Ingénierie des Systèmes d'Information
Vol. 31, No. 5, May, 2026, pp. 1635-1644
Journal homepage: http://iieta.org/journals/isi

VocalCare: Convolutional Neural Network Model for Laryngeal Disorder Detection

Manisha B. Gharde[1*], Vaishali V. Patil[2]

[1] Department of E&TC, AISSMSIOIT, Pune 411001, India
[2] Department of E&TC, International Institute of Information Technology (I²IT), Pune 411057, India

Corresponding Author Email: manishagharde30@gmail.com

https://doi.org/10.18280/isi.310519

Received: 12 February 2026
Revised: 10 April 2026
Accepted: 25 April 2026
Available online: 31 May 2026

Keywords:
laryngeal disease, deep learning, convolutional neural network, speech disorders, speech pathology, prediction

ABSTRACT

Speech disorders caused by laryngeal abnormalities significantly affect communication ability and quality of life. Early detection of pathological speech conditions can support timely clinical intervention and improve patient management. This study proposes Vocal Care, a convolutional neural network (CNN)-based framework for automatic pathological speech classification using acoustic speech features extracted from the Saarbrucken Speech Database. Experimental results demonstrated that the CNN model achieved strong classification performance and reliable detection capability. Most conventional diagnoses are made using invasive procedures called laryngoscopes. In this study, we explore the use of deep learning methods such as CNN, to help Doctors with their efforts to identify patients who may be suffering from laryngeal disease. We will look at how the techniques we used (Mel-frequency cepstral coefficients) together with features like jitter, shimmer and harmonics-to-noise ratio can enable Doctors to develop and train models of their patients' speech, evaluate each model's performance based on metrics such as accuracy, precision, recall and F1-score. The key contribution of this paper is the investigation and comparative analysis of several machine learning classifiers alongside a CNN model for multiclass pathological speech detection. Ultimately, this approach to multiclass analysis to provide for the early identification of laryngeal disorder detection.

1. INTRODUCTION

One of the biggest challenges facing contemporary healthcare systems is the early and precise detection of several diseases. Traditional binary classification algorithms are inadequate in handling real-world clinical circumstances when patients may present symptoms matching to numerous illness categories. In this paper, an automated multiclass classification framework for machine learning-based disease detection and differentiation is presented. The comprehensive analysis of major features was used to catch the differentiating characteristics for different forms of clinic and clinical data. In forecasting various types of diseases, 4 different machine learning models were tested such as support vector machine (SVM), k-nearest neighbors (KNN), random forest (RF) and convolution neural networks (CNNs). The suggested model is proved by experimentations of large numbers of samples that demonstrated high-quality predictions and reliability and indicates the ability for assisting various clients with accurate and fast detection of diseases.

The human larynx is a functional structure with a complex shape located in the neck and plays vital roles in respiration, sound, and airway protection. Potential states of the human glare can greatly impact one's ability to breathe, speak, and sound fully. For example, laryngeal papillomatosis, vocal fold nodules, vocal fold polyps, vocal fold paralysis and/or cancer are potential outcomes of the human glare, and therefore, can have a negative social or psychological effect on the individual as well as the community. Therefore, the timely and accurate detection of laryngeal disorders is paramount to achieving appropriate and correct medical treatment; subsequently improving and enhancing the likelihood of positive medical treatment outcomes and minimizing the long-term effects of laryngeal dysfunction. Laryngeal disorders can be diagnosed using traditional diagnostic methods such as endoscopic, video stroboscopic, and laryngoscope procedures. These traditional diagnostic procedures have been successful however, they are expensive, time consuming, very confidential, and largely subjective based on a medical practitioner's skill set. Additionally, due to restricted access to specialized laryngology services, diagnosis and treatment may be delayed in distant or resource-constrained places. These problems show how important it is to have automated, non-invasive, and affordable ways to find laryngeal disorders early.

Deep learning has made major strides in artificial intelligence, which has greatly increased the ability to analyze acoustic signals. This has altered the field of medical diagnosis. Convolutional neural networks CNNs, which are a sub-class of deep neural network (DN) algorithms, have demonstrated high levels of success in the classification of images and audio signals through their ability to learn structural features from unprocessed input data with regard to the diagnosis of laryngeal diseases, it is common to convert the sound waves collected from a patient's speech, into visual

representation such as a spectrogram, or some other representation in time-frequency space. Within these images are subtle and difficult-to-identify patterns that can be effectively learned and found using the convolution neural network (CNN)-based analysis model. Numerous previous studies have highlighted the significant potential of using AI to support clinical decision making, specifically with regard to the use of machine learning and deep learning to detect speech disorders. In general, these studies have focused on extracting the features that will be used to classify a patient's speech (e.g. pitch, jitter, shimmer, Mel-Frequency Cepstral Coefficients (MFCCs), followed by using traditional classifiers, such as a RF or SVM, to classify the extracted features of a patient's speech.

While there are certainly promising results from the use of these techniques, they are primarily based upon the use of manually defined features and because of this, they may have difficulty generalizing across different patient populations. CNNs on the contrary, provide the ability to automatically extract the required features from the input (i.e. patient's speech) and provide a greater degree of robustness and predictability, thereby improving the accuracy of the predictions.

In this paper we introduce a novel Speech Recognition System based on CNNs for identifying Laryngeal Disease through recorded Speech Data. This method will use spectrogram "visualizations", which are a representation of sound, in conjunction with CNNs to extract the distinguishing features of Speech Data and separate the two types (diseased vs. healthy). This system was designed to be: scalable, non-invasive, allow physicians to diagnose their patient's condition sooner and reduce the need for invasive diagnostic measures, and to expand the reach of high quality medical care.The study discusses the potential benefits of incorporating AI technology into such fields as speech pathology and otolaryngology. This framework will help doctors to make clinical decisions and will provide novel telemedicine applications and remote monitoring for speech health, offering a reliable and automated diagnostic tool for the diagnosis of laryngeal illnesses. The comprehensive testing and analysis of the CNN-based model has demonstrated the usefulness of the model in diagnosing laryngeal illness and can potentially transform present diagnostic techniques and improve patient outcomes. This complete chain makes it possible for CNN to perform reliable automatic acoustic assessment of first vocal disorders.

This paper further highlights the possible benefit of the application of AI in speech pathology and otolaryngology. The proposed architecture supports clinical decision making, and provides a gateway for telemedicine applications, remote monitoring and continuous speech health assessment by providing a reliable automated tool for diagnosis of laryngeal diseases. We show that the CNN-based model can be effectively used for the diagnosis of laryngeal disorders by applying the model to a large number of patients and analyzing the results. It has the potential to change the current diagnostic procedures and enhance patient outcomes. The whole cycle allows the CNN to perform reliable, automated acoustic analysis for preliminary vocal pathology screening. Speech pathology identification has been extensively studied using standard signal processing techniques as well as recent deep learning approaches. Recent approaches have used CNNs to learn discriminative features from the speech input, whereas the former approaches focused mostly on acoustic and spectral properties.

The novelty of the proposed Vocal Care system is the construction of an intelligent CNN based framework for autonomous laryngeal health monitoring using human speech data. The proposed strategy is non-invasive, cost-effective and detects speech abnormalities at an early stage, in contrast to the conventional clinical diagnostic methods based on invasive techniques such as laryngoscopy and manual analysis by experts. The proposed method adopts the advanced machine learning and deep learning methods. The Mel-Frequency Cepstral Coefficient (MFCC) is used to automatically extract the discriminative acoustic features from the speech signals, which minimizes the dependence on the manual feature extraction methods and clinical interpretation. The CNN-based framework learns important speech patterns related to laryngeal anomalies in an efficient manner, resulting in higher accuracy and reliability in the early diagnosis of vocal disorders. A comparison analysis of standard ML methods with the proposed CNN model reveals that the deep learning based framework offers better classification accuracy, robustness and faster detection performance for identification of laryngeal anomalies.

Furthermore, the Vocal Care system can also help improve access to healthcare services by allowing patients to remotely and continuously monitor their speech health especially in rural and underserved areas. Early stage diagnosis is also possible with the suggested approach thereby reducing the chance of severe speech problems by appropriate intervention.

2. RELATED WORK

Speech processing is employed in many different disciplines, such as speech therapy, emotion detection and assessment, and disease diagnosis, and has become widely used in the analysis of medical signals. Recent advances in machine learning have enabled researchers to construct accurate methods for laryngeal pathology identification when paired with signal processing approaches. Verde et al. [1] reviewed several classifiers for identifying dysphonic speech and found that SVM and Decision Trees offered the highest levels of accuracy for identifying dysphonic speech

Kharibam and Devi [2] supported this conclusion and confirmed that MFCCs contained sufficient information necessary for speech and speaker recognition. Miliaresi and Pikrakis [3] Additionally, using deep multimodal neural networks to process multiple types of data medical records and speech produces better, more generalizable models than using only traditional computer vision techniques. Thus, speech processing has a significant impact on the evolving field of speech analysis in relation to biological research. As evidenced by their use of different machine learning models (SVMs and decision trees) to develop effective models for detecting speech pathology (dysphonia), Stewart et al. [4] evaluated the relationship between clinician ratings and patient self-ratings in individuals who have been diagnosed with an adductor focal laryngeal dystonia (AD-FLD) by using standardized perceptual tools, such as the Unified Spasmodic Dysphonia Rating Scale (USDRS) and the Speech Handicap Index (VHI) to complete their evaluations. The authors validated that MFCC is a reliable set of acoustic features for speech classification. Kraxberger et al. [5] used machine learning to model speech using flow-induced techniques; Hlavnička et al. [6] studied vocal vibrations for the purpose of detecting Parkinson's disease (PD); and Balaji and

Sadashivappa [7] reviewed various speech recognition implementations for adults with speech disabilities. Finally, Khara et al. [8] compared different techniques to extract speech features, concluding that MFCC is the most efficient method to analyze speech for medical purposes. Furthermore, Islam et al. [9] utilized the concept of expanding MFCC [10] to deep multi-modal architectures with the help of medical metadata to improve the performance. Gharde and Patil [11] discuss laryngeal disorder using machine learning. Shimbre and Solanki [12] Acoustic measurements included their determination of a fundamental frequency (F_0) and the ratio of subharmonic to harmonic (SHR) measurements [13].

Traditional models of machine learning continue to play an important role in identifying smaller datasets, providing transparency for healthcare providers and having fewer computational requirements [14]. Therefore, for these reasons, three different machine learning classification methods SVM, KNN, and RF were selected to be benchmark classifiers in identifying speech-related pathologies for this study. Ai et al. [15] proposed the classification of speech dysfluencies using MFCC and LPCC featuresWhile laryngeal speech disorder diagnosis has changed from a perceptual approach in the past to a data-driven approach based on acoustic measurements, the majority of clinical studies completed within each area of dysphonia have been based on visual and auditory evaluations up until recently [16].

Their results demonstrated that the two forms of measurement (subjective and objective) were strongly correlated to one another, supporting that clinicians and patients have similar perceptions of severity of dysphonia. However, they noted that both assessment types require the use of invasive visualization techniques, thus indicating a need to develop alternative non-invasive clinical options [17].

MFCC is the most effective way to evaluate speech for medical reasons after comparing several methods to extract speech aspects. Traditional approaches of machine learning continue to play a significant role in detecting smaller datasets, giving transparency for healthcare providers and having lower computing requirements. These two aspects, interpretability and decreased computational requirements, are vital when it comes to successfully adopting machine learning technology in clinical practice. Therefore, for these reasons, three distinct machine learning classification algorithms (SVM, KNN, and RF) were selected to be benchmark classifiers in identifying speech-related disorders for this work. While laryngeal speech problem diagnosis has moved from a perceptual approach in the past to a data-driven approach based on acoustic measures, the bulk of clinical research done within each area of dysphonia have been based on visual and auditory evaluations up until recently. Today, clinical trials are being undertaken employing signal processing and machine learning techniques to produce increased diagnosis accuracy [18]. Using standardized perceptual instruments, such as the VHI and the USDRS assessed the association between clinician ratings and patient self-ratings in patients with an AD-FLD. Their findings supported the idea that patients and doctors perceive dysphonia severity similarly by showing a good correlation between the two types of evaluation (subjective and objective). They did point out that both forms of assessments necessitate the use of invasive imaging techniques, which suggests that alternative non-invasive clinical solutions need to be developed. By developing a specific simulation tool based on, in their research on how humans produce vocal sounds using different parts of their bodies, extended the understanding of this field through the use of simulated vocal sound data and machine learning techniques. Using SVM classification of simulated data, they were able to differentiate between types of glottal closures and levels of subglottal pressure with over 91% accuracy. They also produced a computer based research method that allows for the examination of individuals with disordered vocal production to do so without the need for invasive procedures through the use of acoustic feature extraction parameters like cepstral peak prominence and harmonics-to-noise ratio.

In their research on the neurological system, created a pitch-tracking method applying a Kalman filter technique to detect subharmonic vibrations connected to persons diagnosed with PD and atypical parkinsonian syndromes (APS). Using speech recordings from test subjects producing vowel sounds/no vowels longer than one second, researchers were able to use their system to make clear discriminations of acoustic emission between “normal” vowels and vowels from people with PD or ataxia (APS).

The findings demonstrate that acoustic assessment can be used for evaluating neuromotor problems, thus supporting the use of noninvasive acoustic speech measurements for diagnostic purposes. In addition to documenting the findings of the study reviewed the use of Automatic Speech Recognition (ASR) technologies to evaluate individuals with disorders such as stuttering and dysarthria. Kinahan et al. [19] describe the limitations of traditional ASR systems when input originates from an individual with a speech disorder; hence, they recommend developing adaptive and/or user-specific ASR systems for enhancing accessibility to the ASR.

Their work developed a link between speech technology and clinical practice, which demonstrated that ASR can function as both rehabilitative and diagnostic utility in providing speech and language treatment [20] suggested an automated method for speech pathology diagnosis determined by CNN classifier is used to classify the input data as abnormal or healthy speech signal.

3. METHODOLOGY

3.1 Dataset details

The experiments in this study were conducted using the Saarbrücken Speech Database (SVD), a publicly available pathological speech dataset widely used for speech disorder analysis research. The database contains recordings from healthy individuals as well as patients diagnosed with different pathological speech conditions.

A total of 1,037 speech recordings were utilized in this work. The dataset includes sustained vowel phonations and pathological speech samples collected under controlled recording conditions. Our study used speech recordings obtained from the SVD, which contains sustained vowel sounds (/a/, /i/, /u/) from both healthy people and patients with laryngeal pathologies, such as paralysis and nodules. Each recording is captured at 50 kHz and has a 16-bit resolution. The vowels are /a/, /i/, and /u/ are recorded in this set together with an appropriate sentence. When evaluating a patient's speech quality, vowels are preferred above linguistic artifacts. To conduct the experiments on in our experimental testing, Specifically, 477 pathological speech (392 samples of Vox Senilis & 85 samples of Laryngocele) and 560 healthy samples have been selected. The dataset was divided into training and

testing subsets using stratified sampling to preserve class distribution across both partitions, with 80% used for training and 20% for testing. Experiments were conducted to evaluate the performance of the proposed model in accurately classifying multiple diseases across different categories.

3.2 Feature extraction

MFCC: This is popular feature techniques for extraction in the field of speech recognition and audio processing. MFCCs are especially popular in tasks like as speech recognition, speak identification, and music data retrieval. Following are the steps for Mel-Frequency Cepstral Coefficients.

Pre-processing: The speech signal first undergoes pre-processing to eliminate areas of noise and silence. The signal is often broken into brief frames of 20 to 40 millisecond duration, with only a slight overlap among consecutive frames.

Mel-scale Filter Library: The FFT is used to determine the power spectrum of every frame. The power spectrum is then routed through a bank of filter spaced out on the Mel-scale to better simulate the non-linear perception of frequencies by the human auditory system. The Mel-scale is a pitch perception scale based on the human ear's reaction to various frequencies.

Logarithm: The filter bank energies' logarithm is computed. This stage compresses the spectrum's dynamic range and simulates the non-linear human sense of loudness.

Discrete cosine transform (DCT): The energies of the log filter bank are subjected to a discrete cosine transformation. DCT de-correlates motion coefficients.

Coefficient Selection: After applying the DCT, typically a selection of the resulting coefficients is made, leaving out the higher-frequency coefficients that hold less pertinent information. The selected coefficients then serve as the final MFCC characteristics.

3.3 Multiclass classification models

In this research, we used MFCCs to classify the speech waveform data into three categories; Laryngeal pathologies, Normal speech, and Vox senilis speech. MFCCs are frequently used to analyze speech disorders and efficiently model the perceptual properties of speech sound produced by the human body. Consequently, we developed classification algorithms based on the MFCC characteristics identified to classify the speech waveform data into the three categories: SVM, KNN, RF and CNN were developed, and we analysed their performance in differentiating normal and sick speech through performance measures such as accuracy, sensitivity and specificity, as well as the F1 score, through confusion matrix-based performance analyses. The laryngeal sample and the vox senilis (senior citizens) sample, had very slight misclassifications between both samples; however, there were a large degree of separation between both samples Laryngeal & Normal than vice versa (Normal & Vox Senilis) & thus, they had different classification performances; therefore, the misclassifications could be attributable to the sound spectra and auditory characteristics that were similar; but still the overall results indicate that MFCC acoustic measurements are a valid representation of the speech produced by people with Laryngeal disorders such as those produced in a normal vocal cord or Vox Senilis such as with senior citizens and have been helpful in diagnosing & assessing speech disorders.

Three techniques for machine learning were used:

•SVM: Models non-linear class boundaries using an RBF kernel.

•KNN: This method uses Euclidean feature space similarity to classify samples.

•RF: Reduces overfitting and increases prediction stability by combining decision trees.

CNN: A neural network that uses an MFCC representation for speech samples and is taught via deep learning. This deep learning model is able to use these MFCC representations of sound to learn multiple levels of spectro-temporal features associated with the laryngeal speech, such as normal speech, abnormal speech and senile speech. Ramitha et al. [18] used a CNN model to accurately identify very small sounds associated with the laryngeal speech.

3.4 Evaluation metrics

Performance was measured using various metrics, including accuracy, precision/recall, specificity, and F1-score, while k-fold cross-validation was used to improve generalization. The robustness of the proposed classification method for speech pathology was tested by evaluating the performance of each model using a strategy of 5-fold cross validation. The data set of available speech was divided at random into five semi-equal size subsets; the first 4 subsets were used for training and the remaining subset was kept out for testing (guaranteeing that each speech sample was examined precisely once). By combining all of the results from each fold, the overall performance of the classifier could be assessed without bias on new speech samples.

The following evaluation metrics measured the diagnostic capabilities of the model:

Accuracy: Accuracy is an estimate of how accurate the classifier is based on how a lot of a speech samples had been accurately classified as being diseased or not diseased. Even though accuracy provides a general view of a classifier's performance, it can be significantly affected by the class imbalance (i.e., fewer samples in the speech disorder classes) commonly seen in most speech pathology data sets.

Sensitivity (Recall) indicates how well the model distinguishes between various speech conditions (example: Disease and No Disease) by dividing the number of True Positives for each speech condition. High sensitivity is especially important when evaluating the speech for clinical reasons since high sensitivity indicates that the speech evaluation system has a high probability of identifying pathological conditions therefore reducing the number of missed diagnoses.

Specificity measures how well the classifier distinguishes between Non-pathological speech samples by dividing the number of True Negatives (non-pathological speech samples) by the total number of Non-pathological samples. This is important for lowering the number of false alarms and making sure that normal speech samples aren't wrongly classified as pathological.

F1-Score: The F1-score is a balanced performance evaluation that is especially useful for imbalanced speech datasets, where some vocal disorders may be underrepresented. It is represented by the harmonic mean of precision and recall. The F1-score gives an all-encompassing evaluation by taking false positives and false negatives into account at the same time.

Collectively, these measures give a detailed and clinically meaningful analysis of the efficacy of the proposed approach in categorizing speech pathologies.

3.5 Speech based laryngeal abnormality detection system

Figure 1 shows that this method uses a CNN to analyze acoustic data taken from recorded speech in order to classify speech signals into Normal or Abnormal (Laryngeal Disorder) categories.

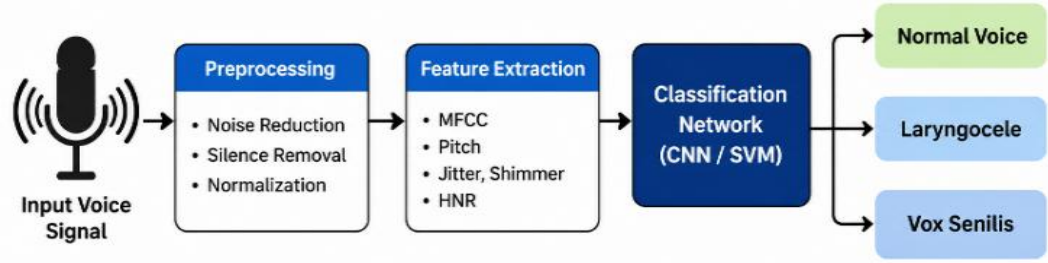

Figure 1. Speech based laryngeal problem identification utilizing convolutional neural network (CNN)

3.5.1 Pre-processing and signal acquisition

A task that may be performed during the experiment is for the participant to extend their vocalization of one of the vowel sounds (a, i, or u) in order to provide a speech sample in a controlled environment. Normalization of each recording takes place, where the amplitude is corrected for any changes in both the volume of the speaker and the distance of the microphone from the speaker, so that the recordings are normalized. Prior to performing the frequency analysis on the continuous one-dimensional (1D) signal, frame blocking is used to separate the continuous 1D signal into a number of segments that overlap one another. An appropriate window function is then applied to each of the segments in order to reduce distortion.

3.5.2 Extracting features with Mel-Frequency Cepstral Coefficients

In order to create the needed two dimensional (2D) input required by the CNN, the MFCCs were used to create a 2D representation of the human ear's hearing perception. Each frame's frequency spectrum was calculated and passed through a Mel-frequency filter bank comprise of filters designed to highlight low frequencies and then the outputs were truncated as perceivable loudnesses. In order to obtain MFCCs, the DCT technique was used on the outputs from the previous stage. The resulting sequence of MFCC vectors thus forms a 2D feature matrix (time × coefficients), which is required for the CNN's inputs.

The CNN extracts and learns discriminative speech patterns through two main stages:

3.5.3 Feature learning

The layered structure of convolutional networks allows for a reduced number of calculations and an improvement in the ability to generalize from images at the same time as allowing for the recognition of localized spectral patterns from filtered input. Convolutional filters identify localized spectral patterns in the input through a set of convolutional (convoluted) networks using Rectified Linear Unit (ReLU) activations, which add non-linearities to the output.

3.5.4 Classification

After being flattened into a 1D vector, the final pooled features are sent to fully connected (FC) layers, which combine the learned features to get a final conclusion. Each class's likelihood scores are produced by the output layer.

3.5.5 Classification results

Based on machine learning and deep learning classification algorithms with probabilities, the system offers a binary categorization:

3.5.6 Normal Speech (Healthy)

The normal class has a higher probability.

3.5.7 Abnormal Speech (Laryngeal Disorder)

Potential vocal pathology is indicated by a higher likelihood assigned to the abnormal class.

This CNN-based system efficiently transforms unorganized speech signals into structured representations. By employing automated acoustic pattern analysis, it enables robust, data-driven diagnosis of laryngeal diseases.

4. CONVOLUTION NEURAL NETWORK-BASED VOCALCARE FRAMEWORK FOR LARYNGEAL HEALTH MONITORING

The proposed CNN-based Vocal Care framework for laryngeal health monitoring is illustrated in Figure 2. Initially, speech signals are provided as input to the system and transformed into MFCC representations for effective acoustic feature extraction. The extracted MFCC features are then processed through multiple convolutional layers combined with ReLU activation functions to automatically learn discriminative pathological speech characteristics. Max-pooling layers are incorporated to reduce feature dimensionality while preserving significant information. To minimize overfitting and improve model generalization, dropout regularization is applied before the fully connected dense layer. Finally, the softmax output layer performs multi-class classification of healthy and pathological speech samples [16].

The detailed configuration of the proposed CNN model is summarized below:

- Number of Convolution Layers: 3
- Kernel Size: 3 × 3
- Activation Function: ReLU
- Pooling Operation: Max Pooling (2 × 2)
- Dropout Rate: 0.5
- Optimizer: Adam
- Learning Rate: 0.001
- Batch Size: 32
- Epochs: 50

The detailed configuration of the proposed CNN model is summarized as follows: the architecture consists of three convolutional layers with a kernel size of 3 × 3 and ReLU

activation functions. Max Pooling operations of size 2 × 2 are applied for dimensionality reduction, while a dropout rate of 0.5 is utilized to reduce overfitting. The model is optimized using the Adam optimizer with a learning rate of 0.001. Training is performed using a batch size of 32 for 50 epochs.

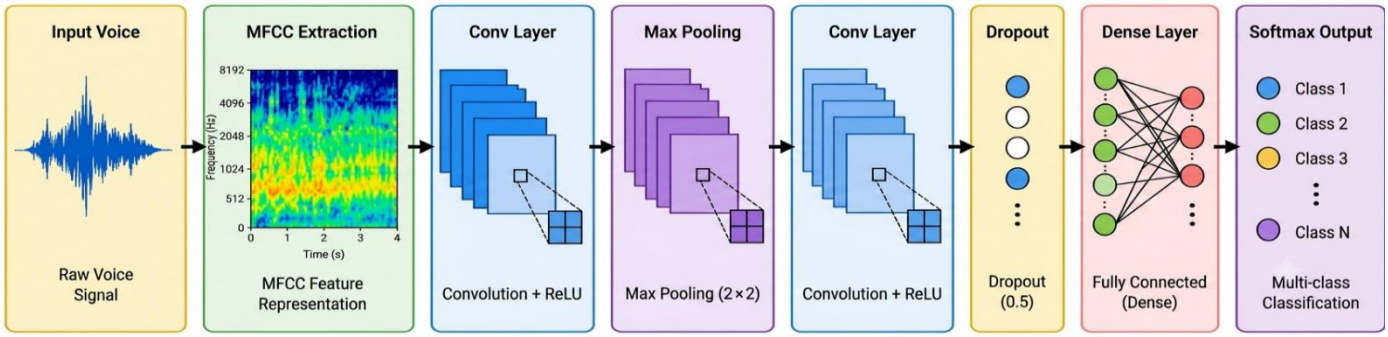

Figure 2. Convolution neural network (CNN)-based VocalCare framework for laryngeal disorder detection

5. RESULTS AND DISCUSSION

5.1 Normal, Vox Sensils, and Laryngocele results using Support Vector Machine, k-Nearest Neighbors, and Random Forest

This study shows that by experimenting with different combinations of vocal fold conditions (Normal, Vox Sensilis, Laryngocele) and using several different supervised learning algorithms (SVM, KNN, RF), it is possible to effectively classify pathological speech disorders. A dataset consisting of 560 healthy, 85 laryngocele, and 395 vox senilis was used in this study. The classifiers were tested using 5-fold cross-validation so that each speech sample was in the test set exactly once. A combination of MFCCs was created to use as input to the classifiers; jitter, shimmer, and HNR features were added to improve the discriminative capabilities of the classifiers. The classifiers will be trained to classify speech as healthy, laryngocele, or vox senilis types. The SVM classifier was tested using 5-fold cross-validation, with each fold used once as the test set and the remaining folds for training. This method allowed every sample in the collected data to contribute to the testing process.

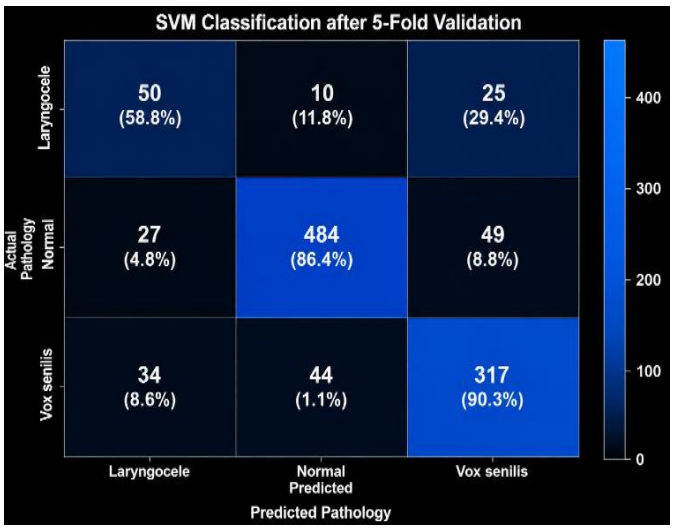

Figure 3. Classification results obtained by Support Vector Machine (SVM) algorithm for Normal, Vox Sensilis, and Laryngocele class

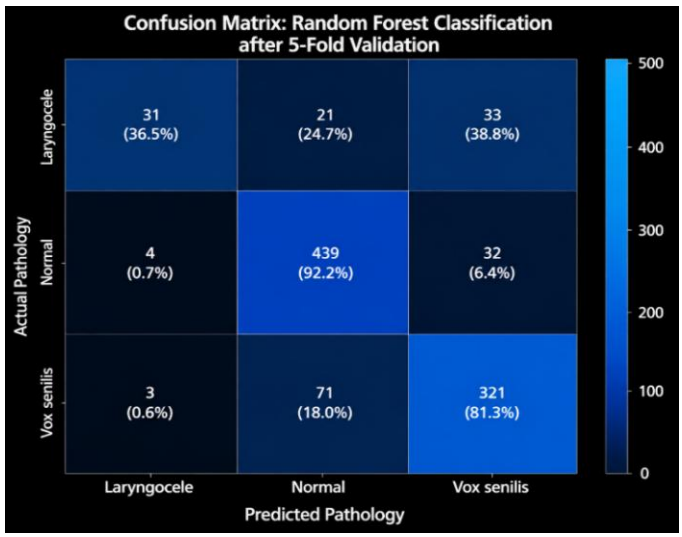

Figure 4. Classification results obtained by random forest (RF) classifier for Normal, Vox sensils and Laryngocele class

Figure 3 indicates that confusion matrix obtained using the SVM classifier shows an overall accuracy of the class-wise accuracy of the SVM classifier after 5-fold cross-validation shows varying performance across the three classes. The Normal class achieved the highest accuracy of 86.4%, indicating that the model is highly effective in correctly identifying normal speech samples. The Vox senilis class was highly accurate, achieving an accuracy rate of 80.3%. There was, however, some misclassification of Vox senilis with other classes due to confusion regarding their characteristics; nonetheless, the Vox senilis class still performed relatively well overall. The Laryngozele class, on the other hand, was not as well detected, obtaining an accuracy of 58.8%. This suggests more difficulty in identifying Laryngozele conditions because it was frequently misclassified as Vox senilis.

Overall, the classifier achieved an excellent overall accuracy of 81.8%, but there was a lot of variation in class-level accuracy. The results suggest that reporting class-level accuracy in addition to total accuracy is important when evaluating clinical speech pathology classification methods, especially when classifying minority classes, which are generally more difficult to detect.

Figure 4 indicates the RF classifier is shown to provide excellent performance on the speech dataset and produces a

strong overall result of 83.75% when tested with 5-fold cross-validation on the speech pathology dataset. This model successfully identifies "healthy" speech with a very high sensitivity of 92.68%. The model also is able to successfully recognize cases of "vox senilis" (elderly, aged) at a rate of 81.27%. It is important to note that there was confusion between the two classes and many of the classified cases were misclassified either as healthy or as vox senilis.

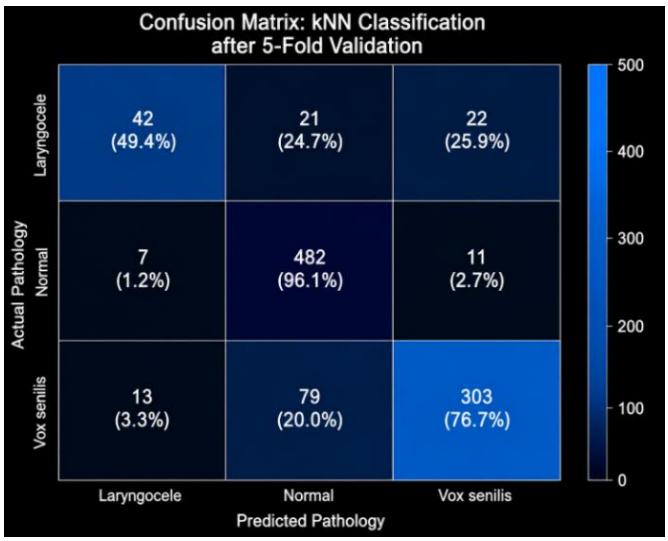

Figure 5. Classification results obtained by k-nearest neighbors (KNN) algorithm for Laryngocele, Healthy and Vox sensils class

The results suggest that the confusion between the two classes may be attributable to a class imbalance and the fact that there is a significant overlap in the acoustic characteristics of the two classes. As illustrated in Figure 3, the RF model exhibit a significant increase in robustness and ability to discriminate between the two classes when compared to the simpler models especially with regard to the healthy and vox senilis. Figure 5 indicates KNN Classifier Solutions for Speech Recording Classification Shows ~80% Accurate Classification of Speech Recordings Overall. The KNN Classifier Shows High Performance in Classifying Healthy Speech, with a High Percentage of Healthy Speech at 86%. The KNN Classifier Shows Moderate Performance in Classifying Vox Senilis at 76.71%, Demonstrating the Overlap between the Normal Speech and Age-Related Speech. At the Lower End of the Classification Percentage (%of 48.91%), Laryngocele Is Not as Easily Classified by KNN than by Other Classes. The Figure 4 Supports the See that KNN Is a Good Example of Classification of Healthy Speech. Results of the Experiments Support that the Feature Set Described Above Will Enable the Modeling of the Characteristics of Pathologic Speech Using the Classifiers That Were Used for This Study. The Classification of Healthy Speech is Very High (and Accurate) Because of It Unique Acoustic Signature and the High Number of Included Healthy Speech in the Data Set and Vox Senilis However, Pathological Conditions Such as Laryngocele Were Not Easily Resolved Because of Limited Samples of Pathological Speech. and overlapping acoustic features with vox senilisThe study shows that using machine-learning based techniques in pathological assessment of speech has potential for being used in practice and highlights the need for appropriate feature selection and case mix in relation to diagnostic accuracy and generalization. It was determined that when using 40 MFCC features, Jitter, Shimmer and Harmonic to Noise Ratio (HNR) the accuracy of the three models improved considerably. The RF classifier had the highest accuracy of 83.17%, which shows its advantage for processing non-linear patterns.

The next highest accuracy was obtained by the KNN method at 78.85% and is considered reliable in identifying speech that have similar acoustic properties. The SVM classifier also demonstrated a moderate accuracy of 76.92% and indicates some difficulty with classifying overlapping feature distributions. Overall, incorporating MFCC, Jitter, Shimmer, and HNR features proved highly effective for representing vocal characteristics, and the RF model demonstrated the most stable and accurate performance for classifying Normal, Vox Senilis, and Laryngocele speech samples.

5.2 Normal, results based on 5-fold cross-validation

This experiment employed five-fold cross-validation to see how well the machine learning models worked and how well they could be used in other situations. The dataset was split into five equal parts, and four of them were used for training and the fifth for testing in each round. Each fold was used as a testing set once over the five iterations of this approach. The resulting calculated performance indicators were also averaged over all five folds to give a precise and consistent evaluation of the models. Stratified splitting of the data was utilized to help preserve the original class distribution across each folds giving an accurate and reliable evaluation of the models' performance. As shown in Figure 6, three different classification methods—SVM, RF, and KNN (k-nearest neighbor)—were used to assess fold-wise accuracy using 5-fold cross-validation. SVM's fold-wise accuracies were 84.62%, 83.65%, 82.69%, 79.33%, and 78.85%, with an average accuracy of 81.83% for all five folds. The fold-wise accuracies for RF were 83.65%, 84.62%, 81.25%, 84.62%, and 84.62%, with an average of 83.75% for all five folds. KNN (k-nearest neighbor) displayed variety in its fold-wise accuracies, reporting: 80.29%, 76.92%, 84.13%, 76.92%, and 79.33%, which gave KNN an average accuracy of 79.52%. While SVM and KNN offered modest accuracy and variability across the folds, RF had the best average accuracy and the most consistent results overall.

The high average accuracy of the RF is due to its ensemble learning approach, which decreases overfitting as well as provides nonlinear representation of the features. Similarly, the variability of the SVM suggests that there is sensitivity with the data distribution and/or kernel selection, however, KNN's accurate results illustrate that local neighborhood learning is consistent. Collectively, these results demonstrate that RF produces more accurate outputs than single classifiers method composed of SVM/RF/KNN classifiers when used to classify speech disorders. The performance of the SVM classifier across the five folds:

83.65% (fold 1), 84.62% (fold 2), 81.25% (fold 3), 84.62% (fold 4), and 84.62% (fold 5), for an average accuracy of 81.83%. The RF produced more consistently across folds with the following average fold accuracies of:

83.65% (fold 1), 84.62% (fold 2), 81.25% (fold 3), 84.62% (fold 4), and 84.62% (fold 5), resulting in an overall average accuracy of 83.75%. Finally, while the KNN exhibited a much greater variance in accuracy values. The five folds achieved accuracies of 80.29%, 76.92%, 84.13%, 76.92%, and 79.33%,

which added up to an overall average accuracy of 79.52%. In general, RF performed the best. SVM and KNN both did okay, but there were notable differences between the folds.

The RF works better than other classifiers due to their ensemble learning (which minimizes the risk of overfitting by creating multiple trees) and its ability to find nonlinear relationships among features; however, the results from SVM show a great deal of variability based on how the algorithm reacts to the data set and how the specific kernel was selected. KNN produces results that are more stable and, therefore, indicate that KNN successfully learns local neighborhoods The results from our study indicate that RF provides more stable and reliable predictions related to speech issues than SVM or KNN when using single classifiers compared to having ensembles of classifiers working together.

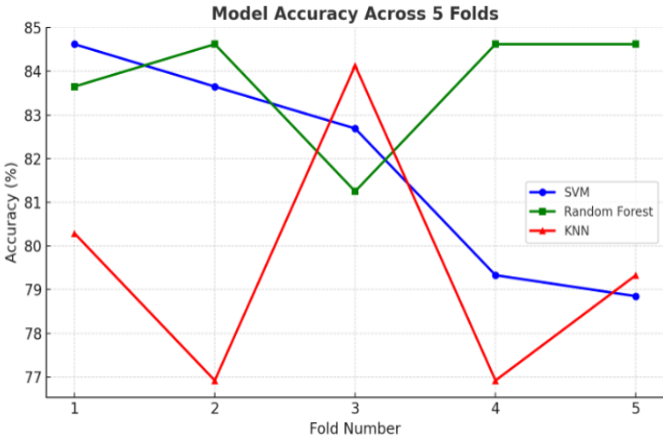

Figure 6. Calculated accuracy results 5 folds cross validation

5.3 Deep learning approach using Convolutional Neural Network

The study involves trials utilizing a CNN to evaluate its capacity to categorize vocal disorders based on audio characteristics. The dataset was made up of .wav files organized into three categories: Normal, Vox Senilis, and Laryngocele. Key speech features, including MFCCs, Jitter, Shimmer, and HNR (Harmonics-to-Noise Ratio), were extracted from these audio files. The CNN model was trained using an 80-20 train-test split, where 80% of the data was used for training and 20% for testing. 1D CNN model was trained for 50 epochs on the extracted features. Below is a summary of its training and validation accuracy.The training accuracy exceeded 98%, and validation accuracy stabilized around 88–91% from Epoch 30 onward and achieved best Validation Accuracy is 90.16%.

Figure 7 shows the training accuracy exceeded 98%, and validation accuracy stabilized around 88–91% from Epoch 30 onward and achieved best Validation Accuracy is 91.19%.

Epochs 1–10: Steady improvement in accuracy (from 68% to 84%)

Epochs 10–30: Model stabilizes (87 to 89% validation accuracy)

Epochs 30–50: validation accuracy peaks 90.16% shown in among all models and shown in Figure 8 in confusion matrix, CNN achieved the highest accuracy of 90.16%, followed by RF with 83.17%, making CNN the most effective model for this classification task. The CNN model good performs all traditional machine learning models, demonstrating its ability to effectively capture complex temporal patterns in speech features.

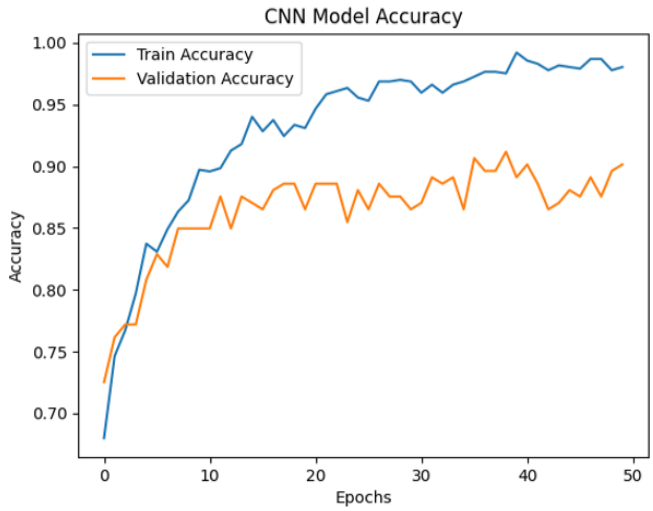

Figure 7. Convolution neural network (CNN) model training and validation accuracy

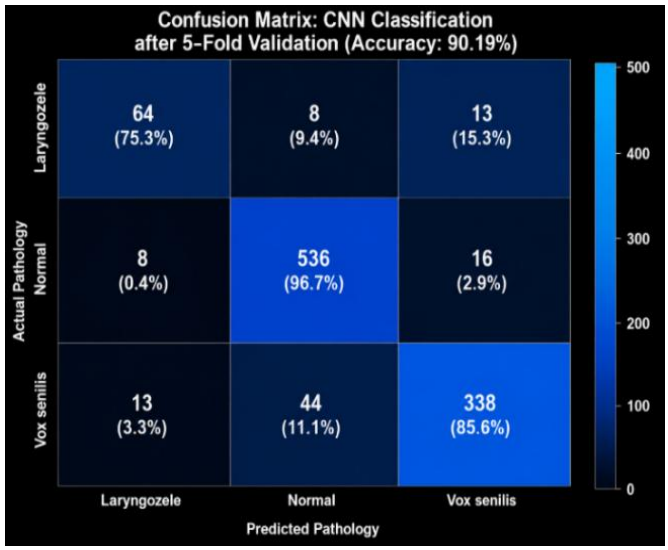

Figure 8. Classification results obtained by using convolution neural network (CNN) for Normal, Vox sensils and Laryngocele class

The results in Table 1 show that while ensemble methods, like RF, give high accuracy for majority classes, the CNN offers the best balanced and strong diagnostic performance across all categories. With an accuracy of 90.16% and an F1-score of 85.80%, the CNN stands out as the most effective tool for automated laryngeal pathology detection in this study.

Successful multiclass classification outcomes were attained from the proposed model for the three classification types of laryngocele, normal, and vox senilis voice samples. The highest recall value (95.71%) was associated with the normal classification in terms of the three classes demonstrating exceptional detection capability. The highest precision value (94%) was associated with the vox senilis class demonstrating that the model generates highly accurate classifications. The combined precision values, recall values, and F1-score values provide additional evidence of the classification systems' robustness.

The proposed CNN framework achieved an overall classification accuracy of 90.16%.

Table 1. Class-wise performance analysis using convolution neural network (CNN)

Class	Precision	Recall	F1-Score
Laryngocele	85.53%	85.33%	85%
Normal	91%	95.71%	93%
Vox senilis	94%	87.79%	91%

5.4 Comparative performance analysis of different model using Features Mel-Frequency Cepstral Coefficient, Jitter, Shimmer and Harmonic to Noise Ratio

The experimental results show in Figure 9 demonstrate that the use of additional acoustic features alongside MFCC improved the overall classification performance of the proposed system. Using MFCC features alone, the SVM, RF, KNN, and CNN classifiers achieved accuracies of 76.25%, 77.00%, 78.00%, and 89.00%, respectively. After incorporating Jitter, Shimmer, and HNR features with MFCC, the classification accuracies increased to 81.83% for SVM, 79.52% for RF, 83.75% for KNN, and 90.16% for CNN. The improvement in performance indicates that the additional acoustic features provide complementary information related to vocal instability and noise characteristics, which enhances the model's ability to distinguish pathological speech conditions more effectively.

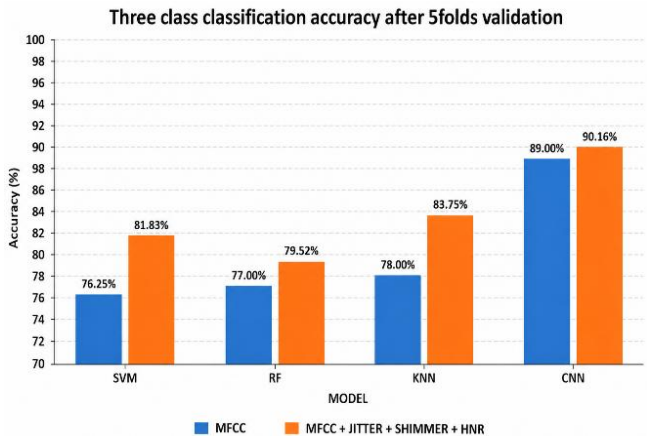

Figure 9. Comparative analysis using Mel-frequency cepstral coefficient (MFCC) and MFCC+Jitter+shimmer +hnr

Table 2. Comparative performance analysis of different models using Mel-frequency cepstral coefficient (MFCC), Jitter, Shimmer, hnr

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score (%)
SVM	81.83	75	72	73
KNN	79.52	71	76	73
RF	83.75	70	84	73
CNN	90.16	85.53	94	85

Note: SVM = support vector machine; KNN = k-nearest neighbor; RF = random forest; CN= convolution neural network.

Table 2 shows that the results indicate RF consistently performed better than both SVM and KNN in overall classification accuracy and specificity. This demonstrates its effectiveness for multi-class pathological speech detection. The high specificity of 94% in the CNN model highlights its strength in correctly identifying healthy samples compared to diseased ones, such as laryngocele and vox sensilis.

5.5 Clinical relevance

According to the experimental evaluation, the suggested CNN-based Vocal Care framework performed reliably in differentiating between healthy and pathological speech samples, achieving a classification accuracy of 90.16%, sensitivity of 85.53%, specificity of 94%, and F1-score of 85%.From a clinical standpoint, the suggested Vocal Care framework offers an automated, non-invasive method for early laryngeal disease diagnosis utilizing speech data. The system's capacity to successfully detect aberrant speech situations while reducing misclassification is demonstrated by the excellent specificity and sensitivity values.Furthermore, the proposed CNN-based system can support healthcare professionals by enabling rapid preliminary screening in remote and resource-limited healthcare environments. The lightweight architecture and automated analysis capability also make the framework suitable for integration into telemedicine and mobile healthcare applications.

6. CONCLUSIONS

The work presents an applied comparative investigation of machine learning versus CNN-based multi classification approaches to speech samples classified . MFCC-based speech features were used, and model performance was assessed using a 5-fold cross-validation approach., the real performance of multiple algorithms/models was evaluated on a clinical dataset that is imbalanced between each of the three speech categories. Results from the study indicate that each of the three different algorithms/methods (ML RF, KNN and SVM). As a result, findings from this study suggest that automated speech analysis can provide reliable and consistent methods of preliminary screener or triage of individuals with speech disorders using data from real world cohort study, regardless of the subjects classification or the subject distribution within speech classification categories. The comparative analysis of the results among the 2 different approaches also show that the CNN-based classifier provided the highest accuracy at 90.16%, while RF provided the second highest accuracy at 83.75%, which demonstrates that despite all 3 models demonstrated adequate performance in terms of accuracy using MFCC features, the CNN based models are superior owing to the capability of learning more complex The research focuses on a 3-class problem approach utilizing a CNN to accurately classify three different types of speech samples, namely; Laryngocele, Vox Senilis, and Healthy speech samples. This research established the effectiveness of the CNN as a means of identifying and categorizing 'complex spectral-temporal patterns' in speech signal data and further establishes the potential for the CNN to become a reliable, non-invasive technology to assist with the diagnostic assessment and ongoing monitoring of laryngeal health.

7. FUTURE SCOPE

Speech characteristics are affected by demographic and environmental factors such as language, age, gender, recording conditions, microphone quality, and pronunciation variability. Although the proposed model shows strong performance on the SVD dataset, its use of a single dataset may limit generalization in real-world clinical settings. Future

work will focus on validating the model using multilingual and diverse datasets recorded under varying conditions to improve robustness and practical applicability. Future extensions of this comparative practical study may include increasing the dataset size to improve class balance and conducting experiments with additional feature extraction techniques. Further practical evaluations can also involve real-time audio recordings, multi-language datasets, and clinical validation with ENT specialists. Developing an integrated software tool and mobile application based on the best-performing model would support broader clinically to speech disabled person.

REFERENCES

[1] Verde, L., De Pietro, G., Sannino, G. (2018). Speech disorder identification by using machine learning techniques. IEEE Access, 6: 1-12. https://doi.org/10.1109/ACCESS.2018.2816338

[2] Kharibam, J., Devi, A.A. (2019). Automatic speaker recognition using MFCC and artificial neural network. International Journal of Innovative Technology and Exploring Engineering, 9(1S): 39-42. https://doi.org/10.35940/ijitee.A1010.1191S19

[3] Miliaresi, I., Pikrakis, A. (2023). A modular deep learning architecture for speech pathology classification. IEEE Access, 11: 80465-80483. https://doi.org/10.1109/ACCESS.2023.3300795

[4] Stewart, C.F., Sinclair, C.F., Kling, I.F., Diamond, B.E., Blitzer, A. (2017). Adductor focal laryngeal dystonia: Correlation between clinicians' ratings and subjects' perception of dysphonia. Journal of Clinical Movement Disorders, 4(20). https://doi.org/10.1186/s40734-017-0066-y

[5] Kraxberger, F., Wurzinger, A., Schoder, S. (2022). Machine learning applied to classify flow-induced sound parameters from simulated human speech. arXiv preprint. https://doi.org/10.48550/arXiv.2207.09265

[6] Hlavnička, J., Čmejla, R., Klempíř, J., Růžička, E., Rusz, J. (2019). Acoustic tracking of pitch, modal, and subharmonic vibrations of vocal folds in Parkinson's disease and parkinsonism. IEEE Access, 7: 150339-150353. https://doi.org/10.1109/ACCESS.2019.2945874

[7] Balaji, V., Sadashivappa, G. (2015). Speech disabilities in adults and the suitable speech recognition software tools: A review. In Proceedings of IEEE International Conference on Computing and Network Communications (CoCoNet'15), Trivandrum, India, pp. 559-563. https://doi.org/10.1109/CoCoNet.2015.7411243

[8] Khara, S., Singh, S., Vir, D. (2018). A comparative study of the techniques for feature extraction and classification in stuttering. In Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, pp. 887-893. https://doi.org/10.1109/ICICCT.2018.8473099

[9] Islam, R., Abdel-Raheem, E., Tarique, M. (2022). Speech pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals. Computer Methods and Programs in Biomedicine Update, 2: 100074. https://doi.org/10.1016/j.cmpbup.2022.100074

[10] Gharde, M.B., Patil Principal, V.V. (2023). Exploring speech disorder detection in laryngeal diseases: A comprehensive examination. In International Conference on Intelligent Computing & Optimization, pp. 185-193. https://doi.org/10.1007/978-3-031-73324-6_19

[11] Gharde, M.B., Patil, V.V. (202). Machine learning classification techniques for the diagnosis of voice disorders: laryngeal conditions. In International Conference on Emerging Technologies and Computing Innovations, pp. 261-267. https://doi.org/10.1007/978-3-031-92854-3_31

[12] Shimbre, N., Solanki, R. (2025). Activation heatmap-guided FT-MultiCNN: Advancing skin cancer classification through transfer learning. Ingénierie des Systèmes d'Information, 30: 1349-1362. https://doi.org/10.18280/isi.300520

[13] Sharma, Y., Singh, B.K. (2020). Classification of children with specific language impairment using pitch-based parameters. In 2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Thiruvananthapuram, India, pp. 42-46. https://doi.org/10.1109/RAICS51191.2020.9332499

[14] Koo, M.W., Choi, J.K., Kim, Y.M. (2008). The development of automatic speech recognition software for portable devices. In First International Conference on Advances in Computer-Human Interaction, Sainte Luce, Martinique, France, pp. 59-62. https://doi.org/10.1109/ACHI.2008.44

[15] Ai, O.C., Hariharan, M., Yaacob, S., Chee, L.S. (2012). Classification of speech dysfluencies with MFCC and LPCC features. Expert Systems with Applications, 39(2): 2157-2165. https://doi.org/10.1016/j.eswa.2011.07.065

[16] Konadath, S., Chatni, S., Lakshmi, M.S., Saini, J.K. (2017). Prevalence of communication disorders in a group of islands in India. Clinical Epidemiology and Global Health, 5(2): 79-86. https://doi.org/10.1016/j.cegh.2016.08.003

[17] Vinay, N.A., Vidyasagar, K.N., Rohith, S., Supreeth, S., Prasad, S.N., Kumar, S.P., Bharathi, S.H. (2024). Dysfluent speech classification using variational mode decomposition and complete ensemble empirical mode decomposition techniques with NGCU-based RNN. IEEE Access, 12: 174934-174953. https://doi.org/10.1109/ACCESS.2024.3502292

[18] Ramitha, V., Chainani, R., Mehrotra, S., Sah, S., Mahajan, S. (2024). Evaluative comparison of machine learning algorithms for stutter detection and classification. MethodsX, 13: 103050. https://doi.org/10.1016/j.mex.2024.103050

[19] Kinahan, S.P., Saidi, P., Daliri, A., Liss, J., Berisha, V. (2024). Electroencephalographic classification reveals atypical speech motor planning in stuttering adults. Journal of Speech, Language, and Hearing Research, 67(7): 2053-2076. https://doi.org/10.1044/2024_JSLHR-23-00635

[20] Nobel, S.N., Swapno, S.M.R., Islam, M.R., Safran, M., Alfarhood, S., Mridha, M.F. (2024). A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method. Scientific Reports, 14(1): 14435. https://doi.org/10.1038/s41598-024-64987-5