocal dose measurement in daily life. In *Proc. IEEE Int. Conf. Consum. Electron. Asia*, 1–4 (2022).
39. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 770–778 (2016).
40. Tan, M. & Le, Q. Efficientnet: rethinking model scaling for convolutional neural networks. In *Proc. Int. Conf. Mach. Learn*. 6105–6114 (2019).
41. Mumuni, A. & Mumuni, F. Data augmentation: a comprehensive survey of modern approaches. *Array* **16**, 100258 (2022).
42. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proc. IEEE Int. Conf. Comput. Vis*., 618–626 (2017).
43. Huang, Y. et al. What makes multi-modal learning better than single (provably). *Adv. Neural Inf. Process. Syst.* **34**, 10944–10956 (2021).
44. Mohammed, A. & Kora, R. A comprehensive review on ensemble deep learning: opportunities and challenges. *J. King Saud. Univ. Comput. Inform. Sci.* **35**, 757–774 (2023).
45. van den Oord, A. et al. Wavenet: a generative model for raw audio. Preprint at https://arxiv.org/abs/1609.03499 (2016).
46. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).

47. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **7**, 21 (2013).
48. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
49. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM Int. Conf. Knowl. Discov. Data Min*. 785–794 (2016).
50. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 3146–3154 (2017).
51. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
52. Hastie, T., Rosset, S., Zhu, J. & Zou, H. Multi-class adaboost. *Stat. Interface* **2**, 349–360 (2009).
53. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Appl.* **13**, 18–28 (1998).
54. Al-Shoshan, A. I. Speech and music classification and separation: a review. *J. King Saud. Univ. Eng. Sci.* **19**, 95–132 (2006).
55. Xiao, Y. et al. The acoustic cough monitoring and manometric profile of cough and throat clearing. *Dis. Esophagus* **27**, 5–12 (2014).
56. Orlandic, L., Teijeiro, T. & Atienza, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Sci. Data* **8**, 156 (2021).
57. Yun, I., Jeung, J., Kim, Y., Song, Y. & Chung, Y. Ultra-low-power wearable vibration sensor with highly accurate embedded classifier. In *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc*. 2451–2454 (2022).
58. Warden, P. Speech commands: a dataset for limited-vocabulary speech recognition. Preprint at https://arxiv.org/abs/1804.03209 (2018).
59. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 60 (2019).
60. Maharana, K., Mondal, S. & Nemade, B. A review: data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **3**, 91–99 (2022).
61. Talkin, D. & Kleijn, W. B. A robust algorithm for pitch tracking (RAPT). *Speech Coding and Synthesis*, 495–518 (Elsevier, 1995).

Acknowledgements

This research was supported by the National R&D Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT (2020M3H2A107804521); by the National Research Foundation grant funded by the Ministry of Science and ICT (2021M3C1C3097512); by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant, funded by the Korean government (RS-2019-II191906, Artificial Intelligence Graduate School Program); and by the Educational Institute for Intelligent Information Integration through the BK21 FOUR project of the Korean government.

Author contributions

Y.S. and I.Y. are co-first authors of this manuscript and contributed equally to this study. Y.S., I.Y., S.G., and Y.C. conceptualized the idea and designed the research studies. Y.S. designed and manufactured the skin-attachable sensor. I.Y. performed the mechanical analysis of the sensor. S.G., C.A.E., and Y.C. designed the human subject studies. Y.S., I.Y., and S.G. performed the data collection. Y.S. and I.Y. developed the deep learning algorithm and performed the data analysis, with assistance from Y.C. Y.C. supervised and validated the entire process as the corresponding author and contributed to the funding acquisition. Y.S. and I.Y. drafted the manuscript, and all authors contributed to the review and editing process. All authors have approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01417-w.

Correspondence and requests for materials should be addressed to Yoonyoung Chung.