ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240118001270
Original scientific paper

Grad-CAM-Based Feature Selection and Dementia Classification Algorithm Using Voice Data

Hansol KO, Bohyun WANG, Joon S. LIM*

Abstract: This study presents a unique methodology for dementia classification that harnesses voice data and integrates transfer learning, feature selection, and attention-based visualization. We examined two deep learning input techniques: one consolidating three Melspectrograms (standard, Harmonic/Percussive average, and delta value) into an integrated image and the other assessing them individually.This study validates the efficacy of melt spectrogram image classification using Gradient-weighted Class Activation Mapping (Grad-CAM) for feature selection. This study exploited the Grad-CAM attention map to pinpoint the Melspectrogram's most impactful features. Evaluations illustrated that the combined synthetic images yielded 1.4% - 9.4% better accuracy than the separate images. Implementing Grad-CAM for feature selection further amplifies accuracy. Models utilizing features identified by Grad-CAM averaged a 4.3% superior accuracy compared with solely fine-tuned models. With integrated mel-spectrograms as input, the classification accuracies for Normal vs. Dementia, Normal vs. Mild Cognitive Impairment, and Dementia vs. Mild Cognitive Impairment were 75%, 67.9%, and 67.8%, respectively, indicating an improvement of up to 13.1% compared to individual images.

Keywords: attention-based visualization; deep learning; dementia classification; feature selection; Grad-CAM; mel-spectrogram; transfer learning; voice data

1 INTRODUCTION

Dementia is a neurodegenerative disorder characterized by a decline in cognitive function and the ability to perform daily activities. The number of people with dementia is rapidly increasing worldwide, and statistical estimates suggest that up to 135 million people will be living with the disease by 2050 [1]. This presents a significant challenge on a global scale, affecting individuals, families, and healthcare systems alike. Early and accurate diagnosis of dementia is critical as it allows for timely intervention and planning of treatment strategies, which can ultimately improve an individual's quality of life. Diagnosing dementia requires evaluation by trained clinicians, which is time-consuming and costly. To reduce these burdens, various methods have been proposed for classifying dementia using artificial intelligence (AI), including approaches based on brainwave signals [2], Magnetic Resonance Imaging (MRI) [3], and voice data [4]. This study focused on the development of an algorithmic model for diagnosis using voice data. Utilizing voice for dementia classification is appealing, as it can substantially reduce the time and resources required compared to other methods. Previous studies have suggested that various features of speech can be useful for dementia classification [5-7]. Language function and its features have been proposed as potential early markers of cognitive decline [8-10]. Recent research has highlighted the potential of Mel spectrograms [11, 12] to capture the frequency content in audio signals and offer discriminative features for voice data classification tasks [13]. Innovative feature extraction techniques, such as those targeting pitch and prosody characteristics, have effectively captured voice features associated with cognitive decline. Deep learning models, including convolutional neural networks (CNNs), have shown impressive performance, with capabilities for learning complex patterns in Mel spectrograms to enable more accurate classification. Therefore, an emerging challenge is to determine how to further improve the accuracy of classification models using Mel spectrograms as input. The interpretability of these classification models is another crucial aspect. While deep learning models showed remarkable performance, their 'black box' nature limits their interpretability. There is a growing need to develop explainable models that can provide insights into the fundamental features and contribute to our understanding of the relationship between voice biomarkers and dementia. Therefore, although substantial progress has been made in leveraging voice data for dementia classification, aided by advancements in machine learning and the exploration of multimodal and end-to-end data, persistent challenges related to feature engineering, model interpretability, and dataset standardization remain. Addressing these challenges and capitalizing on recent developments and improvements can enhance the accuracy and clinical applicability of voice-based dementia classification models. This study proposes feature selection based on Gradient-weighted Class Activation Mapping (Grad-CAM) [14, 15] to address these challenges. Grad-CAM is an attention-based technique [16] for identifying the most relevant regions in an image for a specific classification task. In this study, Grad-CAM was employed to visualize the most relevant features of voice data in dementia classification using heatmaps. This study employed Grad-CAM to understand which vocal characteristics exert the most significant effects on dementia classification by experimenting with different Mel spectrogram image transformations. Our goal was to develop an AI model capable of early diagnosis and classification of dementia by enhancing the interpretability and accuracy of classification models using Mel spectrograms, with Grad-CAM applied to select features. In this study, transfer learning [17] and feature selection techniques [18] were employed for voice data classification tasks using Mel spectrograms. The VGG16 model [19] was utilized as a pretrained model, and its performance was evaluated in tasks involving the classification of voice data using Mel spectrograms. Grad-CAM was adopted for feature selection to identify the most relevant regions in the input image for classification and to selectively train the model based on these regions [20]. Furthermore, experiments were conducted using average spectrograms of Harmonic and Percussive components and delta values of Mel

spectrograms [21] in addition to conventional Mel spectrograms. We compared the model performance using these different images as input and investigated their impact on model accuracy. This study aims to ascertain the influence of feature selection using Grad-CAM on dementia classification accuracy. In addition, the effectiveness of transfer learning and feature selection techniques for voice data analysis using Mel spectrograms is validated. This study endeavors to bridge the gap between traditional audio feature selection methods and modern deep learning approaches, providing a framework for utilizing Mel spectrograms and attention-based techniques in voice data analysis. Our experiments offer a comparative analysis based on input, the effectiveness of Grad-CAM-based feature selection, and a comprehensive evaluation of the performance of the proposed model. This study is expected to contribute to the development of more accurate and efficient models for voice data analysis.

2 RELATED RESEARCH

2.1 Voice-Based Dementia Biomarkers

Research has demonstrated distinct voice characteristics and patterns in dementia patients as compared to healthy individuals [5-7]. Speech rate, pitch variations, pronunciation, and prosody have been identified as potential biological indicators for dementia classification [8-10]. These studies underscore the potential of acoustic features in capturing cognitive decline and lay the foundation for utilizing voice data as a diagnostic tool. Therefore, this study aims to employ these acoustic features as indicators for cost-effective dementia classification.

2.2 Voice Feature Extraction Techniques for Dementia Classification

To capture the acoustic features of speech, techniques such as Mel-frequency cepstral coefficients (MFCC) [11], pitch and energy-based feature extraction, prosodic feature extraction, and moment frequency feature extraction have been widely utilized. These feature-extraction technologies are applied to voice data to capture dementia-related acoustic characteristics. The extracted features served as inputs for the classification models, which could distinguish between patients with dementia and a healthy control group. This study aims to develop robust and accurate classification models for the early detection and diagnosis of dementia based on voice data. In this study, Mel spectrograms were used among various feature extraction techniques because recent studies have shown that Mel spectrograms [12] can offer discriminative features for voice data classification tasks [13]. For one voice data, three types of Mel spectrogram images were extracted in this study [21]: the basic Mel spectrogram, the average spectrogram of Harmonic and Percussive components, and the delta value spectrogram. These three spectrograms were used to extract diverse image features of the voice, and the extent to which each acoustic feature affected the dementia classification was evaluated.

2.3 Deep Learning for Dementia Classification

In dementia classification, various deep learning models have been explored to leverage the power of neural networks for effective feature extraction and classification. These include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), hybrid models (CNN-RNNs), Transformer models, and ensemble models. Deep learning models, particularly CNNs, have demonstrated remarkable performance in various image and audio classification tasks. CNNs have been used to automatically extract discriminative features from voice spectrogram images in dementia classification. Transfer learning [17], which involves the use of pretrained models on large-scale datasets, has been employed to enhance classification accuracy and generalization. In this study, the VGG-16 model [19] was selected as the base architecture for transfer learning. VGG-16 is a widely used CNN model that has demonstrated exceptional performance in various image classification tasks. The model was initialized with pretrained weights obtained from training on the ImageNet dataset [22]. By utilizing these pretrained weights, the model has already learned a set of visual representations that are useful for dementia classification from voice data. In this study, the features were extracted through the final feature map layer of VGG16, and these weights were used to conduct Grad-CAM to select high-impact feature regions for training.

2.4 Feature Selection for Dementia Classification

Feature selection plays a critical role in the development of accurate and interpretable dementia classification models. The performance of the model can be improved by identifying the most relevant acoustic features, and insights can be gained into the underlying biomarkers associated with dementia. Previous research has explored various feature selection methods, including traditional methods such as correlation analysis and backward elimination, univariate feature selection, Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and regularization techniques. Among these feature selection methods, this study employed a technique that selects features based on their weight importance. This study used Grad-CAM [16, 17] as the feature selection technique for dementia classification. Grad-CAM visualizes the regions of the input data that contribute the most to the model's classification decision, thereby enabling the localization of prominent features within the input data. Grad-CAM was initially developed for visual object recognition and has been successfully applied to various domains, including medical imaging and audio analysis. Grad-CAM offers interpretable insights into regions of interest by generating attention maps, thereby enhancing feature selection for improved understanding and classification accuracy. Thus, in this study, Grad-CAM allowed visual confirmation of which features in the voice data had a strong influence on dementia classification. Based on the confirmed results, temporal features of voice, along with rhythm and beat features, emerged as highly influential features in dementia classification. Unlike previous studies, this study explored a combination of transfer learning and feature selection techniques to enhance classification accuracy and interpretability. Furthermore, for a comprehensive comparison of the impact on classification performance, this study employed Grad-CAM, a type of attention mechanism, in feature selection to provide interpretability beyond that offered in existing studies.

3 METHODOLOGY

3.1 Raw Data

This study employed a dataset [23] obtained from AIHub [24] that consisted of audio recordings in the WAV format with a sampling rate of 48000 Hz. The dataset includes 11 scripts and their corresponding responses. The structures of the scripts are listed in Tab. 1. In Tab. 1, questions 1 to 3 are read-along sentences; questions 4 and 5 are picture-description tasks; questions 6 to 8 are word memory tasks; and questions 9 to 11 are free expression tasks. Data were collected from 292 participants, including 121 from the Normal group (N), 89 from the Alzheimer's Disease group (AD), and 82 from the Mild Cognitive Impairment group (SCI). The duration of the voice data ranged from 3 s to approximately one minute. For questions 1 - 4, the duration was within 30 s, and for questions 5 - 11, it was up to one minute.

Table 1 Composition of Voice Data Script

No.	Task	Voice Script	Type
1	Sentence Reading-1	Roses bloomed in the yard	Read Along Sentences
2	Sentence Reading-2	It rained yesterday, so I stayed at home	
3	Sentence Reading-3	Birds hear the day words; mice, the night words	
4	Name the Animal	Say the name of the displayed animal	Picture Description
5	Picture Description	Describe the picture in detail for 1 minute	
6	Language Fluency (Phonemic)	Say a word that starts with the given letter	Word Memory
7	Language Fluency (Semantic)	Say words that belong to the given category	
8	Serial Subtraction	Continuously subtract the same number	
9	Positive Emotion Story	Talk for 1 minute about the happiest event in your life	Free Expression
10	Negative Emotion Story	Talk for 1 minute about the saddest event in your life	
11	Narrative	Talk for 1 minute about what happened yesterday	

3.2 Data Preprocessing

To enhance the performance of the model, the collected data were preprocessed, including removal or manipulation, before feeding the data into the model. We proceed by resizing the existing law data images to 300 × 300.

3.2.1 Mel Spectrogram

The prepared dataset was processed using Mel spectrogram functions to convert audio files into image files. A Mel spectrogram is a representation of a sound signal in which the frequency scale is Mel-scaled. For each audio data, the Mel spectrogram, the average Mel spectrogram of the Harmonic and Percussive components, and the delta (derivative) of the Mel spectrogram were computed (Fig. 1). The concatenation of these three spectrogram images, as well as the individual images, was used for the experiments (Fig 2).

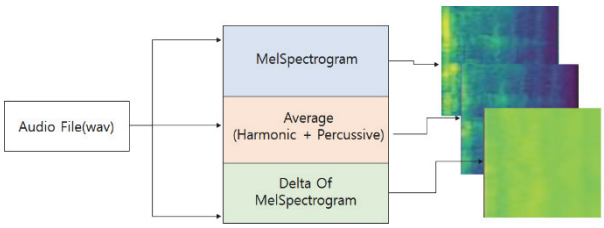

Figure 1 Typical Mel spectrogram and spectrogram of harmonic and percussive component means, spectrogram of delta values

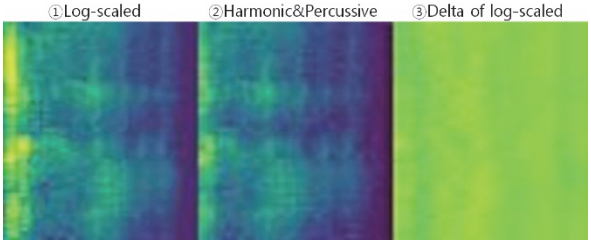

Figure 2 One image that combines ①, ②, and ③

3.3 Transfer Learning Model

3.3.1 Transfer Learning

Transfer learning leverages the knowledge gained during training on a large, general dataset to improve the performance on a specific task or domain. This approach uses learned representations or features extracted by deep neural networks trained on large datasets, such as ImageNet. These features capture general patterns and visual representations that can be transferred to another task. By utilizing transfer learning, the rich knowledge obtained through pre-training on ImageNet can be harnessed to achieve better performance with audio data. This is a practical and efficient solution for research, as it can save computation time and resources.

3.3.2 Transfer Learning Model

In this study, the VGG-16 model was chosen as the base architecture for transfer learning. VGG-16 is a CNN model widely used for various image classification tasks. The VGG-16 model was initialized with pretrained weights obtained from training on the ImageNet image dataset. Using these pretrained weights, the model has already learned a set of general visual representations that can be useful for dementia classification using audio data.

3.3.3 Fine-Tuning

To adapt the VGG-16 model to a specific task, fine-tuning was performed, which involved retraining the model on the target dataset while keeping the initial layers frozen to preserve the general features learned from ImageNet. Using this method, the model can capture patterns and features associated with dementia classification and learn task-specific representations from audio data. Fine-tuning was applied to the weights of the VGG–16 model to classify the mel-spectrogram images generated from audio data. This process includes keeping the earlier layers of the model in a frozen state to preserve the learned image features while unfreezing the last few layers of the model and training them on mel-spectrogram images derived from audio data (Fig. 3). In the experiments conducted in this study, fine-tuning was performed, as depicted in (Fig. 4), where the network was frozen up to the fourth layer,

and only the convolutional block of the fifth layer was unfrozen to update the weights. For this purpose, preprocessed mel spectrogram data were used as inputs during training. Through fine-tuning, the model learns from mel-spectrogram images derived from the dementia audio data, allowing the weights to be updated accordingly.

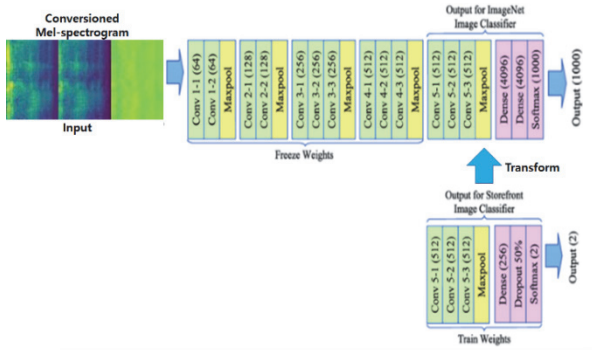

Figure 3 Fine-tuning method of VGG-16 to receive mel spectrogram images as input and learn only the last layer [25]

3.4 K-Fold Validation

To ensure the robustness and generalization of the proposed approach, we employed K-Fold Cross-Validation during the training and evaluation of the deep learning model. This technique allows for the evaluation of the model's performance across multiple subsets of the dataset, addressing potential biases or overfitting issues. In this paper, we utilized 5-fold cross-validation. We divided the preprocessed dataset into five equal-sized subsets, ensuring a balanced distribution of samples for each class in each subset. Subsequently, we performed five rounds of training using four folds for training and one-fold for evaluation (as illustrated in Fig. 3). In each training iteration, we randomly selected four folds as the training set and used the remaining fold as the validation set. This approach mitigates the risk of overfitting and provides insights into the generalization capabilities of the proposed approach. The evaluation results obtained from this cross-validation method enhance the reliability of the research findings.

3.5 Grad-CAM and Feature Selection

3.5.1 Grad-CAM (Gradient-Weighted Class Activation Map)

Grad-CAM is an attention-based visualization technique that enables an understanding of the regions in the input data that contribute significantly to the classification decision. Unlike CAM [26], Grad-CAM generalizes to CNNs without requiring global average pooling layers and instead utilizes fully connected layers (Fig 4).

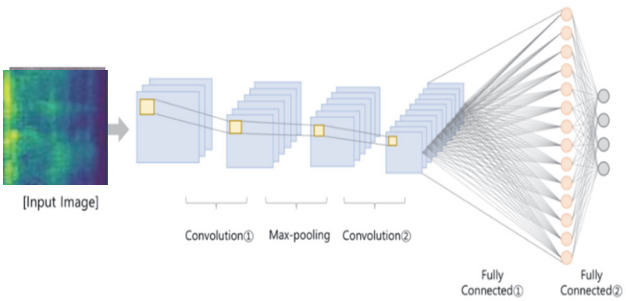

Figure 4 CNN structure using Grad–CAM [27]

The Grad-CAM algorithm generates a heatmap that emphasizes the input regions related to classification, visually highlighting the salient features. The attention map is derived by calculating the gradients of the target class based on the final convolutional feature map of the model. These gradients act as importance weights when multiplied by the feature map to yield class activation maps (Fig. 5).

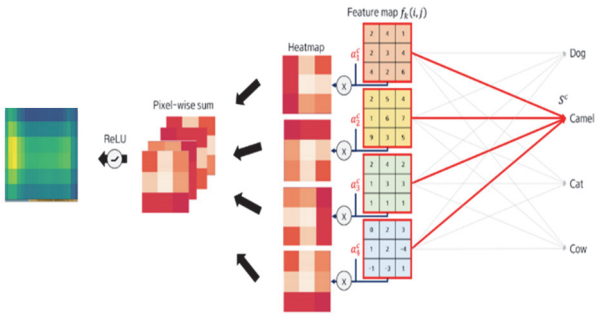

Figure 5 Overall structure for obtaining heat maps using Grad-CAM [27]

The formula for computing the gradients and weights to obtain the heatmap using Grad-CAM, denoted by Eq. (1), is as follows:

$$L^{C}_{GRAD}(i,j) = ReLU\left(\sum_{k} a^{c}_{k} f_{k}(i,j)\right), a^{c}_{k} = \\ = \frac{1}{z}\sum_{i}^{k}\sum_{j}\frac{\partial S^{c}}{\partial f_{k}(i,j)} \quad (1)$$

In Eq. (1), a^{c}_{k} represents the average influence each element i, j from the k-th feature map $f_k(i, j)$ within the Z channels has on the output class matrix value S^c for class c. Essentially the weight of each feature map is given by its gradient. Consequently, a^{c}_{k}, which represents the pixel-wise average of the gradients, was multiplied by each feature map, $f_k(i, j)$, to produce a heatmap. The resulting heatmap was then summed across its pixels and subjected to a ReLU function to selectively highlight regions with positive weights deemed to be of significance. This entire process culminated in what is known as Grad-CAM. The procedure for Grad-CAM is as follows (Tab. 2):

Table 2 Procedure for Grad-CAM

Procedure Grad–CAM
Input: image, model, target_class 1. model = load_pretrained_model() #mapping last convolution layer output 2. feature_maps = model.forward(image) 3. target_class = target_class_number 4. target_score = model.output[target_class] 5. gradient = backward(target_score) #calculate gradient 6. pooled_gradient = reduce_mean_gradients 7. feature_weight = calculate_weights(feature_maps, pooled_gradients) 8. weighted_feature_maps = weight_combination(feature_maps, feature_weights) #calculate importance of feature 9. heatmap = relu_activation(weight_feature_maps)

Grad-CAM uses a pretrained CNN model, an input image, and a target class as input arguments. It loads a pretrained CNN model and employs a forward function to map the input image

to the predictions of the last activated convolutional layer of the called model. The gradients of the target class concerning the activations of the last convolutional layer with respect to the input image are then calculated using a backward function. This gradient was averaged across all channels of the feature map. These average gradients served as weights that revealed the importance of each channel for the target class. These weights were then multiplied by each channel of the feature map array and summed across all channels, followed by the application of a ReLU activation function to produce the heatmap. The procedure for applying Grad-CAM was as follows. Employ the VGG-16 model, trained on either Mel-spectrogram images or concatenated images (Fig. 6), and apply the Grad-CAM algorithm using the model weights to generate heatmaps (Fig. 6). Grad-CAM accentuates the regions of the input image that contribute significantly to the model's predictions by highlighting them in a heatmap. Grad-CAM leverages the gradient information flowing into the final convolutional layer of the model to determine the importance of each spatial location. Through this, we can validate the relevance of the identified acoustic features and better understand the discriminative patterns on which the model relies.

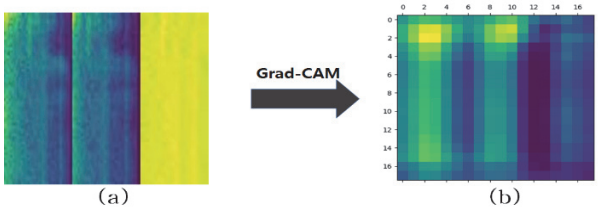

Figure 6 Image (a) combined with three mel spectrograms defined in (Fig. 2) and image (b) after Grad-CAM is applied

Furthermore, we explored the potential of utilizing the Grad-CAM attention maps as an additional input feature for the classification model. This involves creating a composite input representation that combines the original Mel spectrogram images with attention maps, thereby emphasizing the regions of interest identified by Grad-CAM. This aims to selectively train the model on the features highlighted by Grad-CAM with the expectation of achieving higher accuracy and training efficiency.

3.5.2 Feature Selection

Feature selection is a technique aimed at identifying the most relevant and useful features from the original set of input variables to improve classification accuracy while removing redundant or irrelevant features. In this study, two feature selection strategies were explored. The first strategy involves selecting individual Mel spectrograms as input features and comparing them with a method that combines spectrograms and attention maps using Grad-CAM, thereby selecting only high-importance features for training. Both approaches aim to identify the most discriminative acoustic features related to dementia and determine the optimal input representation for the classification model. Initially, we evaluated the impact of individually processing each Mel spectrogram image and using them as inputs, as well as the effect of merging these individual images into a single input by comparing the two approaches. The second strategy involves a feature-selection method using Grad-CAM. The feature values were selected from the regions with the highest activation values in the Grad-CAM-generated heatmap. The selected heatmap was structured as a two-dimensional NumPy array. From this array, the top n features with the highest importance (activation values) were selected, whereas the remaining features were set to zero to prevent them from impacting the learning process (Fig. 7). This model is referred to as the Grad-CAM Feature-selected with Mel-spectrogram (GF-M).

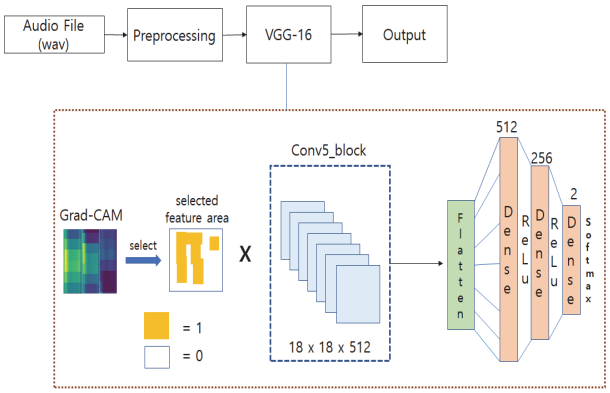

Figure 7 Process of the GF-M model

4 EXPERIMENT

4.1 GF-M Model

The experimentation workflow using the GF-M model when a tripartite mel-spectrogram image was ingested as the input is illustrated in (Fig. 8).

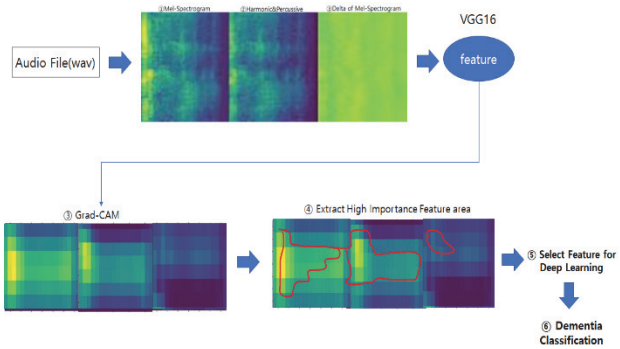

Figure 8 Progress of GF-M when an image combined with three mel spectrograms is received as input

The operational steps of the GF-M are detailed in Tab. 3. Feature Region Definition using Grad-CAM Heatmap: An initial function is defined to extract the top feature regions from the heatmap generated by Grad-CAM (1). This function linearizes the heatmap to one dimension and sorts it (1.1). Subsequently, the top N features were selected (1.2). Any features outside the chosen region are set to zero, ensuring they do not contribute to the learning process, while the selected feature region is set to one (1.3 - 1.5.2). Fine-Tuning the VGG16 Transfer Learning Model:

- A pre-trained VGG16 model is defined (2.1).
- The Grad-CAM method is initialized (2.2).
- Using NumPy, the heatmap produced by Grad-CAM was transformed into a two-dimensional array. This array is then fed to the function designed to extract the top feature regions (2.3).
- The resulting heatmap, highlighting the top features, was

multiplied by the 512 feature maps of the last convolution layer. This ensures that features outside this region do not influence the learning process (2.6).

- The feature-selected 512 maps are flattened and linked to a fully connected layer (2.7).
- Subsequently, they were connected to dense layers to form the transfer learning model. The dense layers consist of 512, 256, and 2 nodes, with activation functions being ReLu, ReLu, and Softmax respectively (2.8 - 2.11).

Training the Model: Finally, the defined transfer learning model is invoked, and the training process is executed (3).

Table 3 Procedure of GF-M

```
Procedure GF-M
Input:
Mel-spectrogram image IMG

1.    feature_area_select(heatmap)
        #make heatmap one dimention
1.1.    one_dimention_heatmap = calculate(heatmap).tolist()
1.2.    sorted_heatmap = one_dimention_heatmap.sort()
          #select top N feature's area in heatmap
1.3.    N = number of parent features to select
1.4.    if:
1.4.1.      heatmap_element > sorted_heatmap[N]
1.4.2.      heatmap_element = 1
1.5.    if:
1.5.1.      heatmap_element < = sorted_heatmap[N]
1.5.2.      heatmap_element = 0
1.6.  returnselected_feature_area_heatmap

2.    transfer_model()
2.1.    load VGG16
2.2.    call Grad-CAM(VGG16, IMG, 0)
2.3.    call feature_area_select(heatmap)
2.4.    initialization VGG16
2.5.    VGG16.trainable = True
2.6.    feature_map_after_selected =
            multiply(selected_feature_area_heatmap,
          VGG16.last_convolution_layer_feature_map)
        #Connect pre-processed feature maps to a FC Layer
2.7.    x = flatten()(feature_map_after_selected)
2.8.    x = dense(256, activation = 'relu' )(x)
2.9.    x = dropout(0.5)(x)
2.10.   x = dense(128, activation = 'relu' )(x)
2.11.   x = dense(2, activation = 'softmax')(x)
2.12.  return transfer_model =
        Model(inputs = VGG16.input, ouputs = x)
3.    while(1) #reapeat learning
3.1.    calltransfer_model()
3.2.    definetrain_data, test_data
3.3.    transfer_model.compile
3.4.    transfer_model.fit
```

4.2 Comparative Experiment Configuration

This study aimed to assess the impact of feature selection using Grad-CAM on dementia classification accuracy. In addition, this research aimed to identify unique characteristics, such as the pitch and rhythm of each voice, to contribute to the development of an enhanced dementia classification algorithm. Thus, a comparative experiment was conducted, as presented in Tab. 4.

An experiment was conducted using four input configurations and two classification algorithms. The classifications were performed under three categories: Normal (N) vs. Alzheimer's Disease (AD), Normal (N) vs. Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD) vs. Mild Cognitive Impairment (MCI), resulting in 24 runs. The decision to conduct each of these as a binary classification rather than a ternary classification stems from previous studies suggesting that the accuracy and reliability of a three-class classification, including MCI, tend to be lower [28-32].

Table 4 Comparative Experiment Configuration

Input	Classification Algorithm
Mel Spectrogram	Using only VGG-16 fine-tuning
	Application of feature selection with Grad-CAM
Average of Harmonic & Percussive values	Using only VGG-16 fine-tuning
	Application of feature selection with Grad-CAM
Delta value spectrogram of the Mel Spectrogram	Using only VGG-16 fine-tuning
	Application of feature selection with Grad-CAM
Image combining the three spectrograms	Using only VGG-16 fine-tuning
	Application of feature selection with Grad-CAM

4.3 Sequential Experimental Comparison

The results were sequentially compared based on the experimental configuration. First, we analyzed the differences in classification accuracy based on the voice data input before feature selection (Fig. 9).

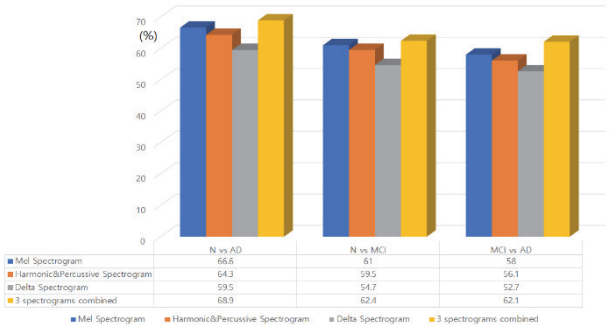

Figure 9 Comparison of classification accuracy based on input prior to feature selection

Upon observing (Fig 9), it is evident that the combined image input showcased the highest classification accuracy. This was followed by the Mel Spectrogram, the Harmonic and Percussive Spectrogram, and then the Delta Spectrogram, in decreasing order of accuracy. This indicates that while the Mel Spectrogram inherently possesses the best features, other transformations also hold valuable attributes for classification. Various vocal characteristics are useful for the classification of dementia. Next, we compared the classification accuracies after feature selection (Fig. 10).

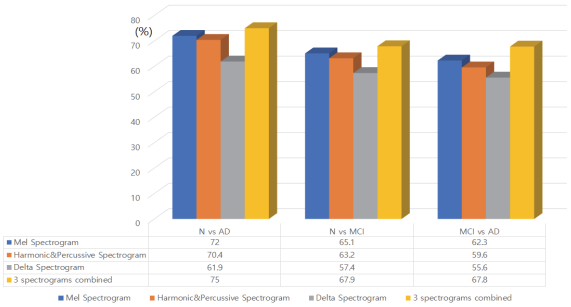

Figure 10 Comparison of classification accuracy based on input after feature selection

The results demonstrated a similar trend: the combined image exhibited the highest accuracy, whereas the delta

showed the lowest accuracy. The efficacy of the feature selection was further validated by comparing the increase in accuracy for each input and classification (Fig 11) and (Fig 12).

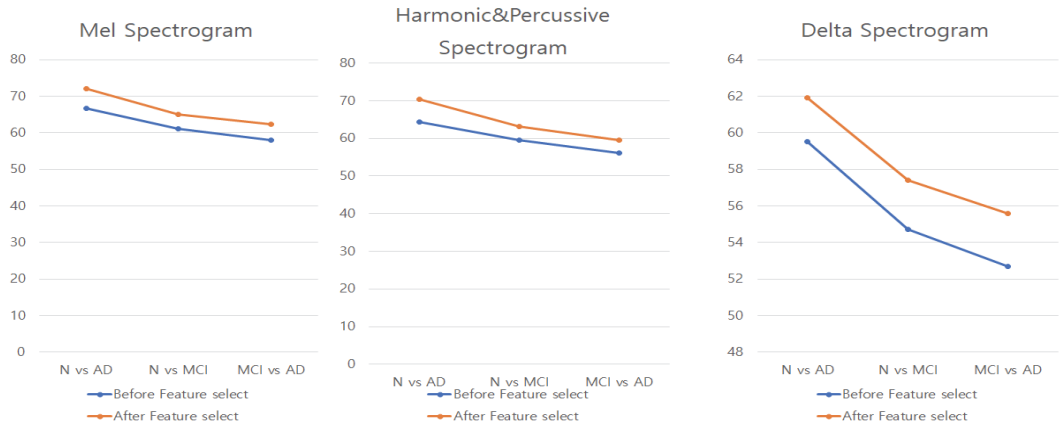

Figure 11 Differences in accuracy before and after feature selection based on three individual inputs

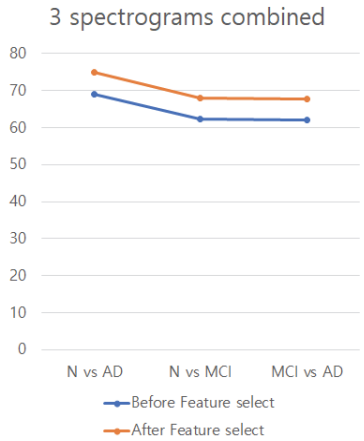

Figure 12 The difference in accuracy before and after feature selection when using a combined image of three spectrograms as an input

After feature selection, every classification showed an increase in accuracy ranging from a minimum of 2.9% to a maximum of 6.1%. This underscores the efficacy of the feature selection using Grad-CAM. In addition, the overall F1 score for the experiments was 0.8552, confirming the reliability of the results.

4.4 Validation against Existing Models

A comparative experiment was conducted within the GF-M framework using transfer learning models other than VGG16. The comparison focused solely on the classification of SCI and AD, which exhibited the highest accuracy. The experiment was divided based on the application of feature selection using Grad-CAM, both before and after, and the results were subsequently compared. Consequently, VGG16 was used because the absolute accuracy level was higher for VGG16.

Table 5 Classification accuracy with Xception and Restnet 50

Model/Classification Accuracy / %	Before Feature Selection	After Feature Selection
Xception	60.84	62.61
Resnet50	60.76	63.2

5 DISCUSSION

The results of this study offer valuable insights into dementia classification based on vocal features and open doors for further exploration and validation. One prominent finding was the superior classification accuracy achieved with the combined image inputs. This suggests that while individual spectrogram types such as Mel, Harmonic & Percussive, and Delta bring their unique attributes to the table, a composite approach might yield a more comprehensive representation of the vocal features. This approach effectively leverages multiple facets of the voice, which can be instrumental in achieving nuanced and precise classifications. It may be worthwhile for future research to explore other potential combinations or transformations that could enhance classification. Feature selection, as observed, was pivotal in boosting the accuracy of the classification. Therefore, the efficacy of Grad-CAM in this regard is commendable. Not only does it streamline the features, making the model focus on the most relevant parts, but it also amplifies the potential of each voice input. This sheds light on the underlying dynamics of voice features in patients with dementia; there could be specific patterns or characteristics that are more prevalent or pronounced in patients with dementia than in others. Another intriguing aspect is the comparative performance of the GF-M framework with other prevailing models, such as Xception and Resnet50. The fact that even these renowned models showed a spike in performance after feature selection is a testament to the universal applicability of Grad-CAM's feature selection methodology across diverse architectures. However, one must proceed cautiously. Although the results are promising, it is essential to consider the diversity of voice samples, potential biases in the data, and variability in the progression of dementia across individuals. A larger, more diverse dataset could offer a more holistic view and help generalize the findings. Additionally, the intricate interplay of factors such as age, gender, ethnicity, and underlying health conditions can influence voice characteristics. Future studies could focus on understanding these nuances and their implications for dementia classification. In conclusion, this discussion emphasizes the potential and challenges of using voice as a diagnostic tool for dementia and highlights avenues for future

exploration. To push the boundaries of technology and medical research, it is important to ensure that findings are grounded, validated, and ethically employed.

6 CONCLUSION

The primary focus of this study was to ascertain the influence of feature selection using Grad-CAM on the accuracy of dementia classification. Through rigorous experimentation, this study aimed to understand the unique characteristics of each voice, such as pitch and rhythm, and their implications for enhanced dementia classification algorithms. From our experiments, it was evident that using combined image inputs yielded the highest classification accuracy, whereas distinct spectrogram types (Mel, Harmonic & Percussive, Delta) showed varying levels of efficacy. The overarching conclusion, however, is that while the Mel Spectrogram inherently contains the most distinctive features, other transformations also significantly contribute to the classification process. This is indicative of the robustness and multifaceted nature of vocal characteristics when distinguishing between dementia and other conditions. A pivotal discovery was the improved accuracy across all classifications after incorporating feature selection using Grad-CAM. The increase in accuracy, ranging from 2.9% to a maximum of 6.1%, and an overall F1 score of 0.8552 emphasized not only the effectiveness of Grad-CAM but also the reliability of the results generated. Moreover, when juxtaposed with existing models, the GF-M framework with its adaptive transfer learning mechanism demonstrated competitive and promising results. The fact that models such as Xception and Resnet50 showed enhanced accuracy after feature selection corroborates the pivotal role of feature selection in enhancing model performance. In summary, this study underscores the importance of vocal characteristics in dementia classification and the transformative potential of feature selection using Grad-CAM. As the medical community continues its endeavors toward the early detection and treatment of dementia, leveraging such advanced methodologies and insights will become indispensable. Future studies will aim to produce figures of absolute accuracy and a greater growth rate, which are the limitations of current studies.

Acknowledgments

This work was partly supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korean government (MOTIE) (20214000000060, Department of Next Generation Energy System Convergence based on Techno-Economics - STEP).

8 REFERENCES

[1] Vrijsen, J., Matulessij, T. F., Joxhorst, T., de Rooji, S. E., & Smidt, N. (2021). Knowledge, health beliefs and attitudes towards dementia and dementia risk reduction among the Dutch general population: a cross-sectional study. *BMC Public Health, 21*, 857. https://doi.org/10.1186/s12889-021-10913-7

[2] Durongbhan, P., Zhao, Y., Chen, L., Zis, P., De Marco, M., Unwin, Z. C., Venneri, A., He, X., Li, S., Zhao, Y., Blackburn, D. J., & Sarrigiannis, P. G. (2019). A Dementia Classification Framework Using Frequency and Time-Frequency Features Based on EEG Signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 27*(5), 826-835. https://doi.org/10.1109/TNSRE.2019.2909100

[3] Bharati, S., Podder, P., Thanh, D. N. H., & Prasath, V. S. (2022). Dementia classification using MR imaging and clinical data with voting based machine learning models. *Multimedia Tools and Applications*, *81*(18), 25971-25992. https://doi.org/10.1007/s11042-022-12754-x

[4] Xue, C., Karjadi, C., Paschalidis, I. C., Au, R., & Kolachalama, V. B. (2021). Detection of dementia on voice recordings using deep learning: a Framingham Heart Study. *Alzheimer's research & therapy*, *13*, 1-15. https://doi.org/10.1186/s13195-021-00888-3

[5] Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., & Christensen, H. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, *58*(2), 373-387. https://doi.org/10.3233/JAD-160507

[6] Ahmed, S., Haigh, A. M. F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, *136*(12), 3727-3737. https://doi.org/10.1093/brain/awt269

[7] Forbes-McKay, K. E. & Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological sciences*, *26*, 243-254. https://doi.org/10.1007/s10072-005-0467-9

[8] Luz, S., de la Fuente, S., & Albert, P. (2018). A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*.

[9] Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease, 49*(2), 407-422. https://doi.org/10.3233/JAD-150520

[10] Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., & Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. *Interspeech*, 1692-1696. https://doi.org/10.21437/Interspeech.2013-32

[11] Chakraborty, K., Talele, A., & Upadhya, S. (2014). Voice recognition using MFCC algorithm. *International Journal of Innovative Research in Advanced Engineering (IJIRAE), 1*(10), 2349-2163.

[12] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *SciPy*, 18-24. https://doi.org/10.25080/Majora-7b98e3ed-003

[13] Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, 1-6. https://doi.org/10.1109/MLSP.2015.7324337

[14] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618-626. https://doi.org/10.1109/ICCV.2017.74

[15] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaise, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[17] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data, 3*, 1-40. https://doi.org/10.1186/s40537-016-0043-6

[18] Chandrashekar, G. & Sahin, F. (2014). A survey on feature selection methods. *Computers & electrical engineering, 40*(1), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024
[19] Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
[20] Gui, N., Ge, D., & Hu, Z. (2019). AFS: An attention-based mechanism for supervised feature selection. *Proceedings of the AAAI conference on artificial intelligence, 33*(01), 3705-3713). https://doi.org/10.1609/aaai.v33i01.33013705
[21] Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, 1-6. https://doi.org/10.1109/MLSP.2015.7324337
[22] https://www.image-net.org/
[23] https://aihub.or.kr/aidata/34103
[24] https://www.aihub.or.kr/
[25] https://www.researchgate.net/ /The-network-architect ure-for-fine-tuning-on-VGG16-The-weights-of-VGG16-are-pre-trained_fig3_327090029
[26] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921-2929. https://doi.org/10.1109/CVPR.2016.319
[27] https://tyami.github.io/deep%20learning/CNN-visualization- Grad-CAM/
[28] Saleem, T. J., Zahra, S. R., Wu, F., Alwakeel, A., Alwakeel, M., Jeribi, F., & Hijji, M. (2022). Deep learning-based diagnosis of Alzheimer's disease. *Journal of Personalized Medicine, 12*(5), 815. https://doi.org/10.3390/jpm12050815
[29] Bozzali, M., Giulietti, G., Basile, B., Serra, L., Spano, B., Perri, R., Giubilei, F., Marra, C., Caltagirone, C., & Cercignani, M. (2012). Damage to the cingulum contributes to Alzheimer's disease pathophysiology by deafferentation mechanism. *Human Brain Mapping, 33*(6), 1295-1308. https://doi.org/10.1002/hbm.21287
[30] Bubb, E. J., Metzler-Baddeley, C., & Aggleton, J. P. (2018). The cingulum bundle: anatomy, function, and dysfunction. *Neuroscience & Biobehavioral Reviews, 92*, 104-127. https://doi.org/10.1016/j.neubiorev.2018.05.008
[31] Kwai, C. K., Subramaniam, P., Razali, R., & Ghazali, S. E. (2019). The usefulness of digital memory album for a person with mild dementia. *Int. J. Adv. Nurs. Educ. Res, 4*, 1-12. https://doi.org/10.21742/ijaner.2019.4.1.01
[32] McCloskey, R., Skerry, L., Keeping-Burke, L., Donovan, A., Donovan, C., & Scheme, E. (2019). Caring for self while caring for others: impact of wearable health monitoring devices on self-care of family/friend caregivers of individuals living with a dementia. *International Journal of Advanced Nursing Education and Research, 4*(3), 69-80. https://doi.org/10.21742/IJANER.2019.4.3.10

Contact information:

Hansol KO, Md
Gachon University, South Korea
E-mail: kohanasol@naver.com

Bohyun WANG, PhD
Gachon University, South Korea
E-mail: bhwang99@gachon.ac.kr

Joon S. LIM, Professor
(Corresponding author)
Gachon University, South Korea
E-mail: jslim@gachon.ac.kr