Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 April 2026
Revised 22 May 2026
Accepted 16 June 2026

Natural Resources for Human Health 2026, Vol. 6 (5s): 903–911

KEYWORDS:
Laryngeal Voice Disorder, Speech Impairment, Severity Classification, Machine Learning, Speech Signal Processing, Mel-Frequency Cepstral Coefficients (MFCC), Acoustic Features, Voice Disorder Detection, Artificial Intelligence, Pattern Recognition

CNN-Based System for Voice Disorder Severity Assessment Using Speech Signals

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

[1]Research Scholar, Dept.of E&TC AISSMSIOIT,Pune,India[0009-0001-1329-703X]
[2]Principal, Dept.of E&TC ,International Institute of Information Technology (I²IT), Pune,India [0000-0003-3922-002X]
Corresponding Author Email: manishagharde30@gmail.com, vvpatil2429@gmail.com

ABSTRACT: Voice disorders in children can affect communication, speech development, and academic performance, making early identification and severity assessment important. This study proposes an Intelligent CNN-Based System for Voice Disorder Severity Assessment Using Speech Signals for children below 12 years of age. The study was conducted to develop an accessible, non-invasive, and cost-effective automated approach for assessing voice-disorder severity using short speech recordings obtained through commonly available mobile devices. Speech samples of 0.2–0.8 seconds were recorded using a mobile phone in a controlled recording room. From each speech signal, a combined set of acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), Jitter, Shimmer, and Harmonics-to-Noise Ratio (HNR), was extracted. The extracted features were used to develop four classification models: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Convolutional Neural Network (CNN). Experimental evaluation demonstrates that the proposed framework achieves an accuracy of approximately 95%, indicating its ability to distinguish different severity levels from voice recordings. Compared with conventional machine-learning approaches. The models make use of three categories for severity classification of voice disorders namely Low, Moderate, and High. Data augmentation was employed to improve the generalization of the model and tackle the class imbalance. The performance of both the proposed and the comparative model was evaluated using several performance metrics including accuracy, sensitivity, specificity, precision, and F1-score. Results from the experiment illustrate the effectiveness of machine learning and deep learning techniques in identifying severity. The proposed CNN-based system provides a promising approach for rapid and automated preliminary voice-disorder severity assessment in children and may support early screening in resource-limited and non-clinical environments.

1. INTRODUCTION

Severity assessment is an important component in the analysis of voice disorders, since the presence of a disorder does not always indicate the level of vocal impairment. Speech is one of the most fundamental forms of human communication, and its quality depends on the proper functioning of the respiratory system, larynx, vocal folds, and vocal tract. Voice disorders can impair one's speech and bring down the quality of sound. A lot of laryngeal disorders, such as vocal fold paralysis, dysphonia, vocal nodules, polyps, cysts, and more, cause the change of the speech sounds' characteristics, like loudness, pitch, voice quality, etc. For this reason, it is crucial to identify problems soon enough and assess the extent of voice problems for proper diagnosis and treatment. Traditional ways of diagnosis often rely on invasive methods, such as laryngoscopy and stroboscope, and subjective assessment by experienced speech pathologists. Although effective, these methods are expensive, time-consuming, and prone to variability from one interpreter to another. Thus, there is a need for objective methods of diagnosis.Recent studies in the area of speech signal processing and artificial intelligence enabled the development of non-invasive systems of computer-

aided diagnosis that work by analysing speech signals. Some acoustic features, such as Mel-Frequency Cepstral Coefficients (MFCCs), jitter, shimmer, Harmonics-to-Noise Ratio (HNR), and prosodic features are capable of detecting the abnormalities related to laryngeal voice disorders. They are taken from the stable phonation of previously recorded speech samples and are analysed using machine learning and deep learning technologies. These advances notwithstanding, automatic classification of severity of laryngeal voice disorders is still a challenging problem. Variability in recording environments, speaker characteristics, class imbalance, background noise, and similarities between different pathological conditions often degrade model generalization. Moreover, many previous studies have been devoted to binary classification (healthy vs. pathological), while severity-based multi-class classification has been scarcely considered. Accurate estimation of severity is of clinical importance as it allows for prioritization of treatment, monitoring of rehabilitation and objective assessment of disease progression.

Recent research has also explored transfer learning, domain adaptation, ensemble learning and pre trained deep learning models to improve robustness and real-world applicability. Promising performance for binary and multi-class voice disorder classification has been demonstrated by hybrid frameworks which combine deep feature extraction and conventional classifiers. Domain-adaptive neural networks and enhance robustness for noisy recording conditions. These advances suggest that combining good acoustic representations with sophisticated machine learning methods can dramatically enhance automated systems for voice pathology assessment. Inspired by these developments, the present work introduces severity-based laryngeal voice disorder classification using machine learning on speech impairment. We aim at developing an automatic framework for accurate classification of different severity levels of laryngeal disorders based on acoustic features derived from speech. The proposed framework is aimed to provide an objective, non-invasive and reliable decision support tool for clinicians by combining effective pre-processing, feature extraction and machine learning techniques. Such a system can potentially improve early diagnosis, reduce the reliance on subjective clinical assessment, enable ongoing surveillance during rehabilitation, and improve the accessibility of voice disorder screening in clinical and remote healthcare system.

2. RELATED WORK

A comprehensive survey on signal-processing-based pathological voice detection techniques is presented by Islam et al. [1]. The study considered several acoustic features such as Mel-frequency cepstral coefficients (MFCCs) , jitter, shimmer, harmonics-to-noise ratio (HNR), pitch, spectrogram, formants, and nonlinear features. Different classification methods such as support vector machine (SVM), Gaussian mixture model (GMM), hidden Markov model (HMM) and artificial neural network (ANN) were studied. The authors reported that no individual acoustic feature or classifier is consistently the best performer for all types of voice pathologies. As pointed out in the study [1], appropriate feature selection and classifier designing are important for reliable pathological voice detection.

Tulics et al. [2] worked on automatic estimation of dysphonia severity based on acoustic parameters of continuous speech. The study considered acoustic features such as jitter, shimmer, harmonics to noise ratio, MFCC related parameters and frequency band ratio features. The reference values were derived from the ratings of perceptual severity by four specialists on the RBH scale. Various statistical and regression methods were investigated for the prediction of severity of dysphonia. The best linear regression model had a correlation of 0.853 and an RMSE of 0.462 and the best RBF regression model had an RMSE of 0.454. The study showed that dysphonia severity could be automatically assessed from acoustic characteristics [2].

For this purpose, Joshy et al. [3] suggested an automatic dysarthria severity classification framework based on acoustic features and deep learning techniques. The authors tested DNN, CNN, GRU and LSTM architectures with MFCC and CQCC features on UA-Speech and TORGO databases. In addition, we examined features specific to speech disorders related to prosody, articulation, phonation and glottal features. The authors also examined i-vector-based representations for severity classification. The best performance on the UA-Speech database reached 93.97% accuracy under the speaker-dependent condition,The results indicate that speaker variability remains a significant challenge for automated dysarthria severity assessment [3].

Kuo et al. [4] investigated voice-disorder classification under real-world recording conditions. The authors identified background noise and domain mismatch between clinical and real-world environments as important challenges for practical voice-disorder classification systems. A factorized convolutional neural network (CNN) combined with domain-adversarial training was proposed to obtain compact and domain-robust representations. The proposed approach improved the unweighted average recall in the noisy real-world domain by 13%, while maintaining approximately 80% unweighted average

recall in the clinical domain. The study demonstrated the importance of developing computationally efficient and robust deep-learning models for practical voice-disorder applications [4].

Rahman et al. [5] proposed a hybrid framework for binary and multi-class voice-disorder classification using pretrained VGGish features and different classifiers. Log-mel spectrograms were processed using the pretrained VGGish model to obtain high-level acoustic embeddings. These representations were subsequently classified using SVM, logistic regression, multilayer perceptron, and ensemble classifiers. For the multi-class classification task, the reported accuracy was 77.81% for male speakers, 63.11% for female. Rahman et al. [5] developed a hybrid deep-learning framework for the classification of voice disorders using pretrained audio representations. The proposed approach employed a VGGish model to extract high-level embeddings from log-Mel spectrograms, followed by SVM, logistic regression, multilayer perceptron, and ensemble classifiers. The framework was evaluated for binary and multi-class voice-disorder classification using the Saarbruecken Voice Database. For multi-class classification, VGGish-SVM achieved an accuracy of 77.81% for male speakers and 63.11% for female speakers, while the combined-speaker accuracy was 70.53%. The study showed that the pretrained representations are effective for voice-disorder classification, but also revealed issues with class imbalance and minority-class recognition. The framework is also validated using controlled datasets and does not investigate real time voice-disorder classification [6].

Yeo et al. [6] investigated the automatic classification of dysarthria severity in different languages. We investigated acoustic characteristics related to phonation, articulation, and prosody to identify speech patterns associated with severity. The authors pointed out that the severity of dysarthria is associated with several speech dimensions, and that the joint use of complementary acoustic features could enhance the automated assessment. The results suggest that articulation and prosodic features are especially helpful for classifying the severity of English dysarthric speech. The study showed the importance of the use of multiple speech dimensions in thedevelopment of automated systems for the assessment of severity of speech disorders [7].

Kuo et al. [4] proposed a robust voice-disorder classification system under real-world and noisy environments. The authors dealt with the mismatch between laboratory or clinical recordings and speech recorded in everyday environments. We combined a factorized convolutional neural network with domain-adversarial training to learn compact and domain-invariant representations [8]-[10].The proposed method improved UAR by 13% in the noisy real world domain while keeping UAR around 80% in the clinical domain. The model also reduced the number of parameters and computation operations by 73.9% and 77.0%, respectively. The study shows that CNN-based architectures are able to provide efficient computational solutions for practical voice-disorder applications [12]- [14].

Recent research has highlighted the need for reliable and standardized annotations for machine-learning based assessment of voice disorders[15]-[17]. The study also found that diagnostic assessment of voice disorders can vary from clinician to clinician as many disorders are characterized by similar symptoms, variable severity, and lack of objective biomarkers. These variations can cause inconsistencies in the machine-learning datasets and have a negative effect on the model generalization[18]-[20]. The authors highlighted the need for standardized diagnostic frameworks and robust annotation procedures to be able to develop reproducible artificial-intelligence systems. These results are especially relevant for severity-based voice-disorder classification, where perceptual ratings are used to define the target classes. Calà et al. [21] proposed a machine-learning framework for severity assessment of adductor spasmodic dysphonia using 48 acoustic and 7 perceptual parameters. The study classified patients into mild, moderate, and severe categories using a KNN classifier and incorporated LIME for interpretation. The proposed KNN model achieved an overall accuracy of 89% for three-class severity classification. However, the study was based on a relatively small patient population and the authors identified the need for larger datasets for improved validation.

3. DATABASE DETAILS

The database was collected from Balvikas Kendra, a Speech, Hearing, and Language, as well as Preschool and Parent Training Center, run by the Ekvira Multipurpose Foundation, Akola.In this database, voice samples of students with impaired speech were collected for research purposes. The total number of participants was 20, including 7 girls and 13 boys. A mobile sound recorder was used to record sound vowels /a/, /e/, /i/, /o/, /u/ for a total of 500 samples. Each child provided 25 samples. The samples were taken with 16 bits resolution. The sampling frequency was 48 kHz and the time duration was between 0.2 seconds and 0.8 seconds. The speech recordings were collected in a separate, soundproof recording room within the organization to minimize background noise, external disturbances, and acoustic interference. All audio recordings were normalized before feature extraction. In preparing the speech recordings, background noise was reduced using the Noise

Reduction tool in Audacity software, leading and trailing silence was removed by trimming, and audio amplitudes were normalized to maintain a consistent recording level across all samples. The speech database consisted of 500 speech samples including 400 High-severity, 75 Moderate-severity, and 25 Low-severity samples. .Additional samples were created using time-stretch augmentation on whole data to balance the classes and increase the robustness of the classification model. The dataset was augmented and approximately balanced to three severity levels: Low, High are same moderate little bit lower. The final dataset had 400 Low-severity, 300 Moderate-severity, and 400 High-severity samples, totaling 1100 speech samples used for model development and evaluation. The processed audio files were exported as WAV files for further analysis. Then, acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), jitter, shimmer, harmonic-to-noise ratio (HNR), pitch, formant frequencies and energy-based features were extracted from each speech sample by using speech analysis tools. These features are effective in characterizing vocal fold vibration and speech quality and therefore are suitable for automatic voice disorder severity classification.

4. PROPOSED METHODOLOGY

The proposed Intelligent Voice Disorder Assessment System processes speech signals recorded via mobile devices to evaluate vocal pathology severity. The end-to-end framework consists of five sequential processing blocks: Audio Signal Capture, Preprocessing, Acoustic Feature Extraction, CNN-Based Classification Engine, and Severity Assessment Output.

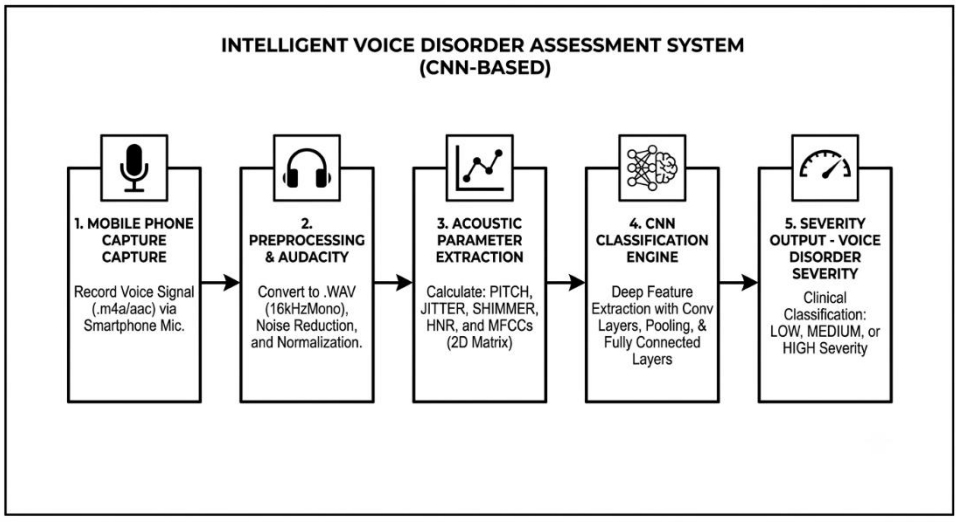

Fig. 1. Intelligent Voice Disorder Assessment System

4.1. Real-Time Voice Data Acquisition and Preprocessing

The first stage involves the collection of voice recordings using a smartphone microphone. The speech signal is acquired from the subject in real time and initially stored in a compressed audio format such as M4A or AAC. The collected recordings are subsequently converted into WAV format with a sampling frequency of 48 kHz to maintain a consistent representation across all samples. The recorded signals may include ambient noise and fluctuations in amplitude of recording. Therefore, preprocessing is the first step before feature extraction. Noise reduction is applied to reduce unwanted background components and amplitude normalization is applied to obtain consistent level of signal. Preprocessed speech signals are then used for the subsequent extraction of acoustic features.Severity annotation is done with a perceptual assessment based on listeners' ratings of voice samples. Based on the ratings obtained, the samples are classified into three classes of severity: Low, Moderate and High. These annotated examples serve as target labels to guide the CNN model development.

4.2. Acoustic Feature Extraction

Then, the relevant acoustic characteristics are extracted from each voice recording after the preprocessing. In the proposed framework the pitch, jitter, shimmer, harmonics to noise ratio (HNR) and Mel frequency cepstral coefficients (MFCCs) are considered.

Pitch is the fundamental frequency attribute of the voice and jitter is the measure of variation of fundamental frequency between consecutive cycles. Shimmer is the variation in the amplitude from the period to period. Shimmer can give information about stability of the voice. HNR is the ratio of harmonic to noise components of speech signal. MFCCs describe the spectral properties of speech and are popularly used to represent speech related information. The extracted acoustic parameters are arranged in a suitable feature representation and are input to the CNN model. This representation allows the network to learn discriminative patterns related to different levels of voice disorder severity.

4.3. CNN-Based Severity Classification

A convolutional neural network (CNN) is employed as the primary classification model. The CNN automatically learns hierarchical representations from the extracted feature representation without requiring all discriminative patterns to be manually defined.The CNN architecture consists of convolutional layers, pooling layers, and fully connected layers. The convolutional layers identify important local patterns in the input representation, while pooling layers reduce the dimensionality of the learned feature maps and retain the most relevant information. The resulting high-level representations are forwarded to fully connected layers for classification.The final classification layer produces the probability of each severity category. The class with the highest predicted probability is selected as the final voice-disorder severity.

4.4. Voice Disorder Severity Prediction

The final stage generates the severity assessment for the input voice sample. The trained CNN classifies each recording into one of three predefined categories are Low Severity Moderate Severity and High Severity

5. EXPERIMENTAL RESULT

Comparative Performance Analysis of SVM, KNN, Random Forest, and CNN using Comparative performance analysis: Accuracy,Sensitivity,specificity, Precision and F1-Score as shown in Table I.

Table I. Comparative Performance Analysis of SVM, KNN, Random Forest, and CNN

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
SVM	93.64	93.47	96.93	93.12	93.25
Random Forest (RF)	91.36	90.69	95.74	90.8	90.74
KNN	85.91	86.53	93.36	86.56	85.19
CNN	95	95	97.5	95.07	95

The performance of the proposed voice-disorder severity classification system was evaluated using three conventional machine-learning classifiers, namely Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN), and one deep-learning model, namely Convolutional Neural Network (CNN). The models were evaluated using accuracy, sensitivity, specificity, precision, and F1-score.

As presented in Table I, the CNN achieved the highest classification accuracy of 95.00%, followed by SVM with 93.64%, Random Forest with 91.36%, and KNN with 85.91%. The results indicate that the CNN was able to learn more discriminative patterns from the voice features and provide better separation among the three severity categories.The CNN confusion matrix further demonstrates its classification capability. The CNN confusion matrix was normalized and presented in percentage form. The test dataset comprised 80 High, 60 Moderate, and 80 Low samples, resulting in a total of 220 test samples. The CNN correctly classified 72 of 80 High samples (90%), 57 of 60 Moderate samples (95%), and all 80 Low samples (100%). Eight High samples were misclassified as Moderate, while three Moderate samples were misclassified as High. No Low samples were misclassified. Overall, 209 out of 220 samples were correctly classified, resulting in an accuracy of 95.00% as shown in Fig 4.

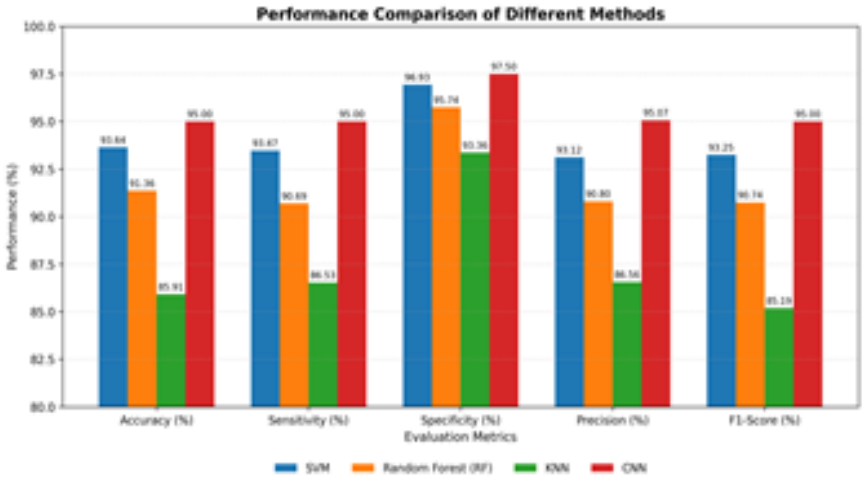

Fig. 2. Performance Comparision of Different Methods

The CNN obtained averaged sensitivity of 95.00% indicating in Fig. 2. Performance Comparisons of Different Methods that the model successfully identified the majority of samples belonging to their respective severity classes. The specificity of 97.50% indicates that the model could reject the samples from wrong severity categories. In addition, the CNN achieved a precision of 95.07% and an F1-score of 95.00%, indicating a good balance between correctly identifying severity classes and minimizing incorrect predictions. The accuracy of the proposed machine learning techniques was 93.64% for SVM, followed by RF and KNN. This demonstrates that SVM was able to build effective decision boundaries between the voice-disorder severities. But the performance was still slightly less than the CNN. RF has accuracy of 91.36% and KNN has 85.91%. The relatively lower performance of KNN could be attributed to its dependency on the similarity of neighboring feature vectors which could be affected by the changes in the voice characteristics of individuals. The better performance of CNN implies that deep learning-based feature learning can be useful for voice-disorder severity classification. CNN can learn hierarchical patterns from the input feature representation without relying solely on pre-defined decision boundaries. This is important for voice signals because subtle differences in spectral, temporal and acoustic features can be involved in pathological changes.

The experimental results generally show that the proposed CNN based approach achieves promising performance for classification of three levels of voice-disorder severity (High, Low, and Moderate). The results also show that voice characteristics can be analysed automatically to potentially allow for objective and efficient severity assessment. More real-time data with larger and more diverse data sets, additional speakers, and independent test data, however, are needed to assess the generalizability and clinical utility of the proposed approach. The results show that the suggested combination of real-time voice data collection, acoustic feature extraction, listener-based severity labelling, and CNN classification is a promising method for automated assessment of voice-disorder severity. In particular, the high recognition rate in the Moderate category indicates that the learned acoustic representation can discriminate this category well among the evaluated samples. However, some confusion was observed between the High and Low categories, suggesting that these severity levels may have overlapping acoustic features. This implies that the generalization ability of the model may be further improved with more voice recordings and more diverse speakers. Overall, the experimental findings demonstrate the potential of the proposed CNN-based framework for automated classification of voice-disorder severity into High, Low, and Moderate categories. The system can potentially support preliminary voice screening and continuous voice monitoring using easily accessible smartphone-based recording. However, the reported results are based on the evaluated dataset, and further validation using a larger, speaker-independent, clinically annotated dataset and recordings collected under different real-world conditions is required before clinical deployment. Thus results demonstrate that the proposed CNN-based voice-disorder assessment framework achieved the highest performance among the evaluated models, with 95.00% accuracy and 95.00% F1-score. The findings indicate that CNN-based deep feature learning can effectively capture severity-related information from acoustic voice representations. The combination of smartphone-based real-time recording, standardized pre-processing, acoustic feature extraction, listener-based annotation, and CNN classification provides a promising framework for automated voice-disorder severity assessment. The experimental results show that the proposed approach can effectively classify the severity of laryngeal voice-disorder into Low, Moderate and High

categories using speech-based acoustic characteristics. The MFCC, Jitter, Shimmer and HNR together provide complementary information about spectral and voice-quality characteristics of impaired speech. MFCCs capture important spectral patterns while Jitter and Shimmer describe variations in the fundamental frequency and amplitude respectively and HNR gives information about the harmonic and noise components of the voice signal. Hence, the combination of these features gives a more complete description of the voice characteristics than using only one type of feature. SVM, KNN and Random Forest showed good classification performance among the evaluated classifiers. However, the highest accuracy of 95% was obtained by the CNN. The better performance of the CNN can be attributed to the fact that it can automatically learn complex and nonlinear patterns from the extracted acoustic representations. Compared to traditional classifiers that rely more on the pre-defined feature–class relationships, CNN can discover higher-level patterns that may be related to the difference in voice-disorder severity. The results also indicate that the classification of severity levels is feasible using relatively informative acoustic parameters obtained from speech signals. The improved performance of CNN suggests that deep learning can effectively exploit the combined information provided by MFCC, Jitter, Shimmer, and HNR.

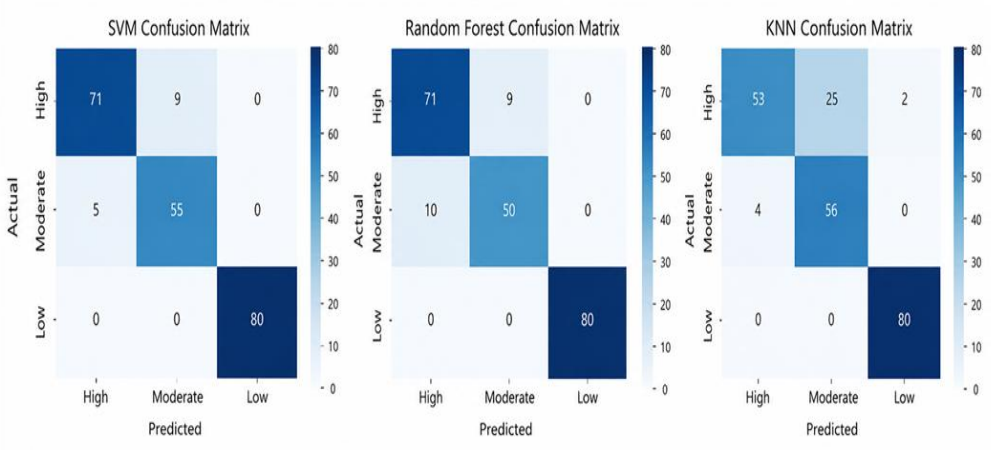

Fig. 3. Confusion matrix for SVM.RF and KNN.

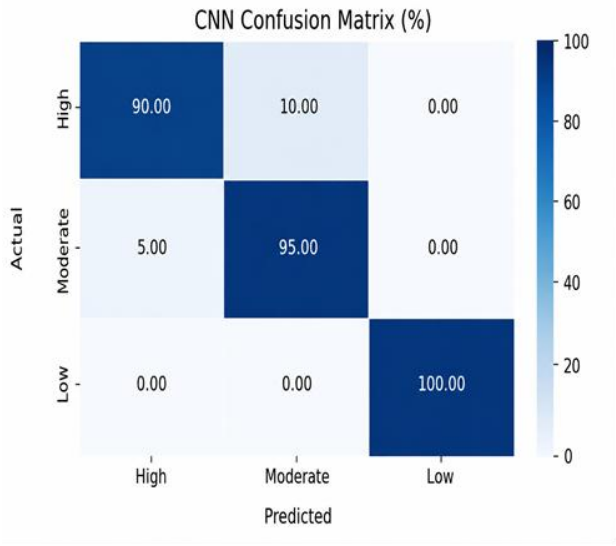

Fig. 4. Confusion matrix for CNN

6. CONCLUSION

This study presented a CNN-based framework for automated voice-disorder severity classification using real-time voice recordings collected through a mobile phone. The proposed framework consists of real-time voice acquisition, pre-processing,

acoustic feature extraction and CNN-based classification. The collected recordings were processed through noise reduction, normalization, and standardization before extracting pitch, jitter, shimmer, HNR, and MFCC features. Based on listener assessment, the voice samples were categorized into High, Moderate, and Low severity classes. The experimental results demonstrated that the CNN achieved an overall accuracy of 95.00%, with a sensitivity of 95.00%, specificity of 97.50%, precision of 95.07%, and F1-score of 95.00%. The CNN also outperformed the evaluated conventional machine-learning approaches, demonstrating its capability to learn discriminative patterns related to voice-disorder severity. The proposed CNN achieved the overall accuracy of 95% for three-level voice-disorder severity classification, evaluated by stratified cross-validation. In the experimental setup, Calà et al. [21] obtained 89% accuracy with KNN in mild, moderate and severe spasmodic dysphonia.

The proposed study involves the use of smartphones for mobile-phone based voice recording in this environmental conditions using smartphones. This makes the framework more practical and potentially accessible for preliminary voice screening and continuous monitoring.. Therefore, robust pre-processing and model generalization are important for reliable real-world application. Overall, the findings indicate that the proposed combination of acoustic features, time-stretching augmentation, and CNN-based classification provides an effective and automated approach for voice-disorder severity assessment and has potential as a decision-support tool for preliminary screening and continuous monitoring.

7. FUTURE SCOPE

Future work will focus on extending the dataset with a larger number of voice recordings collected directly from mobile phones in various real-world environments, like classrooms, homes, hospitals, offices and outdoor environments. Recordings can be collected with different smartphone models and microphone configurations to evaluate the robustness of the proposed system to device-related variations. The dataset can also include speakers with different ages, genders, speaking styles, and levels of voice-disorder severity.

DECLARATIONS

This research received no external funding. The dataset was obtained from the concerned organization upon formal request and is not publicly available due to privacy and organizational restrictions. The authors declare no competing interests. All authors contributed to the study and approved the final manuscript.

ACKNOWLEDGMENT

We gratefully acknowledge the Director, Principal, and staff of Balvikas Kendra, Speech, Hearing and Language, Preschool and Parent Training Centre, operated by the Ekvira Multipurpose Foundation, Akola, for their valuable support, cooperation, and assistance in facilitating the collection of the speech database used in this research. Their guidance, support, and insightful contributions were crucial to the success of this work.

REFERENCES

[1] R. Islam, M. Tarique, and E. Abdel-Raheem, "A Survey on Signal Processing Based Pathological Voice Detection Techniques," IEEE Access, vol. 8, pp. 66750–66773, 2020, doi: 10.1109/ACCESS.2020.2985280.

[2] M. G. Tulics and K. Vicsi, "The Automatic Assessment of the Severity of Dysphonia," International Journal of Speech Technology, vol. 22, pp. 341–350, 2019.

[3] A. A. Joshy and R. Rajan, "Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 1147–1157, 2022, doi: 10.1109/TNSRE.2022.3169814.

[4] H.-C. Kuo et al., "Toward Real-World Voice Disorder Classification," IEEE Transactions on Biomedical Engineering, 2023, doi: 10.1109/TBME.2023.3270532.

[5] M. U. Rahman and C. Direkoglu, "A Hybrid Approach for Binary and Multi-Class Classification of Voice Disorders Using a Pretrained Model and Ensemble Classifiers," BMC Medical Informatics and Decision Making, vol. 25, p. 177, 2025.

[6] E. J. Yeo, K. Choi, S. Kim, and M. Chung, "Cross-lingual dysarthria severity classification for English, Korean, and Tamil," 2022. The study investigated acoustic features associated with phonation, articulation, and prosody for

multilingual dysarthria severity classification.Dolor, C., Sit, D., 2023. Applications of Lorem Ipsum in Scientific Publishing. *International Journal of Placeholder Science* 18(2), 45–60.

[7] H.-C. Kuo, Y.-P. Hsieh, H.-H. Tseng, C.-T. Wang, S.-H. Fang, and Y. Tsao, "Toward real-world voice disorder classification," IEEE Transactions on Biomedical Engineering, 2023.

[8] C. Madill, Z. H. Leong, D. Manoharan, D. Gunjawate, C. Grover, K. Sandham, R. Gupta, C. Jin, D. D. Nguyen, J. J. Johnson, and D. Novakovic, "Classifying voice disorders for machine learning: a pilot study using the USVAC-C2025 diagnostic framework," Frontiers in Digital Health, vol. 8, 2026, Art. no. 1752356, doi: 10.3389/fdgth.2026.1752356.

[9] S. A. Frankford, A. Estrada, and C. E. Stepp, "Contributions of Speech Timing and Articulatory Precision to Listener Perceptions of Intelligibility and Naturalness in Parkinson's Disease," Journal of Speech, Language, and Hearing Research, vol. 67, no. 9, pp. 2951–2963,2024,doi:10.1044/2024_JSLHR-23-00802.

[10] Gharde, M.B., Patil, V.V. (2024). Exploring Speech Disorder Detection in Laryngeal Diseases: A Comprehensive Examination. In: Vasant, P., et al. Intelligent Computing and Optimization. ICO 2023. Lecture Notes in Networks and Systems, vol. 1169. Springer, Cham. https://doi.org/10.1007/978-3-031-73324-6_19

[11] Verde, L., De Pietro, G., Sannino, G. (2018). Voice Disorder Identification by Using Machine Learning Techniques. IEEE Access, 6: 1–12. https://doi.org/10.1109/ACCESS.2018.2816338

[12] Kharibam, J., Devi, A.A. (2019). Automatic Speaker Recognition Using MFCC and Artificial Neural Network. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(1S): 39–42. https://doi.org/10.35940/ijitee.A1010.1191S19

[13] Miliaresi, I., Pikrakis, A. (2023). A Modular Deep Learning Architecture for Voice Pathology Classification. IEEE Access, 11: 80465–80483. https://doi.org/10.1109/ACCESS.2023.3300795

[14] 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

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

[16] 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

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

[18] Gharde, Manisha & Patil, Vaishali. (2026). VocalCare: Convolutional Neural Network Model for Laryngeal Disorder Detection. Ingénierie des systèmes d information. 31. 1635-1644. 10.18280/isi.310519.

[19] 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

[20] Gharde, M.B., Patil, V.V. (2025). Machine Learning Classification Techniques for the Diagnosis of Voice Disorders: Laryngeal Conditions. In: Ghonge, M.M., Liu, H., Khan, M., Tran, T.A. (eds) Advances in Emerging Technologies and Computing Innovations. ICETCI 2025. Sustainable Artificial Intelligence-Powered Applications. Springer, Cham. https://doi.org/10.1007/978-3-031-92854-3_31

[21] Calà, F.; Frassineti, L.; Manfredi, C.; Dejonckere, P.; Messina, F.; Barbieri, S.; Pignataro, L.; Cantarella, G. Machine Learning Assessment of Spasmodic Dysphonia Based on Acoustical and PerceptualParameters. Bioengineering 2023, 10,426.