IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Learning CNN Strategy for Voice Disorder Classification and Detection

Ms. Harshita Bhagwat , Dr.Uttara Gogate

PG student, Associate Professor

Computer Engineering Department,

Alamuri Ratnamala Institute of Engineering & Technology, Mumbai

***Abstract* :** Voice disorder is the abnormal tone of voice which is produced by vocal cord infecting the viruses. If it is not detected in time then it will lead to critical conditions like permanent loss of voice or facing some other problems related to voice. In some patients the voice disorder frequently occurs but due to costly medical policies they ignore it or delay taking treatment. For that various novel techniques are used like SVM, DNN, ANN and CNN. Most of the researchers work on SVM as a deep learning technique to detect the voice disorder, but CNN is also a very efficient method for voice disorder detection. So in this paper we discussed the CNN method which is another option for voice disorder detection.in this paper we will study about the CNN methodology which gives the accuracy 92-95% using Saarbrücken voice database.

Keywords: CNN (Convolutional Neural Network), Voice disorder detection, Deep Learning Technique, voice pathological detection.

I. INTRODUCTION

Pathology of voice has badly impacted on voice functionality which turns to increase the vocal noise.[1] In voice disorder the normal voice changes into a hoarse, weak,tense voice that affects the quality of normal voice. [2] Voice disorder requires the proper and timely treatment otherwise patients will face life threatening consequences.

The disease may be common for those who are involved in teaching, music and related professions. Furthermore, there is less muscle movement at the vocal cord and victims of the deceased.[3] We can also understand diseases using voices because some diseases directlyaffect the human voice. Finding diagnoses such as frontal lobe resection, spasmodic dysphonia and cordectomy from patient voice data has become possible with today's technologies.Voice disorders due to a structural process include papilloma, asthma, laryngopharyngeal reflux caused by a cough, and granuloma. These types of voice disorders can be congenital or acquired. Voice disorders from a neuromuscular process may include cerebral palsy and vocal fold muscular dystrophy. There are other common voice disorders such as vocal fold(s) paralysis,cyst,s paradoxical vocal fold motion, and laryn- goceles Little is known about the neuropsychological outcome after frontal resection.The latest voice pathology detection methods have a biased evaluation based on subjective matters.[4] The voice pathology can be identified with two procedures namely subjective and objective methods. The subjective method deals with oral examination of a doctor with the patient. The voice test will be examined physically and the doctor can identify the level or stages of the disease. The objective method is based on supported or assistive tools with the help of software for identifying the disease.[3]

Sometimes the subjective methods are good but also they are time consuming methods, Because in oral examination patients have to visit the hospital for a test. And in objective methods a patient can send his/her recorded voice to hospital then doctors can apply the computer aided methods on it and send the test result to the patient in less time. The objective method gives better accuracy than the subjective method because it uses computer based algorithms. These methods do not depend on human decisions. Besides, they are easy to apply since the voice recordings can be made available remotely via different internet recording applications.[5] In this paper, we are using CNN algorithm to identify the voice disorder using the Saarbrucken voice database. Here both training and testing data is given to CNN algorithm for further processing.in section V we described our CNN methodology for voice detection.

II. RELATED WORK

Machine learning applications are used in many applications such as online fraud detection, image recognition, self driving cars etc.Many researchers focus on the Saarbrücken voice database (SVD) in their studies. Most researchers use SVD voice records to pathology identification. The features that are frequently extracted are entropy, energy, time, contained Mel-frequency cepstral coefficients, cepstral domains, frequency, harmonics-to-noise ratio, short-term cepstral parameters, normalized noise energy, and others[6,7] According to Martınez et al. [8], the accuracy achieved utilizing 200 records of sustained vowel /a/ represent a high value.In the studies by Souissi et al. in [9] they achieved high accuracy of 87.82% utilizing subset involving four kinds of voice pathologies that include 71 types. Also, Al-Nasheri et al. [10] achieved an accuracy of 99.68% due to their use of a subset involving a few of the pathologies to conduct a test on information that was moreover displayed in other accessible datasets, such as Arabic Voice Pathology Database (AVPD), and Massachusetts Eye and Ear Infirmary Database (MEEI). Another study conducted by Muhammad et al.[11] utilized a subset involving three kinds of voice pathologies that achieved an accuracy of 93.20%..Muhammad et al. propose a system that uses transfer learning and adopts CaffeNet.CaffeNet is a Convolutional Neural Network (CNN) which is powerful on image classifications [12], and input representations such as mel-spectrogram and octave-spectrogram can be treated as image representations of an audio signal.

In the past few years, deep learning has received considerable attention due to its superior performances on various machine learning tasks including voice disorder detection. Fang et al. propose a system that passing MFCCs into a multilayer Deep Neural Network (DNN) with a sigmoid activation function and a softmax layer as the output layer[13] It is clear that each voice disorder produces different frequencies depending on the sort of voice disorder and its area on the vocal folds[14]

III. PROPOSED METHODOLOGY

3.1. Database

For our work we use the Saarbrucken Voice database (SVD) . It is the collection of the voice recordings from more than 2000 speakers. It comprises records of 687 healthy voices (259 men and428 women) and 1356 individuals (629 men and 727 women) with different pathologies [15]. Each recording session contains i)Recordings of vowels produced at normal ,high and pitch voice ii) recordings with rising falling voice concept iii) recordings of the sentences like “Good morning ,how are you”?. For the above voice components the voice signals and the ECG signal have been stored in separate single files. Both signals can be exported in the original file format (NSP and EGG format) as well as in WAV format. If no format is selected for one of the signals, the corresponding signal files are not exported. Depending on the quality of speech where not every vowel is present in every file.The advantage of this voice task is that it is free and other linguistic confusions compared with other language standard tasks such as reading or speaking activities. This identity makes it effective to use the large database for supervised machine learning model.[16]

From the SVD database we selected the list of voices which are shown in table 1.

Table 1: SVD database

ID of recording	Type of record	Sex	Age	Pathologies	Remark
828	P	w	50	Laryngitis, myocarditis	Chronic hyperplastic, monochorditis rechts
1228	P	w	47	laryngitis	chronische
1571	P	w	47	Hyper Funktionelle, dysphonie	Ausgeprägte form, mit ubergang zur spasmodischen dysphonie
1610	P	w	44	Laryngitis, reinke oedema	Laryngitis chronisch, beidseitig, beginnend
2578	P	w	44	laryngitis	Chronisch hyperplastische
4	N	m	22	Normal	no disorder
29	N	m	58	Normal	no disorder
69	N	m	50	Normal	no disorder
103	N	m	21	Normal	no disorder
132	N	m	24	Normal	no disorder

3.2. Proposed Model

We propose a smart healthcare framework using Edge Computing (EC) and cloud computing.As a case study, we propose a voice disorder detection and classification system in the framework. The voice disorder assessment includes the detection and classification of voice.The proposed voice disorder assessment and treatment framework involves several main components: clients, assessors, SLPs, service providers, and network structure. The framework can work in a smart city. The smart city has smart homes, smart schools, a smart transportation system, and smart shopping. The clients can be citizens of the city or schools, where the students are actually the recipients. The assessors are specialists such as laryngologists. The assessors are affiliated with designated clinics, which are registered in the framework. The SLPs can move Voice Disorder Assessment and Treatment using Machine Learning between the clinic and the clients if necessary, or they can provide therapy remotely.

The traffic is controlled by a central system, which is synchronized by vehicular networks. The network structure consists of EC and core cloud computing. A software-defined network (SDN) will provide high-bandwidth flexible and programmable communication. Convolutional neural network (CNN) used to calculate the voice training set to get the result. Edge computing offers the computing resources at the edge of the network to perform smooth and near-real-time operation. In the case of conducting the therapy from a distant location, the treatment needs another automated system that can have a communication framework of speech samples without any distortion. The treatment can also be in the form of a humanoid robot, who will be instructed to interact with the patient for therapy or rehabilitation. It has been observed that patients often do not show up to hospitals for post-diagnosis management purposes if they feel better. A humanoid robot can therefore be engaged for follow-up of the patients.

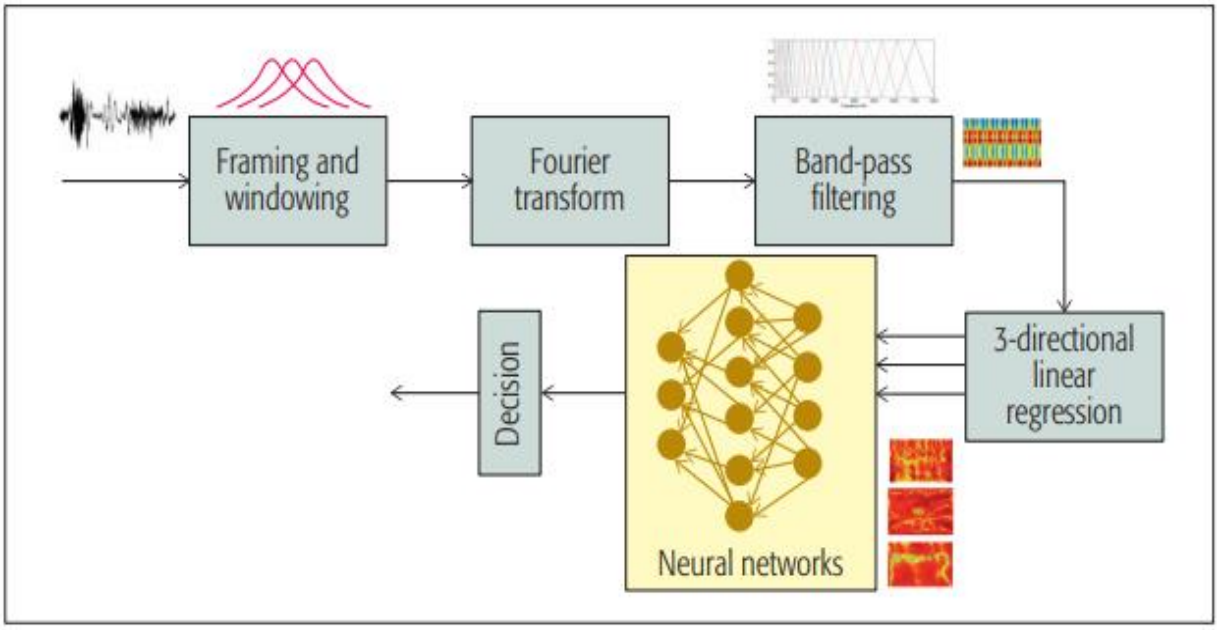

Fig 1: Block diagram of proposed voice disorder detection and classification system.

The figure 1 classifies the patient voice in two forms i.e normal voice or pathological voice using the band pass filtering and applying fourier transform algorithm on it.

3.3. CNN method

For deep diagnosis of the voice disorder we will use Convolutional neural network (CNN) as a deep learning method. Convolutional Neural Network is a widely popular deep learning architecture. A general CNN is comprised of one or more convolutional layers and max pooling layers followed by one or more fully connected layers [17] CNNs have been extensively applied to image classification tasks [18] but also have been increasingly popular for audio analysis such as speech recognition [19] and instrument activity detection [20] Borrowing the idea of passing images into a CNN, a mel spectrogram can be treated as a single channel image. An input will go through multiple convolutional layers, pooling layers, and fully-connected layers to reach a classification result. A softmax function at the output layer of the CNN can predict the probabilities of each class for a data point, and this data point will be assigned to the class with the highest probability. Due to the limited amount of data points, the designed model's structure is relatively simple. The CNN architecture used in this study consists of two convolutional layers, a pooling layer, and multiple fully-connected layers. In this CNN, the convolutional layers and pooling layer are designed to be the feature extractors, and the fully-connected layers are treated as a classifier.

The first 2-dimensional convolutional layer has 9 output channels, and the kernel is chosen to be a 5 × 5 moving window, followed by a (1, 1) stride in both directions. The second convolutional layer has 15 output channels with a 3 × 3 kernel, and it has the same stride as the previous convolutional layer. On the other hand, the following pooling layers apply 2-dimensional average pooling, and the nine fully-connected layers contain 1024, 512, 256, 128, 64, 32, 16, 8, and 4 neurons. Learned representation is expected to be more representative.[20]

The input to each convolution layer can be padded to ensure that the first and last input bands are processed by a suitable number of filters in the convolution layer. In this work, each input is padded by adding half of the filter size of dummy bands before and after the first and last bands so that the number of bands stays the same in both the input and convolution layers. Usually the top layers in CNN are fully connected just like that of a normal forward-feeding NN. These fully connected top layers are expected to combine different local structures extracted in the lower layers for the final recognition purpose.[18]

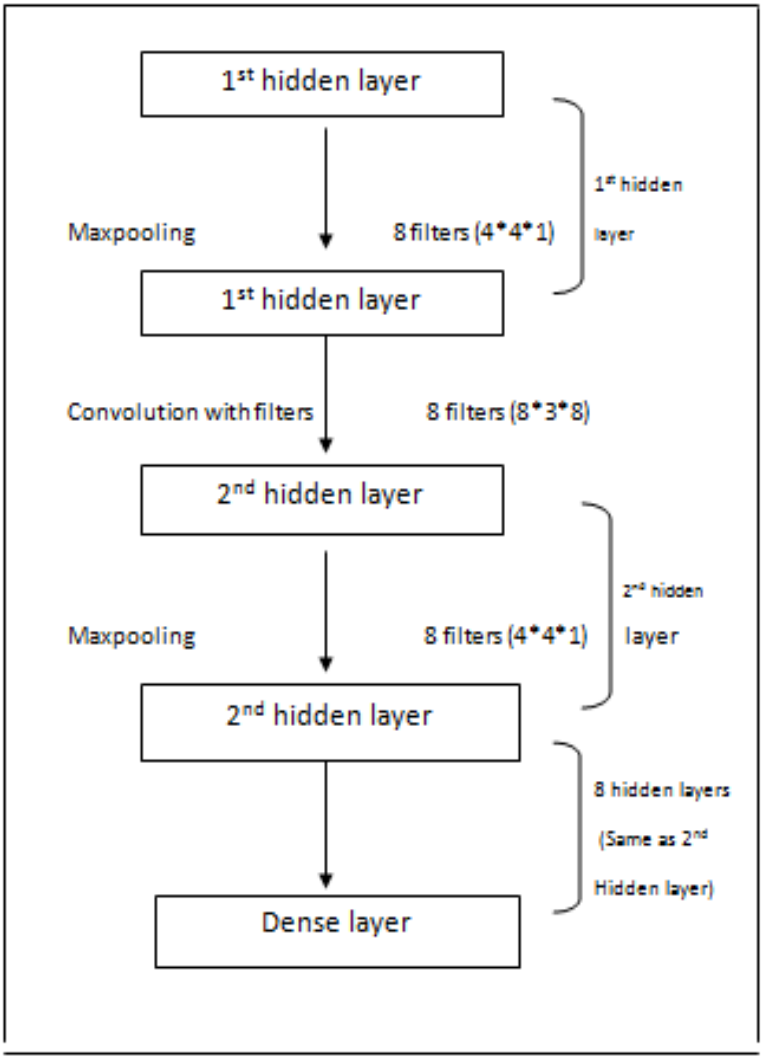

Fig 2: CNN architecture

The fig 2 represents the CNN architecture. In that the two layers are used like max pooling and convolutional filters. Using that we can detect the voice disorder at early stages and diagnose the people from before they lead to critical condition.

3.4 Experimental results

In our work, the dataset consists of a collection of 10 voice samples of both pathological and non-pathological voices which are equally present in its SVD Database.From the database we took 5 pathological and 5 normal voices for the work. So here Two types of experiments were performed: voice disorder detection and voice disorder classification. In the voice disorder detection task, the objection was to determine whether the given sample was produced by a person having a voice disorder or not. In the voice disorder classification task, the objective was to classify the type of disorder. As the number of samples of many disorders was not sufficient in the database, we chose only those disorders that had a good number of samples.

The disorders that we chose were cordectomy, functional dysphonia, hyper- functional dysphonia, Chronic hyperplastic, monochorditis rechts, Massive form mit breiter excavation, subglottisches chondroma, Ausgeprägte form, mit ubergang zur spasmodischen dysphonie, Laryngitis, Laryngitis chronisch, beidseitig, beginnend etc. The used database was divided into three parts, which are the training set, the validation set, and the testing set. 70 percent of the database was used as the training set, 5 percent as the validation set, and 25 percent as the testing set. While dividing the database into parts, we ensured that there was no overlapping between the speakers in the three parts .The proposed system was successfully deployed in the framework. Experimental results show that the proposed system could achieve high detection and classification accuracy. The performance of the selected machine learning classification techniques was evaluated in terms of accuracy, Precision and Recall area by using the following measurements:

• True Positive (TP): the voice sample is pathological and the algorithm recognizes same

• True Negative (TN): the voice sample is healthy and the algorithm recognizes same

• False Positive (FP): the voice sample is healthy but the algorithm recognizes it as pathological;

• False Negative (FN): the voice sample is pathological but the algorithm recognizes it as healthy.

The accuracy, that is the percentage of correctly classified instances, is defined as:

$$\text{Accuracy} = (TP + TN) \,/\, (TP + TN = FP + FN)$$

Precision is the ratio between the True Positives and all the Positives.

$$\text{Precision} = \text{TP}/\text{ Predicted yes}$$

The recall is the measure of our model correctly identifying True Positives

$$\text{Recall} = \text{TP}/\text{ Actual yes}$$

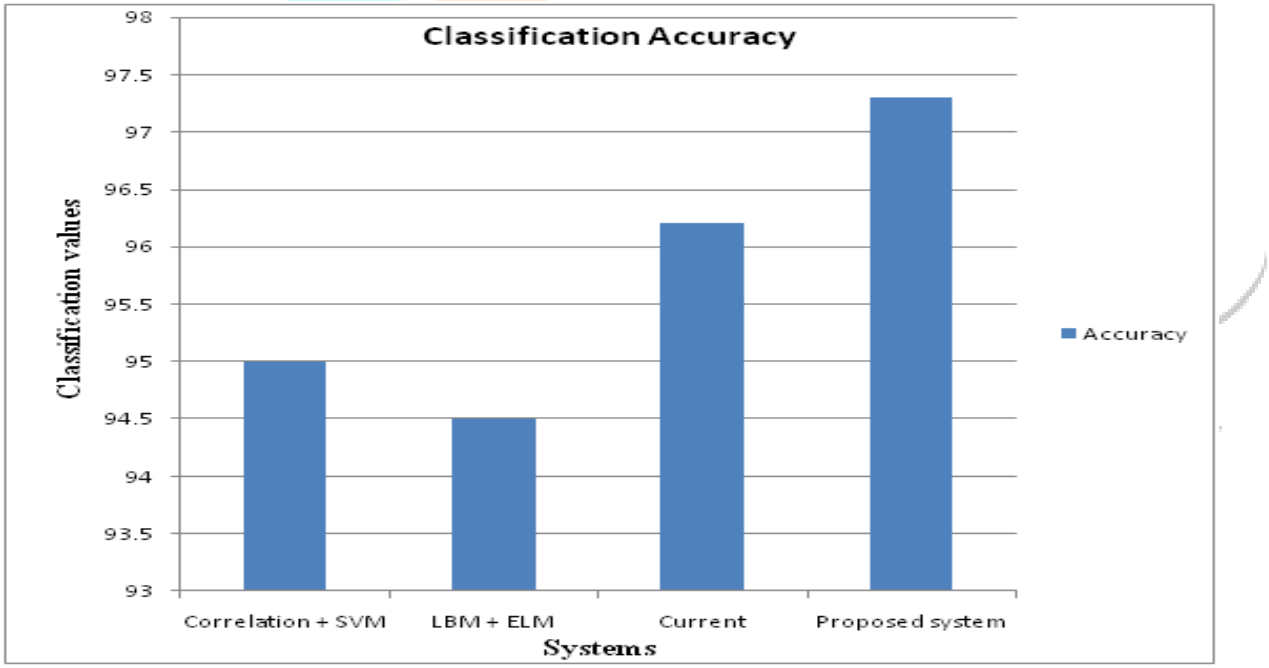

Fig 3: Classification accuracy graph

System	Accuracy
Correlation +SVM	95%
Local BInary Pattern(LBP) + Extreme Learning Machine (ELM)	94.5%
Current	96.2%
Proposed System	97.3%

Table 2: Accuracy Table

Table 2 shows the accuracy between the different machine learning algorithms .When 100 samples are analyzed by the system ,it gives 97.3% accuracy. So from that comparison we conclude that CNN also gives the best accuracy result using the algorithm.

V. CONCLUSION

A healthcare framework using machine learning techniques was proposed. This paper provided an overview of the techniques used by machine learning methods in several healthcare fields, particularly in the voice disorder detection. We can conclude the pathological voices into two forms of observation. One using medical methods by using expensive equipment to check and Second is by computer based system check by using Deep learning approach. In the framework, we develop a voice disorder assessment and treatment system using a deep learning approach. The system successfully distinguishes between pathology and normal voice with the given dataset. A client provides his or her voice sample and the sample goes for initial processing. Once complete the initial process this data forwarded to the convolutional filters and max pooling filters for the detection of voice disability and will get the results. For voice disorder classification we compared the SVM and CNN. In CNN inconsistency problems can be easily solved through the use of max-pooling. The CNN algorithm works well in the given data set and identifies the pathology and non-pathology disease. The results have shown that the best accuracy in voice pathology detection is achieved using the Convolutional Neural Network This technique classifies a voice as pathological or healthy with an accuracy equal to about 97-97.3% using all parameters.Moreover, in this work we focus on identifying appropriate voice signal features by using the comparative study of different classifiers. All analyses are performed on a wide dataset from the Saarbruecken Voice Database.

So as a result CNN is also a great option for detection of voice samples. To enhance the classification rate we will improve the classification phase by developing a hybrid system using a combination of several machine learning techniques.

REFERENCES

[1]Titze, I.R.; Martin, D.W. Principles of Voice Production; the Journal of the Acoustical Society of America. Acoust. Soc. Am. 1998, 104, 1148.
[2]. Teager, H. Some observations on oral air flow during phonation. IEEE Trans. Acoust. Speech Signal Process. 1980, 28, 599–601

[3]Voice Pathology Detection Based on Deep Neural Network Approach Jeberson Retna Raj1, J Jabez1, S Senduru Srinivasulu1, S Gowri1 and J S Vimali1
Published under licence by IOP Publishing Ltd(2020)

[4]Hillenbrand, J.; Houde, R.A. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. J. Speech Lang. Hear. Res. 1996, 39, 311–321

[5]Mehta, D.D.; Hillman, R.E. Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. Curr. Opin. Otolaryngol. Head Neck Surg. 2008, 16, 211

[6]Dubuisson, T.; Dutoit, T.; Gosselin, B.; Remacle, M. On the use of the correlation between acoustic descriptors for the normal/pathological voices discrimination. EURASIP J. Adv. Signal Process. 2009, 2009, 173967.

[7] Fredouille, C.; Pouchoulin, G.; Bonastre, J.F.; Azzarello, M.; Giovanni, A.; Ghio, A. Application of automatic speaker recognition techniques to pathological voice assessment. In Proceedings of the International Conference on Acoustic Speech and Signal Processing (ICASSP 2005), Philadelphia, PA, USA, 23 March 2005.

[8]Martínez, D.; Lleida, E.; Ortega, A.; Miguel, A.; Villalba, J. Voice pathology detection on the Saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In Advances in Speech and Language Technologies for Iberian Languages; Springer: Berlin/Heidelberg, Germany, 2012; pp. 99–109.

[9]Souissi, N.; Cherif, A. Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine. In Proceedings of the 2015 7th International Conference on Modelling, Identification and Control (ICMIC), Sousse, Tunisia, 18–20 December 2015; pp. 1–6.

[10]Al-nasheri, A.; Muhammad, G.; Alsulaiman, M.; Ali, Z. Investigation of voice pathology detection and classification on different frequency regions using correlation functions. J. Voice 2017, 31, 3–15.

[11]Muhammad, G.; Alhamid, M.F.; Hossain, M.S.; Almogren, A.S.; Vasilakos, A.V. Enhanced living by assessing voice pathology using a co-occurrence matrix. Sensors 2017, 17, 267

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 675–678.

[13]Guan, H., & Lerch, A. (2019). Learning Strategies for Voice Disorder Detection. 2019 IEEE 13th International Conference on Semantic Computing (ICSC). doi:10.1109/icosc.2019.8665504

[14]Pouchoulin, G.; Fredouille, C.; Bonastre, J.F.; Ghio, A.; Révis, J. Characterization of the Pathological Voices (Dysphonia) in the Frequency Space; International Congress of Phonetic Sciences (ICPhS): Saarbrücken, Germany, 2007; pp. 1993–1996.

[15]Mohammad M. , Abdulkareem H. (2020) Voice Pathology Detection and Classification Using Convolutional Neural Network Model. doi.org/10.3390/app10113723

[16]Arunkumar, N.; Mohammed, M.A.; Ghani, M.K.A.; Ibrahim, D.A.; Abdulhay, E.; Ramirez-Gonzalez, G.; de Albuquerque, V.H.C. K-means clustering and neural network for object detecting and identifying abnormality of brain tumor. Soft Comput. 2019, 23, 9083–9096.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[18] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[19] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2012, pp. 4277–4280.

[20] S. Gururani, C. Summers, and A. Lerch, "Instrument activity detection in polyphonic music using deep neural networks," in International Society for Music Information Retrieval (ISMIR), 2018.