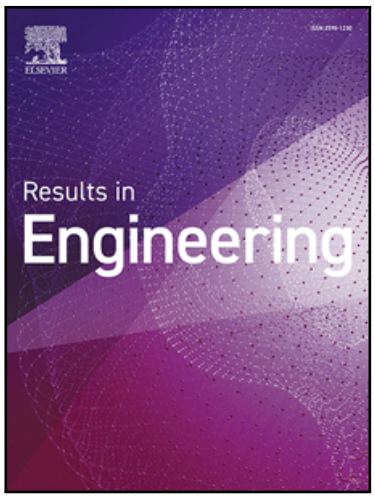


Advanced Grad-CAM Extensions for Interpretable Aphasia Speech Keyword Classification: Bridging the Gap in Impaired Speech with XAI

Gowri Prasood Usha , John Sahaya Rani Alex

PII: S2590-1230(24)01666-9
DOI: https://doi.org/10.1016/j.rineng.2024.103414
Reference: RINENG 103414

To appear in: *Results in Engineering*

Received date: 15 September 2024
Revised date: 5 November 2024
Accepted date: 13 November 2024

Please cite this article as: Gowri Prasood Usha , John Sahaya Rani Alex , Advanced Grad-CAM Extensions for Interpretable Aphasia Speech Keyword Classification: Bridging the Gap in Impaired Speech with XAI, *Results in Engineering* (2024), doi: https://doi.org/10.1016/j.rineng.2024.103414

Highlights

- Aphasic speech poses challenges in speech recognition.
- Standard Grad-CAM has some limitations when it comes to processing Aphasic speech.
- Proposed four enhanced Grad-CAM techniques, including hierarchical feature mapping.
- Achieved focused, class-specific heatmaps using multi-scale, directional, and dropout methods.
- Techniques improve interpretability and hold potential for clinical use in Aphasia therapy.

Advanced Grad-CAM Extensions for Interpretable Aphasia Speech Keyword Classification: Bridging the Gap in Impaired Speech with XAI

Gowri Prasood Usha[1], John Sahaya Rani Alex[1a]

[1]*School of Electronics Engineering, Vellore Institute of Technology Chennai, India - 600127*

[a]Corresponding author. *School of Electronics Engineering, Vellore Institute of Technology Chennai, India - 600127*

E-mail address: jsranialex@vit.ac.in

Abstract

Aphasia, a language disorder caused by brain injury, presents significant speech recognition and classification challenges due to irregular speech patterns. While the standard Grad-CAM (Gradient-weighted Class Activation Mapping) technique is widely used for model interpretation, its application to impaired speech remains largely unexplored. To address this gap, we introduce a set of extension studies of enhanced Grad-CAM techniques, namely Enhanced Directional Grad-CAM (ED-GCAM), Multi-Scale Channel-wise Grad-CAM (MSCW-GCAM), Stochastic Gradient-Dropout Integrated Grad-CAM (SGD-GCAM), and Enhanced Hierarchical Filtered Grad-CAM (EH-FCAM) to improve interpretability and performance in aphasia speech keyword classification. When applied to attention-based CNN models, these techniques generate more focused, class-specific heatmaps, providing a deeper understanding of model behaviour, particularly in noisy and impaired speech. Our results demonstrate that these enhanced Grad-CAM methods outperform the standard Grad-CAM by offering more detailed and meaningful explanations, which is critical for interpreting models applied to aphasia speech. We compare our approach using qualitative and perturbation-based trustworthiness, infidelity and sufficiency scores as quantitative metrics. Among the techniques, ED-GCAM outperformed all others. The proposed methods significantly improve the accuracy and transparency of speech processing models, with potential suggestions for clinical applications.

Keywords: Aphasia, Grad-CAM, Explainable AI, XAI, Speech Recognition, Spoken keyword Classification, Impaired Speech, Deep Learning

1. Introduction

A brain injury or stroke can affect a person's brain, leading to a language impairment called aphasia, which hinders communication. Individuals with aphasia often struggle with writing, reading, speaking, and understanding language[1]. Speech recognition and classification can be challenging in patients with aphasia, as speech patterns vary significantly from those of unaffected people[2]. This challenge has led to a surge in use of machine learning and deep learning techniques in the automatic recognition and classification of aphasia speech, particularly in speech analysis and keyword spotting tasks[3].

Even though feature engineering, rule-based algorithms and statistical techniques perform satisfactorily,[4] traditional models for aphasia speech analysis sometimes fall short because of the complexity and variety of aphasic speech patterns [5,6]. Possessing the ability to automatically extract features from raw data, deep learning models, particularly those that employ convolutional neural networks (CNNs), have demonstrated great promise in the processing and analysis of speech [7]. Deep learning models are better equipped to deal with aphasia, where speech abnormalities and disruptions are widespread, by collecting complex patterns in the data without requiring much manual intervention [8]. Spoken keyword detection and Automatic Speech Recognition (ASR) have become essential tasks in speech processing. ASR systems seek to translate spoken words into text, whereas keyword spotting concentrates on detecting particular words or phrases within speech being spoken continuously. However, spoken keyword classification models that categorise speech into pre-defined keyword groups are significant for aphasia applications as they make it possible to identify important keywords that support communication and therapy.

Several related studies have investigated the application of deep learning models for recognising speech in aphasia[9,10] emphasising speech-to-text systems specially designed to identify and evaluate the features of aphasic speech. Recurrent Neural Networks (RNNs), specifically Gated Recurrent Units (GRUs), are designed to handle aphasia's complexity by utilising architectures. The researchers sought to improve the assessment accuracy and performance of the voice recognition systems, through better representation of the temporal relationships in sequential data. According to these findings, deep learning models, especially those that incorporate attention mechanisms, are better equipped to adjust to the irregular characteristics of aphasia speech [11]. In this context, models incorporating attention layers offer significant advantages over traditional models. Some studies use deep learning models to handle significant variability in speech patterns among aphasics[12][13]. Researchers aimed to enhance the accuracy and resilience of automatic speech recognition systems for aphasia assessment by creating models specifically tailored to the unique characteristics of aphasic speech, such as paraphrastic errors and neologisms. When dealing with the irregular and noisy character of aphasia speech, attention mechanisms enable the model to focus on the most relevant portions of the input, such as particular time-frequency regions in the Mel-spectrogram. Therefore, attention-based models are significantly more successful than traditional models at precisely identifying crucial aspects of speech impairment [8].

Few studies focussed on spoken keyword spotting, which involves identifying or detecting specific keyword presence in a continuous speech or audio[14][15]. This falls within the broader field of speech recognition, where it can be challenging to analyse spontaneous speech in aphasia because of phonological and semantic abnormalities, frequent pauses, hesitations, and grammatical errors. Literature shows that adoption of hybrid HMM/MLP and Bidirectional GRU for keyword spotting for aphasia [16], [17]. However, when it comes to the spoken keyword classification for aphasia, it remains minimal.

Application domains, such as object recognition, picture classification, text classification, and audio classification, have demonstrated significant success with deep learning models. However, many of these sophisticated models are black-box models, which makes it challenging to understand how they make decisions, particularly in the medical field. This is essential for categorising spoken keywords for people with aphasia, a language impairment that affects speech production and comprehension. In these situations, explainable AI (XAI) becomes crucial, allowing clinicians and patients to understand the logic behind the model for dependable and credible results.

Understanding the reasons behind model decision is as important as the decision themselves, XAI plays a critical role in applications. Although they are restricted to healthy speech, studies have merged XAI techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) in audio processing [18,19] for sound classification and speech recognition. By integrating XAI into audio classification for impaired speech aphasia, improving interpretability can shed light on the model's emphasis areas and make it easier for clinicians to comprehend and accept the model's predictions. While current XAI methods can direct the creation of more tailored and efficient treatment interventions, which will enhance patient outcomes, it is advised that user interfaces be used for this purpose [20]. However, there is still a need to develop interpretable models that provide light on the model's decision-making process, which is a crucial aspect of therapeutic and clinical contexts.

Gradient-weighted Class Activation Mapping, or Grad-CAM, is one of the methods that is currently in use and has gained popularity for its ability to visualise the input data and show which sections of the data influence the choice made by the model, making neural networks more accessible to read [21]. Grad-CAM is widely used in image classification or object detection [22] but its use in impaired speech processing especially for speech aphasia is mainly unexplored. Grad-CAM allows users to comprehend which portions of the speech signal are most crucial for categorisation by producing heatmaps that emphasise significant areas of the input. This can be especially helpful in speech applications for aphasia, where decision-making in the model must be transparent.

Several enhancements have been suggested to improve Grad-CAM's interpretability and accuracy. These include techniques like Grad-CAM++[23], which provides improved localisation of class discriminative zones, and Guided Grad-CAM, which sharpens the gradients to provide more insightful explanations. These techniques have been modified for use with time-frequency representations in the context of audio processing, such as Mel-spectrograms, to aid in visualising the audio signal segments that are most important to a model's conclusion. However, these extensions haven't been thoroughly investigated for aphasia speech, where abnormalities in speech patterns provide difficulties. Our work builds on these extensions, introducing further innovations like Enhanced Directional Grad-CAM and Multi-Scale Channel-wise Grad-CAM to improve interpretability and performance in aphasia speech keyword classification.

This work investigates how well models intended for aphasia speech keyword categorisation can utilise Grad-CAM and its improved extensions. We show that the attention based models integrated with enhanced Grad-CAM methods, yield

more comprehensive and insightful explanations, eventually enhancing model's interpretability and accuracy in the difficult field of aphasia speech processing.

Our work introduces four advanced modifications: Enhanced Directional Grad-CAM, Multi-scale Channel-wise Grad-CAM, Stochastic Gradient-Dropout Integrated Grad-CAM and Enhanced Hierarchical Filtered Grad-CAM.

Aphasia patients' speech differs significantly from that of healthy individuals. It is less intelligible and often includes mispronounced phonemes. Class Activation Maps (CAMs) can capture these speech characteristics, making them more reliable and interpretable for therapists. This, in turn, enhances their utility in therapeutic practice. For example, focused diagnoses are made possible by Enhanced Directional Grad-CAM, which enables healthcare providers to isolate particular phonemes or words that most favourably contribute to accurate classifications. By using this strategy, a speech therapist could, for instance, point out mispronounced phonemes in a patient's speech and modify therapy sessions to address the mistakes, thus expediting the therapeutic process.

The mispronunciation of phonemes in between or in the beginning of the words reduces the speech clarity. In this kind of scenario, the usage of CAM provides a more thorough examination of speech at several linguistic levels—phonemes, syllables, and words. As an example, Multiscale-Channel-wise CAM enables clinicians to identify small problems that are inapparent at higher language levels. Its application can also help therapists to pinpoint and address a patient's difficulties with particular phonemes that impact word pronunciation, resulting in more targeted treatment regimens.

Individuals with aphasia frequently show inconsistencies in their speech production, which can vary based on the type and severity of their condition. Factors such as fatigue, stress, and cognitive demands can exacerbate this variability, resulting in more erratic speech over time. Nevertheless, focused speech therapy and external prompts can enhance consistency, enabling patients to produce words and phrases more reliably across different sessions. For instance, evaluation of the consistency of speech production over time for aphasia speech can be done by adding variability to the gradient computation, Stochastic Gradient-Dropout Integrated Grad-CAM. This technique allows clinicians to monitor a patient's speech patterns over several sessions to monitor their stability and progress. While fluctuating CAMs may signal ongoing challenges that require therapy changes, steady highlights throughout time indicate solid improvement.

Aphasia speech often reveals difficulties across various linguistic dimensions, including inconsistencies in phonemes, challenges in word retrieval, and interruptions in sentence structure. To accurately capture these intricate patterns, a hierarchical CAM technique, one that can analyze each linguistic dimension—phonemic, lexical, and syntactic—both separately and as a whole, is necessary. For example, the Enhanced Hierarchical Filtered Grad-CAM (EH-F Grad-CAM) is an appropriate method. It initially captures detailed phoneme production by examining localized areas in spectrograms at the lower levels, facilitating in-depth phonemic analysis. Next, it compiles these phoneme-level activations to uncover lexical patterns at a more advanced level, emphasizing words and phrases. Finally, this approach integrates these findings at a syntactic level, offering a comprehensive perspective on sentence structure and linguistic context, thus effectively tackling the multi-layered intricacies of aphasia speech.

Together, these Grad-CAM methods offer a comprehensive and structured insight into speech issues, supporting the development of personalized treatment plans. They give therapists the ability to track patients' development over time, evaluate how consistently improvements occur, and modify therapies in response to unbiased, data-driven input. Furthermore, these methods can be easily incorporated into automated systems or remote therapy, providing patients who might not be able to attend in-person sessions with ongoing, individualized care. These tools can be used by healthcare practitioners to give visual feedback during telemedicine appointments, making therapy more accessible and successful. The suggested Grad-CAM approaches greatly improve speech diagnostics' precision, resilience, and interpretability, enabling medical professionals to provide aphasia patients with more successful therapies and enhancing patient outcomes. The versions mentioned are basic and aim to implement GRAD-CAM-based XAI methods for classifying speech-impaired individuals, which can assist clinicians in understanding black-box decisions.

The contributions are as follows:

- We introduced Multi-layer, **Enhanced Directional Gradient CAM (ED-GCAM)**. This advanced multi-scale visualization method combines guided gradients, gradient directionality, and median smoothing to generate highly interpretable class activation maps for neural network layers that are more profound than the outermost layer.
- The Multi-layer, **Enhanced Hierarchical Filtered Grad-CAM (EH-FCAM)** has been extended to provide a multilevel visualization of how information is processed through different stages of the model, from input to output.

- Multi-layer, **Multi-scale Channel-wise CAM** which highlights the importance of individual channels within the model's layers and offers a more granular view of model decisions, has been presented.
- For a more comprehensive feature importance analysis, we enhanced Multi-layer, **Stochastic Gradient-Dropout Integrated Grad-CAM (SGD-GCAM)** by incorporating stochastic dropout and averaging Grad-CAM results across many layers. This enhances uncertainty quantification, interpretability, and robustness.
- Applied these extended Grad-CAM techniques to the proposed attention based spoken keyword classification models, enabling a comparative analysis of their interpretability.
- Finally, compared these extensions with normal Grad-CAM, offering valuable insights into their interpretability strengths and limitations.

2. Materials and methods

2.1. Models

In this study, we experiment with 2 model structures of neural networks: the Single attention-based CNN and the Multi attention-based Parallel CNN network for classification purposes.

The modelling process includes three stages. First, we apply data augmentation techniques to the audio data to handle the data scarcity issues and improve the classification accuracy. Second, we extract one of the dominant features inherent in the audio data, i.e., the Mel spectrogram. Third, we feed the feature vector into the specially designed convolutional neural network models.

A. Single Attention-based CNN Model

The CNN proposed [24] 1989 was classified based on the features extracted through the convolutional operation. Any combination of the convolutional and pooling layers is called the CNN model. A fully connected layer is positioned behind the CNN model to classify targets. The CNN is separated into two main sections: one for feature extraction and the other for classification. Usually, the feature extraction process comprises a convolutional layer, a pooling layer, and the input data's retrieved features. The categorisation part includes a fully connected (FC) layer.

The CNN model can detect the critical features of the audio in the audio message [25]. The Single attention-based CNN model architecture used in this study is shown in Figure. 1. below, and the core model is explained below.

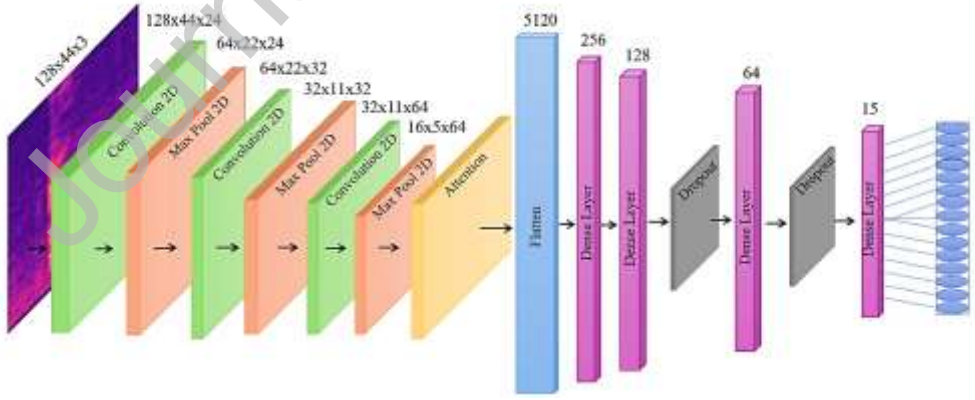


Fig. 1. Architecture of the Single attention-based CNN model

This model is a purely convolutional architecture integrating a spatial attention mechanism to enhance its interpretability and performance. The spatial attention layer focuses the model's attention on the most important regions in the input Mel-spectrogram, ensuring that relevant features are highlighted during the classification process.

Initially, Mel spectrograms are extracted from the audio and used as the CNN model input. The CNN model uses the 2D convolution layers to retain the original feature arrangement and obtain local patterns, some essential features such as frequency and temporal changes from the spectrogram. All convolution layers adopt a rectified linear unit (Relu) to shave off the eigenvalues less than 0 at the site to speed up model training, followed by a pooling layer. It reduces the spatial

dimensions of the feature maps while retaining the most essential information. Then, the spatial attention mechanism is applied to generate an attention map. This will highlight the important region of the input based on the learned features from the CNN layers. The spatial attention layer computes the importance of each region by considering both the frequency and time dimensions of the Mel-spectrogram. The attention map is then multiplied with the feature maps from the CNN layers, effectively reweighting the features to focus on the most informative areas for classification.

The output of the attention-weighted feature map is passed through the flattened layer to facilitate the subsequent use of the fully connected layer (FC) called the dense layer. FC layers are often placed at the end of neural network to increase computational power. Finally, the last dense layer with the softmax as activation function is connected to the classification output.

B. Multi Attention-based Parallel CNN Model

The model proposed is a parallel CNN with multi-attention, designed to improve the classification of spoken keywords, especially for aphasia speech. This model takes a Mel-spectrogram with specific dimensions as input, representing an audio signal's time-frequency characteristics. Its architecture includes three separate CNN branches that process the input individually through convolutional and max-pooling layers. These branches capture various multi-scale features from the Mel-spectrogram, enabling the model to extract a more comprehensive representation of the audio data. The Multi attention-based parallel CNN model architecture adopted in this study is shown in Fig. 2. below.

Attention mechanisms are added to each branch to enhance the focus on important areas of the Mel-spectrogram. The first branch improves the model's capacity to highlight pertinent time-frequency regions by employing a spatial attention mechanism that creates attention maps to highlight important spatial regions in the input. Graph attention, incorporated in the second branch, helps the model grasp context and interactions across distinct areas of the input by modelling the links between different regions of the spectrogram as a graph. To capture the sequential nature of spoken words, the third branch uses temporal attention, concentrating on the audio's most important time segments.

The outputs of the three attention mechanisms are concatenated to create a single feature representation that combines spatial, temporal, and contextual data after the parallel branches have processed the input. A sequence of completely connected layers passes through this concatenated feature vector.

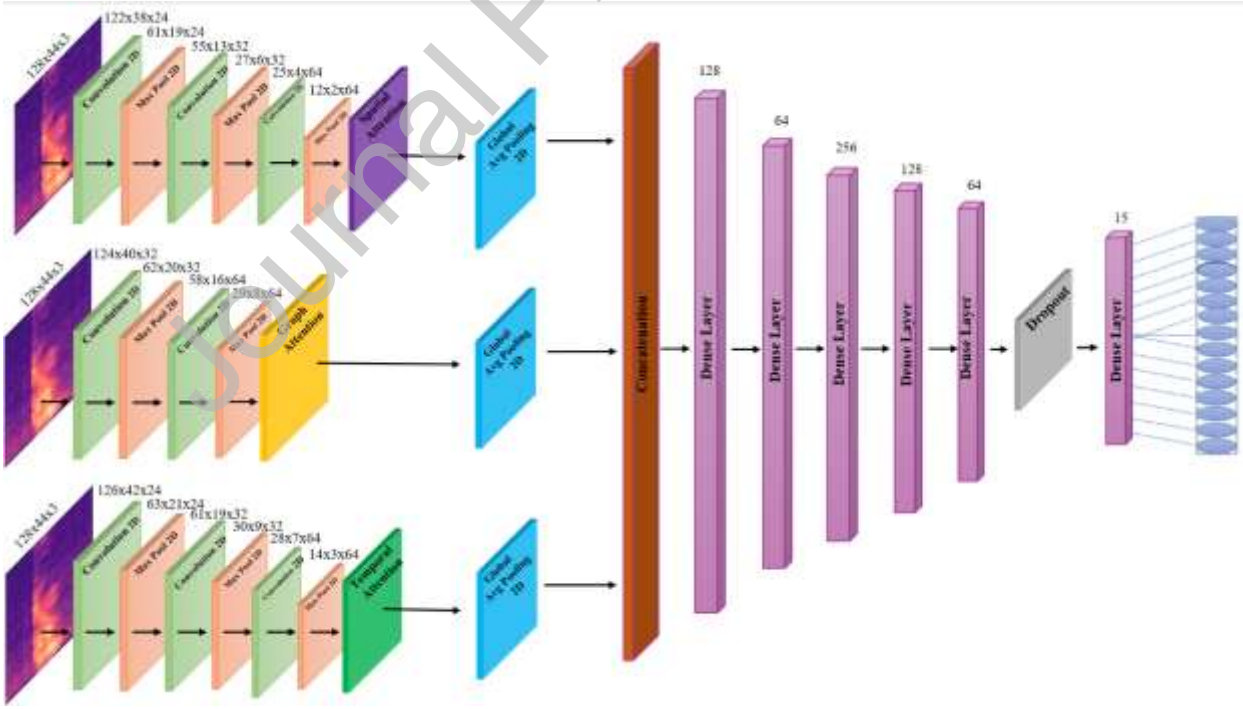


Fig. 2. Architecture of the Multi attention-based parallel CNN model

Non-linear interactions between the characteristics are captured by the first dense layer, which has 128 units. The second dense layer, which has 64 units, further refined these representations. By preventing overfitting, dropout regularisation ensures that the model performs appropriately when applied to unseen data.

The model consists of two classification heads: a main classification output layer that predicts the spoken keyword among 15 possible classes and an auxiliary classification output to aid in the primary task. The attention maps that were previously developed are also used to improve the final prediction by emphasising the most significant areas of the input and reweighing the features based on their significance. Overall, the model's ability to classify speech with aphasia is enhanced by this multi-attention-based parallel CNN architecture, which also increases interpretability by offering attention-based insights.

2.2. Data

The dataset used in this study includes spoken keywords from individuals with aphasia. Initially, we validated our model using the Google Speech Commands dataset [26], which contains a diverse set of spoken keywords from many speakers. No extra segmentation was needed as isolated spoken keywords exist in the aphasia datasets. The audio waves were converted into Mel spectrograms and fed into the CNN models. The spectrograms were normalised to guarantee that the neural network's input ranges were constant. We used various data augmentation techniques, including time stretching, pitch shifting, and time shifting, to improve the dataset size and variability due to the scarcity of aphasia speech data. Fig. 3. shows the effect of different types of waveform augmentation on the waveform of a sample audio.

The Aphasia dataset used in the study is taken from the Aphasia Bank data [27], which contains recordings of aphasic patients performing 15 isolated spoken keywords (15 different output classes) from speakers with 1758 samples. After the augmentation, the number of samples increased to 6822.

A. Time stretch

We simulated variations in speech speed by compressing and expanding the duration of the audio signal without affecting the pitch through the use of the data augmentation technique Time stretch. Theoretically, this would strengthen generalization by increasing the model's independence from the speaking rate. The stretching factor is usually represented by γ; a higher γ value ($\gamma > 1$) suggests a faster audio stream, while a lower value ($\gamma < 1$) suggests a slower stream. The stretching factor is evenly distributed with $\gamma \sim U$ (0.8 and 1.2)[28]. Figures 2g and 2h show the augmentation outcomes of this transformation on the original waveform.

B. Pitch shift

By slightly shifting the pitch of the speech, we created examples that varied in vocal tone. This helped the model generalize across speakers with different vocal ranges. This is particularly useful in aphasia speech, where pronunciation can vary widely. We assume that when using the pitch shifting factor as, the amount of artificial training data generated is $Naug$ times greater than the original data. The length of the audio samples remains constant. In our experiments, the range of semitone changes was [$-as$, as] for each signal. The pitch shift factors are $n \in \{-4 \text{ and } 4\}$. The effect of this can be seen in both fig. 3c and 3d[28].

C. Time shift

We shifted the audio to the left and right by a random number of seconds. When shifting the audio to the left (fast forward/negative shift) by x seconds, the initial x seconds will be considered as 0 (i.e., silence). When shifting the audio to the right (back forward/positive shift) by x seconds, the last x seconds will be considered as 0 (i.e., silence). This transformation's effect is shown in Fig. 3e and 3f.

D. Gaussian Noise

The addition of stochastic noise from a standard Gaussian N (0, 1) to each data point when it is presented to the model is a common practice, making the data point different from its original form. The hyperparameter for the noise amplitude, σ, is uniformly distributed within the range $\sigma \sim U$ (0.002)[28]. The result of this data augmentation is evident in Fig. 3b.

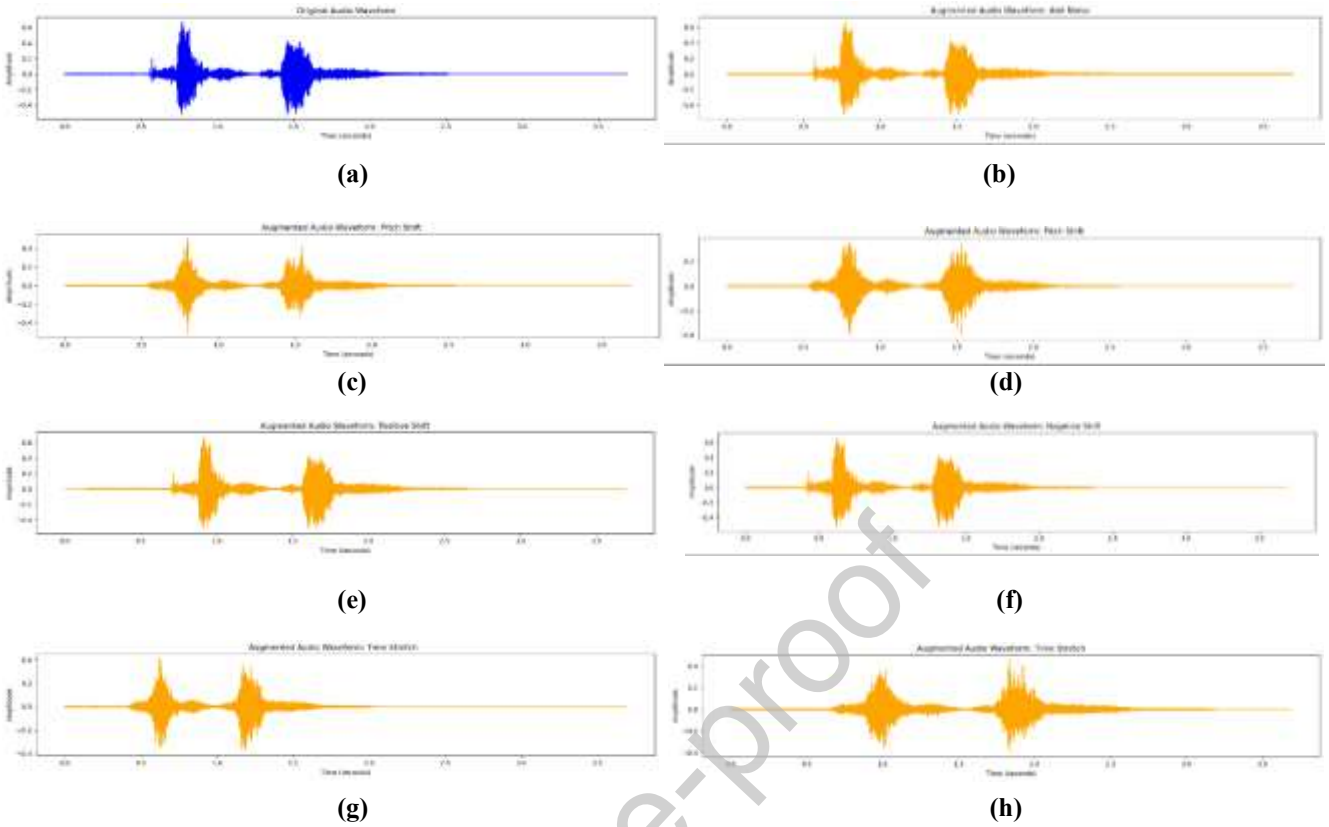


Fig. 3. Effect of Data Augmentation on Audio Waveform. (a) Original waveform, (b) Noise addition, (c) Pitch shift positive, (d) Pitch shift negative, (e) Time shift positive, (f) Time shift negative, (g) Time stretch 1, (h) Time stretch 2

2.3. Proposed Grad-CAM Extensions

A. Enhanced Directional Gradient CAM (ED-GCAM)

The Enhanced Directional Grad-CAM (ED-GCAM) improves the standard Grad-CAM method by integrating gradient direction, guided gradients, and median smoothing. It is optimised explicitly for classifying spoken keywords in aphasia. Within ED-GCAM, only the positive activations and gradients contributing to the target class prediction are preserved, effectively filtering out irrelevant information and noise. This approach ensures that the Class Activation Map (CAM) emphasises the most important areas of the input Mel-spectrogram, which is crucial for speech classification tasks involving individuals with aphasia. In such cases, subtle differences in speech patterns and phonemes are highly informative.

Furthermore, median filtering improves the resilience of the heatmap by reducing distortions while maintaining clear boundaries. This marks a significant advancement from the conventional Grad-CAM method, which frequently generates blurry and noisy heatmaps, resulting in less understandable outcomes. ED-GCAM offers more distinct insights into how the model detects and categorizes phonemic and syllabic elements in impaired speech by providing more accurate identification of crucial areas in the time-frequency domain. This increased interpretability is especially beneficial for analysing intricate speech data. It makes ED-GCAM particularly suitable for aphasia spoken keyword classification, where understanding the model's focus is essential for clinical and research purposes.

For an input image I_n, The feature maps are extracted from a convolutional layer in a deep neural network. The model's prediction score $y_c(\mathrm{I}_n)$ for the target class c is given as Eq. 1

$$y_c(\mathrm{In}) = \mathrm{Q}_c^T f(\mathrm{I}_n) \tag{1}$$

Where, $f(\mathrm{I}_n)$ is the feature representation of the input image In at layer l and Q_c is the set of weights for class c in the final fully connected layer.

The gradient of the class score y_c with respect to the feature map activations V^k is calculated as:

$$\frac{\partial y_c}{\partial V_{ij}^k} \tag{2}$$

where V_{ij}^k is the activation at spatial location *(i,j)* in the *k*-th feature map. This gradient quantifies how the activations at each spatial location affect the class score y_c .

After the gradient calculation, only positive activations and positive gradients that contribute to the class prediction are retained. The guided gradients $\widehat{G_{ij}^k}$ are computed as:

$$\widehat{G_{ij}^k} = \frac{\partial V_{ij}^k}{\partial y_c} \cdot I\left(V_{ij}^k > 0\right) \cdot I\left(\frac{\partial V_{ij}^k}{\partial y_c} > 0\right) \tag{3}$$

Where I (·) is an indicator function only considering positive activations and gradients. To ensure the gradients are scaled appropriately, they are normalised by using:

$$\widehat{G_{ij}^k} = \frac{\widehat{G_{ij}^k}}{\sqrt{\frac{1}{M}\sum_{i,j}\widehat{G^{k^2}_{ij}}+\epsilon}} \tag{4}$$

where M is the total number of spatial locations in the feature map, and ϵ is a small constant to prevent division by zero. The importance weights q^k of each feature map V^k is determined by computing a spatial average of the normalised gradients as:

$$q^k = \frac{1}{H' \times W'}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\widehat{G_{ij}^k} \tag{5}$$

Where H' and W' are the height and width of the feature map. The raw Class Activation Map (CAM_{r}) is then computed as the weighted sum of the feature maps:

$$CAM_{\mathrm{r}} = \sum_k q^k V^k \tag{6}$$

Where q^k is the weights. To reduce noise, the raw CAM is smoothed using a median filter:

$$CAM_{\mathrm{s}} = \mathrm{MedianFilter}(CAM_{\mathrm{r}}, k) \tag{7}$$

Where k is the kernel size of the median filter. The smoothed CAM is normalised and clipped to ensure values lie between 0 and 1 as

$$CAM_{\mathrm{EDG}} = \frac{CAM_{\mathrm{s}} - \min(CAM_{\mathrm{s}})}{\max(CAM_{\mathrm{s}}) - \min(CAM_{\mathrm{s}}) + \epsilon} \tag{8}$$

$$Final\ CAM_{\mathrm{EDG}} = \max(0, CAM_{\mathrm{EDG}}) \tag{9}$$

Where $Final\ CAM_{\mathrm{EDG}}$ represents the ED-GCAM, CAM_{EDG} is the normalised and CAM_{s} is the normalised smoothed CAM. After applying the normalisation and clipping step, the final heatmap CAM_{EDG} represents the ED-GCAM, highlighting the important regions for the target class prediction.

Algorithm 1 represents the proposed ED-GCAM extension steps.

ALGORITHM 1: ENHANCED DIRECTIONAL GRADIENT CAM (ED-GCAM) APPROACH STUDY

Input*: Pretrained model M, Input image* I_n*, Target class c, Set of feature maps* V^k *from convolutional layer l, kernel size k for median filter, Small constant* ϵ *to prevent division by zero*

Output*: Final Class Activation Map* CAM_{EDG}

Step 1*: Perform a forward pass to extract feature maps* V^k *and prediction score for the target class c.*

Step 2*: Compute the gradient of the target class score with respect to the feature map activations.*

Step 3*: Retain only positive activations and gradients to focus on regions contributing to the class prediction as* $\widehat{G_{ij}^k}$

Step 4*: Normalise the guided gradients to ensure appropriate scaling as* $\widehat{G_{ij}^k}$

Step 5*: Calculate the importance of each feature map by averaging the normalised gradients as* q^k

Step 6*: Compute the raw Class Activation Map (CAM) as a weighted sum of the feature maps as* CAM_{r}

Step 7*: Apply a median filter to reduce noise in the raw CAM as* CAM_{s}

Step 8*: Normalise and clip the CAM values to ensure they lie between 0 and 1 as* CAM_{EDG}

Step 9*: Return the final* CAM_{EDG}

B. Enhanced Hierarchical Filtered Grad-CAM (EH-FCAM)

The proposed Enhanced Hierarchical Filtered Grad-CAM (EH-FCAM) framework expands the Grad-CAM methodology by integrating guided gradient filtering and hierarchical feature representations. This method works very well for classifying spoken keywords in aphasia. It involves computing Class Activation Maps (CAMs) over several convolutional layers. Higher-level layers capture abstract properties, and lower-level layers that capture finer details are subsequently integrated.

The gradient information is filtered at each layer to guarantee that only positive activations and positive gradients are considered in the final heatmap. By decreasing the influence of irrelevant areas, this filtering increases the focus on regions that positively contribute to the target class prediction. To ensure that both broad, abstract patterns and localised, fine-grained features are represented in the final visualisation, the CAMs are then hierarchically integrated using element-wise maximum procedures.

Input image X is given to the CNN model M. For each convolutional layer l, the filtered Grad-CAM for the predicted class c is computed as follows:

Let $A^l \in R^{h_l \times w_l \times d_l}$ be the output feature map of the convolutional layer l, where h_l, w_l, and d_l are the feature map's height, width, and number of channels, respectively. Let y^c be the score of the predicted class c. The gradient of y^c with respect to the feature map A^l is:

$$g_{ijk}^l = \frac{\partial y^c}{\partial A_{ijk}^l} \tag{10}$$

where i *and* j are spatial indices, and k is the channel index. Guided gradients $\widehat{g_{ijk}^l}$ will be applied, which are to retain only positive activations and gradients:

$$\widehat{g_{ijk}^l} = \mathrm{ReLU}\left(A_{ijk}^l\right) \cdot \mathrm{ReLU}\left(\frac{\partial y^c}{\partial A_{ijk}^l}\right) \tag{11}$$

This ensures that only positive gradients and activations contribute to the final Grad-CAM heatmap. The importance (weight) α_k^l for each feature map channel k, is computed by averaging the guided gradients spatially:

$$\alpha_k^l = \frac{1}{h_l \times w_l} \sum_{i=1}^{h_l} \sum_{j=1}^{w_l} \widehat{g_{ijk}^l} \tag{12}$$

The Grad-CAM heatmap L_{ij}^l for layer l is a weighted combination of the feature maps, followed by ReLU activation to retain only positive values:

$$L_{ij}^l = \mathrm{ReLU}\left(\sum_{k=1}^{d_l} \alpha_k^l A_{ijk}^l\right) \tag{13}$$

This results in a heatmap $L^l \in R^{h_l \times w_l}$.

The heatmap L^l is resized to match the size of the input image as height H and width W:

$$\widetilde{L^l} = \mathrm{Resize}(L^l, H, W) \tag{14}$$

The key idea of hierarchical integration is combining heatmaps from multiple layers, starting from the deepest layer and progressively combining them with shallower layers $L = \{l_1, l_2, \ldots, l_n\}$. Let the layers be indexed from the deepest l_n to the shallowest. l_1.

The combined heatmap CAM_{combined} initially starts with the deepest layer as

$$CAM_{\text{combined}} = \widetilde{L^{l_n}} \tag{15}$$

Where $\widetilde{L^{l_n}}$ represent a resized version of $\widetilde{L^l}$. For each layer l=*n-1,n-2,…,*l_1, the current combined heatmap CAM_{combined} is resized to match the shape of the next layer's heatmap $\widetilde{L^l}$. Then, the heatmap from layer l is combined with the resized combined heatmap using the element-wise maximum as

$$CAM_{\text{combined}} = \max\left(\mathrm{Resize}(CAM_{\text{combined}}, h_l, w_l), \widetilde{L^{l_n}}\right) \tag{16}$$

This process ensures that the fine details from lower layers and the abstract features from deeper layers are combined hierarchically.

After combining the heatmaps from all the layers, the final combined heatmap CAM_{EHF} is normalised to the range [0,1] for visualisation as

$$CAM_{\text{EHF}} = \frac{\max(CAM_{\text{combined}}) - \min(CAM_{\text{combined}})}{CAM_{\text{combined}} - \min(CAM_{\text{combined}})} \quad (17)$$

where, CAM_{EHF} represents EH-FCAM.

This hierarchical technique offers substantial benefits over the conventional Grad-CAM for aphasia spoken keyword classification, requiring recording minor phonetic and auditory variations. Individuals suffering from aphasia frequently experience complex distortions in their speech that are difficult for a single layer of neural representations to capture fully. The suggested approach can emphasise complex speech patterns and the subtle aspects of phoneme articulation by merging data from several layers. These features are essential for comprehending and categorising speech disorders. Moreover, guided gradient filtering strengthens the method's ability to analyse speech impairment by minimising noise and improving the CAMs' interpretability by highlighting the most important speech characteristics. The interpretability and performance of the classification task are enhanced by this more-centred approach, which makes it possible to identify the critical input regions that contribute to model predictions more precisely.

Algorithm 2 represents the proposed EH-FCAM extension steps.

ALGORITHM 2: ENHANCED HIERARCHICAL FILTERED GRAD-CAM (EH-FCAM) EXTENSION STUDY

Input*: Pretrained model M, Input image X, Predicted class c, Set of convolutional layer* $L = \{l_1, l_2, \dots, l_n\}$*, Image size height H and width W*

Output*:* CAM_{EHF}

Step 1*: Initialize and Forward Pass Through the Model*

Feed the input image X through the model M and perform a forward pass to obtain the predictions for the target class c.

Extract the predicted class score y^c *from the model's output.*

Step 2: *Compute Grad-CAM for Each Convolutional Layer*

for *each convolutional layer* $L_l \in \{L_1, L_2, \dots, L_n\}$ ***do***

Compute the Gradients, g_{ijk}^{l}

Apply Guided Gradient Filtering, $\widehat{g_{ijk}^{l}}$

Compute the Importance Weights, α_k^l

Generate the Grad-CAM Heatmap, L_{ij}^{l}

Resize the Heatmap, $\widetilde{L^l}$

end for

Step 3*: Hierarchical Combination of Heatmaps*

Initialise the combined heatmap with the heatmap from the deepest layer L_n

for *each subsequent layer l=n-1,n-2,…,* l_1 ***do***

Resize CAM_{combined} *to match the spatial dimensions of* $\widetilde{L^{l_n}}$

end for

Step 4*: Normalise the Final Combined Heatmap*

Normalise the final combined heatmap CAM_{combined} *to the range [0,1],* CAM_{EHF}

C. Stochastic Gradient-Dropout Integrated Grad-CAM (SGD-GCAM)

The Stochastic Gradient-Dropout Integrated Grad-CAM (SGD-GCAM) is an enhanced extension of the traditional Grad-CAM framework, designed to improve robustness and interpretability in spoken keyword classification tasks, particularly for aphasia speech analysis. By incorporating dropout during inference and averaging Grad-CAM results over multiple forward passes, the approach captures the inherent variability in feature importance induced by stochastic neuron activations. This technique results in more reliable and consistent heatmaps, addressing the challenges of speech variability present in aphasia, where phoneme distortions and omissions can lead to inconsistent acoustic representations. The averaged Grad-CAM maps highlight the most robust features that consistently contribute to classification decisions, mitigating the effects of variability caused by impaired speech production.

Moreover, SGD-GCAM introduces uncertainty quantification by calculating the variance across the Grad-CAM heatmaps generated during the multiple stochastic forward passes. This variance map identifies regions in the Mel-spectrogram where the model's predictions exhibit uncertainty, offering additional insights into the reliability of the feature attributions. Such uncertainty estimation is critical in aphasia, where speech impairments lead to varying levels of predictability in speech patterns. The dual benefit of robust feature identification and uncertainty quantification makes SGD-GCAM a valuable tool for clinicians and researchers, providing more reliable and interpretable model explanations in clinical applications of speech impairment analysis. This enhanced interpretability is especially useful in high-stakes environments where understanding the decision-making process of AI models is critical.

For each layer L and target class c, the gradients of the class score y^c with respect to the feature maps $\mathrm{A}_{\mathrm{ij}}^{\mathrm{k}}$ are computed. However, this extension modifies the standard Grad-CAM by focusing only on positive gradients and normalising them to improve stability. The guided directional gradient is computed as

$$\text{guided\_grads}_{ij}^{k} = \mathbb{1}\left(A_{ij}^{k} > 0\right) \cdot \mathbb{1}\left(\frac{\partial A_{ij}^{k}}{\partial y^{c}} > 0\right) \cdot \frac{\partial A_{ij}^{k}}{\partial y^{c}} \tag{18}$$

Where $\frac{\partial A_{ij}^{k}}{\partial y^{c}}$ is the gradient of the output score for the activation $\mathrm{A}_{\mathrm{ij}}^{\mathrm{k}}$. These gradients are normalised to avoid large values dominating the Grad-CAM as

$$\text{guided\_grads}_{ij}^{k} = \frac{\text{guided\_grads}_{ij}^{k}}{\sqrt{\text{mean}(\text{guided\_grads}^{2}) + \epsilon}} \tag{19}$$

For each feature map k, the importance weights α_k are computed by averaging the normalised guided gradients across the spatial dimensions:

$$\alpha_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \text{guided\_grads}_{ij}^{k} \tag{20}$$

where H is the height, and W is the width of the feature map.

The Grad-CAM heatmap $L_{\text{Grad-CAM}}^{c}$ is computed as a weighted sum of the feature maps, followed by a ReLU to ensure only positive contributions as

$$L_{\text{Grad-CAM}}^{c} = \text{ReLU}(\sum_{k} \alpha_k A^{k}) \tag{21}$$

Where α_k is the importance weights of the activation A^k for the k-th feature map, and c is the target class. Dropout is enabled during inference to introduce stochasticity, resulting in different feature maps A^k and gradients for each forward pass. For each inference run t, a different Grad-CAM heatmap $L_{\text{Grad-CAM}}^{c}(t)$ is computed. After performing T such runs, the stochastic Grad-CAM is calculated as the average of these heatmaps using

$$L_{\text{SGD-GCAM}}^{c} = \frac{1}{T} \sum_{t=1}^{T} L_{\text{Grad-CAM}}^{c}(t) \tag{22}$$

To quantify uncertainty, the variance of the Grad-CAM heatmaps across the T runs is computed as

$$\sigma^{2}(L_{\text{Grad-CAM}}^{c}) = \frac{1}{T} \sum_{t=1}^{T} (L_{\text{Grad-CAM}}^{c}(t) - L_{\text{SGD-GCAM}}^{c})^{2} \tag{23}$$

where, $L_{\text{SGD-GCAM}}^{c}$ the averaged heatmap is more robust and smoother, capturing consistent features across multiple dropout-enabled runs. Variance heatmap $\sigma^{2}(L_{\text{Grad-CAM}}^{c})$, which provides insight into the uncertainty in feature importance. Compute the final combined CAM by averaging the averaged CAMs across all layers. $L_1, L_2, \ldots, L_n$ as

$$L_{\text{final}} = \frac{1}{n} \sum_{i=1}^{n} L_{\text{SGD-GCAM}}^{c}(L_i) \tag{24}$$

Algorithm 3 represents the proposed SGD-GCAM extension steps.

ALGORITHM 3: STOCHASTIC GRADIENT-DROPOUT INTEGRATED GRAD-CAM (SGD-GCAM)

Input*: Pretrained model M, Input image I, Target class c, convolutional layers* $L = L_1, L_2, \ldots, L_n$*, number of stochastic runs*

Output*: Final combined Grad-CAM heatmap* $L_{\text{SGD-GCAM}}^{c}$ *and variance heatmap* $\sigma^{2}(L_{\text{Grad-CAM}}^{c})$.

Step 1*: For each layer in the model MMM, ensure that dropout layers remain active during inference to introduce stochastic behaviour in the activations.*

***Step 2**: Initialize Storage for Grad-CAMs*
Step 3**: **for** t = 1 to T, **do
* **for** each convolutional layer L_i, **do***
* Compute the Grad-CAM heatmap $L^c_{\text{Grad-CAM}}$ for the i-th layer*
* Store the Grad-CAM heatmap $L^c_{\text{Grad-CAM}}(t)$ for each layer*
* **end for***
* **end for***
Step 4**: **For** each layer L_i, **do
* Compute the averaged Grad-CAM across all runs, $L^c_{\text{SGD-GCAM}}$*
* Compute the variance heatmap for uncertainty quantification, $\sigma^2(L^c_{Grad\text{-}CAM})$*
* **end for***
***Step 5**: Compute Final Combined CAM, L_{final}*
***Step 6**: Output the Combined and Variance Heatmaps*

D. Multi-Scale Channel-Wise Grad-CAM (MSCW-GCAM)

The Multi-Scale Channel-Wise Grad-CAM (MSCW-GCAM), a new extension of the standard Grad-CAM framework, aims to improve the interpretability of deep learning models by merging multi-scale analysis and channel-wise heatmap generation. This approach overcomes the limitations of traditional Grad-CAM by producing Class Activation Maps (CAMs) at various spatial resolutions, capturing fine details and broader patterns necessary for complex data representations. Furthermore, MSCW-GCAM decomposes the input data channel by channel, generating independent CAMs for each channel, such as RGB channels in images or frequency bands in Mel-spectrograms. These CAMs are combined to produce a more comprehensive and interpretable heatmap, revealing the relative importance of different channels and scales to the model's prediction.

This method is well suited for classifying spoken keywords in aphasia due to the complex nature of speech patterns and the variability in time-frequency representations found in Mel-spectrograms. Aphasia, a condition affecting speech production, often causes subtle changes in speech characteristics, occurring at different temporal and spectral resolutions. By examining Mel-spectrograms at multiple scales, MSCW-GCAM ensures that it captures high-frequency phonetic details and broader speech patterns at lower frequencies, which is crucial for accurately understanding speech from individuals with aphasia. Additionally, analysing individual channels enables the model to focus on specific frequency bands that may significantly distinguish aphasic speech, thereby improving the model's ability to identify relevant speech features across different frequency bands.

Let $I \in R^{H\times W\times C}$ represent the input image (Mel-spectrogram) where H, W, and C are the height, width, and number of channels, respectively. The model predicts a score y_c for class *c* after passing I through the convolutional layers. For each channel $k \in \{1, 2..., C\}$ in the input I, we extract a single-channel image $I_k \in R^{H\times W}$ by setting all other channels to zero.

Now, define the output of a convolutional layer for the input channel I_k as $A_{k,l}(x,y) \in R^{H_l\times W_l\times D_l}$, where $A_{k,l}$ is the activation map for layer l, with dimensions $H_l \times W_l$, and D_l is the number of filters at that layer. The Grad-CAM heatmap for channel *k* can be expressed as

$$L_{c,k}(x,y) = \sum_{d=1}^{D_l} \alpha_{c,k,d} A^d_{k,l}(x,y) \quad (25)$$

Where $A^d_{k,l}(x,y)$ is the activation of the *d*-th filter at spatial location *(x, y)* and $\alpha_{c,k,d}$ is the importance weight for the *d*-th filter with respect to class *c*. The importance weight $\alpha_{c,k,d}$ is computed by taking the global average of the gradients as

$$\alpha_{c,k,d} = \frac{1}{H_l W_l} \sum_{i=1}^{H_l} \sum_{j=1}^{W_l} \frac{\partial y_c}{\partial A^d_{k,l}(i,j)} \quad (26)$$

Where $\frac{\partial y_c}{\partial A^d_{k,l}(i,j)}$ is the gradient of the class score y_c with respect to the activation map $A^d_{k,l}$.

For each channel I_k, the input is resized to different scales $s \in S = \{s_1, s_2..., s_n\}$, where s_i denotes a scaling factor. At each scale s_i, the input image is resized as

$$I_k^{(s_i)} = \text{Resize}(I_k, s_i) \quad (27)$$

And the corresponding Grad-CAM heatmap for that scale is as in

$$L_{c,k}^{(s_i)}(x,y) = \sum_{d=1}^{D_l} \alpha_{c,k,d}^{(s_i)} A_{k,l}^{(s_i),d}(x,y) \quad (28)$$

Where $\alpha_{c,k,d}^{(s_i)}$ are the important weights for the scale s_i.

For each scale s_i, the channel-wise Grad-CAMs are averaged to form a combined heatmap as

$$L_c^{(s_i)}(x,y) = \frac{1}{C}\sum_{k=1}^{C} L_{c,k}^{(s_i)}(x,y) \quad (29)$$

This captures the importance of all channels at the given scale s_i. Finally, the heatmaps from different scales are combined to produce the final multi-scale CAM as

$$L_c(x,y) = \frac{1}{n}\sum_{i=1}^{n} L_c^{(s_i)}(x,y) \quad (30)$$

This combined heatmap $L_c(x,y)$ represents the final MSCW-GCAM for class c, capturing important features across multiple scales and channels.

In the realm of aphasia spoken keyword classification, speech patterns are intricate and showcase significant characteristics at different time and frequency levels. Mel-spectrograms, representing time and frequency, inherently encompass information at multiple resolutions. The multi-scale analysis in MSCW-GCAM captures detailed phonetic nuances and broader speech patterns by processing the Mel-spectrogram at varying resolutions. Furthermore, the channel-wise decomposition ensures that the impacts of specific frequency ranges are assessed independently, which is especially relevant when certain frequency bands convey more crucial information in aphasia speech. The combined multi-scale and multi-channel approach allows for a more detailed understanding of how the model identifies speech impairments, enhancing the interpretability of model predictions in this demanding field.

Algorithm 4 represents the proposed MSCW-GCAM extension steps.

ALGORITHM 4: MULTI-SCALE CHANNEL-WISE GRAD-CAM (MSCW-GCAM) APPROACH STUDY

***Input**: Input Mel-spectrogram image with height H, width W, and C channels $I \in R^{H\times W\times C}$, Trained model M, Convolution layer L, Target class c, Set of scaling factors $s \in S = \{s_1, s_2..., s_n\}$*

***Output**: Final Multi-Scale Channel-Wise Grad-CAM heatmap for class c, capturing features across scales and channels.*

***Step 1**: Extract individual channels $I \in R^{H\times W\times C}$ for each channel $k \in \{1, 2..., C\}$*

for** each channel k, **do

create I_k by setting all other channels in I to zero

end for

Step 2**: **for** each scale $s_i \in S$, **do

resize each channel-specific input I_k by scaling factor s_i

end for

Step 3**: **for** each channel $k \in \{1, 2..., C\}$ and each scale $s_i \in S$, **do

Perform a forward pass through the model M to obtain the activations $A_{k,l}^{(s_i)}$ at layer L and the class score y_c for class c

end for

Compute the gradient of the class score y_c with respect to the activation map $A_{k,l}^{(s_i)}$ for each filter $d \in \{1, 2, ...D_l\}$

***Step 4**: Compute importance weights $\alpha_{c,k,d}^{(s_i)}$ for each filter d at scale s_i by taking the global average of the gradients*

Step 5**: **for** each channel k and scale s_i, **do

compute the Grad-CAM heatmap $L_{c,k}^{(s_i)}(x,y)$

end for

***Step 6**: Aggregate Grad-CAMs across channels for each scale s_i by averaging the heatmaps over all channels k*

***Step 7**: Aggregate Grad-CAMs across scales by averaging the heatmaps across all scales s_i*

***Step 8**: Return the final Multi-Scale Channel-Wise Grad-CAM heatmap $L_c(x,y)$ for class c, the combined importance is represented across all scales and channels.*

In the context of aphasia spoken keyword classification, speech patterns are complex and exhibit important features at different temporal and frequency resolutions. Mel-spectrograms, which are time-frequency representations, inherently contain multi-resolution information. The multi-scale analysis in MSCW-GCAM

captures fine-grained phonetic details and broader speech patterns by processing the Mel-spectrogram at various resolutions. Additionally, the channel-wise decomposition ensures that the contributions of specific frequency bands are independently evaluated, which is particularly relevant when certain frequency bands carry more diagnostic information in aphasic speech. This combined multi-scale and multi-channel approach enables a more nuanced interpretation of how the model detects speech impairments, improving the interpretability of model predictions in this challenging domain.

3. Results

The models discussed in Section 2.1 were compiled using the categorical cross-entropy loss function to classify 15 different classes. The Adam optimizer was utilized for the research. The deep learning models were validated, pre-trained with the Google Speech Command dataset and subsequently fine-tuned using the aphasia dataset. To prevent overfitting, k-fold cross-validation and data augmentation techniques were employed. The experiments were conducted on a system with a 3.40 GHz Intel® Xeon(R) E-2236 CPU, and an NVIDIA Corporation TU10104GL [Quadro RTX 4000] 2.3 TB.

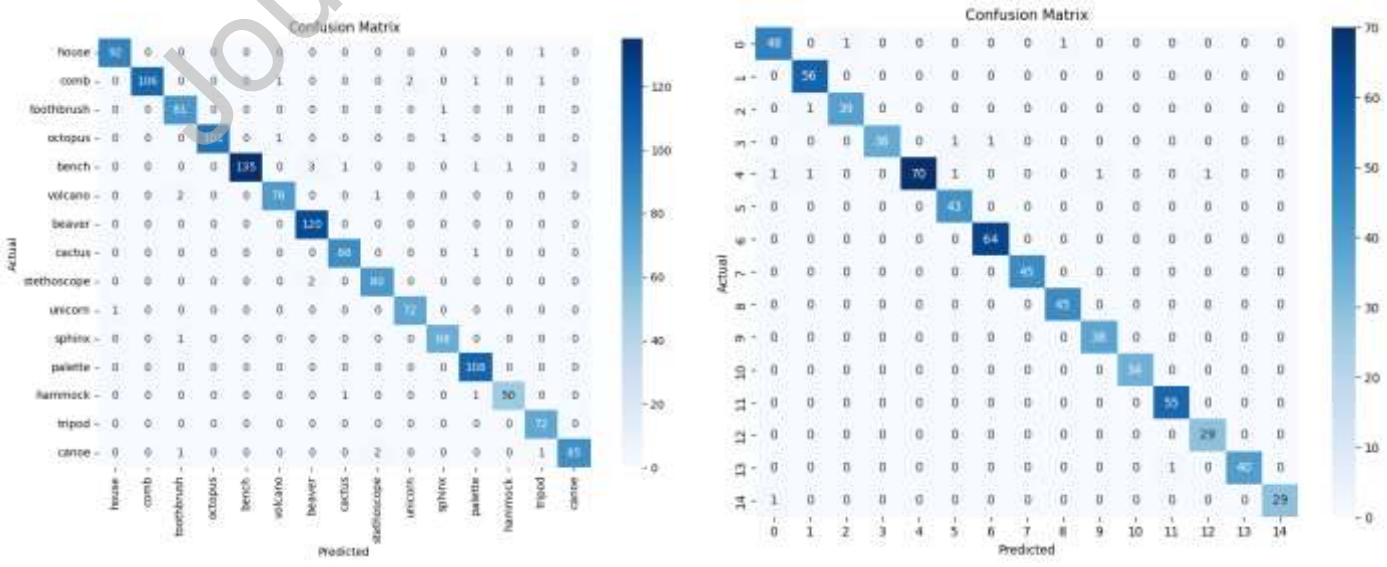
The evaluation results of spoken keyword classification models using with and without data augmentation are shown in Table 2 and Table 1, respectively. The confusion Matrices of the models are shown in Fig. 4.

TABLE 1
Evaluation Metrics of the Classification Models without Augmentation

Model	Accuracy	Precision	Recall	F1-Score
Single attention-based CNN	80.26	83.27	80.26	80.88
Multi-attention-based parallel CNN	83.55	83.78	83.55	83.26

TABLE 2
Evaluation Metrics of the Classification Models with Augmentation

Model	Accuracy	Precision	Recall	F1-Score	MCC	ROC	PRC
Single attention-based CNN	97.73	97.77	97.73	97.73	0.9756	0.9992	0.9935
Multi-attention-based parallel CNN	98.24	98.28	98.24	98.23	0.9827	0.9997	0.9980



(a) (b)

Fig. 4. Confusion matrix of (a) Single attention-based CNN (b) Multi-attention-based parallel CNN model.

After the data augmentation, the model accuracy increased significantly: For single-attention-based CNN, accuracy improved from 80.26% to 97.73%, and in the case of multiple-attention-based parallel CNN, accuracy improved from 83.55% to 98.24%. These improvements show that augmentation reduced overfitting by increasing the dataset size and introducing variations that improved both models' ability to generalize to test data that had not yet been encountered. The models were better able to manage loud and damaged speech because of the varied, enhanced dataset, which is especially difficult when it comes to aphasia speech. We validated our proposed extension methods with traditional Grad-CAM by using different metrics, and we also included a LIME explanation map for a comprehensive comparison as well.

3.1. Quantitative Evaluation

Three different quantitative variables were adopted in this study to evaluate our approach with the traditional Grad-CAM method. We provide a table summarising the important quantitative metrics, Perturbation Score (Pt), Infidelity Score (Is), and Sufficiency Score (S_s), to assess the efficacy of the suggested modifications. The indicators are integrated to produce a trustworthy and comprehensible summary of the model's overall performance. This facilitates efficient results sharing and makes comparing and evaluating models across all categories more accessible. These measures provide a comprehensive view of our method's performance compared to the traditional Grad-CAM.

A. Perturbation-based trustworthiness

The perturbation Score is one of the important metrics for evaluating the accuracy and reliability of explanations generated by methods like Grad-CAM [29]. The perturbation-based trustworthiness (P_t) is defined as

$$P_t(Z) = P(y \mid I) - P(y \mid I') \tag{31}$$

where $P(y \mid I)$ represents the model's confidence before perturbation and $P(y \mid I')$ denotes the confidence after perturbing the identified regions *Z*. Higher $P_t(Z)$ values support the validity of the CAM's explanation by indicating that the perturbed regions are, in fact, important to the model's conclusion.

B. Infidelity Score

The Infidelity Score (I_s) is calculated as (18)

$$\text{Infidelity} = E_\delta\left[\left(f(x) - f(x+\delta)\right) \cdot \phi(x, x+\delta)\right] \tag{32}$$

where $\phi(, x+\delta)$ indicates the variation in the explanation before and after the perturbation, and $f(x)$, is the model output given the input x, δ as a minor perturbation applied [30]. A lower score indicates better alignment. It gauges how well the explanation matches the model's sensitivity to input changes.

C. Sufficiency Score

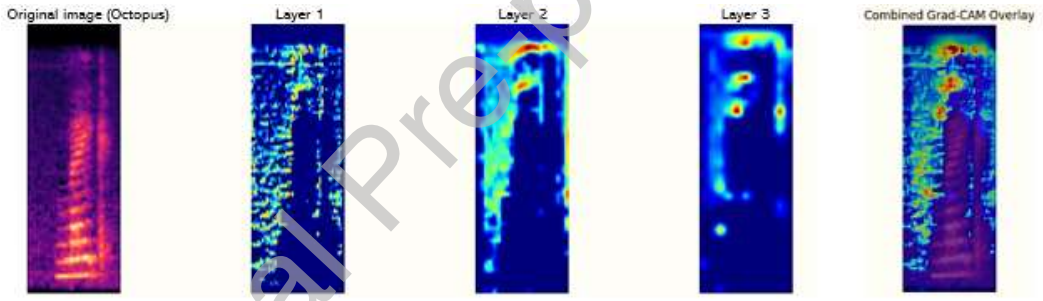
The sufficiency score (S_s) evaluates the faithfulness of the explanation approach. [31]. It evaluates how much important information a Class Activation Map (CAM) identifies to maintain the model's prediction confidence for a given class. It measures how sufficient the highlighted regions are for preserving the original prediction when the rest of the input is masked or perturbed.

$$S_s = \frac{OC}{OC - C_{IR}} \times 100 \tag{33}$$

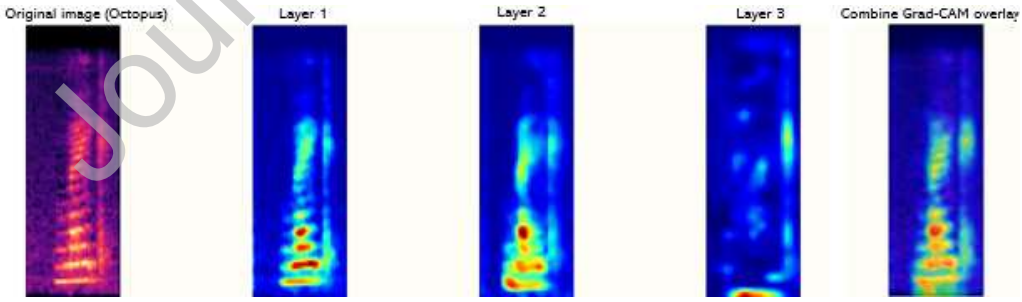
where *OC,* original confidence, is the model's confidence score for the target class when given the full input image. C_{IR} is the model's confidence score for the target class when only the important regions identified by the CAM are retained, and the rest of the input is masked or perturbed. Table 3 represents the quantitative metrics comparison scores with Grad-CAM variants.

Table 3
Results of Quantitative Analysis

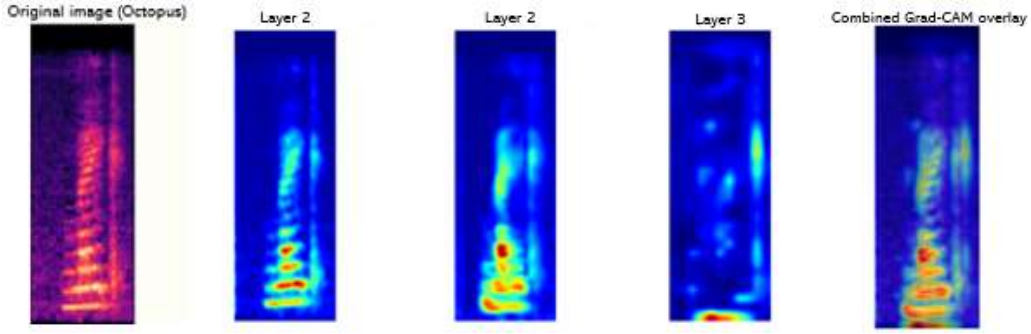
Approach		Metrics		
		Pt↑	I_s↓	S_s↑
Single attention-based CNN	Grad-CAM	.53±.04	3.56	77.19±.03
	ED-GCAM	.88±.04	0.79	82.50±.04
	EH-FCAM	.84±.04	0.85	80±.04
	SGD-GCAM	.83±.04	1.90	79.36±.04
	MSCW-GCAM	.79±.04	1.25	77.31±.04
Multi-attention-based parallel CNN	Grad-CAM	.62±.03	1.93	71±.04
	ED-GCAM	.87±.03	0.94	79.02±.04
	EH-FCAM	.85±.03	1.52	73.22±.03
	SGD-GCAM	.82±.03	1.07	72.02±0.4
	MSCW-GCAM	.77±.03	1.58	71±0.4



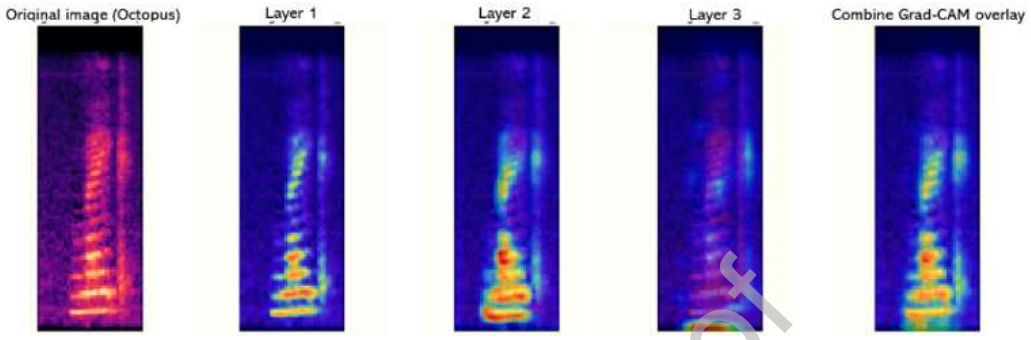
(a)



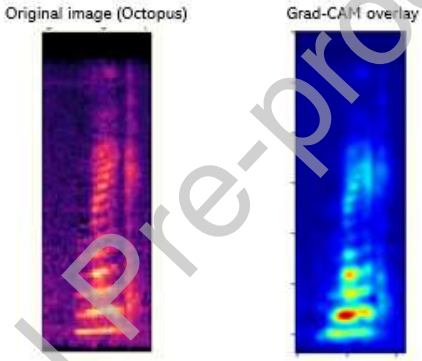
(b)



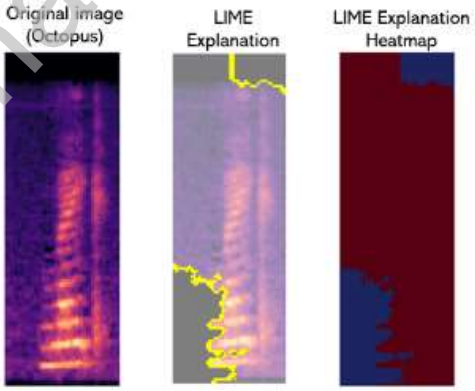
(c)



(d)

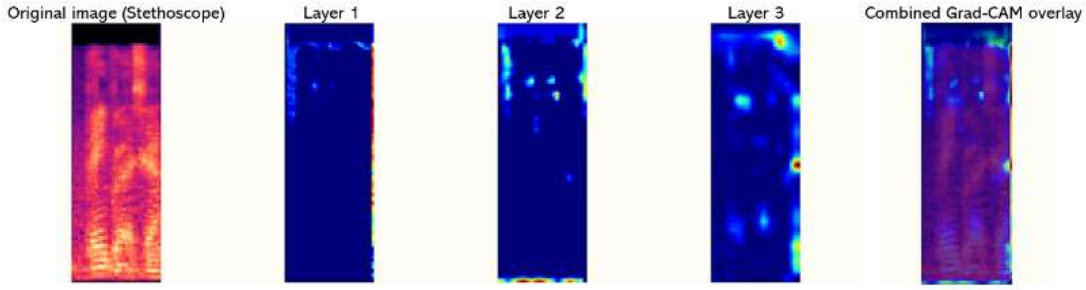


(e)

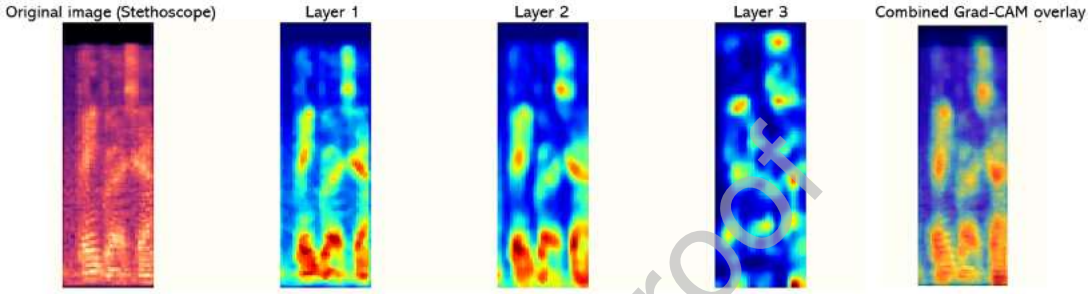


(f)

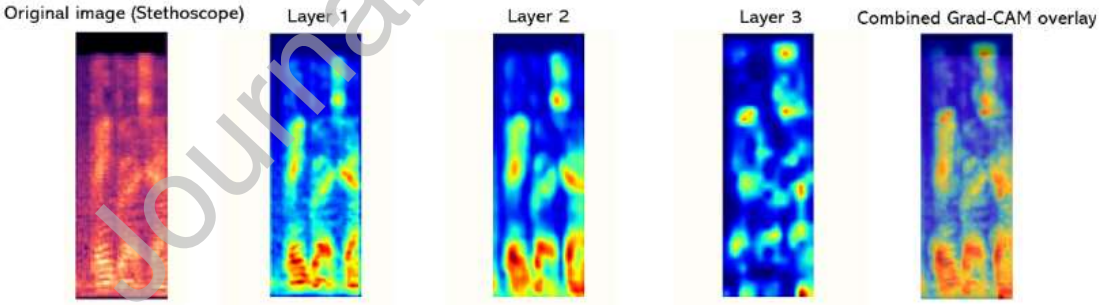
Fig.5. Grad-CAM results for the keyword 'Octopus' from Single attention-based CNN model (a) Traditional Grad-CAM (b) ED-GCAM (c) EH-FCAM (d) SGD-GCAM (e) MSCW-GCAM (f) LIME



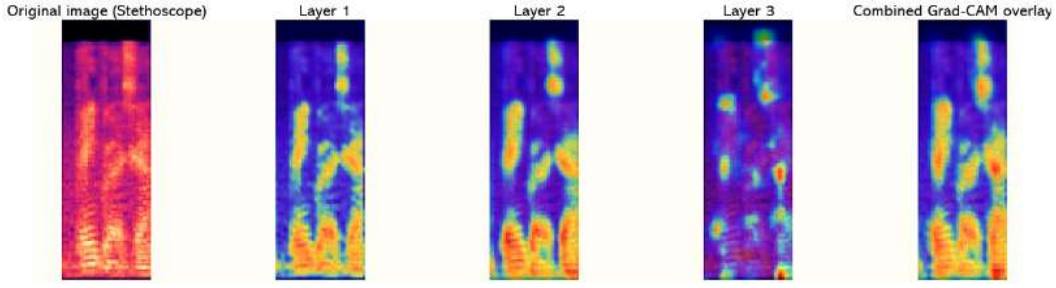
(a)



(b)



(c)



(d)

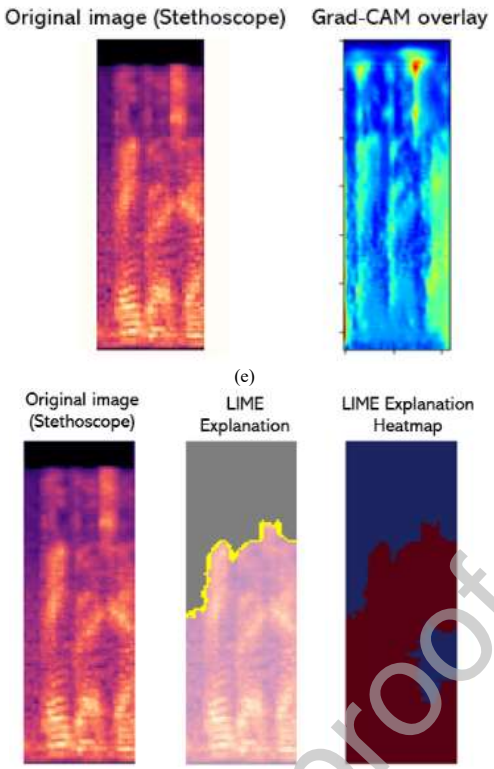
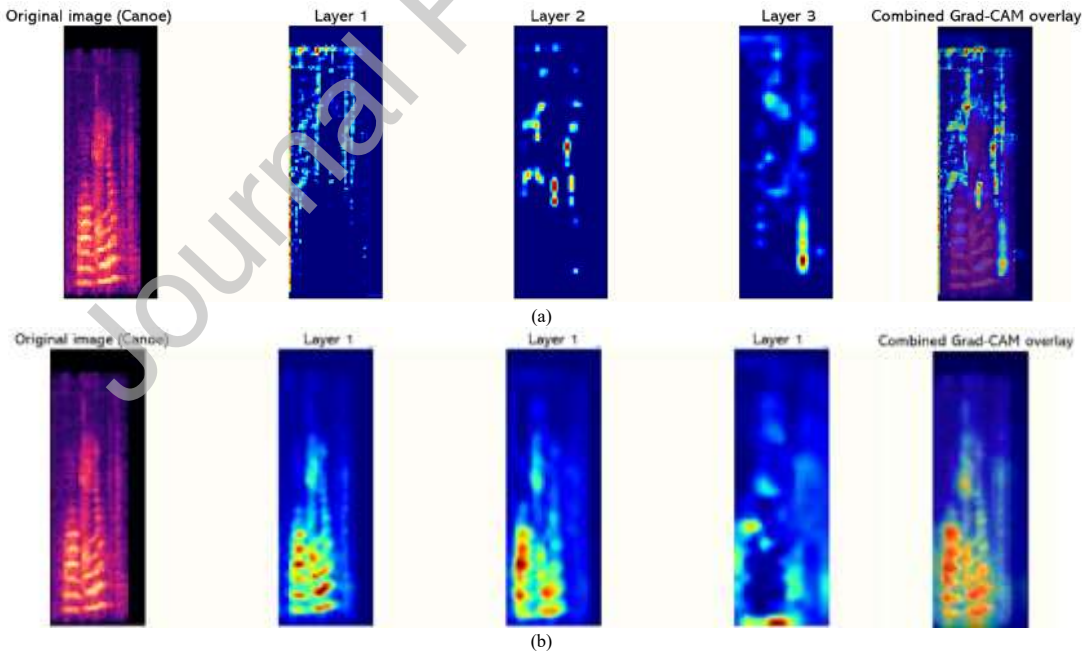
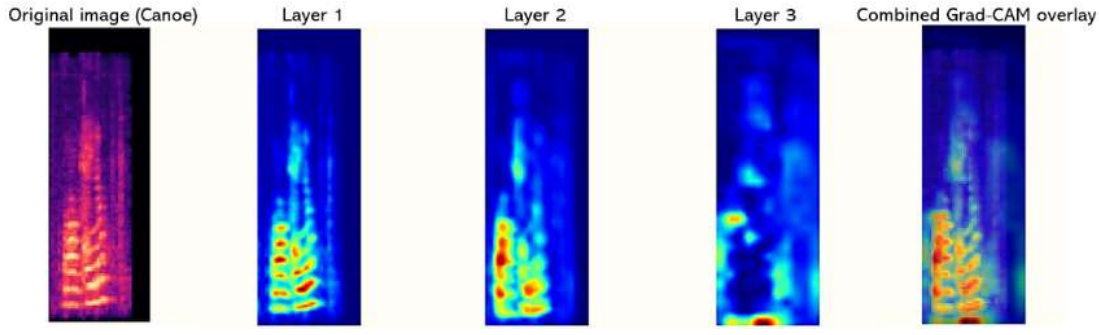
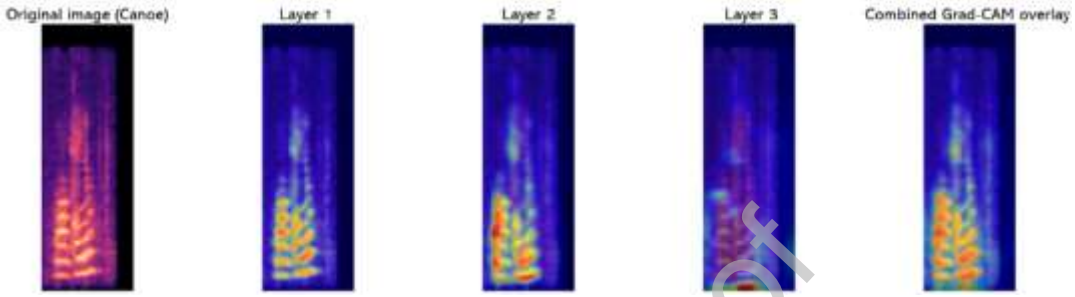


Fig. 6. Grad-CAM results for the keyword 'Stethoscope' from Single attention-based CNN model (a) Traditional Grad-CAM (b) ED-GCAM (c) EH-FCAM (d) SGD-GCAM (e) MSCW-GCAM (f) LIME

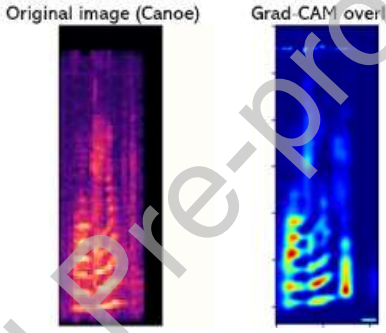




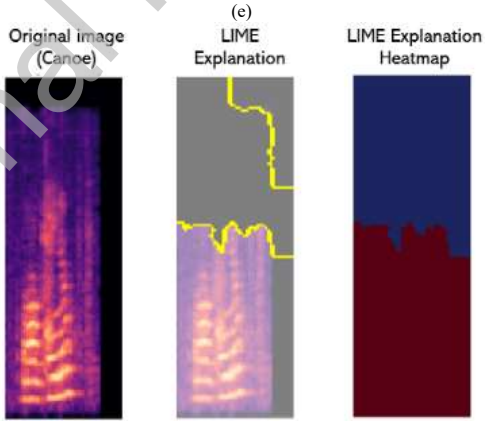
(c)



(d)

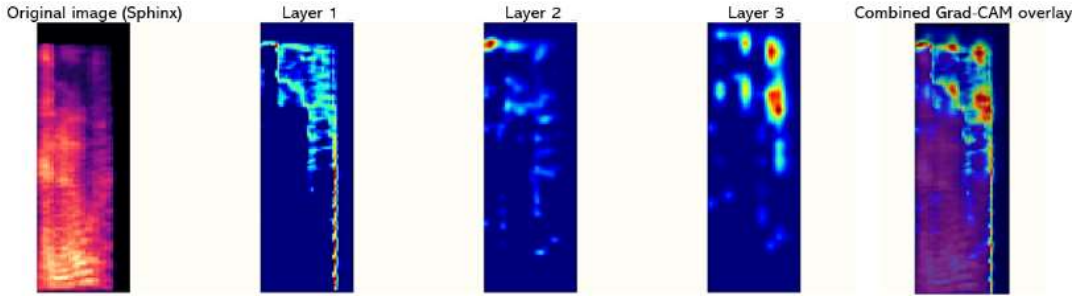


(e)

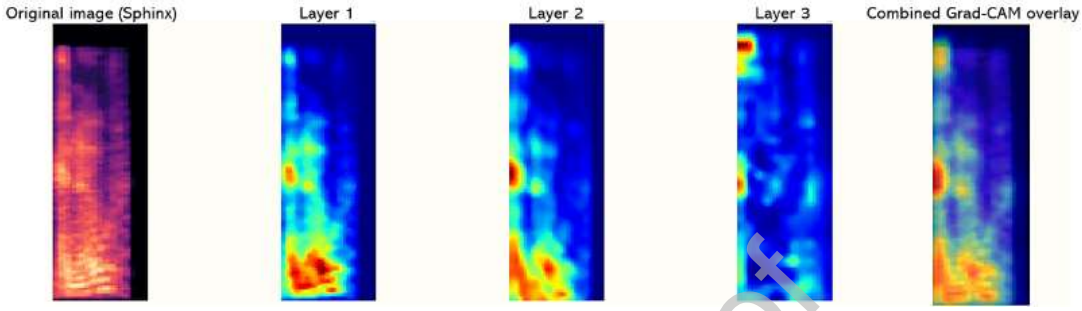


(f)

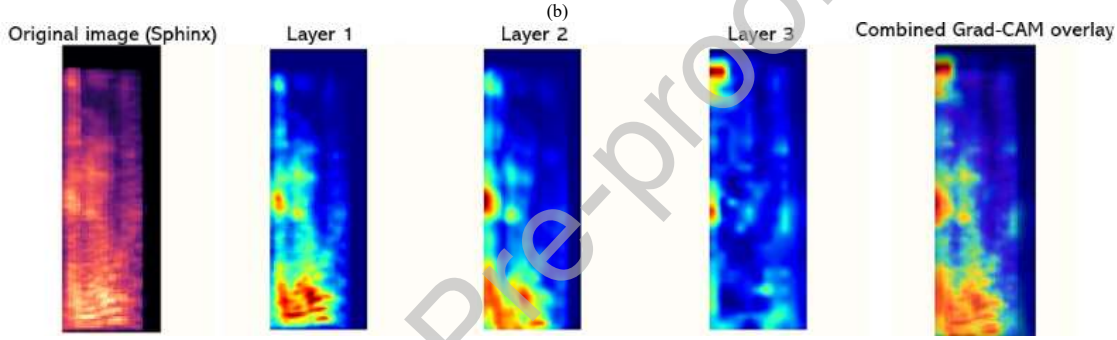
Fig. 7. Grad-CAM results for the keyword 'Canoe' from Single attention-based CNN model (a) Traditional Grad-CAM (b) ED-GCAM (c) EH-FCAM (d) SGD-GCAM (e) MSCW-GCAM (f) LIME



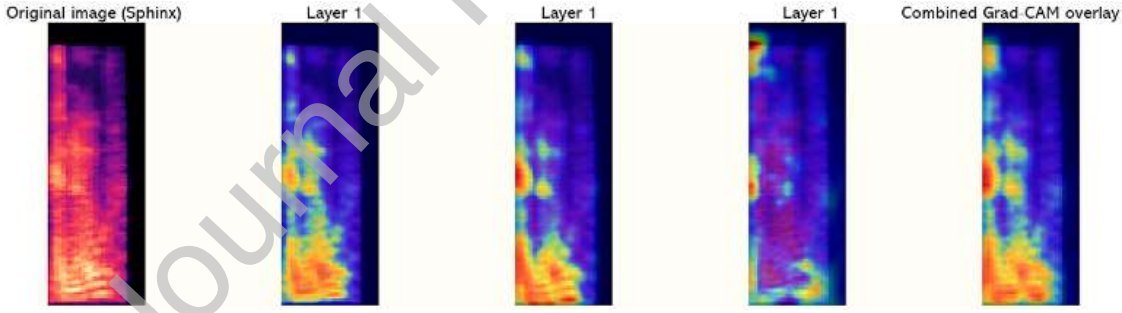
(a)



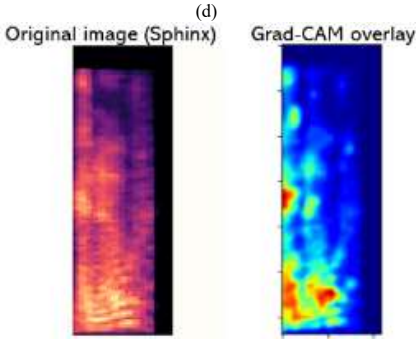
(b)



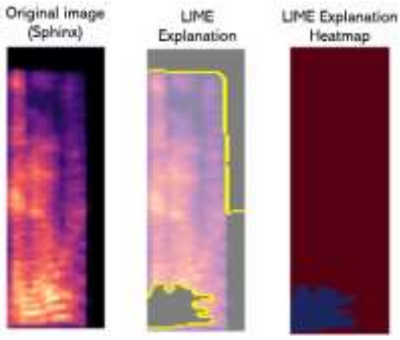
(c)



(d)

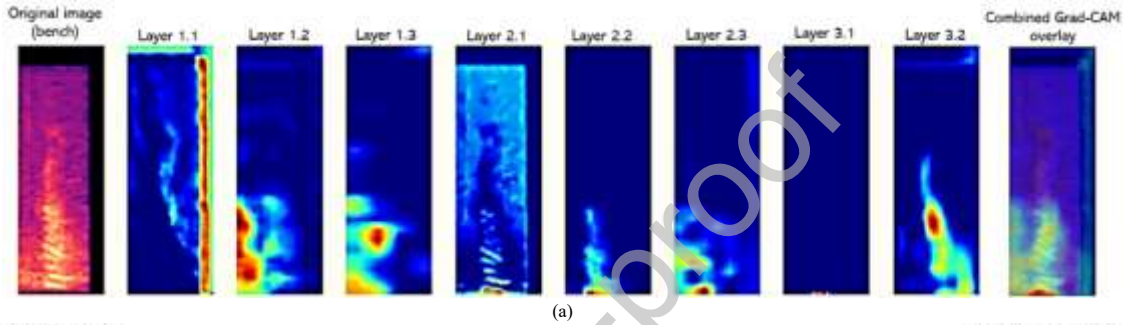


(e)

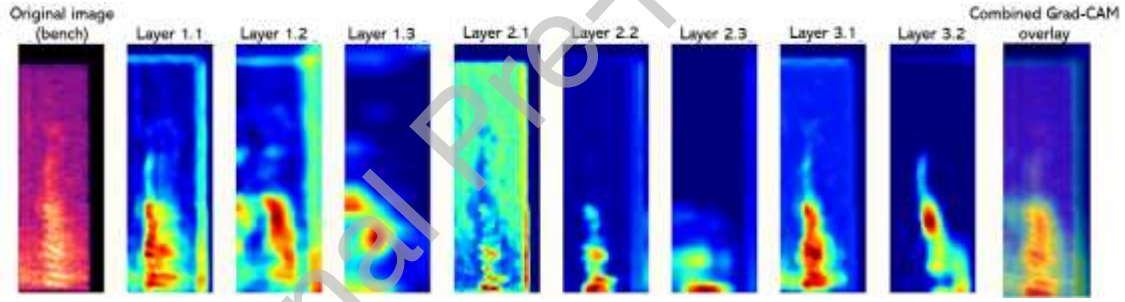


(f)

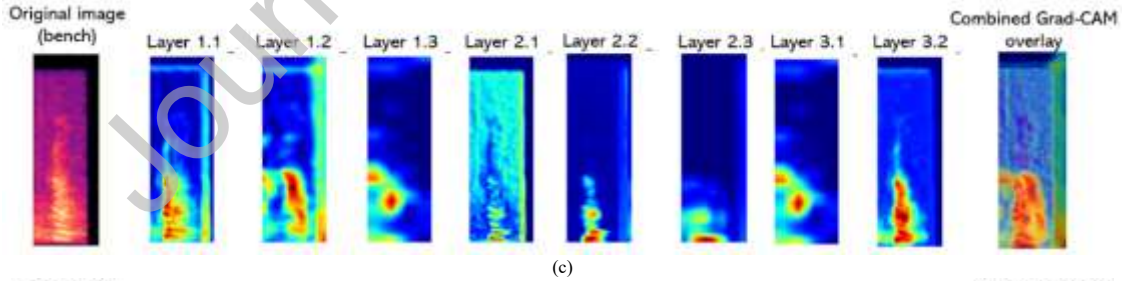
Fig.8. Grad-CAM results for the keyword 'Sphinx' from Single attention-based CNN model (a) Traditional Grad-CAM (b) ED-GCAM (c) EH-FCAM (d) SGD-GCAM (e) MSCW-GCAM (f) LIME



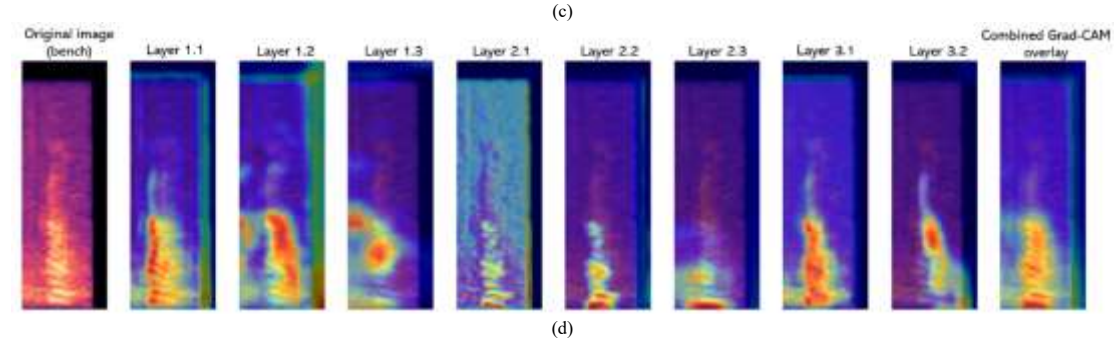
(a)



(b)



(c)



(d)

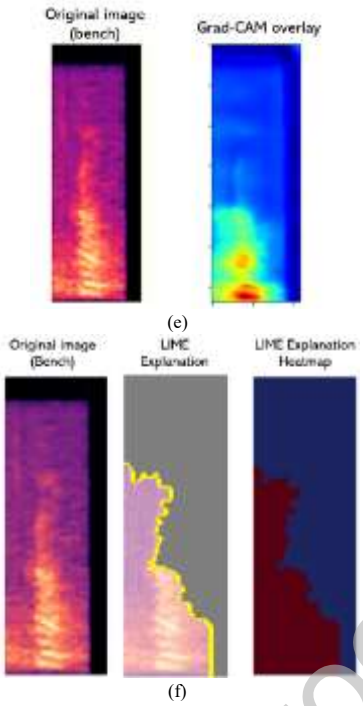
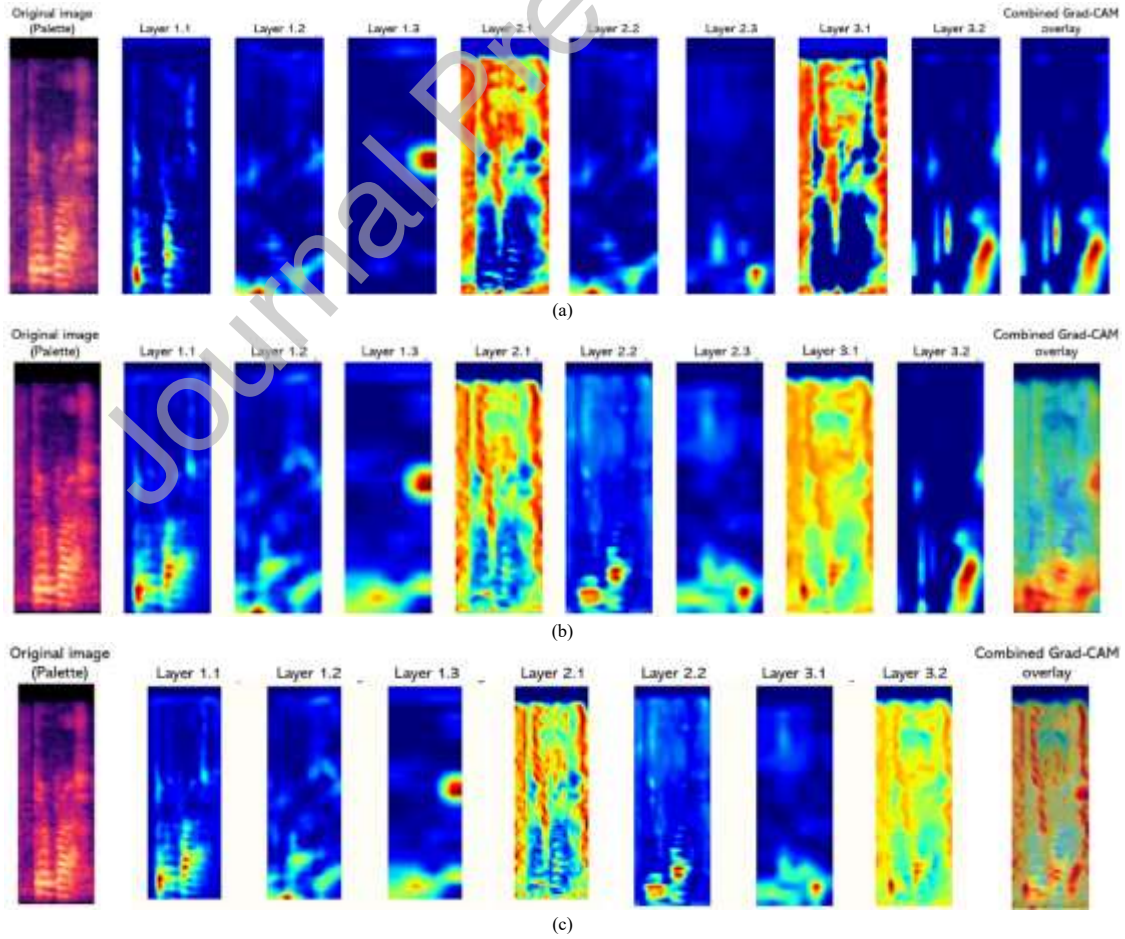
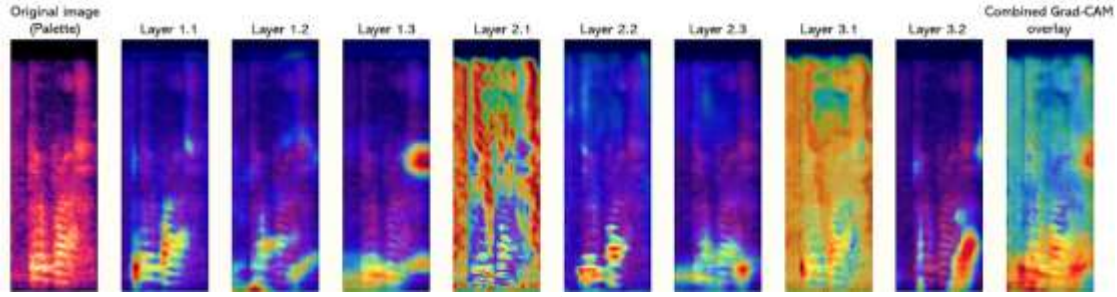
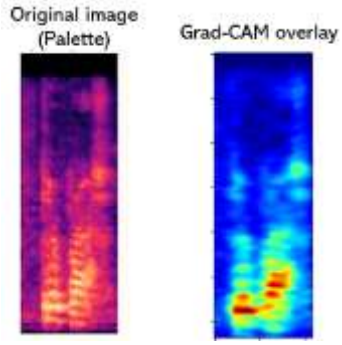


Fig.9. Grad-CAM results for the keyword 'Bench' from the Multi attention-based CNN model (a) Traditional Grad-CAM (b) ED-GCAM (c) EH-FCAM (d) SGD-GCAM (e) MSCW-GCAM (f) Lime

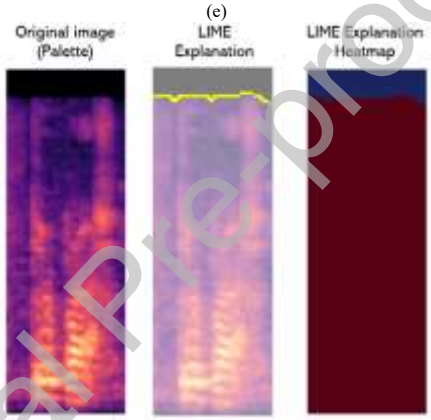




(d)

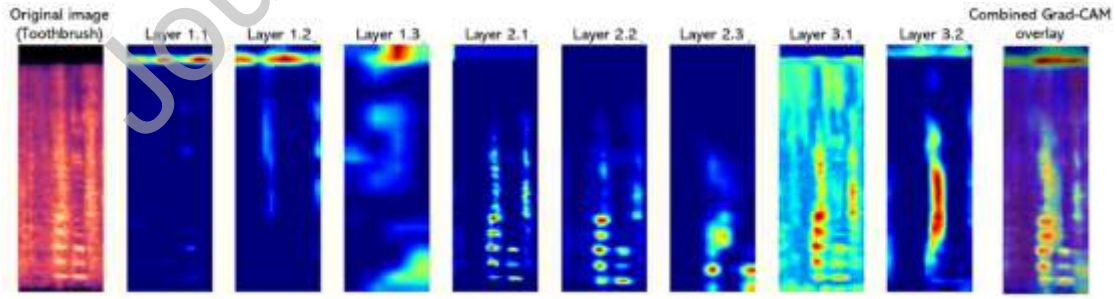


(e)

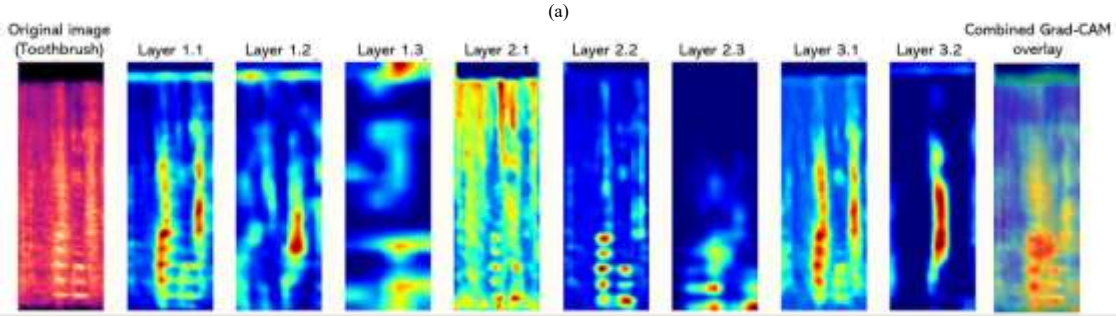


(f)

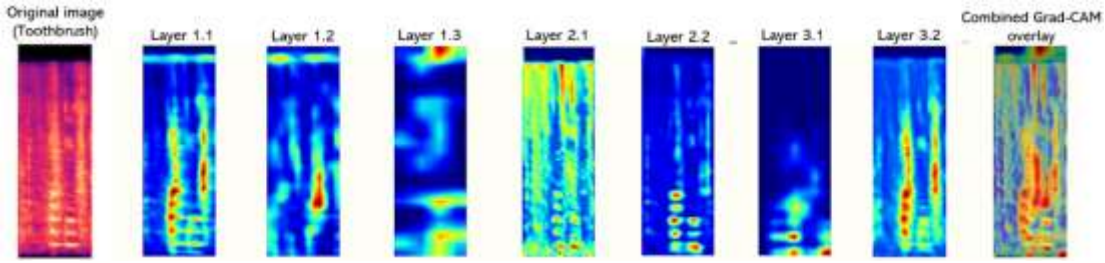
Fig.10. Grad-CAM results for the keyword 'Palette' from the Multi attention-based CNN model (a) Traditional Grad-CAM (b) ED-GCAM (c) EH-FCAM (d) SGD-GCAM (e) MSCW-GCAM (f) Lime



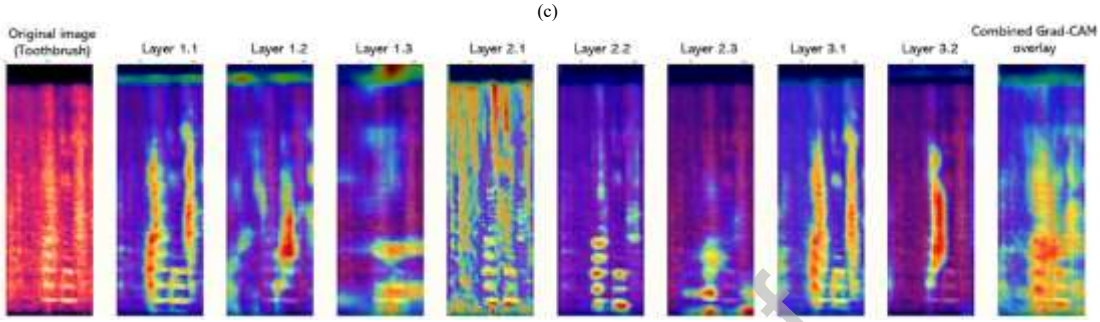
(a)



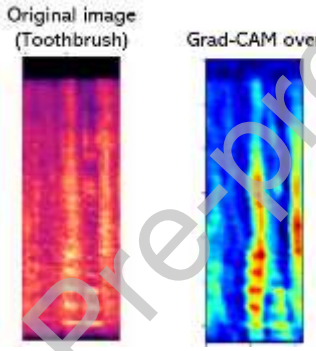
(b)



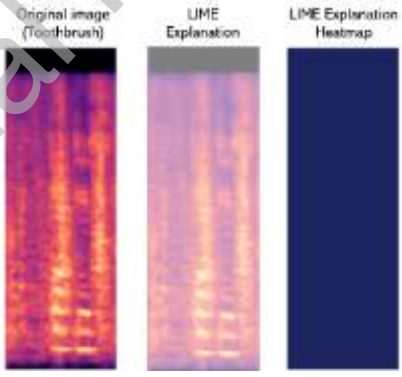
(c)



(d)

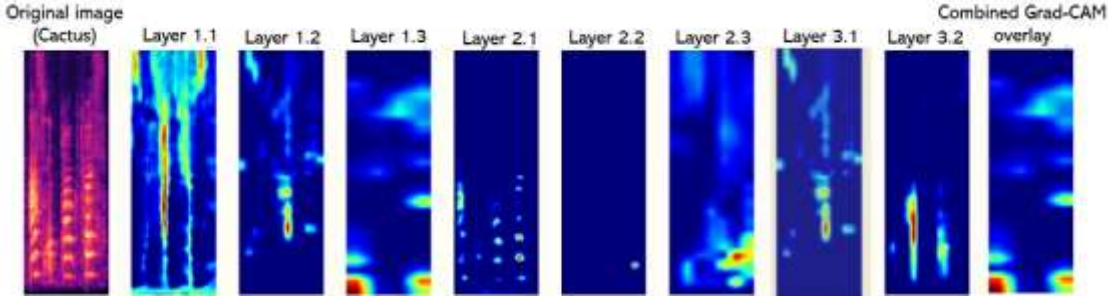


(e)



(f)

Fig.11. Grad-CAM results for the keyword 'Toothbrush' from the Multi attention-based CNN model (a) Traditional Grad-CAM (b) ED-GCAM (c) EH-FCAM (d) SGD-GCAM (e) MSCW-GCAM (f) LIME



(a)

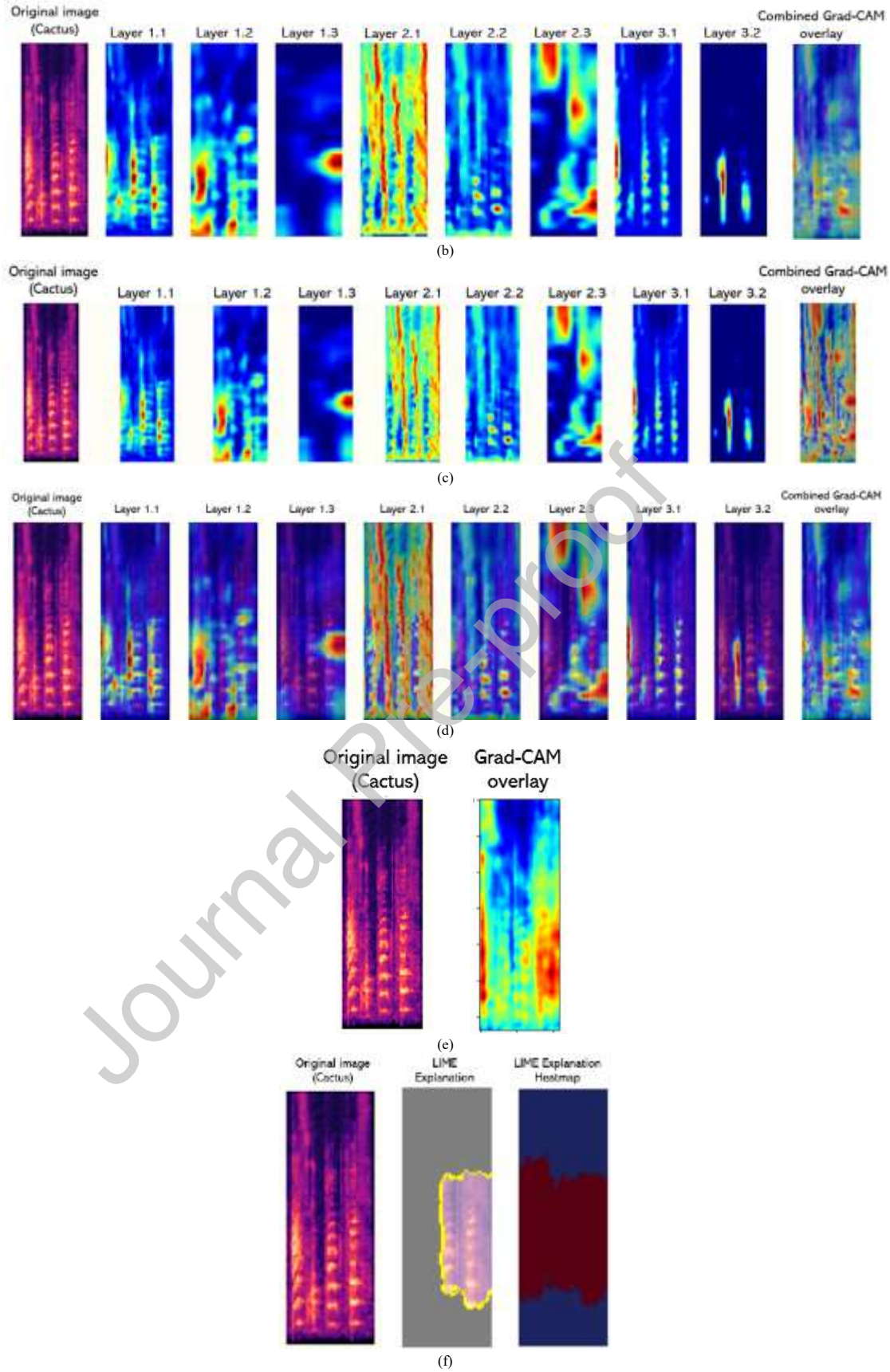


Fig.12. Grad-CAM results for the keyword 'Cactus' from the Multi attention-based CNN model (a) Traditional Grad-CAM (b) ED-GCAM (c) EH-FCAM (d) SGD-GCAM (e) MSCW-GCAM (f) LIME

4. Discussion

Our results show that using our advanced Grad-CAM approaches over the traditional Grad-CAM lead to considerable improvements in model behaviour and interpretability in quantitative and qualitative evaluation metrics. These improved methods offer more detailed justifications for model selections, especially in the problematic area of aphasia speech keyword categorisation, where conventional approaches frequently fail to capture the complex nature of speech impairment patterns.

Enhanced Directional Grad-CAM improves the model's ability to identify significant regions in the Mel-spectrogram by enabling the model to concentrate on gradients that positively impact the target class. By eliminating unnecessary gradients, our technique guarantees that the heatmaps produced are sharply focused on the crucial regions that support accurate classification. This is particularly helpful for aphasic speech, as distracting or loud elements can trick the model. Enhanced Directional Grad-CAM differs from Guided Grad-CAM and Guided Backpropagation, eliminating irrelevant gradients and only concentrating on positive gradients that enhance the target class. This produces more accurate and targeted heatmaps highlighting only the areas crucial for categorisation. These heatmaps are especially useful for complex data, such as speech aphasia. Guided Grad-CAM, on the other hand, creates broader, less focused representations by combining gradients from all neurons, including those that are not directly related to the target class. Regardless of their significance to the prediction accuracy, all neuron activations are further highlighted by guided backpropagation, which frequently results in less concentrated heatmaps. Enhanced Directional Grad-CAM is a helpful tool for addressing noisy or irregular input as it isolates the essential features, impacting the model's choice and providing more explanations that are unique to a class. In Fig.5. (a) to Fig.12. (a), you can see the technique produced well fine-grained results, which was not present in the traditional Grad-CAM result from both speech classification models.

Channel-wise multi-scale Grad-CAM further assesses the model's behaviour at various Mel-spectrogram scales. This method captures high-level and fine-grained characteristics, that traditional Grad-CAM can overlook, by creating Grad-CAMs for specific channels at several resolutions. The channel-wise breakdown provides an excellent grasp of how various feature maps contribute to the final prediction. At the same time, the multi-scale approach helps the model discover minor fluctuations in the speech signal, making it more sensitive to the distinct distortions inherent in damaged speech. The result of the combined final CAM of the method of both classification models are shown in Fig.5. (e) to Fig.12. (e).

Stochastic Gradient-Dropout Integrated Grad-CAM averages the results across several runs and applies dropout during inference to integrate uncertainty into the interpretability framework. This technique produces more stable and resilient heat maps by quantifying uncertainty in model predictions and highlighting consistent regions of relevance. The technique also evaluates the consistently important features activated over time during the speech. As this strategy is stochastic, it can handle noisy inputs (typical in aphasia speech) better because it concentrates on regions that consistently affect model decisions. Fig.5. (e) through Fig.12. (e) display the integrated final CAM of both the classification model's results.

The Enhanced Hierarchical Filtered Grad-CAM technique improves on earlier approaches by creating Grad-CAMs for each smaller, hierarchical region into which the input is segmented. The model can now concentrate on specific regions of the Mel-spectrogram, which is important for detecting subtle aspects frequently distorted in aphasic speech, such as fluctuations in phoneme level. Compared to ordinary Grad-CAM, hierarchical filtering offers a more granular picture by presenting a multi-layered understanding of how various input components contribute to the multi-attention model classification results from Fig.9. (d) to Fig.12. (d) and the single-attention classification model results from Fig.5. (d) to Fig.8. (d).

Evaluating all the proposed extension studies, we observed that most performed better than the traditional Grad-CAM variants in metrics such as perturbation, infidelity, and sufficiency scores. While few showed the identical range scores as the traditional Grad-CAM, some metrics scores showed noticeable performance variability across different sample sets. For instance, specific test batches scored higher on infidelity and trustworthiness than others, especially when it involved speech impairment. We averaged the values for each statistic across the several sample sets to account for this. The observed diversity underscores the complex nature of the data and the inherent challenges in implementing these expansions to a range of speech patterns, particularly in aphasic speech. Our open reporting of this variability has pointed out how important it is to consider these variations when analysing the data.

For a more comprehensive evaluation of these Grad-CAM extensions, we included the LIME explanation map along with the traditional and proposed extensions in the result. LIME's explanation for the Mel-spectrogram image classification task can be found in Fig.5. (f) to Fig.12. (f) for both the models. The overlaid boundary around the regions of the image is identified as the most important for the model predictions. These regions represent the areas that positively contributed to the prediction, meaning the models found these areas to indicate the specific class most. It perturbs the image by masking or altering different super pixels and observes the model response to each altered version. This process builds a linear approximation of the model decision boundaries around the image, highlighting the super pixels that had the most impact on it. But from the results of LIME, we can say that it may not align perfectly with time-frequency features like Grad-CAM due to the reliance on segmenting the spectrogram into super pixels.

In general, the expanded Grad-CAM techniques yielded more comprehensive and significant explanations compared to the normal Grad-CAM. This enhanced the model's capacity to manage intricate and irregular speech inputs and furnished crucial insights into how the model interprets speech impairment.

These enhanced Grad-CAM techniques can improve the architecture of interpretable models for spoken keyword categorisation, where interpretability plays a critical role. Using directional filtering, multi-scale analysis, and stochastic dropout methods, model designers can better comprehend how various features influence the conclusion. This improves models' robustness, refinement, and interpretability, especially when dealing with complex datasets such as aphasia speech. Furthermore, these techniques are beneficial in fields like healthcare, where speech models could be used to diagnose or support treating patients with speech problems. These applications depend heavily on the transparency of model judgments. More trust in AI systems can be fostered by their capacity to produce precise, dependable, and interpretable explanations of model behaviour. This is especially important in delicate fields such as these.

The proposed techniques introduce higher computational complexity due to the nature of impaired speech and the advanced Grad-CAM extensions used. Impaired speech, such as aphasia, presents unique challenges compared to healthy speech, making it inherently more difficult to interpret model decisions, especially with black-box models. These techniques aim to reduce the gap in using CAM-based explainable AI methods to clarify how models arrive at decisions in impaired spoken keyword classification. Techniques like Enhanced Directional Grad-CAM help highlight important phonemes or words for therapeutic purposes. Enhanced Hierarchical Filters aid in making the model's decision-making process more understandable at different linguistic levels. Additionally, Multi-Scale Channel-wise Grad-CAM provides better insights into the features extracted from spectrograms, and Stochastic Gradient-Dropout CAM helps address uncertainty in model decisions, adding robustness.

Despite their interpretability and usefulness, some performance variability remains across different sample sets. This may be attributed to the inherent complexity of impaired speech data, where variability in speech patterns is significant. Overfitting may also arise due to the limited size of clinical datasets. This can be mitigated by data augmentation, regularization techniques and cross-validation. Scalability can pose a challenge, particularly in practical clinical applications, because of the higher computational demands of multi-scale and stochastic methods. However, these models represent foundational steps toward more reliable, interpretable AI in diagnosing and treating speech impairments, such as aphasia. We continue to explore ways to optimize these techniques for clinical scalability while maintaining interpretability and robustness. The mentioned editions are fundamental and intended to incorporate XAI methods based on GRAD-CAM for classifying speech-impaired individuals. This can help healthcare professionals comprehend opaque decisions. Nevertheless, these fundamental additions have constraints. We can accomplish our desired objectives by enhancing these methods in the future.

5. Conclusion and Future work

In this study, we introduced enhanced Grad-CAM techniques, such as Enhanced Directional Grad-CAM, Multi-Scale Channel-wise Grad-CAM, Stochastic Gradient-Dropout Integrated Grad-CAM, and Enhanced Hierarchical Filtered Grad-CAM, to improve interpretability and reliability in aphasia speech keyword classification. These techniques highlighted the most important areas in Mel-spectrograms, producing more accurate and targeted heatmaps. Combined with attention-based deep learning models, the suggested method showed notable interpretability and accuracy in categorising the challenging, noisy speech patterns associated with aphasia. Our results showed that these Grad-CAM extensions performed better than standard Grad-CAM in providing meaningful explanations and important information for understanding how models' decision-making process in impaired speech. This enhanced interpretability not only tried to bridge the gap in explainable AI methods for

speech impairments but also aims for practical implications in applications where model transparency is needed. However, we noticed variation in performance between sample sets, especially in the data on speech impairment. This highlights how complicated the data is and how carefully the results must be interpreted.

Future research could strengthen the suggested Grad-CAM techniques' resilience to manage different kinds of noise frequently encountered in real-world environments. This could involve strengthening the Grad-CAM framework's resistance to background noise and artifacts typical of aphasia and other speech problems by adding sophisticated noise reduction algorithms or adaptive filtering. These improved Grad-CAM techniques could be modified to treat additional speech problems like dysarthria or stuttering, broadening its application beyond aphasia. A broader range of therapeutic scenarios may find the framework more applicable and interpretable if modified to account for the unique aspects of these conditions that impact speech patterns.

Additional research could focus on optimizing the multi-scale, directed, and hierarchical methods for real-time processing, considering the computational complexity they impose. Enabling use in real-time clinical situations or mobile health applications may entail improving model structures, lowering computational overhead, or utilising parallel processing. Grad-CAM's diagnostic capabilities may also be improved by combining it with other data sources, such as physiological or visual inputs. More research may investigate multimodal methods integrating sensor or facial expression data with audio analysis to offer more thorough and insightful information.

Finally, validating and improving the model's adaptability may require evaluating the generalizability of these improved Grad-CAM approaches across larger and more varied aphasia speech datasets or those with different linguistic and cultural characteristics. This could ensure the approaches' robustness and applicability to various clinical contexts and the population.

References

[1] M.C. Brady, H. Kelly, J. Godwin, P. Enderby, P. Campbell, Speech and language therapy for aphasia following stroke, Cochrane Database Syst Rev 2016 (2016). https://doi.org/10.1002/14651858.CD000425.PUB4.

[2] D.S. Barbera, M. Huckvale, V. Fleming, E. Upton, H. Coley-Fisher, C. Doogan, I. Shaw, W. Latham, A.P. Leff, J. Crinion, NUVA: A Naming Utterance Verifier for Aphasia Treatment, Comput Speech Lang 69 (2021) 101221. https://doi.org/10.1016/J.CSL.2021.101221.

[3] J. Wade, B. Petheram, R. Cain, Voice recognition and aphasia: Can computers understand aphasic speech?, Disabil Rehabil 23 (2001) 604–613. https://doi.org/10.1080/09638280110044932.

[4] M.R. Akbarzadeh-T, M. Moshtagh-Khorasani, A hierarchical fuzzy rule-based approach to aphasia diagnosis, J Biomed Inform 40 (2007) 465–475. https://doi.org/10.1016/j.jbi.2006.12.005.

[5] M. Danly, B. Shapiro, Speech prosody in Broca's aphasia, Brain Lang 16 (1982) 171–190. https://doi.org/10.1016/0093-934X(82)90082-7.

[6] S. Ash, C. McMillan, D. Gunawardena, B. Avants, B. Morgan, A. Khan, P. Moore, J. Gee, M. Grossman, Speech errors in progressive non-fluent aphasia, Brain Lang 113 (2010) 13–20. https://doi.org/10.1016/J.BANDL.2009.12.001.

[7] N. Jamal, S. Shanta, F. Mahmud, M. Sha, Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review □ Automatic Speech Recognition (ASR) based Approach for Speech Therapy of Aphasic Patients: A Review, AIP Conf. Proc 1883 (2017) 20028. https://doi.org/10.1063/1.5002046.

[8] J. Tang, W. Chen, X. Chang, S. Watanabe, B. MacWhinney, A New Benchmark of Aphasia Speech Recognition and Detection Based on E-Branchformer and Multi-task Learning, (2023). http://arxiv.org/abs/2305.13331.

[9] M. Day, R.K. Dey, M. Baucum, E.J. Paek, H. Park, A. Khojandi, Predicting Severity in People with Aphasia: A Natural Language Processing and Machine Learning Approach, Annu Int Conf IEEE Eng Med Biol Soc 2021 (2021) 2299–2302. https://doi.org/10.1109/EMBC46164.2021.9630694.

[10] A. Adikari, N. Hernandez, D. Alahakoon, M.L. Rose, J.E. Pierce, From concept to practice: a scoping review of the application of AI to aphasia diagnosis and management, Disabil Rehabil 46 (2024) 1288–1297. https://doi.org/10.1080/09638288.2023.2199463.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, Adv Neural Inf Process Syst 2017-December (2017) 5999–6009. https://arxiv.org/abs/1706.03762v7 (accessed July 25, 2024).

[12] Y. Qin, T. Lee, Y. Wu, A.P.H. Kong, An end-to-end approach to automatic speech assessment for people with aphasia, 2018 11th International Symposium on Chinese Spoken Language Processing, ISCSLP 2018 - Proceedings (2018) 66–70. https://doi.org/10.1109/ISCSLP.2018.8706690.

[13] K. Jothi, V.M.-2020 3rd I.C. on, undefined 2020, A systematic review of machine learning based automatic speech assessment system to evaluate speech impairment, Ieeexplore.Ieee.OrgKR Jothi, VL Mamatha2020 3rd International Conference on Intelligent Sustainable, 2020•ieeexplore.Ieee.Org (2020) 175–185. https://doi.org/10.1109/ICISS49785.2020.9315920.

[14] I. Lopez-Espejo, Z.H. Tan, J.H.L. Hansen, J. Jensen, Deep Spoken Keyword Spotting: An Overview, IEEE Access 10 (2021) 4169–4199. https://doi.org/10.1109/ACCESS.2021.3139508.

[15] C. Shan, J. Zhang, Y. Wang, L. Xie, Attention-based End-to-End Models for Small-Footprint Keyword Spotting, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2018-September (2018) 2037–2041. https://doi.org/10.21437/Interspeech.2018-1777.

[16] A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, I.P. Martins, Automatic word naming recognition for an on-line aphasia treatment system, Comput Speech Lang 27 (2013) 1235–1248. https://doi.org/10.1016/J.CSL.2012.10.003.

[17] D.S. Barbera, M. Huckvale, V. Fleming, E. Upton, H. Coley-Fisher, I. Shaw, W. Latham, A.P. Leff, J. Crinion, An Utterance Verification System for Word Naming Therapy in Aphasia, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2020-October (2020) 706–710. https://doi.org/10.21437/INTERSPEECH.2020-2265.

[18] X. Wu, P. Bell, A. Rajan, Explanations for Automatic Speech Recognition, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2023-June (2023). https://doi.org/10.1109/ICASSP49357.2023.10094635.

[19] A. Akman, B.W. Schuller, Audio Explainable Artificial Intelligence: A Review, Intelligent Computing 3 (2024). https://doi.org/10.34133/ICOMPUTING.0074.

[20] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain?, (2017). https://arxiv.org/abs/1712.09923v1 (accessed July 25, 2024).

[21] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, Int J Comput Vis 128 (2016) 336–359. https://doi.org/10.1007/s11263-019-01228-7.

[22] E. Kim, G.S. Dahiya, S. Løset, R. Skjetne, Can a computer see what an ice expert sees? Multilabel ice objects classification with convolutional neural networks, Results in Engineering 4 (2019). https://doi.org/10.1016/j.rineng.2019.100036.

[23] A. Chattopadhyay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks, Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018 2018-January (2017) 839–847. https://doi.org/10.1109/WACV.2018.00097.

[24] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation Applied to Handwritten Zip Code Recognition, Neural Comput 1 (1989) 541–551. https://doi.org/10.1162/NECO.1989.1.4.541.

[25] P. Dumane, B. Hungund, S. Chavan, Dysarthria Detection Using Convolutional Neural Network, Techno-Societal 2020 (2021) 449–457. https://doi.org/10.1007/978-3-030-69921-5_45.

[26] P. Warden, Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, (2018).

[27] B. MacWhinney, D. Fromm, M. Forbes, A. Holland, AphasiaBank: Methods for Studying Discourse, Aphasiology 25 (2011) 1286. https://doi.org/10.1080/02687038.2011.589893.

[28] B. Moëll, J. O'regan, S. Mehta, A. Kirkland, H. Lameris, J. Gustafsson, J. Beskow, Speech Data Augmentation for Improving Phoneme Transcriptions of Aphasic Speech using wav2vec 2.0 for the PSST Challenge, n.d. https://github.com/iver56/.

[29] X. Li, H. Xiong, X. Li, X. Wu, Z. Chen, D. Dou, InterpretDL: Explaining Deep Models in PaddlePaddle, Journal of Machine Learning Research 23 (2022) 1–6. http://jmlr.org/papers/v23/21-0738.html (accessed September 11, 2024).

[30] C.K. Yeh, C.Y. Hsieh, A.S. Suggala, D.I. Inouye, P. Ravikumar, On the (In)fidelity and Sensitivity for Explanations, Adv Neural Inf Process Syst 32 (2019). https://arxiv.org/abs/1901.09392v4 (accessed September 11, 2024).

[31] S. Dasgupta, N. Frost, M. Moshkovitz, Framework for Evaluating Faithfulness of Local Explanations, Proc Mach Learn Res 162 (2022) 4794–4815. https://arxiv.org/abs/2202.00734v1 (accessed September 11, 2024).

APPENDIX I

I. Case study – Keyword Classification for Aphasia Patient

A. Case Study Example 1

Patient Information:
- Patient: X
- Severity Level: Y aphasia
- Speech Data: Pronouncing the keyword ‘Octopus'

Objective:
- Analyze which parts of the spectrogram ED-GCAM highlights as important for correctly classifying the word 'Octopus'.

Steps:

- Input Data: A Mel-spectrogram of the patient's pronunciation of the word 'Octopus' is fed into the model.
- ED-GCAM Output: The ED-GCAM heatmap highlights regions of the spectrogram that contribute most to the classification of 'Octopus'.
- Key Observations:
    - The ED-GCAM identified crucial phoneme/pronunciation transitions as key regions for the model’s classification.
    - Compared to standard Grad-CAM, ED-GCAM provided more focused and captured even fine-grained details also in the heatmap, which standard CAM failed to capture, enhancing the interpretability of impaired speech signals.

- *Conclusion:*
    - ED-GCAM successfully identified key areas in the speech signal, showing its applicability in aphasia speech analysis for understanding how a speech impairment affects specific phonetic regions. Results can be found in fig in the result section.

B. Case Study Example 2

Patient Information:
- Patient: X, Y yrs
- Severity level: Y aphasia
- Speech Data: Audio recordings of speech therapy sessions were collected, focusing on 15 specific keywords, such as "cactus" and "canoe."

Objective:
- The goal was to assess the consistency of speech production over time using Stochastic Gradient-Dropout Integrated Grad-CAM (SGD-GCAM). This method aimed to visualize how the importance of speech features evolved, helping clinicians monitor progress and identify areas requiring targeted interventions.

Steps:

- Input Data: Mel-spectrograms of the patient generated from the recordings.
- Model Setup: A pre-trained CNN model with dropout enabled during inference was used for keyword classification.
- SGD-GCAM Output: The heatmap generated with stochastic dropout reveals variations in feature importance over multiple inference runs.

- Temporal analysis: The consistency of speech patterns was analyzed by comparing averaged Grad-CAM heatmaps across different sessions.
- Key Observations:
    - Improved Consistency: Heatmaps for words like "canoe" became more focused over time, indicating clearer pronunciation.
    - Identified Challenges: Keywords such as "cactus" displayed high variability in heatmaps across sessions, highlighting persistent difficulties.
    - Uncertainty Insights: Variance in heatmaps revealed areas of inconsistency, guiding therapists to adjust therapy approaches.

- *Conclusion:*
    - Indeed, extension can be used to monitor the speech patterns of aphasia patients over time. Practitioners can assess the consistency of key features being activated over time by using Stochastic Gradient-Dropout Integrated Grad-CAM (SGD-GCAM) on several speech sessions of a patient. This lets you see if the same areas (such as spectrogram segments) are consistently identified as being important for predicting the target keyword or phrase.

    - Significant variances may indicate changes in speech production or articulation, whereas consistency over time in the indicated Grad-CAM regions would suggest that the patient's speech pattern stays consistent. Therefore, the improved SGD-GCAM is a useful diagnostic tool for evaluating temporal changes in the speech characteristics of aphasia patients in addition to being a tool for robust feature importance analysis. The results of one session can be found in fig. [6 – 13]. (d), in the result section.

[Note: All the case studies mentioned are examples for the understanding to show how these proposed models and extensions can be applied to actual purpose. None of these are applied clinically or tested but used actual aphasia patient datasets for the research study. As previously mentioned in the manuscript, these extensions aim to adapt GRAD-CAM for impaired speech treatment and diagnostics. Additionally, they seek to bridge the gap in the application of CAM XAI techniques in this field. Further enhancements are necessary to better meet the specific needs of these applications.]

Cover_Letter

31/10/2024

Francisco Martínez-Álvarez, Ph.D.
Editor
Results in Engineering

Dear Editor

Kindly refer to your email dated 11th October 2024 regarding our manuscript entitled "Advanced Grad-CAM Extensions for Interpretable Aphasia Speech Keyword Classification: Bridging the Gap in Impaired Speech with XAI" (Ref.: Ms. No. **RINENG-D-24-05467**). First, we would like to thank the honorable reviewers and associate editor for their suggestive comments, which were of great help in improving the quality and strength of the current manuscript. We have read each reviewer's comments carefully and, accordingly, made utmost efforts to address each comment point-by-point. We much appreciate the reviewer's time, effort, and detailed explanations. We believe that the revised version of our manuscript has incorporated all the suggestions from reviewers and is now more understandable, and better arranged to suit the publication standards. The changes made in the manuscript are in red colour.

Regards
Corresponding author
John Sahaya Rani Alex

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: