A Pre-trained Neural Network to Predict Alzheimer's Disease at an Early Stage

Ragavamsi Davuluri, Ragupathy Rengaswamy
Department of Computer Science and Engineering
Annamalai University, Chidambaram, Tamil Nadu - 608001

***Abstract*—Alzheimer's disease (AD), which is a neuro associated disease, has become a common for past few years. In this competitive world, individual has to perform lot of multi tasking to prove their efficiency, in this process the neurons in the brain gets affected after a while i.e., "Alzheimer's Disease". Existing models to identify the disease at early stage has taken the individuals speech as input then they are converted into textual transcripts. These transcripts are analyzed using neural network approached by integrating them with NLP techniques. These techniques failed in designing the model which can process the long conversation text at faster rate and few models are unable to recognize the replacement of the unknown words during the translation process. The proposed system addresses these issues by converting the speech obtained into image format and then the output "Mel-spectrum" is passed as input to pre-trained VGG-16. This process has greatly reduced the pre-processing step and improved the efficiency of the system with less kernel size architecture. The speech to image translation mechanism has improved accuracy when compared to speech to text translators.**

Keywords—Mel-spectrum; VGG-16; ADAM optimizer; softmax; flatten layers; ReLU

I. Introduction

Alzheimer's disease is a progressive brain disease that gradually deteriorates memory and thinking abilities, as well as the ability to do even the most fundamental tasks. The majority of people with late-onset type symptoms are in their mid-60s when they get the disease [6]. Early-onset Alzheimer's disease is extremely rare and occurs between the ages of 30 and 60. The most prevalent cause of dementia in elderly people is Alzheimer's disease [20]. Memory issues are usually one of the early signs of Alzheimer's disease, though the severity of the symptoms varies from person to person [7]. Other areas of thinking, such as finding the proper words, vision/spatial difficulties, and impaired reasoning or judgments, may also indicate Alzheimer's disease in its early stages. There is no cure or treatment for Alzheimer's disease that affects the disease process in the brain. Complications from severe loss of brain function, such as dehydration, malnutrition, or infection, result in death in advanced stages of the disease [13]. Alzheimer's disease can be detected using a machine learning approach, which involves the use of various machine learning algorithms [18]. Furthermore, the patient's severity level will be predicted in percentages, and the percentage levels will be divided into several categories. The importance of early detection in Alzheimer's disease management cannot be overstated [14]. Convolutional neural network (CNN) is one of the deep-learning algorithms that have been used to detect structural brain alterations on magnetic resonance imaging (MRI) because of its high efficiency in automated feature learning [15]. Many additional deep learning methods are being utilized to diagnose Alzheimer's disease [17] and even pre-trained models can be used in the detection of Alzheimer's disease [19].

The proposed model uses the audio dataset to improve the accuracy of the model by analyzing the live streaming data in case of Alzheimer's disease prediction [8]. The model produces a spectrogram from audio files, which produces visualization of signal strength in the form of 2D graph. The model needs to do two pre-processing steps before converting it into image [23].

The audio data is stored in the digitized format. Any machine learning algorithm is difficult to work with this digital form. So machine needs sampling mechanism to convert the digital data into analog data [9]. Sampling technique transforms the signal with respect to time into numerical values by identifying the difference between two consecutive samples of audio segments [10].

The obtained sample may contain noisy values i.e., at few intervals the obtained time signals may have amplitude has zero, which represents the state of silence [21]. The model address this issue by performing quantization that helps them to replace silence with nearest precision value and then it normalizes the data to have values in between -1 to 1 [24].

The model needs frequencies to identify the voice modulation, so it applies Fourier transformations to convert the time signals into individual frequencies. The frequencies can be produced into two ways namely FFT & STFT. The model employs STFT because it can efficiently convert the 1-dimensional data into 2-dimensional data where horizontal axis represents time and vertical axis represents frequency [11]. In short note, the sound waves are segmented into smaller chunks and few of them might over lap. In the final stage, each frequency amplitude in decibels is stored as a "pixel" value of the image. This conversion process is known as "Mel Scale" [12]. These pixels need to be stored in the form of vectors. The spectrogram represented in Mel Scale is a Mel Spectrogram [22].

II. Literature Survey

Rohanian et al. proposed Multi-modal fusion on sequential modeling using LSTM technique. These fusion models are good in handling the lexical information. Since this context changes with time, the model uses gating concept. This receives inputs from two different models i.e., (audio and text) and combines them into single unit by eliminating the noises from the audio data acquired. All the higher layers utilize non-linear functions but the lower layers use linear functions. The non-linear is transformed into linear by attaching the carriers integrated with self repairing system. The model hyper tunes the LSTM to identify the error co-efficient at early stage [1].

Raghavendra et al. designed embedded techniques related to speech and fine tune those using BERT models. The model extracts the essential features by constructing the x-vectors, which is a deep neural vector. In the next step, Mel-frequency coefficient is computed by embedding all the global pooling layers. The encoder of the ResNet stores the signals in the encrypted format but at decoder side it implements a pre-trained VoxCeleb1. The performance of this model is estimated using the MMSE because most of the characteristics are relevant to prosody features. These features are hard to maintain, so BERT model using its self attention layer which converts the linguistics elements into embeddings [2].

Ning Wang et al. implemented Attention Network by collecting real time data from Google Speech Recognizer API. The model extracts four different features using four different networks. The model for extracting the frame level features of the linguistic, it implements VGG network because it is good in handling the embossed features. FREQ commands convert the received audio into transcript form. NLTK takes care of the converted text but the context of each sentence is extracted using the Universal Sentence Embedding technique [3]. The model implements dilated CNN layers instead of 2D-CNN layers because multi head component associated with the embedded layers can represent the encoded representation very efficiently [16].

Amish et al. designed a framework BERT integrated CNN to extract the embedding text associated with the context. The audio dataset is initially segmented into shorter clips and then Mel-spectrum is generated for those segments. The model utilized Fast Text-CNN to generate transcripts and clusters are formed based on the common word vectors. In this model, instead of encoding all the sentences, it encrypts only Out-of-Vocabulary words using sentence BERT technique. During the text processing phase, the probability of segment is compared against the probability of entire transcript and then combined to form a fusion. Finally, a concatenation embedding model is designed to treat the each text segments separately [4].

Zhaoci Liu et al. worked with bottle neck features generated from augment images. The model doesn't convert the audio signals into transcripts because the designed extractor creates temporary intermediate form of representation. The audio signals are divided into set of window frames, from which both local and global context information is extracted. This information is stored as a sequence in LSTM and attention pooling is applied to perform the classification. The validation of the model is performed using query system, which is designed as feed forward network with sigmoid activation function and the input is obtained in the form of key-value pair [5].

III. Proposed Methodology

Most of the researchers analyzed the Alzheimer's disease either from images or from the .csv file extracted from the images. Few researchers, which are stated in Table I, worked on predicting system using audio dataset, which is converted into textual format. The accessing and processing of text using NLP techniques takes a lot of time, which makes the system to take late decisions. This issue is resolved in this paper by transforming the speech into images and processing them using pre-trained model. The entire process is illustrated in Fig. 1.

TABLE I. Identification of Limitations from Existing Model

S.No	Author	Algorithm	Merits	Demerits	Accuracy
1	Rohanian et al. [1]	Multi-modal fusion	The feed-forward helps the model to transfer the data quickly	The model can be extended by introducing the bio-markers	79.20
2	Raghavendra et al. [2]	BERT	Embedding BERT and X-vector has solved the modulation frequencies very effectively during speech recognition	Adaptable BERT integrated with LM interpolation will refine the predictions associated with target values	84.51
3	Ning Wang et al. [3]	Attention Network	Instead of single feature extraction, the model extracted multiple features from different sources. So, this model achieves less misclassification rate	The model doesn't apply any segmentation technique to process the speech with respect to time	80.28
4	Amish et al. [4]	Multi Modal Sequence	The FastText CNN has the capability to generate transcripts for unknown words also	The model has implemented late fusion technique to combine the text segments which consumes more memory to activate the cells	85.30
5	Zhaoci Liu et al. [5]	Masked data augmentation	Since the bottleneck features extracts the low level values, it requires less space and time to encrypt	Usage of query system for validating the model made the process complicated	82.59

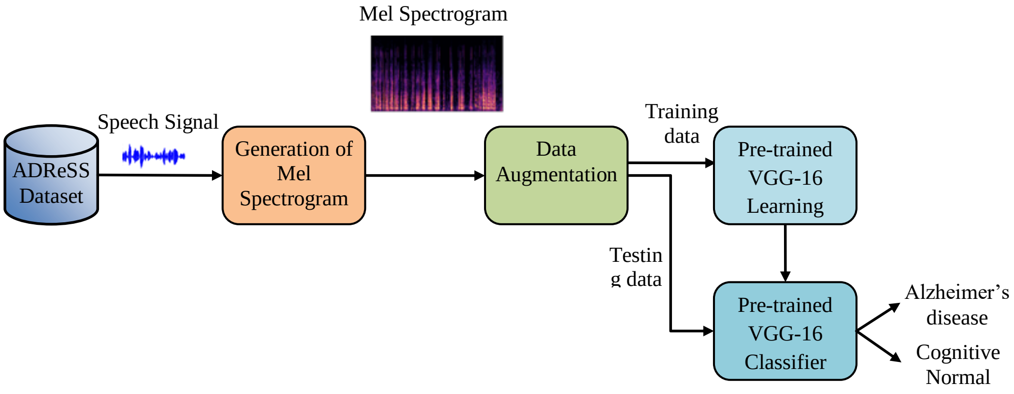


Fig. 1. Overall Architecture of Proposed System.

A. Conversion of Speech Signals to Mel-spectrum Images

The proposed research focused to analyze the disease from the speech, which are basically classified into Alzheimer's disease (AD) and Cognitive normal (CN). Initially the model has constructed the Mel spectrogram from the speech files as represented in Fig. 2.

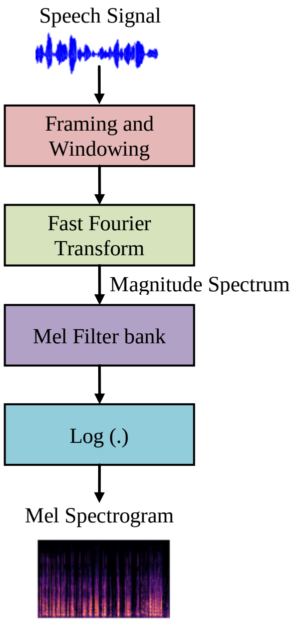


Fig. 2. Generation of Spectrogram Images from Audio Dataset.

The approach employs samples of air pressure over time to digitally represent a speech signal. The voice signal is transferred from the time domain to the frequency domain using the fast Fourier transform, and the system uses overlapping windowed portions of the speech stream. The technology converts the y-axis (frequency) to a log scale and the color dimension (amplitude) to decibels to create the spectrogram. The y-axis (frequency) was mapped onto the Mel scale to generate the Mel spectrogram. The model applies a traditional Short Time Fourier Transformation (STFT); this is a very flexible class to represent time and frequency distribution from the processed speech signals. The identification of pitch variation is the major hyper tuning point. The computation is presented in equation (1).

$$STFT_m(input) = \sum_{n=\infty}^{-\infty} input_n * [H(n) - R_m] * e^{-j} \quad (1)$$

where, $input_n$ denotes input signal recognized at time 'n' H (n) denotes N-length Hamming function applied on sliding window

R_m Represents hop size in between the m-size sliding window

e^{-j} is a threshold multiplication function

Humans do not perceive frequencies on a linear scale, according to research. Lower frequency differences are easier to notice than higher frequency variances. Humans can readily distinguish between 500 and 1000 Hz, but we will struggle to distinguish between 10,000 and 10,500 Hz, despite the fact that the distance between the two pairs is same. As a result, we'll utilize the Mel scale, which is a logarithmic scale based on the idea that equal lengths on the scale correspond to the same perceptual distance. Conversion from frequency (f) to Mel scale (m) is given in equation (2).

$$m = 2595.log(1 + \frac{f}{500}) \quad (2)$$

A Mel Spectrogram is a spectrogram that converts frequencies to Mel scale. For the creation of log-Mel spectrograms, we chose 40 Mel filter banks. The availability of 40 Mel filter banks allows us to use pre-trained models such as VGG16 in the future. A Hamming window with a size of 2048 samples is used. With a hop-length of 1024 samples (the amount of samples between subsequent frames) and a sampling rate of 44.1 kHz. In the FFT calculation, the number of points is also 2048. The log-Mel spectrograms are separated into overlapping chunks of 100 frames each encompassing a length of 23ms to achieve this.

Algorithm for Construction of Mel Spectrum

```
Input: Audio Dataset with binary class labels, AD_Binary
Output: Mel-spectrum Generation
Begin
1. Set the path of the output folder to store the images of spectograms
2. ad_binary_labels←[“AD”, ”CN”]
3. for i in ad_binary_labels: for j in path:
        a. audio_time_series,audio_srate←load(j,sr=None)
        // load the speech data in time series format with the specified sampling rate
        b. speech_frequency←stft(speech_time_series)
        c. speech_magintude, speech_phase ← magphase (speech_frequency)
        d. mel_scale_speech ← melspectogram (S=speech_magnitude,sr=speech_srate)
        e. mel_speech← amplitude_to_db (mel_scale_speech,ref=”minimum”)
        f. specshow(mel_speech,sr=speech_srate)
End
```

B. *Disease Classification using VGG-16*

The proposed system applies VGG-16 pre-trained model on generated spectrogram images to identify the disease. The model has chosen VGG network because of its simplicity nature even though it has huge parameter to hyper tune. The basic thumb rule for any deep learning algorithms is “More the training data more the accuracy”, but the training dataset has got fewer images from the speech signals. This issue is resolved by the model initially by applying basic image manipulation operations to increase the size of the generated spectrogram images as shown in Fig. 3.

The model has customized the few operations in the generator module to minimize the error rate. The basic customized operation is “pre-processing” unit because it de-noises the images, which is a basic challenge faced by any of the computer vision applications. In this model, images are generated from the speech signals; therefore there are high chances to get noisy images due to sudden voice modulation from the external factors. The remaining customization operations are elaborated in Table II.

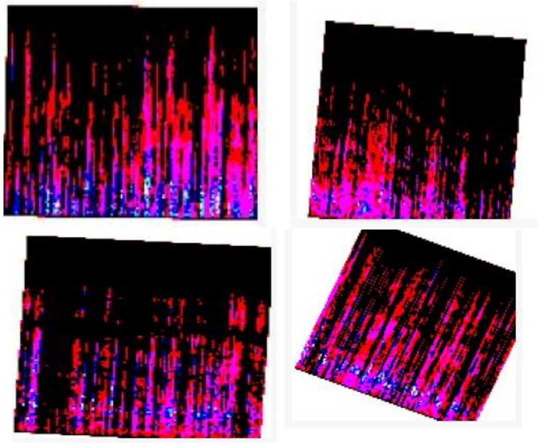


Fig. 3. Synthetic Data Created using the Image Data Generator Module.

The main advantages of the generator module lies in passing the multiple operations performed on the image are sent directly to the neural network instead of storing them in a temporary buffer or memory. Finally, the images are rescaled to 150 and divided them into 32 batch size. These synthesized images are passed as input to the VGG-16 neural network, whose task is achieved in two stages. In the first stage, it identifies the objects from the generated signals from the pre-defined classes available. In the second stage, the model contains 1000 class labels, so it has to classify the generated image from the customized class label. The model implemented its validation across Image Net dataset. The model implements only a kernel filter with size 3×3. The overall layers and their configurations for the VGG-16 are presented in Table III.

The entire architecture of the model is divided into 5 blocks of Ensemble 3-Dimenisional CNN layers and 1 block of Fully Connected (or) Dense Layer. It takes standard input of size 224×224 with 3 dimensions. The first two blocks contains 2 layers of CNN with max pooling layer then remaining blocks contains 3 layers of CNN with max pooling. The final block contains two fully connected layers and one softmax layer.

TABLE II. DESCRIPTION OF IMAGE MANIPULATION CUSTOMIZATION OPERATIONS IN GENERATOR MODULE

S.No	Parameter Name	Description	Initialized Value
1	preprocessing_function	It removes the noise from generated images	Processed images from VGG16
2	Rotation_range	Some random images are rotated to 40 degrees angle	40
3	Width_shift	The images are translated 20% horizontally based on total width	0.2
4	Height_shift	The images are translated 20% vertically based on total height	0.2
5	Shear_range	It transforms the point to a particular	0.2
6	Zoom_range	The inside image contents are zoomed to 20%	0.2
7	Horizontal_flip	It randomly flips the half images of the dataset	True
8	Fill_mode	It fills the newly created pixels with nearest pixel values	Nearest

TABLE III. VGG-16 CONFIGURATION

S.No	Layer Name	Dimensions	Activation Function	Count
1	Convolution Layer	Initially it starts with 64 then it enhances up to 512	ReLU	13
2	Max Pooling Layer	From 75 × 75 × 64 to 4 × 4 × 512	-	5
3	Dense Layer	1×1×4096 is converted into 1×1×1000	ReLU	2

The customization of the pre-trained model is represented in the below section:

1) The model acquired the input images in the 2-Dimensional but VGG-16 requires in three dimensional. So it changes the input shape from 2-D to 3-D by passing additional parameters.

2) The model adjusted the weights of the neural network based on the "ImageNet" dataset.

3) The augment images have the shape of 150×150 but the VGG-16 accepts 224×224. So, they include top attribute has assigned with false value.

4) In general VGG-16 requires softmax layer for multi classification but the model dataset has binary class labels. So, the model implements the softmax layer by flattening the layer. The overall architecture of VGG-16 is presented in Fig. 4.

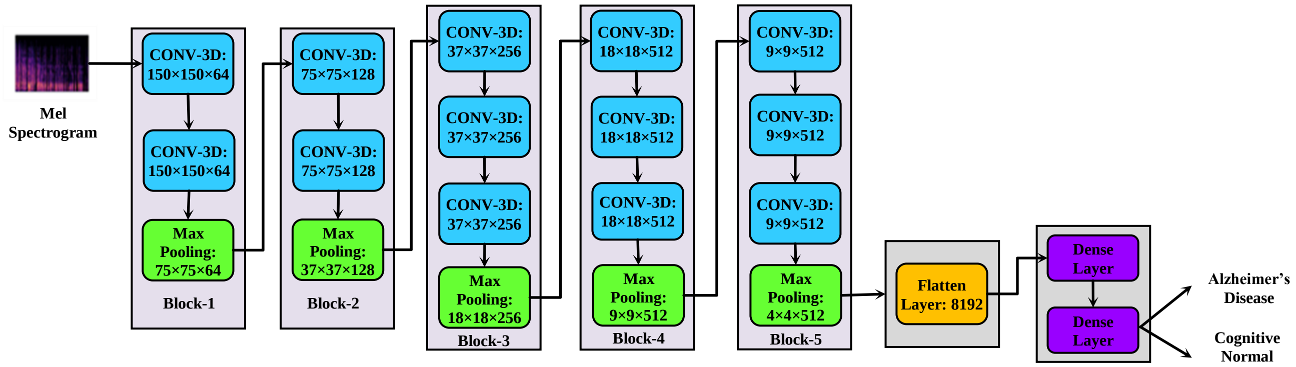


Fig. 4. Pre-trained Architecture of VGG-16.

IV. Result and Discussion

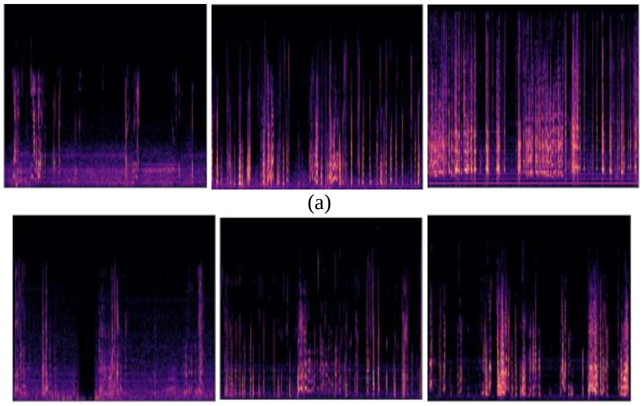
This model considers dataset from the publicly available repositories and uses Google Laboratory as the execution environment because the model needs GPU's to work with speech signals. The speech signals are labeled as "AD" and "CN". Alzheimer's is a group of neurodegenerative disorders characterized by a steady and long-term decline in cognitive function. Because age is the most important risk factor for AD, it affects the elderly the most. Because of the global severity of the situation, institutions and researchers are putting significant resources towards Alzheimer prevention and early detection, with an emphasis on disease progression. Cost-effective and scalable approaches for detecting Alzheimer disease in its most mild manifestations are needed. While several studies have investigated speech and language features for Alzheimer's disease and proposed various signal processing and machine learning methods for this task, the field still lacks balanced and standardized data sets on which these different approaches can be systematically compared. The ADReSS challenge dataset includes CN and AD patients' speech recordings, transcripts, and metadata (age, gender, and MMSE score). The dataset is balanced in terms of age, gender, and the number of CN vs. AD patients, with 78 patients in each class. The speech data is converted into Mel spectrograms. The outputs for the individual modules are represented in the below section.

Fig. 5(a) and 5(b) represents the generation of spectrogram from the speech signals for both the class labels. The graphs are represented by using the voice frequency obtained from the speech signals.

Fig. 6 represents the trainable and non-trainable parameters of VGG-16 by customizing the necessary parameters and layers. In general, neural networks need more parameters to train the model but due to usage of pre-trained models, the trainable parameters got reduced.

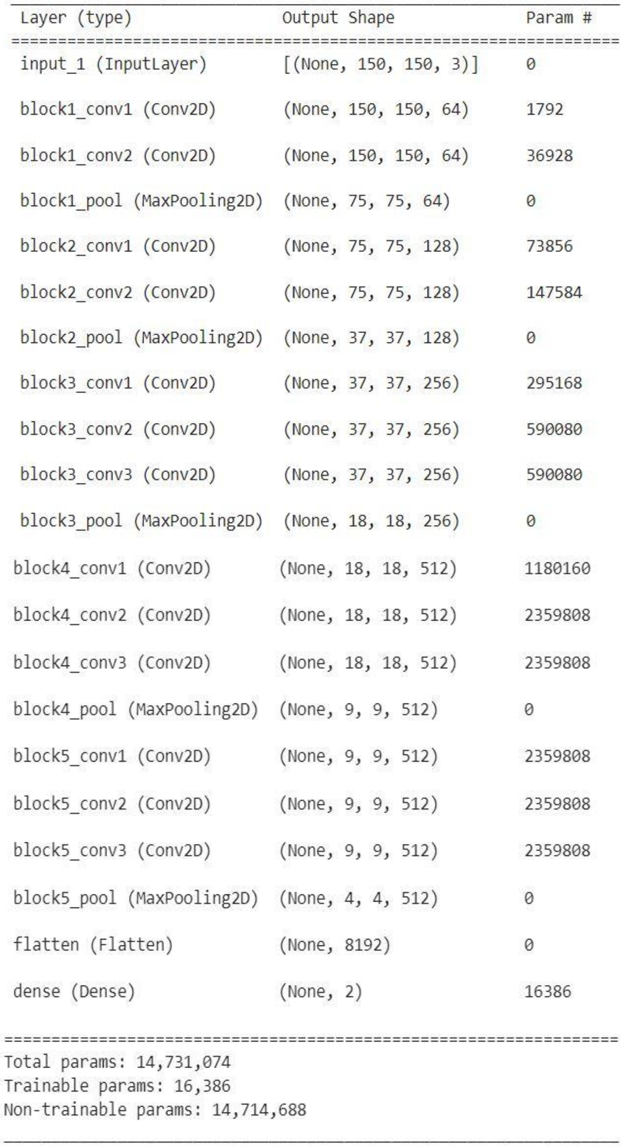
Fig. 7 represents the training phase of the model in pre-defined Epochs values. With the increase of Epoch, there might be increase or decrease of the accuracy. So, the model saves the highest accuracy as the best model. It updates the checkpoints if and only if the model gets better accuracy than the previous value. The model wants to prove the state of accuracy by comparing the standard CNN with VGG-16 pre-trained model. So, it represented the initial output training results of CNN in Fig. 8.

Fig. 9 represents the accuracy values obtained by VGG-16 at different Epochs levels. From Fig. 10, the model can clearly state that the VGG-16 has performed very well than CNN by comparing from the initial Epochs.


(a)

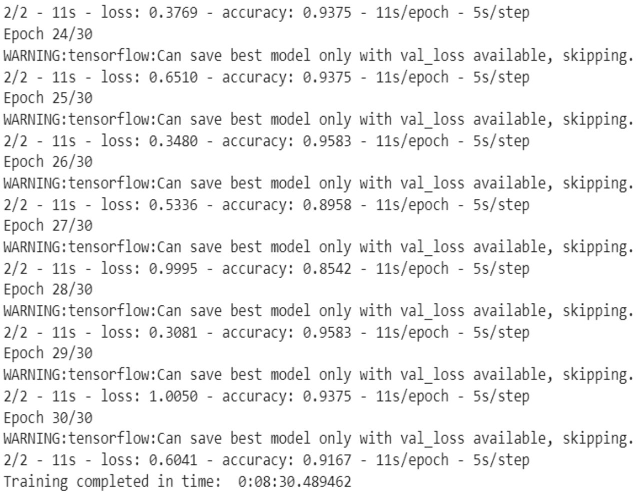
(b)

Fig. 5. Mel Spectrum. (a) AD Class Label. (b) CN Class Label.



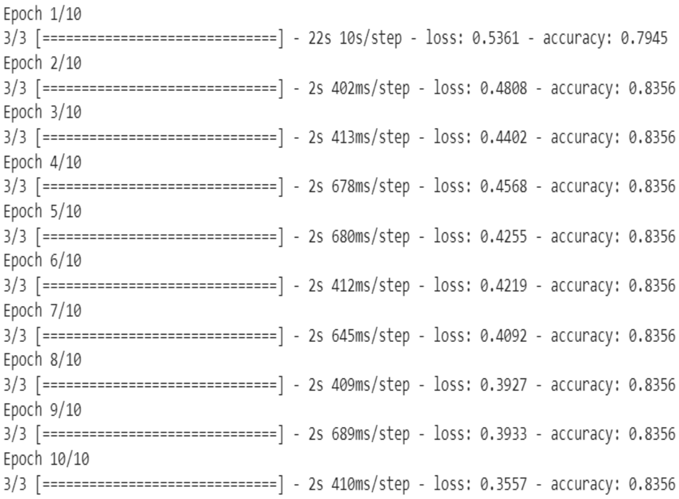
```
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 150, 150, 3)]     0

 block1_conv1 (Conv2D)       (None, 150, 150, 64)      1792

 block1_conv2 (Conv2D)       (None, 150, 150, 64)      36928

 block1_pool (MaxPooling2D)  (None, 75, 75, 64)        0

 block2_conv1 (Conv2D)       (None, 75, 75, 128)       73856

 block2_conv2 (Conv2D)       (None, 75, 75, 128)       147584

 block2_pool (MaxPooling2D)  (None, 37, 37, 128)       0

 block3_conv1 (Conv2D)       (None, 37, 37, 256)       295168

 block3_conv2 (Conv2D)       (None, 37, 37, 256)       590080

 block3_conv3 (Conv2D)       (None, 37, 37, 256)       590080

 block3_pool (MaxPooling2D)  (None, 18, 18, 256)       0

 block4_conv1 (Conv2D)       (None, 18, 18, 512)       1180160

 block4_conv2 (Conv2D)       (None, 18, 18, 512)       2359808

 block4_conv3 (Conv2D)       (None, 18, 18, 512)       2359808

 block4_pool (MaxPooling2D)  (None, 9, 9, 512)         0

 block5_conv1 (Conv2D)       (None, 9, 9, 512)         2359808

 block5_conv2 (Conv2D)       (None, 9, 9, 512)         2359808

 block5_conv3 (Conv2D)       (None, 9, 9, 512)         2359808

 block5_pool (MaxPooling2D)  (None, 4, 4, 512)         0

 flatten (Flatten)           (None, 8192)              0

 dense (Dense)               (None, 2)                 16386

=================================================================
Total params: 14,731,074
Trainable params: 16,386
Non-trainable params: 14,714,688
_________________________________________________________________
```

Fig. 6. Summary of VGG-16.



```
2/2 - 11s - loss: 0.3769 - accuracy: 0.9375 - 11s/epoch - 5s/step
Epoch 24/30
WARNING:tensorflow:Can save best model only with val_loss available, skipping.
2/2 - 11s - loss: 0.6510 - accuracy: 0.9375 - 11s/epoch - 5s/step
Epoch 25/30
WARNING:tensorflow:Can save best model only with val_loss available, skipping.
2/2 - 11s - loss: 0.3480 - accuracy: 0.9583 - 11s/epoch - 5s/step
Epoch 26/30
WARNING:tensorflow:Can save best model only with val_loss available, skipping.
2/2 - 11s - loss: 0.5336 - accuracy: 0.8958 - 11s/epoch - 5s/step
Epoch 27/30
WARNING:tensorflow:Can save best model only with val_loss available, skipping.
2/2 - 11s - loss: 0.9995 - accuracy: 0.8542 - 11s/epoch - 5s/step
Epoch 28/30
WARNING:tensorflow:Can save best model only with val_loss available, skipping.
2/2 - 11s - loss: 0.3081 - accuracy: 0.9583 - 11s/epoch - 5s/step
Epoch 29/30
WARNING:tensorflow:Can save best model only with val_loss available, skipping.
2/2 - 11s - loss: 1.0050 - accuracy: 0.9375 - 11s/epoch - 5s/step
Epoch 30/30
WARNING:tensorflow:Can save best model only with val_loss available, skipping.
2/2 - 11s - loss: 0.6041 - accuracy: 0.9167 - 11s/epoch - 5s/step
Training completed in time:  0:08:30.489462
```

Fig. 7. Few Training Iterations of the Proposed Model.



```
Epoch 1/10
3/3 [==============================] - 22s 10s/step - loss: 0.5361 - accuracy: 0.7945
Epoch 2/10
3/3 [==============================] - 2s 402ms/step - loss: 0.4808 - accuracy: 0.8356
Epoch 3/10
3/3 [==============================] - 2s 413ms/step - loss: 0.4402 - accuracy: 0.8356
Epoch 4/10
3/3 [==============================] - 2s 678ms/step - loss: 0.4568 - accuracy: 0.8356
Epoch 5/10
3/3 [==============================] - 2s 680ms/step - loss: 0.4255 - accuracy: 0.8356
Epoch 6/10
3/3 [==============================] - 2s 412ms/step - loss: 0.4219 - accuracy: 0.8356
Epoch 7/10
3/3 [==============================] - 2s 645ms/step - loss: 0.4092 - accuracy: 0.8356
Epoch 8/10
3/3 [==============================] - 2s 409ms/step - loss: 0.3927 - accuracy: 0.8356
Epoch 9/10
3/3 [==============================] - 2s 689ms/step - loss: 0.3933 - accuracy: 0.8356
Epoch 10/10
3/3 [==============================] - 2s 410ms/step - loss: 0.3557 - accuracy: 0.8356
```

Fig. 8. Training Process using Standard CNN.

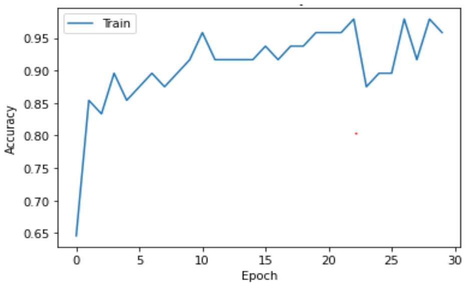


Fig. 9. Accuracy Graph Obtained by the VGG-16.

With reference to Table I and from the above figures, the Fig. 10 plots the accuracies obtained by different existing models along with proposed and standard CNN architecture to project the performance of the model. X-axis represents the approaches and Y-axis represents the accuracy obtained.

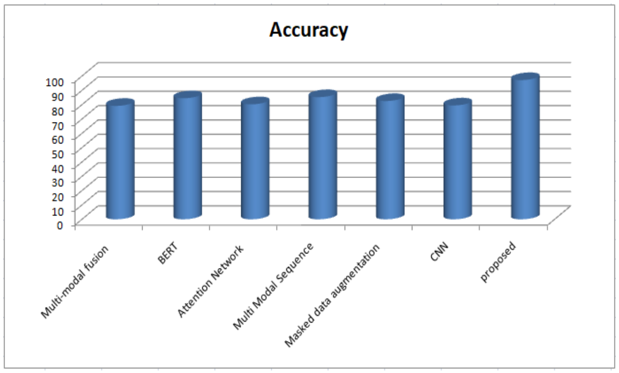


Fig. 10. Comparison Analysis.

V. Conclusion

In this paper, the model has recognized Alzheimer’s disease at early stage effectively by creating the synthetic dataset of Mel spectrum images then these are acted as input for the ImageNet trained VGG-16 system. The model has opted pre-trained model instead of CNN because the dataset contains different modulation signals with noise. The design becomes complicated because to extract essential features from the different modulations, more number of layers is required. It requires an efficient back propagation system to update the weights accurately. The numbers of trainable parameters are 8,53,145 in CNN where as the number of trainable parameters are 16,386 in proposed network. The misclassification rate and accuracy are also got affected because of the transfer learning process. As a conclusion remarks, it can be stated that variations in CNN model work well on the images because of its implicit feature extraction process.

REFERENCES

[1] M. Rohanian, J. Hough, and M. Purver, "Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech," Interspeech, pp. 2187–2191, October 2020.

[2] P. Raghavendra, J. Cho, S. Joshi, L. Moro-Velazquez, P. Żelasko, J. Villalba, and N. Dehak, "Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios," Interspeech, pp. 3825–3829, June 2021.

[3] N. Wang, Y. Cao, S. Hao, Z. Shao, and K. P. Subbalakshmi, "Modular Multi-Modal Attention Network for Alzheimer's Disease Detection Using Patient Audio and Language Data," Interspeech. August 2021.

[4] A. Mittal, S. Sahoo, A. Datar, J. Kadiwala, H. Shalu, and J. Mathew, "Multi-Modal Detection of Alzheimer's Disease from Speech and Text," arXiv, July 2021.

[5] Z. Liu, Z. Guo, Z. Ling, and Y. Li, "Detecting Alzheimer's Disease from Speech Using Neural Networks with Bottleneck Features and Data Augmentation," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7323–7327, June 2021.

[6] D. Ragavamsi, and R. Ragupathy, "A Survey of Different Machine Learning Models for Alzheimer Disease Prediction," International Journal of Emerging Trends in Engineering Research, vol.8, pp. 3328–3337, July 2020.

[7] D. Ragavamsi, and R. Ragupathy, "Identification of Alzheimer's Disease Using Various Deep Learning Techniques—A Review," Smart Innovation Systems and Technologies, vol. 265, pp. 485–498, December 2021.

[8] I. Vigo, L. Coelho, and S. Reis, "Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review," Bioengineering, vol. 9, 2022.

[9] Y. Qin, W. Liu, Z. Peng, S. Ng, J. Li, H. Hu, and T. Lee, "Exploiting Pre-Trained ASR Models for Alzheimer's Disease Recognition Through Spontaneous Speech," arXiv, vol. 9, January 2022.

[10] J. Laguarta, and B. Subirana, "Longitudinal Speech Biomarkers for Automated Alzheimer's Detection," Frontiers in Computer Science, vol. 3, April 2021.

[11] S. Al-Shoukry, T. H. Rassem, and N. M. Makbol, "Alzheimer's Diseases Detection by Using Deep Learning Algorithms: A Mini-Review," IEEE Access, vol. 8, pp. 77131–77141, April 2020.

[12] Q. Zhou, J. Shan, W. Ding, C. Wang, S. Yuan, F. Sun, H. Li, and B. Fang, "Cough Recognition Based on Mel-Spectrogram and Convolutional Neural Network," Frontiers in robotics and AI, vol. 8, May 2021.

[13] T. Jo, K. Nho, and A. J. Saykin, "Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data," Frontiers in aging neuroscience, vol. 11, August 2019.

[14] S. Kaur, S. Gupta, S. Singh, and I. Gupta, "Detection of Alzheimer's Disease Using Deep Convolutional Neural Network," International Journal of Image and Graphics, January 2021.

[15] Y. Wang, X. Liu, and C. Yu, "Assisted Diagnosis of Alzheimer's Disease Based on Deep Learning and Multimodal Feature Fusion," Complexity, April 2021.

[16] J. Islam, and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," Brain Informatics, May 2018.

[17] R. R. Janghel, "Deep-Learning-Based Classification and Diagnosis of Alzheimer's Disease," Deep Learning and Neural Networks, pp. 1358–1382, 2020.

[18] D. Ragavamsi, and R. Ragupathy, "Neuro-imaging-based Diagnosing System for Alzheimer's Disease Using Machine Learning Algorithms," Innovations in Computer Science and Engineering, vol. 385, pp. 501–509, March 2022.

[19] Y. Qin, W. Liu, Z. Peng, S. Ng, J. Li, H. Hu, and T. Lee, "Exploiting Pre-Trained ASR Models for Alzheimer's Disease Recognition Through Spontaneous Speech," arXiv, October 2021.

[20] L. Ilias, D. Askounis, and J. Psarras, "Detecting Dementia from Speech and Transcripts using Transformers," arXiv, October 2021.

[21] A Mallikarjuna Reddy, Vakulabharanam Venkata Krishna, Lingamgunta Sumalatha and Avuku Obulesh, "Age Classification Using Motif and Statistical Features Derived On Gradient Facial Images", Recent Advances in Computer Science and Communications, vol.13,2020.

[22] A. Meghanani, C. S. Anoop, and A. G. Ramakrishnan, "An Exploration of Log-Mel Spectrogram and MFCC Features for Alzheimer's Dementia Recognition from Spontaneous Speech," IEEE, January 2021.

[23] Y. Zhu, X. Liang, J. Batsis, and R. M. Roth, "Exploring Deep Transfer Learning Techniques for Alzheimer's Dementia Detection," Frontiers in Computer Science, Vol. 3, May 2021.

[24] M. S. Syed, Z. S. Syed, E. Pirogova, and M. Lech, "Static vs. Dynamic Modelling of Acoustic Speech Features for Detection of Dementia," International Journal of Advanced Computer Scicence and Applications, vol. 11, October 2020.