Advancing Voice Health: A Hybrid ML Approach for Early Detection of Voice Disorders

CH. LAKSHMI KUMARI[1], ARRAVAPULA SIDDARTHA REDDY[2], KABIR SRINIDH[3], MALGARI SUPRIYA[4]

[1] *Asst. Professor, IT, MGIT*

[2, 3, 4] *Student, MGIT*

Abstract*— In our study, we propose an innovative method for early detection and intervention of vocal disorders. Our comprehensive dataset consists of voice samples from healthy individuals and those with voice pathologies. We consider acoustic features like fundamental frequency, jitter, shimmer, and Mel-frequency cepstral coefficients, which are analyzed using tree-based machine learning algorithms. Additionally, we extract modes from audio signals through Variational Mode Decompo- sition (VMD) and convert them into Mel spectrograms. These spectrograms are then processed by a Vision Transformer archi- tecture. With a focus on multi-class classification, we combine the outputs of the tree-based algorithms and Vision Transformer into an ensemble model to enhance predictive accuracy across all classes. The method yields good results, which achieves an overall accuracy of 93% along with strong performance on other metrics, demonstrating its potential for improving early detection techniques for voice disorders.

Index Terms*- Voice Disorders, tree based, machine learning, classification model, acoustic features, Mel-Frequency Cepstral Coefficients, Variational Mode Decompostion, VMD modes.

I. INTRODUCTION

Voice disorders, such as Laryngitis and Reinke Odem, have a profound impact on individuals' quality of life and commu- nication abilities [1]. Mobile health systems have emerged as promising tools for detecting and monitoring voice disorders due to their widespread use [3]. However, accurate methods for early detection and intervention are crucial to mitigate the negative consequences of these disorders [4]. In this context, a novel approach utilizing tree based machine learning (ML) and deep learning (DL) techniques for developing a classification model for the advanced detection and intervention of voice disorders is proposed. This study leverages a comprehensive dataset of recorded voice samples, encompassing both healthy individuals and those with pathological voices [5]. Various acoustic features, including Jitter, Shimmer, Fundamental fre- quency, Mel-Frequency Cepstral Coefficients (MFCCs), and Variational Mode Decomposition (VMD) modes capturing the acoustic properties of pathological voices are extracted [2] [3]. These features have shown effectiveness in accurately classifying voice disorders and providing valuable observations into the presence and characteristics of disorders. Tree based machine learning techniques, such as XGBoost Gradient Boosting and Extra tree classifier [10], and deep learning models like Vision Transformers are employed in our research [25].

DL models, particularly Vision Transformer, demonstrate the ability to capture intricate patterns within the mel-spectrogram representation of VMD modes derived from voice signals. This enables the identification of subtle changes that serve as indicators of voice disorders. Furthermore, the classification problem is extended from binary (healthy vs. pathological) to multi-class classification, which allows for differentiation between different types of voice disorders and facilitating tailored treatment plans. The main objective of this study is to develop an ensembled model which combines deep learning-based classification models with machine learning models [14] [16] [30]. The ensembled model is assessed using various performance metrics such as accuracy, precision, recall, and F1-score [2]. This research tries to contribute to the advancement of reliable and accurate mobile health systems. Ultimately, these systems have the potential to support early detection and intervention of voice disorders, leading to enhanced patient outcomes and enhanced quality of life.

1.1Objective:
The main intention of the Paper is to detect the vocal disorders as early as possible with help of the Ensemble model for improved results learning algorithms, and neural network architectures. The primary aim is to create a dependent and useful system capable of early detection and intervention of voice disorders, thereby improving patient outcomes and advancing the section of voice pathology diagnostics."

1.2 Motivation:
The reason behind this system is to take precautions to help people prevent suffering from harmful vocal disorders which if not detected could lead to serious consequences in future

1.3 Problem Statement:
Voice disorders pose significant challenges in timely diagnosis and effective intervention, leading to potential long-term complications and diminished standard of life for affected individuals. Current diagnostic methods often rely on subjective assessments and lack the precision needed for early detection. Furthermore, there is a requirement for more comprehensive and efficient approaches that can precisely differentiate between healthy voice patterns and those indicative of various pathologies. This project aims to address these issues by developing an innovative methodology leveraging the signal processing techniques, machine timely diagnosis and effective intervention, leading to potential long-term complications and diminished quality of life for affected individuals.

II. EXISTING SYSTEM

Several studies have used the application of traditional machine learning techniques and deep learning approaches in vocal disorder detection and classification. SupportVector Machines (SVM) have demonstrated promising results in identifying voice impairments [2] [21]. Linear Discriminant Analysis (LDA) has been used for precisely classifying pathological and normal voices [4] [24]. Various classifiers such as Gaussian Mixture Model (GMM) [5], Convolutional Neural Networks (CNN) [6] [7] [10] [11] [15] [26] [27], Random Forest (RF) [9], Recurrent Neural Networks (RNN) [17] [28], and K-Nearest Neighbor (KNN) [1] have employed for binary and multi-class classification tasks, achieving accuracies ranging from 71 % to 98.94%. Noteworthy contributions include the successful utilization of VGG-16, CaffeNet, and VGG19 CNN architectures, which achieved high accuraciesof 93.9%, 94.40%, and 97.80%, respectively [10] [15] [26]. Studies have emphasized the significance of feature selection and classifier optimization in enhancing classification accuracy [5] [6] [7] [21]. Deep learning techniques have beenextensively investigated in the area of voice disorder detection. A voice disorder identification model utilizing Hilbert-Huang Transform (HHT) and K-nearest neighbor (KNN) achieved improved accuracy [21]. Convolutional neural networks (CNN) have been shows to be effective in accurately classifying pathological voice changes, particularly in cases associated with laryngeal cancer [10] [26]. Parallel deep models has been proposed for automatic voice pathology monitoring, leveraging the potential of deep learning to enhance accuracy

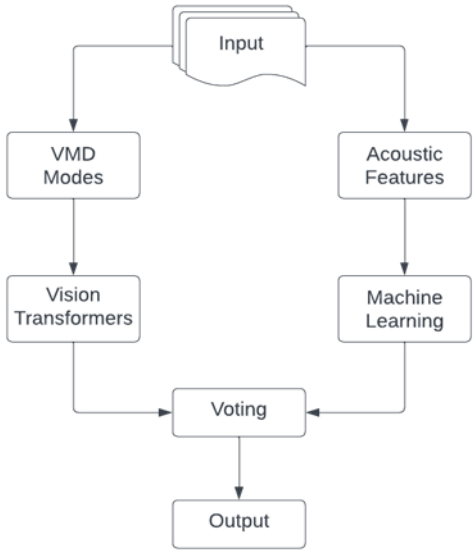

Fig. 1: Proposed Methodology

In this study, an ensembled model Fig.1 is developed to accurately distinguish between healthy voices and pathological voices. The methodology involves the utilization of acoustic features along with Variational Mode Decomposition (VMD) modes for audio to catch the distinctive acoustic properties exhibited by pathological voices [8]. The VMD modes are extracted and transformed into mel-spectrograms,

which offer diverse and informative features. These extracted mel-spectrograms are given as input to vision transformer and features such as jitter, shimmer , MFCC, fundamental frequency are given as input to machine learning (ML) models to achieve classification of voice samples. The primary aim is to establish a robust classification framework by leveraging the strengths of DL and ML algorithms.

A. Feature Extraction

(i) Fundamental Frequency (F0): The fundamental frequency (F0) is a metric of how quickly the vocal cords vibrate when producing sound. It is a important indicator of the health and performance of the larynx, which is part of body responsible for producing voice.

(ii) Jitter: Jitter is a metric of how much the frequency of vocal folds changes from one cycle to the next. It reflects the instability or difference in the oscillation pattern of the vocal cords.

(iii) Shimmer: Shimmer refers to the variability or fluctuations in the amplitude of vocal folds' oscillation pattern from cycle to cycle. It is a metric of the instability of the vocal fold or cord vibration and quantifies the cycle-to-cycle changes in extent of vibration.

(iv) Mel-Frequency Cepstral Coefficients (MFCC): Mel- Frequency Cepstral Coefficients (MFCC) are a group of features that aim to analyze all spectral characteristics of a speech signal while minimizing the influence of vocal cord damage or pathologies. By focusing on the vocal tract's properties, MFCCs can provide insights into the acoustic characteristics of speech signals that are not dependent on vocal fold variations. MFCCs are created by projecting the speech signal's power spectrum onto the mel-scale, then using a discrete cosine transform (DCT) to produce a group of coefficients that represent the speech signal's spectral envelope. The coefficients are designed to capture important information about the spectral features of the speech signal while minimizing the effects of vocal cord damage or pathologies.

(v) VMD modes: In a pathological voice, various conditions that can affect the vocal tract and/or vocal folds, such as laryngitis,Reinke Odem and other voice disorders, may cause the acoustic properties of a pathological voice to differ from those of a healthy voice. These conditions can affect the frequencies and amplitudes of the various modes of the VMD decomposition.

(vi) Mel-Spectrogram : A 2D illustration of an audio sig- nal's frequency content is called a mel-spectrogram. It converts the time-varying audio signal into a visual representation where, on a perceptual mel scale, the x-axis stands for time and the y-axis for frequency. By matching the power or magnitude of frequency components onto a logarithmic scale, the mel- spectrogram provides a compact and informative feature for analyzing and classifying audio signals.In this study VMD modes are represented as mel-spectrograms.

B. Classifiers Used

Classifiers such as XGBoost ,Gradient Boosting algorithm, Extra Trees Classifier as seen in Fig.3 and Vision Trans- formers as seen in Fig.2 were used to dinstinguish healthy and pathological voices based on clinical features taken from voice recordings.

1) Vision Transformers: The Vision Transformer (ViT) as shown in Fig.2 is primarily used for image recognition tasks. It divides the image into patches, treating them as a series of flattened patches rather than pixels. ViTs employ self-attention mechanisms, like the transformer architecture, to capture global relationships between the patches, enabling long-range dependency learning and holistic understanding the image. With stacked self-attention layers, ViTs excel at recognizing complex patterns and extracting essential features. To prevent the loss of spatial information, positional encodings are incorporated into the input embeddings. These encodings provide the ViT with knowledge about patch locations. Addi- tional techniques, such as learnable class tokens and normal- ization methods, are often utilized to enhance performance and stability during training. Finally, a classification head, usually an MLP, is attached to the ViT. It takes the learned features from the self-attention layers and performs classification .

2) XGBoost: XGBoost, short for "Extreme Gradient Boost- ing," is a highly efficient and scalable machine learning algorithm. Its architecture revolves around the concept of gradient boosting, combining multiple decision trees to form a powerful ensemble model. Decision tree serve as the fun- damental building blocks in XGBoost.

Through the gradient boosting technique, each subsequent tree attempts to rectify the errors made by its predecessors, progressively improving the overall predictive capability of the ensemble. XGBoost employs gradient-based optimization to train these decision trees effectively. By computing the gradients of a user-defined loss function with due respect to the predicted values from previous trees, XGBoost updates the model parameters in an iterative manner. This optimization process minimizes the loss function and enhances the model's predictive performance XGBoost incorporates regularization techniques to prevent overfitting .

3) Gradient Boosting : Gradient boosting is an exceptional algorithm that builds an ensemble of weak learners, typi- cally decision trees, to generate predictions grounded on a loss function. The algorithm repetetively adds decision trees that fit the negative gradient of loss function, which indicates how to improve the ensemble's predictions. The learning speed controls the contribution of each decision tree to the ensemble and helps prevent overfitting. To avoid overfitting, regularization approaches such as restricting the tree depth, introducing a shrinkage parameter, and applying early stopping are utilized. The algorithm stops adding decision trees when a maximum number of iterations is reached, a predefined threshold for the loss improvement is met, or overfitting occurs. The ensemble's ultimate forecast is derived by merging the predictions from all decision trees. In essence, gradient boosting represents a sequential and collaborative approach to acquiring knowledge, employing decision trees as its apprentices and refining its skills by optimizing a loss function through the application of gradient descent. It adjusts the predictions of each decision tree using a learning rate and applies regularization to prevent overfitting.
4) Extra Trees Classifier: : The Extra Trees Classifier em- ploys the technique of ensemble learning to derive predictions by utilizing numerous decision trees. These trees are based on random subsets of features. At every juncture of decision- making, the algorithm exercises its penchant for serendipity by haphazardly cherry-picking a subset of features, thus spawning

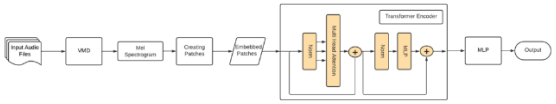

Fig. 2: Vision Transformer

numerous decision trees that employ disparate subsets of training data. Each decision tree ventures forth on its own, making predictions with unbridled autonomy. The final verdict is then delivered through the hallowed process of majority voting.

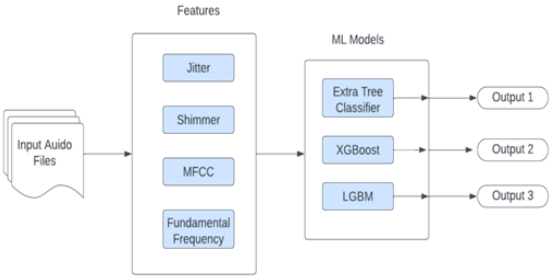

Fig. 3: ML Models

III. RESULTS & DISCUSSIONS

TABLE II: Evaluation metrics

Model	Accuracy	Precision	Recall	F1 Score
XG Boost	86.046	86.259	86.046	85.964
Gradient Boosting	89.147	89.194	89.147	89.147
Extra Tree Classifier	89.924	90.019	89.924	89.934
Vision Transformers	89.152	89.330	89.147	89.127
Ensembled	93.02	93.02	93.02	93.02

TABLE III: Confusion matrix for Ensembled model

Classes	Healthy	Laryngitis	Reinke Odem
Healthy	38	2	1
Laryngitis	1	44	2
Reinke Odem	2	1	38

Assessment of the ML models and vision tranfomers revealed good outcome in the classification of audio samples into the three classes: healthy, laryngitis, and Reinke's edema, utilizing various acoustic features

and VMD modes. Several machine learning models such as XGBoost, Gradient Boosting, and Ex- tra Tree Classifier, were employed, additionally to vision trans- formers ML models were evaluated on this dataset.Among ML models, XGBoost achieved a relatively lower evaluation metrics measured agianst to the other models indicating a relatively higher percentage of false positives and false negatives which is shown in Table II, while gradient boosting exhibited better per- formance. The most impressive performance was observed with the Extra Tree Classifier. This model surpassed both XGBoost and Gradient Boosting, achieving an accuracy of 89.924%. It displayed the highest precision, recall, and F1 scores among the evaluated models, indicating its superior ability to correctly classify audio samples into their respective classes. These results shows that all three ML models are effective in classifying between healthy and pathological voice samples.

Another employed approach incorporated vision transform- ers for audio classification, utilizing mel spectrograms gen- erated through VMD. The vision transformers showed better results in aspect of accuracy, precision, recall, and F1 scores compared to XG Boost and Gradient Boosting. This highlights the effectiveness of vision transformers in processing audio signals transformed into visual representations.

Finally, we employed a probabilistic voting ensemble ap- proach, considering the accuracies of all four models (XG- Boost, Gradient Boosting, Extra Tree Classifier, and vision transformers). The ensembled model achieved an impressive accuracy of 93.02%, with recall, precision, and F1 score of 93.02%. The confusion matrix for the ensembled model further confirms its robust performance, with minimal misclassifica- tions. The major portions of samples were correctly classified , with only a few instances of misclassification observed.

Overall, our study demonstrates the effectiveness of both of traditional machine learning models and deep learning models in classifying audio samples. The blend of acoustic features and visual representations derived from audio sig- nals improved classification performance. The ensembled model, incorporating the strengths of all models, achieved the highest accuracy, highlighting the significance of combining multiple approaches for enhanced classification outcomes in pathological voice analysis.

CONCLUSION

In the study, we proposed an ensembled model for the classification of pathological speech based on the combination of acoustic and spectrographic features. The model integrated XGBoost, Gradient Boosting, Extra Tree Classifier, and a vision transformer model, leveraging their individual strengths to achieve robust classification results. The ensembled model exhibited exceptional performance with an accuracy, precision, recall, and F1 score of 93.02 % .

This method's effectiveness was confirmed by the confusion matrix as shown in Table III, which demonstrated its ability to accurately classify the audio into healthy, laryngitis, and reinke's edema. These results underscore the potential of combining acoustic and spectrographic features in an ensem- bled framework for precise and reliable pathological voice classification.

The results of this study have major implications for clinical applications in the assesment and treatment of voice disorders. The ensembled model's high accuracy and comprehensive performance metrics demonstrate its potential to assist healthcare professionals in accurately identifying and categorizing pathological speech conditions. Further research can focus on exploring additional feature representations and ensemble strategies to further enhance the classification model's performance in real-world scenarios.

REFERENCES

[1] Chen, Lili, et al. "Voice disorder identification by using Hilbert-Huang transform (HHT) and K nearest neighbor (KNN)." Journal of Voice 35.6 (2021): 932-e1.

[2] Godino-Llorente, Juan Ignacio, et al. "Support vector machines applied to the detection of voice disorders." Nonlinear Analyses and Algorithms for Speech Processing: International Conference on Non-Linear Speech Processing, NOLISP 2005, Barcelona, Spain, April 19-22, 2005,

Revised Selected Papers. Springer Berlin Heidelberg, 2005.

[3] Souissi, Nawel, and Adnane Cherif. ”Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coeffi- cients and support vector machine.” 2015 7th international conference on modelling, identification and control (ICMIC). IEEE, 2015.

[4] Lee, Ji-Yeoun, SangBae Jeong, and Minsoo Hahn. ”Classification of pathological and normal voice based on linear discriminant analy- sis.” Adaptive and Natural Computing Algorithms: 8th International Conference, ICANNGA 2007, Warsaw, Poland, April 11-14, 2007, Proceedings, Part II 8. Springer Berlin Heidelberg, 2007.

[5] Cordeiro, Hugo, et al. ”Voice pathologies identification speech signals, features and classifiers evaluation.” 2015 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA). IEEE, 2015.

[6] El Emary, I. M. M., M. Fezari, and F. Amara. ”Towards developing a voice pathologies detection system.” Journal of Communications Technology and Electronics 59 (2014): 1280-1288.

[7] Amara, Fethi, Mohamed Fezari, and Hocine Bourouba. ”An improved GMM-SVM system based on distance metric for voice pathology detection.” Appl. Math 10.3 (2016): 1061-1070

[8] Dragomiretskiy, Konstantin, and Dominique Zosso. ”Variational mode decomposition.” IEEE transactions on signal processing 62.3 (2013): 531-544.

[9] Ritchings, R. T., Mark McGillion, and Christopher J. Moore. ”Patholog- ical voice quality assessment using artificial neural networks.” Medical engineering physics 24.7-8 (2002): 561-564.

[10] Kim, HyunBum, et al. ”Convolutional neural network classifies patho- logical voice change in laryngeal cancer with high accuracy.” Journal of Clinical Medicine 9.11 (2020): 3415.

[11] Alhussein, Musaed, and Ghulam Muhammad. ”Automatic voice pathol- ogy monitoring using parallel deep models for smart healthcare.” Ieee Access 7 (2019): 46474-46479.

[12] Wang, Jianglin, and Cheolwoo Jo. ”Vocal folds disorder detection using pattern recognition methods.” 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2007.

[13] Al-Nasheri, Ahmed, et al. ”Voice pathology detection and classifica- tion using auto-correlation and entropy features in different frequency regions.” Ieee Access 6 (2017): 6961-6974.

[14] Leite, Danilo Rangel Arruda, Ronei Marcos de Moraes, and Leonardo Wanderley Lopes. ”Different Performances of Machine Learning Models to Classify Dysphonic and Non-Dysphonic Voices.” Journal of Voice (2022).

[15] Zakaria, Salman, S. Thanush, and M. Mugilan. ”Voice Disorder iden- tification Using Convolutional Neural Network.” 2022 1st International Conference on Computational Science and Technology (ICCST). IEEE, 2022.

[16] Reid, Jonathan, et al. ”Development of a machine-learning based voice disorder screening tool.” American Journal of Otolaryngology 43.2 (2022): 103327.

[17] Syed, Sidra Abid, et al. ”Comparative analysis of CNN and RNN for voice pathology detection.” BioMed Research International 2021 (2021): 1-8.

[18] J. I. Godino-Llorente and P. Gomez-Vilda, ”Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors,” in IEEE Transactions on Biomed- ical Engineering, vol. 51, no. 2, pp. 380-384, Feb. 2004, doi: 10.1109/TBME.2003.820386.

[19] J. Nayak and P. S. Bhat, ”Identification of voice disorders using speech samples,” TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, Bangalore, India, 2003, pp. 951-953 Vol.3, doi: 10.1109/TENCON.2003.1273387.

[20] Mohammed, Mazin Abed, et al. ”Voice pathology detection and clas- sification using convolutional neural network model.” Applied Sciences 10.11 (2020): 3723.

[21] Lee, Ji-Yeoun. ”Experimental evaluation of deep

learning methods for an intelligent pathological voice detection system using the saarbruecken voice database.” Applied Sciences 11.15 (2021): 7149.

[22] Wu, Huiyi, et al. ”Convolutional neural networks for pathological voice detection.” 2018 40th annual international conference of the ieee engineering in medicine and biology society (EMBC). IEEE, 2018.

[23] Al-Nasheri, Ahmed, et al. ”An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification.” Journal of Voice 31.1 (2017): 113-e9.

[24] Wu, Huiyi, et al. ”A deep learning method for pathological voice detection using convolutional deep belief networks.” Interspeech 2018 (2018).

[25] Dosovitskiy, Alexey, et al. ”An image is worth 16x16 words: Transform- ers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).

[26] Wu, Yuanbo, et al. ”Investigation and evaluation of glottal flow wave- form for voice pathology detection.” IEEE Access 9 (2020): 30-44.

[27] Al-Nasheri, Ahmed, et al. ”Investigation of voice pathology detection and classification on different frequency regions using correlation func- tions.” Journal of Voice 31.1 (2017): 3-15.

[28] Muhammad, Ghulam, et al. ”Voice pathology detection using interlaced derivative pattern on glottal source excitation.” Biomedical signal pro- cessing and control 31 (2017): 156-164.

[29] Alhussein, Musaed, and Ghulam Muhammad. ”Voice pathology detec- tion using deep learning on mobile healthcare framework.” IEEE Access 6 (2018): 41034-41041.

[30] Tembhurne, Jitendra V., et al. ”Skin cancer detection using ensemble of machine learning and deep learning techniques.” Multimedia Tools and Applications (2023): 1-24.