Comparative Analysis of Deep Learning Models for Dysarthric Speech Detection

Shanmugapriya Padmanaban (shanmugapriya-ece@saranathan.ac.in)
Saranathan College of Engineering
V Mohan
Saranathan College of Engineering

Research Article

Keywords: Deep learning, Dysarthria detection, Wavelet Transformation, Pre-trained CNNs

Posted Date: June 1st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-1916239/v1

Version of Record: A version of this preprint was published at Soft Computing on November 8th, 2023.
See the published version at https://doi.org/10.1007/s00500-023-09302-6.

Comparative Analysis of Deep Learning Models for Dysarthric Speech Detection

*Shanmugapriya P[1], Mohan V[2]
[1,2]*Associate Professor, Department of Electronics and Communication Engineering,*
Saranathan College Of Engineering, Venkateswara Nagar, Panjappur,
Tiruchirappalli-620012, Tamil Nadu, India.
** Corresponding Author*
shanmugapriya-ece@saranathan.ac.in Ph.No:91-9994539389
mohan-ece@saranathan.ac.in Ph.No: 91-9994338212

Abstract: Dysarthria is a speech communication disorder that is associated with neurological impairments. In order to detect this disorder from speech, we present an experimental comparison of deep models developed based on frequency domain features. A comparative analysis of deep models is performed in the detection of dysarthria using scalogram of Dysarthric Speech. Also, it can assist physicians, specialists, and doctors based on the results of its detection. Since Dysarthric speech signals have segments of breathy and semi-whispery, experiments are performed only on the frequency domain representation of speech signals. Time domain speech signal is transformed into a 2-D scalogram image through wavelet transformation. Then, the scalogram images are applied to pre-trained convolutional neural networks. The layers of pre-trained networks are tuned for our scalogram images through transfer learning. The proposed method of applying the scalogram images as input to pre-trained CNNs is evaluated on the TORGO database and the classification performance of these networks is compared. In this work, AlexNet, GoogLeNet, and ResNet 50 are considered deep models of pre-trained convolutional neural networks. The proposed method of using pre-trained and transfer learned CNN with scalogram image feature achieved better accuracy when compared to other machine learning models in the dysarthria detection system.

Keywords: Deep learning; Dysarthria detection; Wavelet Transformation; Pre-trained CNNs

1. Introduction

A speech disorder that happens due to muscle weakness is dysarthria. It results from damage to the nervous system which controls certain muscles. Especially, when the muscles used to produce speech sounds are damaged, speech impairment is caused. The severity level of speech impairment can be described based on the degree to which the speech muscles are affected [1]. Mainly, Tongue, Larynx, and surrounding muscles are affected because of damage to speech organs which is named Peripheral dysarthria [2]. The symptoms of this type of dysarthria are slow or quick delivery of the speech, slurred or choppy speech, and difficulty [3] in moving the lips, jaw, or tongue during speaking. Since dysarthria is a frequent symptom of speech disorder, it has to be identified and diagnosed in the early stage itself. Assessment and classification of dysarthria gained significant importance because of the need to understand the variety of impairment results in speech disorder [4] and to develop speech and language therapy which can be used to encourage the patients to improve their communication skills. Many works had been carried out to identify the dysarthria speech and the level of severity in the dysarthria by extracting features like MFCC [5], LPC [6], Log RASTA PLP [7], Centroid Formants [8], glottal features [9], Perceptually Enhanced Single Frequency Cepstral Coefficients [10], and Spectro –temporal sparsity features extracted through STFT with Mel warping and Single Frequency Filtering [11]. However, in addition to all these features which represent the spectro-temporal characteristics, the continuous wavelet transformation of the dysarthric speech signal which has the ability to extract the spectral characteristics at multiple scales is proposed in this work. In the theory of wavelet transformation, the scalogram is the time-frequency representation of the signal which is used to indicate the coefficient values with brightness or color at different locations of the time-frequency scale.

Though several CNN architectures are used for dysarthric speech detection and classification, like the Interpretable Deep Learning Model for the detection of dysarthric speech [12], CNN with EMDH for improving dysarthric speech quality [13] Transfer Learning with convolutional NN for dysarthric speech detection [14] Automatic dysarthric speech recognition system using deep learning [15], the proposed architecture with the scalogram features as input performs better.

2. Related Work

Recently, the development of Deep Learning breaks the limitations of existing Machine Learning Techniques. Hence, many new methods using Deep Learning with different architectures are proposed by researchers. Deep learning is a multi-layered structure that has heterogeneous layers with non-polynomial activation

functions. Several CNN architectures are designed for the automatic classification of dysarthric speech with various acoustic and spectro- temporal features [16] as input to CNN. Unfortunately, those approaches are not robust and time-consuming when the training data is large. Usage of pre-trained deep models [17] with transfer learning for a large-scale dataset is the best method to avoid this difficulty. According to the design and construction of CNN, one of the best ways of giving input to CNN for the speech data is the representation of speech in terms of image. That is the method of conversion of speech into scalogram and the scalogram images are resized to the image size requirement of pre-trained CNN fed as input. Since the convolutional neural network extracts the features from the input image through convolutional filters, there is no need to extract the features from speech after being transferred into an image.

The transformation of the signal from the time domain to the frequency domain through wavelet transformation performs the analysis of the signal on the multiresolution scale[18]. Hence it becomes possible for using the scalogram image of dysarthric speech as a frequency domain representation of the one-dimensional speech signal. In this work, the possibility of usage of this scalogram image of dysarthric speech as input to pre-trained CNN with deep learning is investigated and it is proven that the performance of the deep models is optimal for this scalogram images. There is no work carried out so far for the comparison of the performance of pre-trained deep models with two-dimensional scalogram images extracted from the dysarthric speech signal through wavelet transform. Thus, we developed the most frequently used pre-trained CNNs such as Alexnet, Googlenet, and Resnet 50 for the scalogram images by transfer learning. The performance of the dysarthric speech detection system using these networks is analyzed and compared with the simple CNN also. The novelty lies in the usage of a scalogram to represent the characteristics of the dysarthric speech signal and testing its strength in the classification of dysarthric speech using various pre-trained CNNs. TORGO database is used to evaluate the performance of these networks.

In this paper, a comparative analysis of deep models in dysarthric speech detection using scalogram of the speech signal is performed. Section 3 describes the proposed methodology using deep models of convolutional neural networks. Section 4 presents the experimental results and discussions, while section 5 concludes the paper.

3. Proposed Methodology

3.1. Speech to Image Conversion

Since dysarthria speech is a signal with speech disorders like stuttering, abnormal pauses, and imprecise articulation, it is necessary to analyze the signal in the frequency domain. Wavelet transformation of the signal provides high resolution of low-frequency components and low resolution of high-frequency components of the dysarthric speech signal which indicates the spectral characteristics of the signal with multi resolution scale. Generally, the speech signal is a non-stationary signal. To localize the transients in speech signal wavelet transformation is better than ordinary transformation like Short Time Fourier Transform (STFT) and Single Frequency filtered –Fourier Transform (SFF-FT). Hence the possibility and benefit of utilizing this effective representation of dysarthric speech signal in terms of wavelet transformed scalogram images and using them as input for pre-trained CNNs are investigated in this work.

Dysarthric and non-dysarthric speech signals from the TORGO database are preprocessed and transformed into scalogram images through continuous wavelet transformation. After preprocessing such as noise removal and normalization, the scalogram of each signal is computed by taking Wavelet transformation. In our work, the Morlet wavelet is used for taking wavelet transformation. The analysis of dysarthric speech signals is carried out on a multiscale time-frequency plane using the Morlet wavelet. Morlet Wavelet is a Gaussian-windowed complex sinusoid that provides good time and frequency localization [20]. It is given by

$$W_x(s,\tau) = \frac{1}{\sqrt{s}} \int x(t) \psi^* \left(\frac{t-\tau}{s}\right) dt \quad (1)$$

where $\psi(t) = \pi^{-1/4} e^{j\omega_0 t} e^{\frac{-1}{2}t^2}$, x is the signal to be analyzed and s is the scale in which the signal is analyzed. The correlation between the signal and the mother wavelet is performed with the scale 's' and the position 'τ' where is the wavelet coefficient which represents the signal and its Fourier transform in the time-frequency region. CWT analysis of a single plot spectrum produces a wavelet power scalogram. Each component of the scalogram is the wavelet power, which denotes the magnitude of each wavelet coefficient. The wavelet coefficients describe the correlation between a subset of the input spectrum and a scaled, shifted version of the mother wavelet. However, the visual representation of the spectral characteristics of the dysarthric signal at different scales and frequencies through CWT is efficient in the morphological analysis of the complex signal. Figure 1 shows the scalogram of dysarthric speech and non-dysarthric speech with their corresponding speech waveforms. Since the wavelet transformation is good at localizing transients in non-stationary signals, it provides a detailed time-frequency representation of a speech sample from a person having dysarthria.

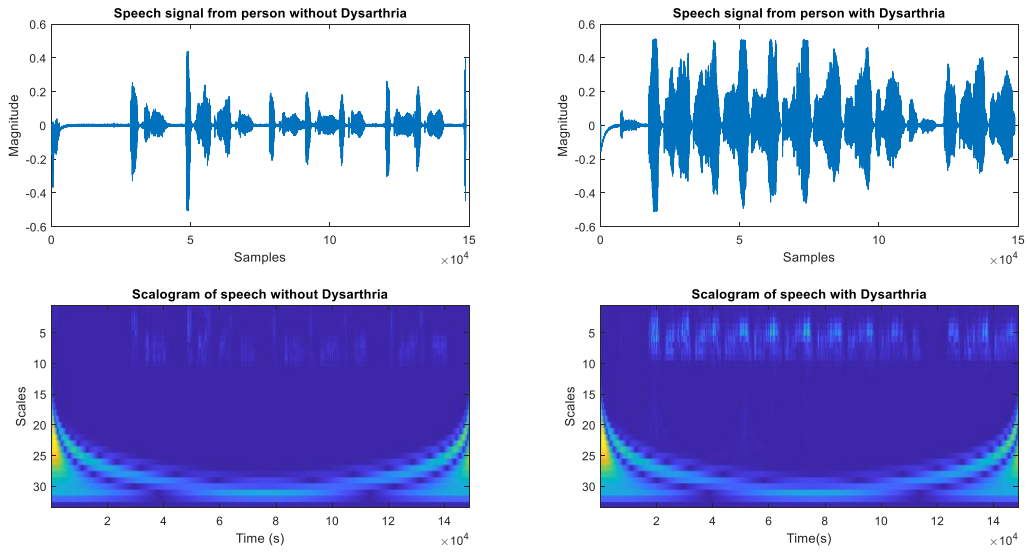

Figure 1. Comparison of Speech waveforms and their scalograms with Dysarthria and without Dysarthria

3.2. Architecture of Pre trained CNNs

Since the scalogram forms the basis of the input as discussed earlier and since CNNs are the best suited for image classifications, different CNN-based architectures are chosen for this work. By transfer learning, pre-trained CNNs are transferred to our dysarthria speech classification task for extracting the deep scalogram features. For the pre-trained CNNs, we choose AlexNet, GoogLeNet, and Resnet 50, since they have proven to be successful in a large number of natural image classification tasks.

AlexNet is one of the pre-trained CNN which has five convolutional layers, three pooling layers, and three fully connected layers. It is trained for classifying images of size 227x227x3 into 1000 classes. Special features of Alexnet are the introduction of nonlinearity through the ReLU layer, allowing training through multi-GPU, and overlapping in pooling to reduce the error rate. It also avoids overfitting by dropout layers and data augmentation. The architecture of Alexnet used for dysarthric speech detection is shown in figure 2.

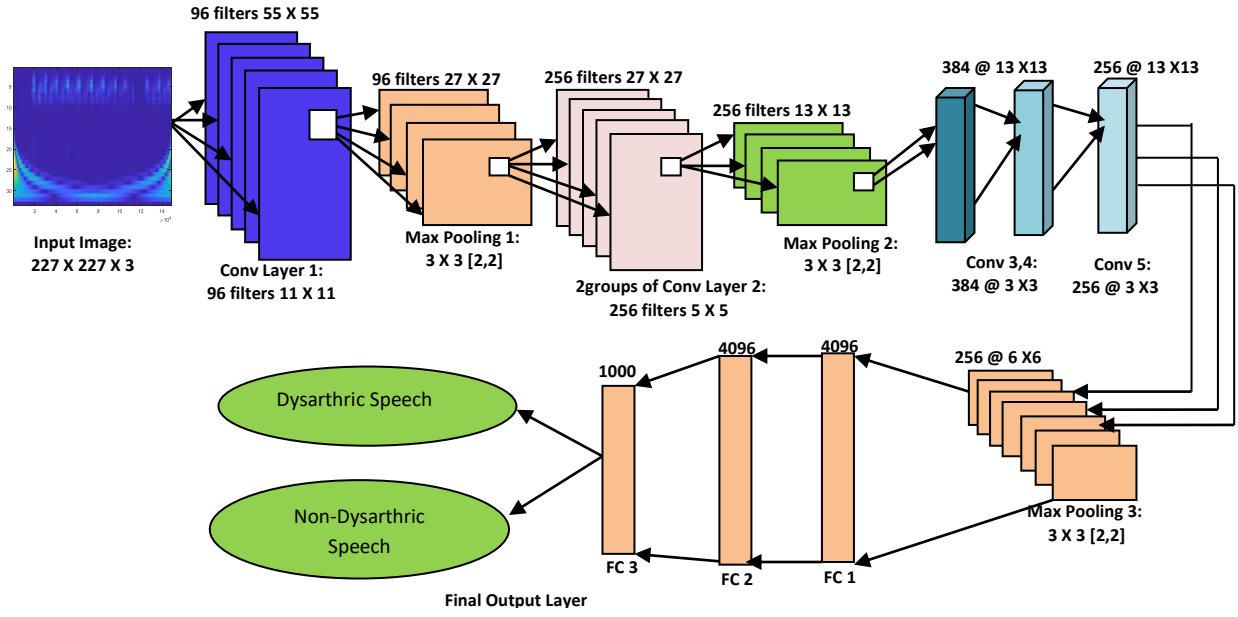

Figure 2. Architecture of Alexnet

GoogLeNet is also a 22-layer deep pre-trained CNN that uses 1x1 convolutions in the architecture and global average pooling. The depth of this architecture is increased by decreasing the number of parameters by inception convolution layers. The performance of this architecture is better than the AlexNet because of the inception method convolution which provides regularization. Also, it uses intermediate classifiers during training which helps to avoid the gradient vanishing problem. Figure 3 illustrates the architecture of GoogLeNet used for dysarthric speech detection work.

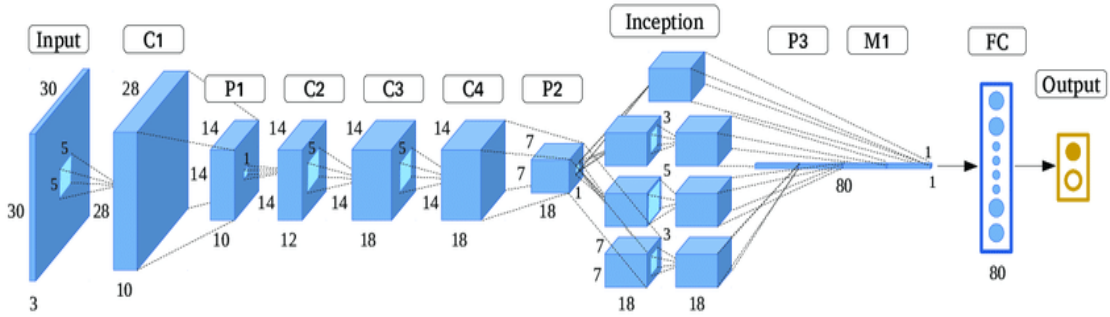

Figure 3. Architecture of GoogLeNet

Resnet 50 is a pre-trained CNN model that is stacked deeply with 50 layers. It has been trained on more than a million images to classify images into 1000 classes. It outperforms AlexNet and GoogLeNet in terms of learning efficiency. It achieves good performance by residual learning which means reusing the upper layer features in order to avoid overfitting and vanishing gradient. The architecture of Resnet 50 used in this work is shown in figure 4. We obtained the pre-trained network, AlexNet, GoogLeNet, and Resnet 50 from the MATLAB R2021a.

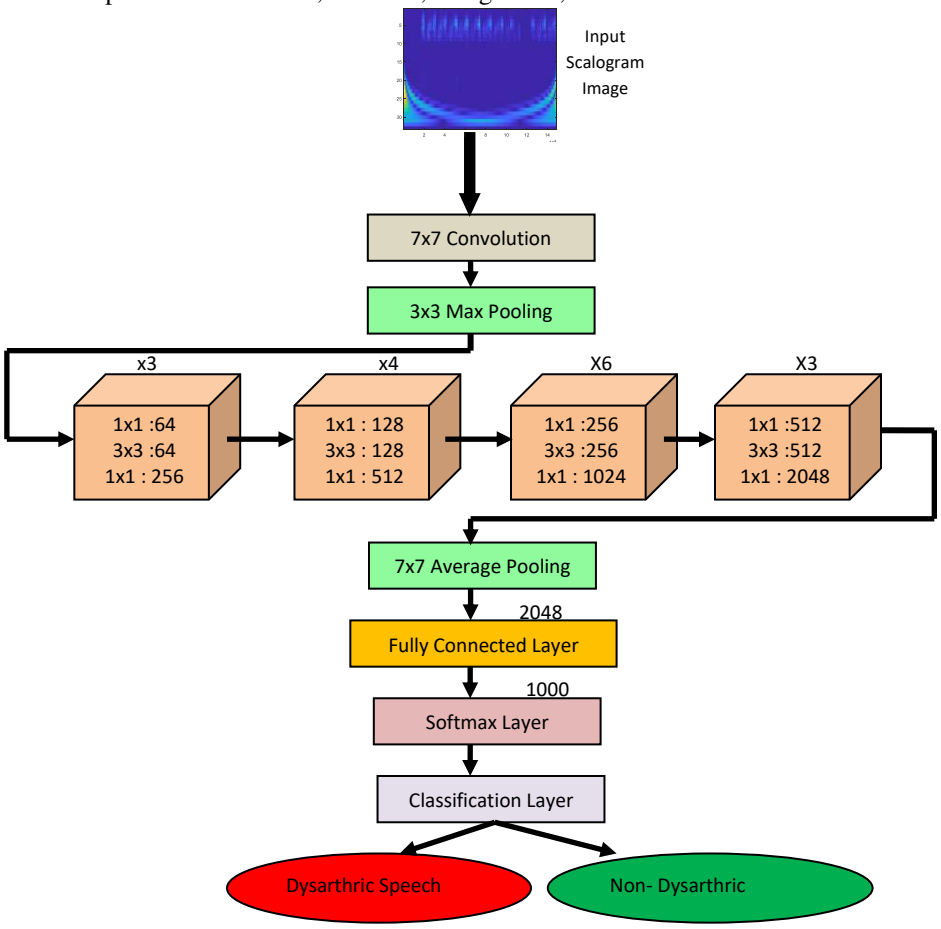

Figure 4. Architecture of Resnet 50

In all these three pre-trained networks the layers are modified under transfer learning. The last three layers of all these three networks are removed and newly designed layers based on the classes in our work are inserted. Then the new transfer learned network is trained for the set of scalogram images of dysarthric and non-dysarthric speech data. Figure 5 illustrates the processes involved in transfer learning.

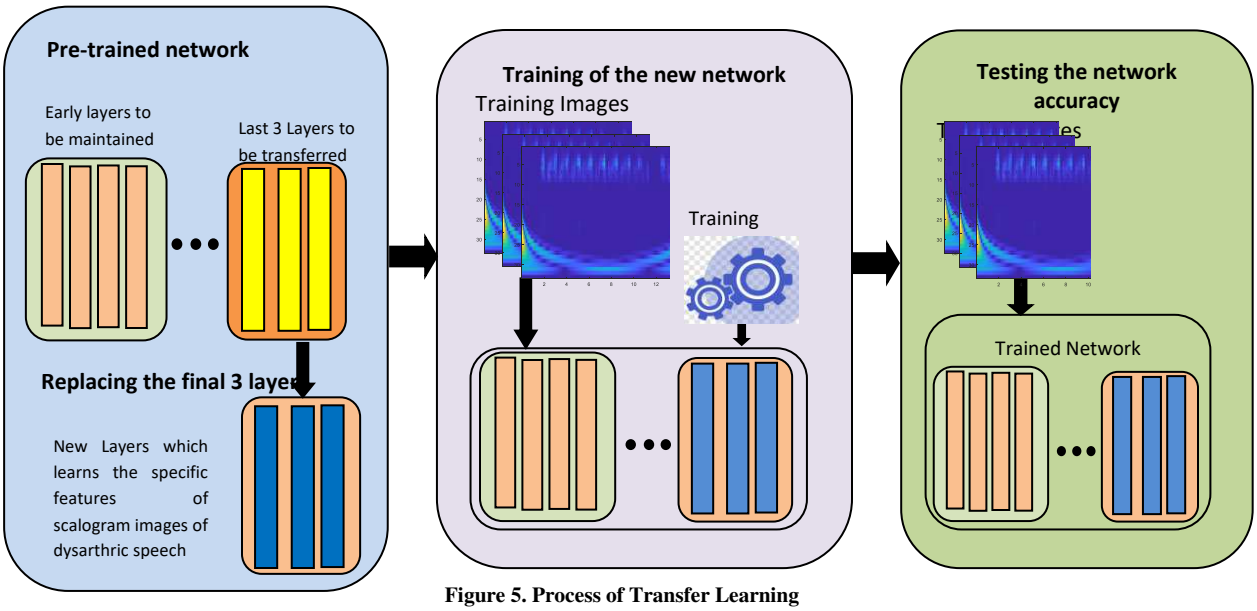

Figure 5. Process of Transfer Learning

For comparing the performance of pre-trained networks with state-of-the-art, the general structure of CNN is used for dysarthric speech detection with a scalogram as input. Configuration details of the proposed CNN are listed in Table I.

Table I. Description of CNN architecture

Layers of CNN	Size of the layers with description
Input Layer	Size:128x128x3 images with 'zerocenter' normalization
Convolution 1	12 filters of size 3 x 3 with stride [1,1] and padding 'same'
Norm 1, Norm 2, Norm3 & Norm 4	Mini-batch : 128 Batch normalization
ReLU1/ ReLU2/ ReLU3/ ReLU4	ReLU
Pool 1/Pool 2/Pool 3	Pool size: [3 3], stride [2 2]
Convolution 2	24 filters of size 3 x 3 with stride [1 1] and padding 'same'
Conv3/Conv4/Conv5	48 filters of size 3 x 3
Pool 4	Pool size: [1 16] max pooling with stride [1 1] and padding [0 0 0 0]
Dropout Layer	Dropout Probability: 20%
Fully Connected FC1, FC2	Nodes: 2 = Number of classes
Softmax Layer	Softmax function
Classification output	Weighted cross entropy

4. Results and Discussion

4.1. Database

Our proposed approach is evaluated on the TORGO database. TORGO database consists of speech data recorded from 8 dysarthric speakers suffering from Cerebral Palsy (CP) and Amyotrophic lateral sclerosis (ALS).

Head-mounted microphones and an array of directional microphones are used to record the speech of dysarthric patients. The sound recordings in this database are sampled at 16 kHz. Only the speech signals in the form of .wav format alone were used. A total of 9416 sentences are used in this work. The dataset is divided into training data and validation data according to the split up required and it is illustrated in Table II.

Table II Speakers and sentences used in training, validation and testing

Process	Gender	Speakers affected with dysarthria	Control Group	Number of sentences used
Training	Male	M02	MC03, MC 04	4564
	Female	F01,F03	FC02	
Validation	Male	M01,M03	MC02	1752
	Female	F04	FC01	
Test	Male	M04, M05	MC01	3100
	Female	F04	FC03	

4.2. Evaluation parameters

The performance of the proposed system is evaluated using the metric accuracy, which is the ratio of correct predictions and the total number of test/validation samples. The accuracy of the system is evaluated in two modes: Evaluation mode – in this mode, the accuracy is defined in terms of Validation Accuracy which is the recognition rate of the system which defines the level of learning during training. Testing Mode – in this mode the accuracy of the system is defined in terms of testing accuracy which is the recognition rate of the system for the new data. For an unbalanced dataset, the performance is to be verified not only by accuracy but also by measuring the parameters such as sensitivity, specificity, FPR, FNR, and EER.

The parameters sensitivity and specificity are defined as

$$Sensitivity = \frac{TP}{TP+FN} \; ; \; Specificity = \frac{TN}{TN+FP} \qquad (2)$$

where TP is True Positive which represents the correctly classified samples & FP is False Positive which represents the speech samples which are misclassified. False Positive Rate and False Negative Rate are also defined as

$$FPR = \frac{FP}{TN+FP} ; \quad FNR = \frac{FN}{TP+FN} \qquad (3)$$

4.3. Performance of the proposed system

In this section, the performance of the dysarthric speech detection system is investigated with the input feature as wavelet transformation of speech signal as image applied to pre-trained CNNs. The wavelet-transformed scalograms were resized to 227x227x3-Alexnet / 224 x 224 x 3 –GoogLeNet & Resnet 50 for normalization. Then classification was performed by using these three models of deep learning. The proposed system of dysarthria speech detection is shown in figure 6.

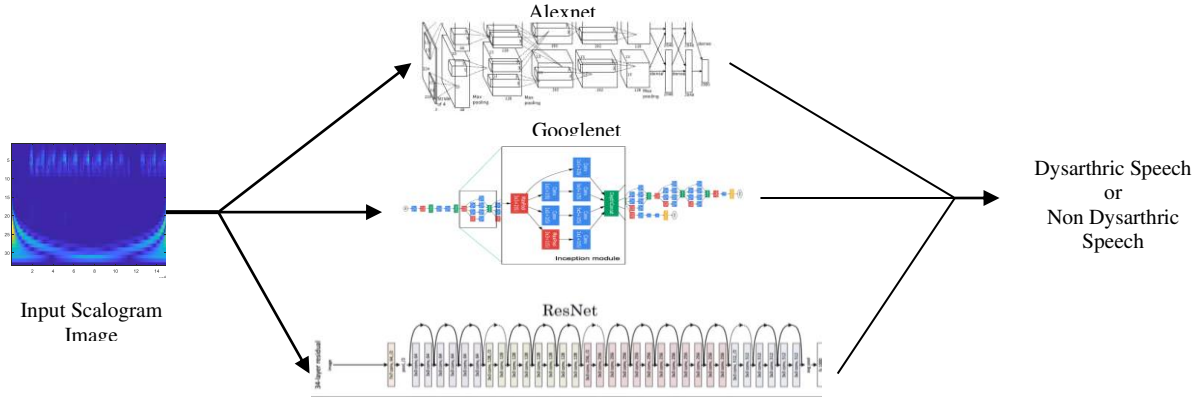

Figure 6. Dysarthria Detection System

The effect of scalogram images of the speech signal in the classification performance of various pre-trained networks is analyzed with different learning algorithms, different batch size, and various numbers of epochs. The learning algorithms used during the training of pre-trained networks are the Stochastic Gradient Descent Method (SGDM), RMS Prop, and Adaptive Moment Estimation (Adam). In the learning method, adam performs better than the other two, because it combines the advantages of both these methods. It updates the parameters with a moderate learning rate initially and modifies the learning rate according to the convergence. The performance of the CNNs for various minibatch size and the number of epochs is compared in table III for Alexnet, table IV for GoogLeNet, and table V for Resnet 50. Table VI illustrates the performance of proposed CNN architecture.

Table III. Performance of Transfer Learned Alexnet on TORGO Database

Method of Learning	Training : Validation	Batch size	Epoch	Validation Accuracy	Test Accuracy
SGDM	80:20	50	2	83.64	82.83
SGDM	80:20	100	2	82.66	81.85
SGDM	90:10	300	5	82.83	82.54
SGDM	90:10	300	10	84.22	84.97
SGDM	90:10	300	15	86.42	85.61
RMSProp	90:10	300	5	86.65	86.69
RMSProp	90:10	300	15	93.58	93.41
Adam	90:10	300	10	93.56	93.15
Adam	90:10	300	15	95.48	**96.94**

Table IV. Performance of Transfer Learned GoogLeNet on TORGO Database

Method of Learning	Training : Validation	Batch size	Epoch	Validation Accuracy	Test Accuracy
SGDM	80:20	50	2	84.85	84.89
SGDM	90:10	100	2	71.33	71.04
SGDM	90:10	100	5	87.83	87.54
SGDM	90:10	300	10	88.12	88.94
SGDM	90:10	300	15	89.22	88.61
RMSProp	90:10	100	5	91.25	91.56
RMSProp	90:10	300	10	92.24	92.44
Adam	90:10	50	2	89.13	89.88
Adam	90:10	300	5	94.45	93.99
Adam	90:10	300	10	97.45	**97.54**

Table V. Performance of Transfer Learned Resnet50 on TORGO Database

Method of Learning	Training : Validation	Batch size	Epoch	Validation Accuracy	Test Accuracy
SGDM	90:10	100	10	89.92	89.33
SGDM	90:10	300	15	95.23	95.74
RMSProp	90:10	100	10	96.12	96.25
RMSProp	90:10	300	15	98.21	98.10
Adam	90:10	100	10	98.56	98.01
Adam	90:10	300	15	98.72	**98.21**

Table VI. Performance of Proposed CNN on TORGO Database

Method of Learning	Training : Validation	Batch size	Epoch	Validation Accuracy	Test Accuracy
SGDM	80:20	50	2	84.63	82.99
SGDM	80:20	100	2	86.65	85.12
SGDM	90:10	300	10	89.92	88.51
SGDM	90:10	300	15	90.32	89.91
RMSProp	90:10	300	10	95.21	90.23
RMSProp	90:10	300	15	96.22	92.32
Adam	90:10	300	10	98.55	97.58
Adam	90:10	300	15	98.99	**97.78**

For the TORGO database, Alexnet with transfer learning produces the maximum accuracy of 96.94% for test data, GoogLeNet produces the maximum accuracy of 97.54% for the test data and Resnet50 produces the maximum accuracy of 98.21% for test data. From this analysis, it is understood that the performance of Resnet50 is better than

the other two networks for scalogram images as input. It is also observed that the scalogram is also efficient in classifying dysarthric speech from non-dysarthric speech when compared to other frequency domain representations of speech signals such as STFT-based spectrogram and Single Frequency Filtering (SFF) based Spectrogram. The result shows that the wavelet transformation of the dysarthric speech signal as input to the pre-trained network outperforms the other spectral features. Figure 7 shows the training progress of Alexnet for the scalogram image as input for minibatch size:300 and epoch=15 with the "adam" learning method. The resultant accuracy of the model is 96.94% and the corresponding confusion matrix is also shown nearby.

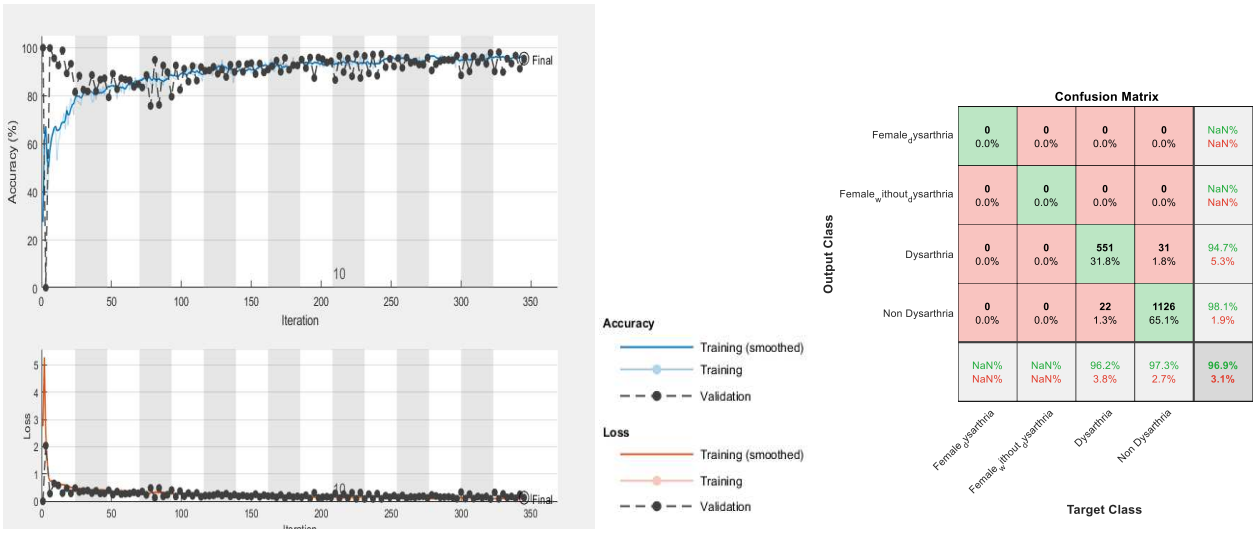

Figure 7. Training progress of Alexnet for the scalogram image

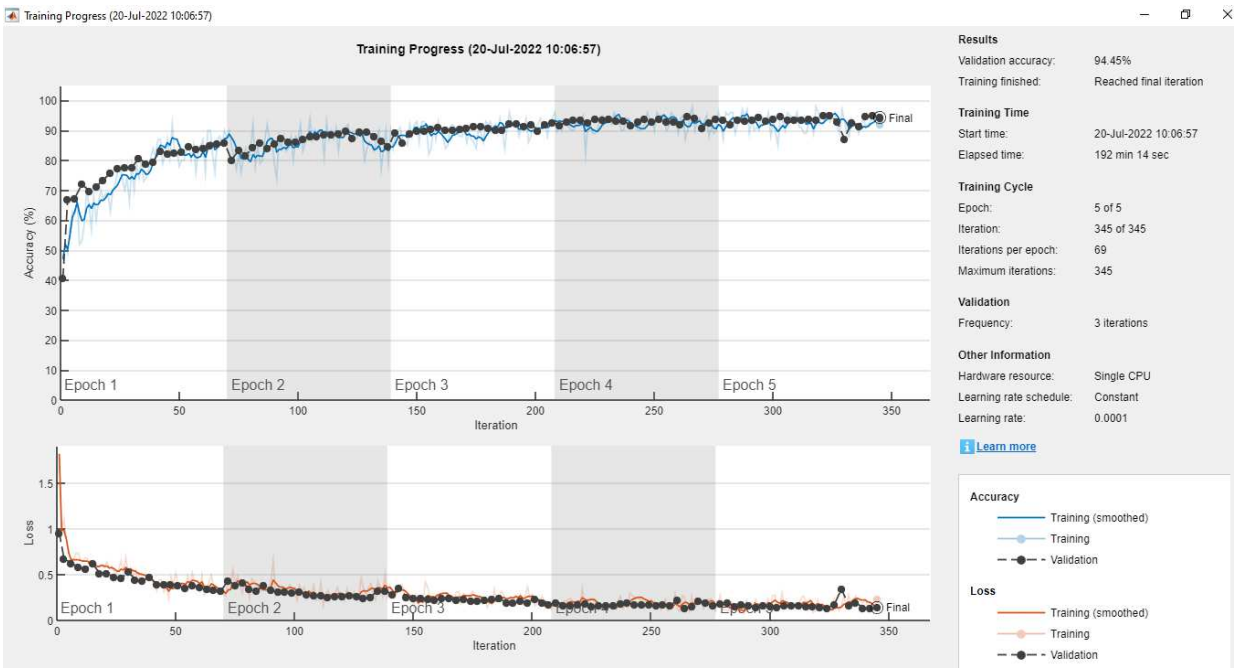

Figure 8. Training progress of GoogLeNet for the scalogram image

Figure 8 shows the training progress of GoogLeNet for the scalogram image as input for minibatch size: 300 and epoch=5 with “adam” learning method. The resultant accuracy of the model is 94.45% which is better than Alexnet.

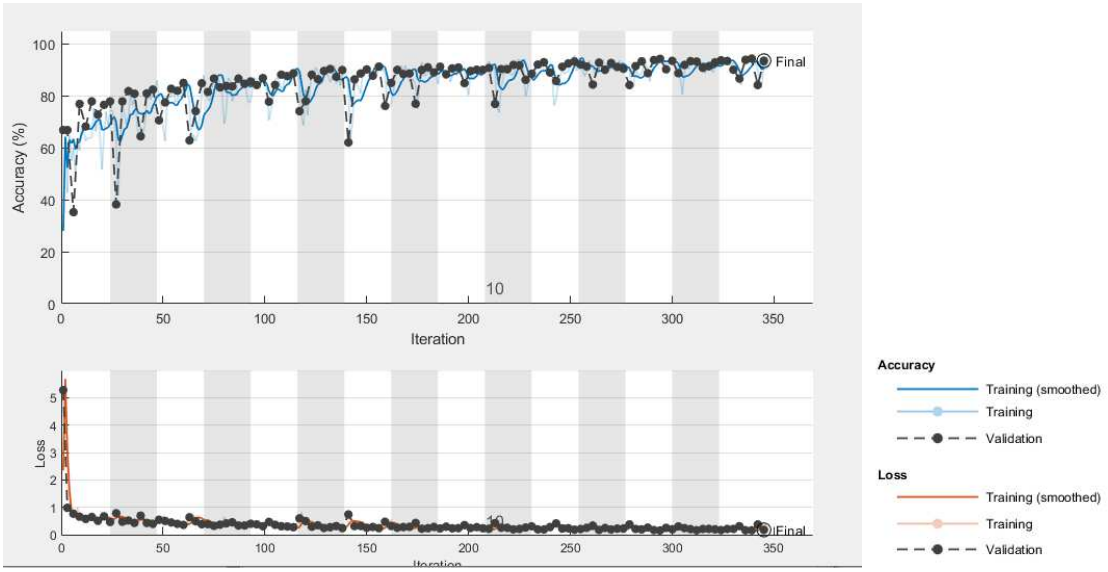

Figure 9. Training progress of Renet50 for the scalogram image

Figure 9 shows the training progress of Resnet for the scalogram image as input for minibatch size: 300 and epoch=15 with “adam” learning method. The resultant accuracy of the model is 98.21% which is the maximum accuracy obtained for the Torgo database for the test data.

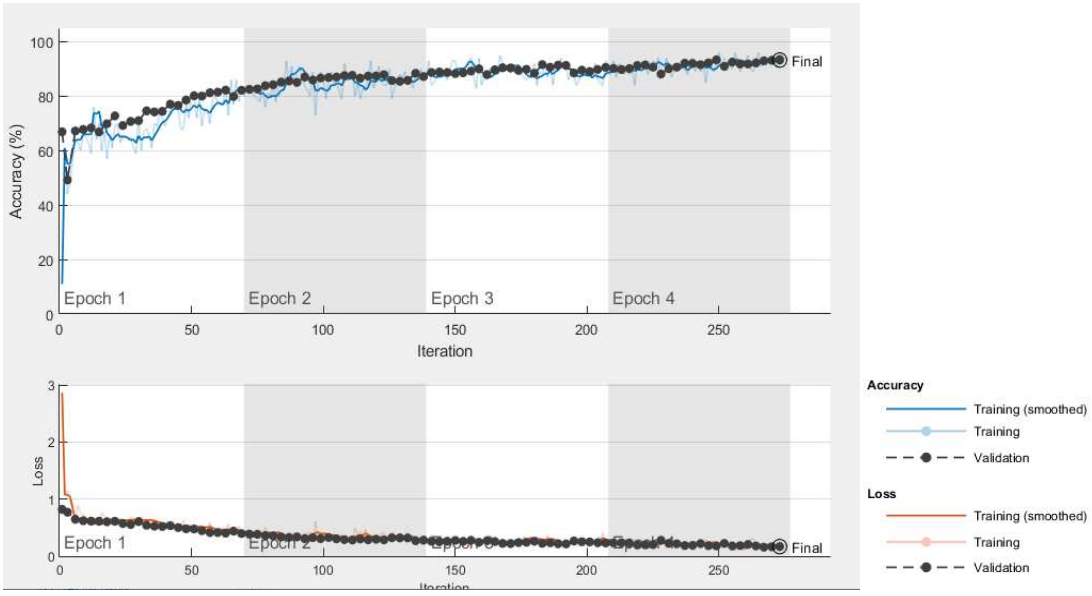

Figure 10 Training and validation processes of the proposed CNN model

Figure 10 shows the training and validation processes of the proposed CNN deep learning model based on the scalogram images. Accuracy and cross-entropy loss were plotted against the training step during the length of the

training of the classifier. The red and black lines represent the training and validation processes, respectively. The cross-entropy loss was close to 0.5, while the final validation accuracy was 97.78%.

Table VII. Comparison of Computational Time

Model	Training time	Testing Time	Max Epoch	Batch size
Alexnet	71 m 55 sec	16.18 sec	15	300
GoogLeNet	214 m 56 sec	30.23 sec	15	300
Resnet50	650 m 22 sec	1m 10 sec	10	300
Proposed CNN	405 m 15 sec	1m 15 sec	10	300

Table VIII. Performance of CNNs on TORGO Database

Model	Accuracy %	Sensitivity	Specificity	FPR	FNR	EER %
Alexnet	95.48	0.75	0.98	0.01	0.20	0.01
GoogLeNet	97.45	0.95	0.82	0.20	0.05	0.02
Resnet50	99.22	1.00	0.92	0.30	0.23	0.01
Proposed CNN	96.32	1.00	0.85	0.25	0.04	0.01

Performance comparisons of the pre-trained networks and the proposed CNN are also made in terms of computational time required for training and testing. Training time is calculated for one epoch and testing time is calculated for a batch of test signals. Based on the comparison as shown in Table VII, it is observed that the training time and testing time required for the network resnet50 is larger than the other two pre-trained networks. This may be due to the fine-tuning of network parameters required for producing high accuracy. In Table VIII, the performance of all the networks is compared in terms of various parameters which indicates that Resnet 50 performs better when compared to other networks. Also, the performance of the proposed CNN is comparable with the pre-trained networks for the recognition of dysarthric speech using the scalogram image feature.

5. Conclusions

Comparative analysis of pre-trained as well as proposed deep models in dysarthria speech detection system using scalogram of the speech signal is performed in this work. Since speech signals from persons having dysarthria are affected due to the slow rate of speaking and large variations in fundamental frequency (F0) range across utterances, studies have been conducted by transforming signals into the frequency domain. By transforming the one-dimensional speech signal into two-dimensional images, the analysis of the speech signal can be performed on a multi-resolution platform. Hence the possibility of using the scalogram as input to CNN for the detection of dysarthria is investigated in this work. Experiments are performed on Torgo Database for performance evaluation of the pre-trained CNNs - AlexNet, GoogLeNet, and Resnet50. The performance of pre-trained networks is also compared with the proposed CNN with the image feature as input. The performance of Resnet 50 is better than the other networks for dysarthria detection using a scalogram.

Author's Contribution

The authors confirm sole responsibility for the following: study conception and design, analysis and interpretation of results, and manuscript preparation. The novelty lies in the usage of a scalogram image to represent the characteristics of the dysarthric speech signal and testing its strength in the classification of dysarthric speech using various pre-trained CNNs.

Compliance with ethical standards

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Funding Details: This work was not supported by any grant. This work was not carried out under any research program.

Conflict of interest: The authors declare that they have no conflict of interest.
Informed consent: None.

References

[1] S. D. Barreto and K. Z. Ortiz, "Speech intelligibility in dysarthrias: Influence of utterance length", Folia Phoniatrica Logopaedica, Vol. 72, no. 3, pp. 202–210, 2020.

[2] K. P. Connaghan and R. Patel, "The impact of contrastive stress on vowel acoustics and intelligibility in dysarthria", J. Speech, Lang., Hearing Res., Vol. 60, no. 1, pp. 38–50, Jan. 2017.

[3] E. K. Hanson and S. K. Fager, "Communication supports for people with motor speech disorders", Topics Lang. Disorders, Vol. 37, no. 4, pp. 375–388, 2017.

[4] I. Calvo, P. Tropea, M. Vigano, M. Scialla, A. B. Cavalcante, M. Grajzer, M. Gilardone, M. Corbo, "Evaluation of an automatic speech recognition platform for dysarthric speech", Folia Phoniatr Logop, 2020, doi: 10.1159/000511042.

[5]N. Souissi and A. Cherif, "Dimensionality reduction for voice disorders identification system based on Mel Frequency Cepstral Coefficients and Support Vector Machine", in 7th International Conference on Modelling, Identification and Control (ICMIC), pp. 1-6, 2015.

[6]U. N. Wisesty, Adiwijaya, and W. Astuti, "Feature extraction analysis on Indonesian speech recognition system", 3rd International Conference on Information and Communication Technology (ICoICT 2015), pp. 54-58, 2015.

[7]Megha Rughani and D. Shivakrishna, "Hybridized Feature Extraction and Acoustic Modelling Approach for Dysarthric Speech Recognition",2015.

[8]T. B. Ijitona, J. J. Soraghan, A. Lowit, G. Di-Caterina and H. Yue, "Automatic detection of speech disorder in dysarthria using extended speech feature extraction and neural networks classification", IET 3rd International Conference on Intelligent Signal Processing (ISP 2017), London, pp. 1-6, 2017. doi: 10.1049/cp.2017.0360,

[9]N P Narendra, Paavo Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences", Interspeech, 2018.

[10]Krishna Gurugubelli, Anil Kumar Vuppala, "Perceptually Enhanced Single Frequency Filtering For Dysarthric Speech Detection And Intelligibility Assessment", International Conference on Acoustics, Speech, and Signal Processing, 2019.

[11]H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria", IEEE Journal of Selected Topics in Signal Processing, vol. 14, no.2, pp. 390- 399, Feb. 2020.

[12]Daniel Korzekwa, Roberto Barra-Chicote, Bozena Kostek, Thomas Drugman, Mateusz Lajszczak, "Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech", Electrical Engineering and Systems Science , Audio and Speech Processing, arxiv: https://arxiv.org/abs/1907.04743

[13]Mohammed Sidi Yakoub1, Sid-ahmed Selouani, Brahim-Fares Zaidi and Asma Bouchair, "Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network", EURASIP Journal on Audio, Speech, and Music Processing, pp:1-7, 2020. https://doi.org/10.1186/s13636-019-0169-5

[14] S R Mani Sekhar, Gaurav Kashyap, Akshay Bhansali, Andrew Abishek A., Kushan Singh, "Dysarthric-speech detection using transfer learning with convolutional neural networks", ICT Express, Science Direct, 2021

[15] Seyed Reza Shahamiri, "Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 29, pp. 852-861, 2021.

[16] Bassam Ali Al-Qatab , Mumtaz Begum Mustafa, "Classification of Dysarthric Speech According to the Severity of Impairment: an Analysis of Acoustic Features", IEEE Access, Vol. 9, pp. 18183-18194, 2021.

[17] Amlu Anna Joshy, Rajeev Rajan, "Automated Dysarthria Severity Classification Using Deep Learning Frameworks", EUSIPCO 2020, pp. 116-120, 2020.

[18]Yeong-Hyeon Byeon, Sung-Bum Pan and Keun-Chang Kwak, "Intelligent Deep Models Based on Scalograms of Electrocardiogram Signals for Biometrics", Sensors 2019, Vol. 19, 935, pp. 1-25, 2019.