IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 March 2025, accepted 12 April 2025, date of publication 18 April 2025, date of current version 19 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3562232

RESEARCH ARTICLE

Multi-Feature Fusion-Based Speech Disorder Classification Using MobileNetV3-EfficientNetB7, Linformer-Performer, and SHAP-Aware XGBoost

ABDUL RAHAMAN WAHAB SAIT[1], (Member, IEEE), SURESH SANKARANARAYANAN[2], (Senior Member, IEEE), AND P. GOUTHAMAN[3]

[1]Department of Archives and Communication, Center of Documentation and Administrative Communication, King Faisal University, Al-Ahsa, Al Hofuf 31982, Saudi Arabia
[2]Department of Computer Science, College of Computer Sciences and Information Technology (CCSIT), King Faisal University, Al-Ahsa, Al Hofuf 31982, Saudi Arabia
[3]Department of Networking and Communications, Faculty of Engineering & Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India

Corresponding author: Abdul Rahaman Wahab Sait (asait@kfu.edu.sa)

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant KFU251485).

ABSTRACT Traditional speech disorders (SD) detection relies on subjective analysis, resulting in inconsistent outcome. Direct voice classification lacks effective approaches to capture temporal dependencies. Machine learning (ML) models face challenges in extracting the complex temporal and spectral variations in speech signals. Advanced deep learning (DL) and transfer learning techniques offer a foundation for early screening of SD. However, the lack of interpretability reduces the generalization capabilities of these models. Addressing these shortcomings is essential in order to improve the accuracy of SD detection and clinical trustworthiness. Thus, the proposed study introduces a novel image-based SD classification model to classify healthy and pathological speech with high accuracy and robustness. The raw speech signals are transformed into Mel-Spectrograms to overcome the limitations of direct voice classification. To facilitate the model's interpretability, the statistical and handcrafted acoustic features are extracted from the raw speech signals. Hybrid MobileNet V3-EfficientNet B7and Linformer-Performer are employed to extract diverse features from the Mel-Spectrograms. An attention-based feature fusion is used to identify critical features indicating the SD patterns from the extracted features. The XGBoost classifier is optimized using Bayesian Optimization and HyperBand (BOHB) to classify the healthy and pathological speech. SHapley Additive exPlanations (SHAP) values is employed to offer valuable insights into the model's decisions. The proposed model obtains an exceptional performance on two benchmark datasets. On the Saarbruecken Voice Database (SVD), it achieves an accuracy of 98.1% with loss of 0.13. It yields a remarkable generalization accuracy of 98.2% on the VOICE dataset, outperforming the state-of-the-art models. In addition, it contributes a significant advancement in SD detection, setting the stage for future research endeavors.

INDEX TERMS Image classification, Mel-spectrogram, speech disorder, handcrafted acoustic features, deep pre-trained model, transformers, machine learning, deep learning, model interpretability.

I. INTRODUCTION

Humans rely on speech as a primary mode of communication. Speech conveys ideas, feelings, and intentions, establishing relationships [1]. Speech disorders (SD) provide considerable challenges to day-to-day communication for millions of individuals across the globe. SD, including stuttering and aphasia can lead to social isolation, reducing individuals' quality of life [2]. These disorders entail a significant emotional and psychological burden, affecting an individual's

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

self-esteem and their social interactions. Early detection can empower individuals to improve their communication abilities and enhance their quality of life [2]. Moreover, automated SD identification may promote awareness of speech impairments, promoting social empathy. The early SD diagnosis with a high degree of accuracy is essential for successful intervention and therapy. Thus, academics and practitioners are exploring novel approaches to diagnose SD with optimum accuracy.

Traditionally, clinical evaluations are widely used for diagnosing SD [3]. These evaluations are subjective and resource intensive [3]. Access to expert treatment is typically restricted in rural or underprivileged locations, exacerbating the challenges faced by the individuals with SD. The emerging artificial intelligence (AI)-based applications can be a promising alternative to identify SD in the early stages [4]. The potential of AI-driven image-based approaches, converting complicated voice signals into visual patterns has attracted a substantial amount of interest in SD detection [4]. Mel-spectrogram is an effective frequency and time-domain representation of sound [5]. By visualizing speech signals, Mel-spectrograms allow deep learning (DL) models, including convolutional neural networks (CNNs) and vision transformers (ViTs) to extract diverse SD features [5]. By identifying these intricate patterns, the performance of the SD detection models can be improved.

The potential of CNNs and ViTs in learning complex patterns overcome the limitations of traditional SD detection approaches [6]. CNNs can detect anomalies in Mel-spectrograms using the pitch irregularities, energy shifts, and frequency modulations. For instance, the identification of irregular pauses or distorted articulation can support healthcare professionals to detect disorders like stuttering or dysarthria [7]. Recurrent neural networks (RNNs) can capture pitch trends or rhythmic irregularities in speech. These models are well-suited for analyzing speech signals over time, identifying subtle speech impairments [8]. In recent years, ViTs gained prominence for their self-attention mechanisms to extract complex interdependencies in medical images [9]. These models can capture long-range dependencies in voice samples. Due to their non-sequential architecture, these models outperformed RNNs. Integrating CNNs with RNNs or ViTs may improve the efficiency of the SD detection [10].

DL models require large amounts annotated data to achieve high accuracy [11]. Obtaining such data for SD detection is challenging due to the lack of annotated recordings. The class imbalance in SD datasets may skew model predictions, reducing the overall performance of the SD detection model [12]. The limited model interpretability poses a significant challenge in healthcare applications. Clinicians demand explainable or interpretable predictions to make informed decisions [13]. Researchers focusses on CNNs and ViTs for developing SD classification model using image classification techniques [14], [15], [16]. However, CNNs and ViTs models are heavily rely on learned features, neglecting well-established handcrafted features. As a result, these models may produce suboptimal diagnostic information.

There is a demand for effective approaches to overcome the limitations in order to build a reliable system for detecting SD using images. The complexity of voice samples vary by language, dialect, and individual's accent, affecting the quality of Mel-spectrograms [17]. Background noise, recording quality, and emotional tone may influence the feature representation. Maintaining a diverse and high-quality set of voice samples is crucial in generating Mel-spectrograms. In order to improve the quality of Mel-spectrograms, robust preprocessing techniques are essential. Designing a feature fusion technique is essential in order to identify the key features associated with SD. Conventional feature fusion techniques, including early or late fusion face challenges in capturing the complex relationships among features [18]. The advanced fusion techniques, such as attention-based and graph neural networks can dynamically prioritize SD features, improving detection accuracy and reliability.

There is a lack of research studies on incorporating handcrafted acoustic and learned features to improve the model's generalizability. The real-word settings with noise, overlapping speech or poor-quality recordings reduce the performance of the existing SD classification models. Developing interpretable models, aligning with clinical knowledge remain a challenge in classifying the speech signals. There is insufficient focus on building lightweight feature extraction models for deploying SD classification models on edge-devices or real-time applications. These knowledge gaps highlight the demand for robust pre-processing techniques and effective feature extraction approaches. The motivation for proposing a robust SD classification model stems from these knowledge gaps and limitations. The proposed study offers an interpretable and generalizable SD classification model. The study contributions are as follows:

1. Extraction of diverse features using CNNs and ViTs architectures to improve the classification accuracy.

The statistical and handcrafted acoustic features are integrated with the features extracted through CNNs and ViTs. The hybrid feature extraction architecture captures intricate patterns of SD. This complementary approach achieves a tradeoff between computational efficiency and reliable feature representation. It supports the proposed classification model deployment on the resource-constrained environment.

2. A novel attention-based feature fusion technique for identifying crucial features associated with SD.

The proposed attention-mechanism-based feature fusion minimizes the influence of redundant or irrelevant features. It enhances the model's interpretability by providing the valuable features.

3. An Interpretable XGBoost-based SD classification.

The XGBoost classifier is fine-tuned using BOHB algorithm. The optimized classification model generates interpretable outcomes using the SHAP values. The recommended approach enhances the XGBoost performance, reducing overfitting and improving generalization to novel data.

The remaining part of the study is organized as follows: The existing SD detection techniques and knowledge gaps are presented in section II. Section III presents the proposed methodology for extracting the classifying the speech signals. It outlines the process of generating Mel-Spectrograms, extracting and fusing diverse features, and classifying the extracted features. The experimental settings and outcomes are highlighted in section IV. Section V reveals the significance of the study findings. Lastly, the study contributions, limitations, and future directions are outlined in section VI.

II. LITERATURE REVIEW

The advancement of DL techniques has led to substantial developments in SD detection [19]. These techniques replaced conventional rule-based and statistical models, enabling automated feature extraction, classification, and pathological speech interpretation. The pathological variations in speech signals introduce computational and diagnostic complexities [20]. For instance, individuals with conditions such as dysarthria or vocal fold paralysis exhibit different levels of severity, prosodic changes, and speech intelligibility, causing challenges for machine learning (ML) models to establish robust decision boundaries [21]. In addition, healthy and early-stage SD share similar acoustic characteristic, influencing ML models to generate false negatives [22]. DL models overcome the shortcomings of ML models by identifying subtle variations in speech signals. CNNs have been widely used for image-based speech signal analysis. These models utilize Mel-spectrograms in order to differentiate healthy and pathological speech signals.

Sindhu and Sainin [22] highlight the efficiency of CNNs over traditional ML models. The feature learning capabilities enable CNNs and ViTs architecture to classify speech signals with optimal accuracy. Mohammed et al. [23] develop a CNNs model to classify fluent and stuttered speech using spectrograms. The local frequency features are extracted from Mel-spectrograms in order to classify the voice samples. Mohaghegh and Gascon [24] use multiple modalities for classifying healthy and pathological voices. Feature fusion technique is used for identifying the significant SD features. Verde et al. [25] employ DL approaches for voice disorder detection. Abdulmajeed et al. [26] build SD detection model to identify voice pathology. Ribas et al. [27] integrate used self-supervised representations for detecting SD. Park et al. [28] use adversarial continual learning technique for voice pathology detection. Hemmerling et al. [29] employ ViTs for classifying Parkinson's disease. Islam and Tarique [30] use a CNN model to differentiate healthy and pathological voices based on the spectral features. Rubio et al. [31] investigate the significance of the data augmentations in improving the performance of the SD detection techniques. Lau et al. [32] outline the application of pre-trained ViTs-based automatic SD assessment, demonstrating improved model interpretability and classification performance. Sayadi et al. [33] emphasize the role of handcrafted features in improving the SD classification performance.

Several studies explore hybrid approaches integrating ML and DL techniques to classify speech signals. Memari et al. [34] employ gradient boosting technique with DL-based feature extraction to enhance SD detection and classification. Bindas and Oniuiri [35] propose a SD detection model by combining pre-trained CNNs with gradient boosting technique.

Despite significant progress, CNNs and ViTs face challenges, including class imbalance and lack of model's interpretability. CNNs are intended to capture local patterns in speech signals. However, these models encounter difficulties in handling long-range dependencies. They demand substantial computational resources for extracting crucial patterns associated with SD. ViTs reduce the dependency on large convolutional layers, leading to computationally efficient architectures. Nonetheless, these architectures use image patches without prioritizing specific frequency bands, potentially losing disorder-specific acoustic features.

To enhance the SD classification and address the shortcomings of individual architectures, a hybrid model integrating CNNs and ViTs can be explored. The CNNs component can extract localized frequency-domain features. Capturing global phonetic disruptions using the ViTs can learn contextual relationships between different phonation regions. By optimizing feature extraction using the lightweight CNNs and ViTs, the hybrid models can achieve real-time performance without sacrificing classification precision. This strategy can enable faster and accurate SD diagnosis. The extraction of key features with attention scores can improve model's interpretability, motivating us to develop a hybrid SD classification with improved model interpretability.

III. MATERIALS AND METHODS

In this study, the advanced techniques, including MobileNet V3 [36], EfficientNet B7 [37], Linformer [38], Performer [39], and XGBoost [40], are used to build a robust and interpretable SD classification model. Figure 1 presents the proposed research methodology for classifying SD using individual speech signals. MobileNet V3 and EfficientNet B7 models are ideal for processing Mel-Spectrograms due to their potential in extracting formants, pitch variations, and abrupt frequency shifts. Linformer and Performer architectures address the computational complexities of standard ViTs. These models identify speech rhythm or subtle articulating patterns. XGBoost algorithm can handle high-dimensional data. It guarantees reliable performance using its gradient-boosting approach. By leveraging fine-tuning and interpretability enhancement, XGBoost can deliver optimal classification accuracy in real-time settings.

TABLE 1. Features of SVD and VOICED datasets.

Features	VOICED	SVD
Speech types	Sustained vowels	Sustained vowels, continuous speech, and phoneme speech
Pathologies	Dysphonia, vocal nodules, and Polyps	Laryngitis, Parkinson's disease, dysphonia, vocal nodules, and functional dysphonia
Noise level	High signal-to-noise ratio	High signal-to-noise ratio
Meta data integration	Lifestyle, demographic, and acoustic data	Acoustic and physiological data
SHAP value applicability	SHAP value can explain the significance of acoustic features in the predictions	SHAP value can identify the influence of acoustic, spectral, and physiological features on predictions

A. DATASET ACQUISITION

Two public repositories are utilized to train and test the proposed SD classification model. Saarbruecken Voice Database (SVD) [41] is one of the comprehensive collections of 2043 individual's voice recordings. It comprises of 687 healthy and 1356 pathological voice samples. It is based on German phonetics, containing sustained vowel sounds. The sustained vowel sounds (/a/, /e/, /i/, /o/, and /u/) are phonetically universal, including German and English languages. The features like pitch, jitter, shimmer, and harmonic-to-noise ratio, render the dataset suitable for SD detection and classification across the globe. The dataset encompasses healthy and pathological speech samples. It covers a wide range of vocal conditions, such as laryngitis, Parkinson's disease, and vocal fold paralysis. The high-quality recordings with minimal noises support researches to extract reliable features, including spectral properties, formants, and temporal variations. VOICED [42] dataset contains 208 voice recordings, covering 150 samples of individuals' with pathological speech and 58 healthy individuals. It provides detailed background information, such as age, gender, and lifestyle habits of the individuals. The sustained vowel sounds are ideal for detecting subtle speech irregularities.

Table 1 outlines the features of these datasets.

B. MEL-SPECTROGRAM GENERATION

In order to overcome the background noise, silent sections, and variations in amplitude, the speech data are preprocessed. They employed Wiener filtering technique to remove unwanted background mix. This process preserves the integrity of the speech signal. Voice activity detection was used to segment the data to eliminate silent or irrelevant sections. Additionally, the speech data are normalized to ensure uniformity across the entire samples.

To enrich the data and improve the robustness of the proposed model, data augmentation techniques are applied. Pitch shifting was used to alter the pitch of the speech signal. It simulates diverse vocal characteristics. Time stretching was applied to adjust the speech speed without influencing pitch. It introduces variations in temporal patterns. An additive noise technique is used to include mild background noise

TABLE 2. Details of augmented dataset.

	Healthy	Pathological
Original Data	687	1356
Augmented Data	8777	9216

The audio waveform is transformed into a time-frequency representation by generating Mel-Spectrograms. Mel-Spectrogram maps the frequencies of the speech signal to the relevant Mel-scale. Short-Time Fourier Transform (STFT) was employed to analyze the speech signal in time and frequency domain. Equation 1 presents the mathematical expression for the STFT.

$$S\left(t,f\right) = STFT\left(x\left(t\right)\right) = \sum\nolimits_{n=0}^{N-1} x\left(t+n.w\right).e^{\frac{-j2\pi fn}{N}} \qquad (1)$$

where $S\left(t,f\right)$ is the Spectrogram with frequency (f) over time (t), $x\left(t\right)$ is the speech signal, N is the resolution of the frequency bias, w is the window length, n is the time index, and $e^{\frac{-j2\pi fn}{N}}$ is the complex exponential for the Fourier bias.

Dynamic spectrogram resolution is employed to enhance the ability of Mel-Spectrogram in visualizing SD features. Using this technique, they fine-tuned the Mel-Spectrogram images. Continuous wavelet transform (CWT) is employed to enhance the temporal resolution of speech representations. It enriches the Mel-Spectrogram by revealing intricate speech patterns. It improves the temporal precision of spectral components, allowing the proposed model to identify pathological speech patterns. It decomposes the speech signals into wavelet coefficients in order to offer fine-time resolutions for high-frequency components and better frequency resolution for low-frequency elements. A finer temporal dimension is added to capture variations in jitter, shimmer, and subharmonic structures. By preserving transient acoustic characteristics in Mel-Spectrograms, the proposed model can differentiate healthy and pathological speech samples. Generative adversarial networks (GANs) was employed to enhance the dataset through the generation of realistic augmented spectrograms. SpecGAN [43] was used to generate spectrogram data. It was trained on the primary dataset containing labeled spectrograms. The generator G(z) learns the frequency and temporal patterns associated with healthy and pathological speech data. Equation 2 shows the augmented data.

$$x_{augment} = x_{real} \cup x_{GANs} \qquad (2)$$

where x_{real} is the primary dataset, x_{GANs} is the generated spectrograms, and $x_{augment}$ is the augmented dataset. Table 2 highlight the details of the original and augmented dataset.

Figure 2 (a) and (b) reveal the healthy and pathological Mel-Spectrograms generated through the proposed model. Figure 3 (a) and (b) highlight the original and augmented Mel-Spectrograms. Direct voice classification approaches

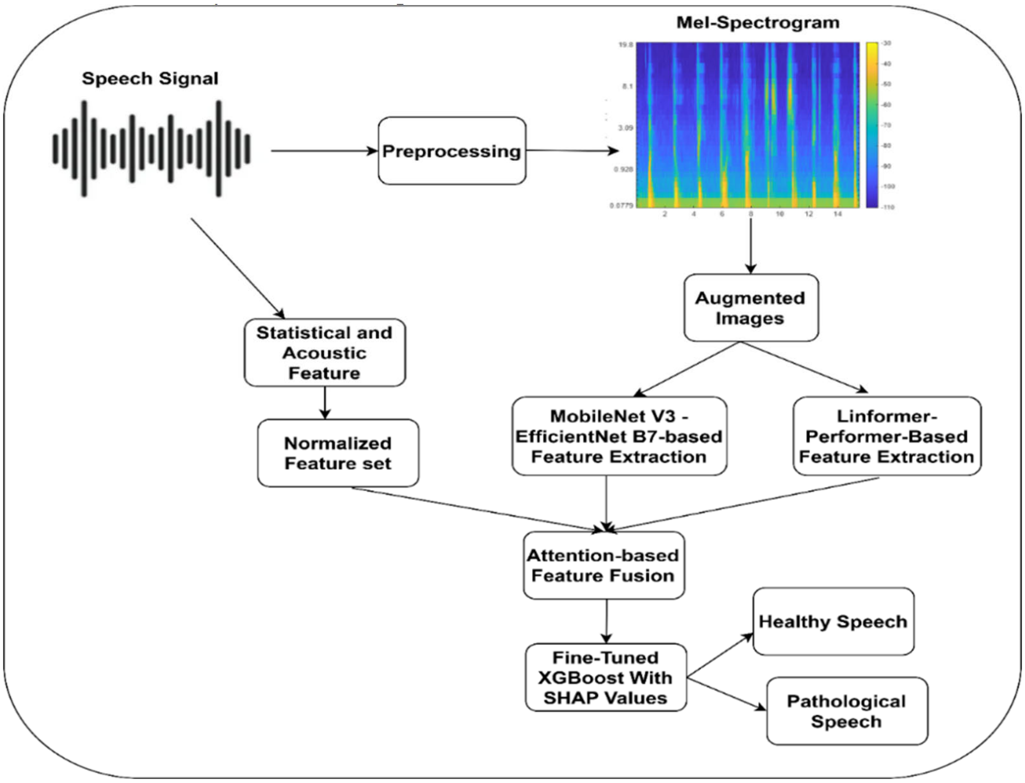
Mel-Spectrogram
19.8
8.1
3.09
0.928
0.0779
2
4
6
8
10
12
14
-30
-40
-50
-60
-70
-80
-90
-100
-110
Speech Signal
Preprocessing
Augmented Images
Statistical and Acoustic Feature
Normalized Feature set
MobileNet V3 - EfficientNet B7-based Feature Extraction
Linformer-Performer-Based Feature Extraction
Attention-based Feature Fusion
Healthy Speech
Fine-Tuned XGBoost With SHAP Values
Pathological Speech

FIGURE 1. The proposed SD classification model.

face challenges in capturing temporal dependencies, requiring computationally intensive recurrent models. Compared to direct voice classification, the proposed image-based SD classification offer a rich, detailed depiction of the speech signal. For instance, Mel-Spectrogram represents spectral and temporal patterns, enabling the detection of irregular harmonics, unstable formants, and noise artifacts.

C. STATISTICAL AND ACOUSTIC FEATURE EXTRACTION

The statistical and acoustic features are extracted prior to the Mel-Spectrogram generation. They compute statistical and handcrafted acoustic features from the preprocessed speech data. Statistical features present the speech signal's distribution and dynamics. These features were extracted over the entire signal in order to present the overall structure and variability of the voice samples.

In parallel, acoustic features are derived, capturing specific aspects of speech associated with disorders. Jitter and shimmer were computed frame-by-frame to guarantee temporal resolution. Harmonics-to-noise ratio is used to measure the vocal clarity. Liner predictive coding is used to extract formant frequencies. Unlike preprocessing, the speech signals are transformed into a structured set of numerical features. These features were normalized in order to ensure compatibility with learned features. This feature engineering approach enriches the overall feature set, contributing to proposed model's effective performance.

D. MOBILENET V3-EFFICIENTNET B7-BASED FEATURE EXTRACTION

To extract SD features, a sequential architecture integrating MobileNet V3 and EfficientNet B7 is employed. This hybrid approach extract detailed features. By capitalizing the unique strengths of MobileNet V3 and EfficientNet B7, the proposed SD model capture frequency variations and transitions within the spectrogram. MobileNet V3 captures edge-level features using its Squeeze-and-Excitation block. EffcientNet B7 detects intricate relationships across the entire time-frequency space. This complementary approach produces an effective feature set with limited computational resources.

MobileNet V3 uses depth-wise separable convolutions for extracting localized features. Depth-wise convolution applies a single convolutional filter per input channel to capture spatial information. Point-wise convolution uses 1 × 1 convolution to integrate the information from the entire channels. Equation 3 shows the mathematical expression integrating point-wise and depth-wise convolutions.

$$F_m = \sigma\left(Conv_{depthwise}\left(x\right).Conv_{pointwise}\right) \tag{3}$$

where x is the Mel-spectrogram, σ is the hard swish function, F_m is the MobileNetV3-features, $Conv_{depthwise}$ and $Conv_{pointwise}$ are the convolution operations.

EfficientNet B7 processes the extracted feature map (F_m) in order to extract higher-order features. The compound scaling function is used to maintain scale, depth, width,

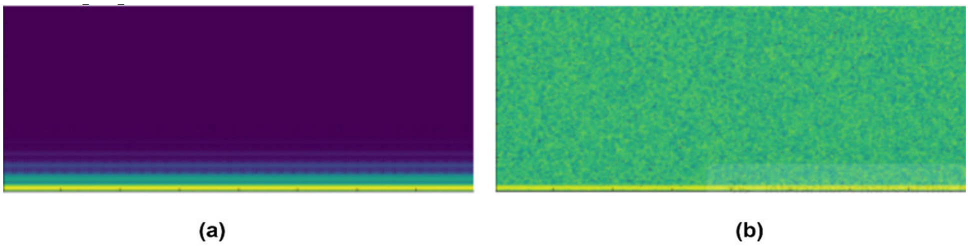

FIGURE 2. Sample Mel-Spectrograms (a) Healthy speech (b) Pathological speech.

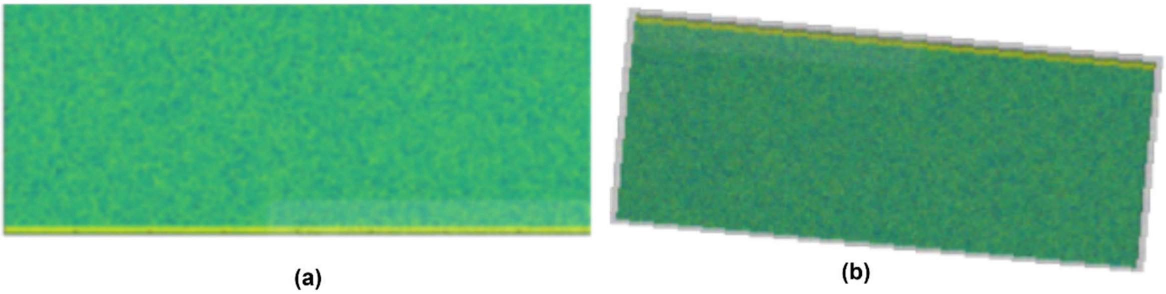

FIGURE 3. Sample generated Mel-Spectrograms (a) Original pathological speech (b) Augmented pathological speech.

and resolution. The mobile inverted bottleneck convolutions applies expanded filters and skip connections to preserve information flow. Equation 4 outlines the extraction process of EfficientNet B7.

$$F_E = MBConv(F_m) \tag{4}$$

where F_E is the EffiientNet B7 features and $MBConv$ is the mobile inverted bottleneck convolutions.

E. LINFORMER-PERFORMER-BASED FEATURE EXTRACTION

The proposed study builds a hybrid feature extraction approach using Linformer and Performer architectures. Unlike traditional ViTs, Linformer and Performer apply approximations to reduce the computational complexities of attention mechanism while retaining accuracy. Linformer captures long-range rhythmic patterns or sustained frequency shifts in the Mel-Spectrogram. It uses a projection matrix ($P \in \mathbb{R}^{n \times k}$) to approximate the self-attention mechanism. Equation 5 outlines the projection matrix.

$$K_{proj} = P.K;\, V_{proj} = P.V \tag{5}$$

where K is the reduced dimensionality, which is significantly smaller than the original sequence ($K \ll n$).

Equation 6 reflects the extraction of global dependencies using the self-attention mechanism.

$$F_L = Softmax\left(\frac{Q\left(K_{proj}\right)^T}{\sqrt{d}_k}\right)V_{proj} \tag{6}$$

where $F_L \in \mathbb{R}^{n \times d}$ represents the feature map, Q is the query, k is the key, and d_k is the dimension, and T is the transpose.

Furthermore, Performer refines the extracted features to ensure a multi-scale analysis of the Mel-Spectrogram. It approximates the softmax attention mechanism using kernalized random feature maps. This hybrid approach focusses on long-term temporal patterns and short-term spectral variations, identifying subtle SD patterns. Equation 7 shows the kernel mappings for queries and keys.

$$\emptyset\,(Q)\,,\emptyset\,(K) \tag{7}$$

where $\emptyset\,(.)$ transforms Q and k into a lower-dimensional feature space.

Equation 8 presents the linearized attention for extracting features.

$$F_P = \emptyset\,(Q)\,.\left(\emptyset\,(K)^T\,V\right) \tag{8}$$

where $F_P \in \mathbb{R}^{n \times d}$ represents the feature map highlighting localized temporal and spectral variations.

F. ATTENTION-BASED FEATURE FUSION

An attention-based feature fusion mechanism is used to integrate diverse feature sets. It identifies the critical features contributing to the final classification, reducing redundancy and improving interpretability. The features, including statistical (F_s) and acoustic (F_a) are normalized in order to support the feature fusion process. Equation 9 presents the concatenation process for combining the features.

$$F_{combined} = F_s + F_a + F_E + F_P \tag{9}$$

where $F_{combined}$ is the concatenated features.

An attention mechanism is applied to assign importance scores (W_i) to each feature dimension based on its significance. Eqn. 10 presents the attention mechanism.

$$F_{attention} = \sum_{i=1}^{d} W_i . F_{combined,i} \tag{10}$$

where $F_{combined,i}$ is the ith feature in the combined feature matrix.

Using this approach, high importance features are amplified and irrelevant features are down-weighted or discarded. The dynamic weighting focusses on pitch variability or specific spectral patterns. Additionally, principal component analysis mitigates the curse of dimensionality. Equation 11 highlights the process of reducing feature dimensions while preserving the most informative features.

$$F_{final} = PCA\left(F_{attention}\right) \tag{11}$$

where $F_{final} \in \mathbb{R}^m$ and $m \ll$ d is the final set of features.

G. XGBoost-BASED SD CLASSIFICATION

To classify the features, XGBoost algorithm is employed. XGBoost uses decision trees as base learners for classifying the features. It builds an ensemble of weak learners and rectify their errors iteratively. It employs regularization, shrinkage, and column subsampling to prevent overfitting, ensuring the classification performance using BOHB algorithm. BOHB fine-tunes the XGBoost's parameters, including learning rate, maximum tree depth, and regularization co-efficient. It integrates the potential of Bayesian optimization and Hyperband techniques. It builds the objective function and accelerates the search process by focusing on promising hyper-parameters to achieve optimal outcomes.

To address the lack of interpretability of the XGBoost classification, SHAP values are incorporated. The SHAP values quantify the contribution of each feature to the outcome. It enhances transparency by identifying the features, including jitter, shimmer, or specific spectrogram embedding. Equation 12 presents the classification of the extracted features.

$$\hat{y} = \sum_{n=1}^{k} W_k T_k \left(F_{final} : h^*\right) \& \hat{y} = \emptyset_0 + \sum_{x=1}^{d} \emptyset_i \tag{12}$$

where $\hat{y}$ is the final prediction, h^* is the best hyperparameters, k is the total number of trees, W_k is the weight of the kth decision tree, T_k is the prediction made by kth decision tree, $\emptyset_0$ is the baseline value, and d is the total number of trees, and $\emptyset_i$ is the SHAP value of ith feature.

H. EVALUATION METRICS

A comprehensive evaluation framework is implemented to assess the proposed model's performance and robustness. Multiple metrics, including accuracy, precision, recall, and F1-score, to evaluate the model's ability in making correct predictions, avoiding false positives, and ensuring balanced performance across the entire classes. Specificity were used to identify the model's capability to correctly identify SD patterns while minimizing the misclassification. Standard deviation and confidence intervals were used to determine the reliability and variability of the model's predictions. The loss values were computed to indicate the degree of error during training and testing phases. The area under receiver operating characteristics (AUROC) and area under precision-recall curve (AUPRC) are used to identify the model's ability in distinguishing healthy and pathological speeches. They calculated the number of parameters and floating-point operations (FLOPs) to determine the computational efficiency of the proposed SD classification model.

I. MODEL IMPLEMENTATION

The experimental setup is carefully designed in order to ensure robust and efficient implementation of the proposed SD classification. The system is configured on a Windows 11 environment, leveraging the capabilities of 32 GB RAM and NVIDIA GEFORCE RTX 3060 Ti with CUDA 11.2. This hardware configuration offered sufficient memory and processing power to preprocessing voice samples and training CNNs, VITs, and XGBoost models. The model development was performed using Python 3.10.6, supporting wide range of libraries required for DL models. The key libraries, including TensorFlow, Librosa, and Pandas, were utilized for feature extraction. Table 3 reveals the key configurations of DL models for implementing the model.

IV. RESULTS

To ensure a rigorous and unbiased evaluation of the model's performance, a five-fold cross-validation strategy with controlled data augmentation is used [44], [45]. The SVD dataset is systematically partitioned into five subsets. The four sets contribute to model training and the remaining set is used to test the model's performance. To enhance the diversity of training data, data augmentation is applied to four training folds. To prevent data leakage, data augmentation is not applied to the fifth fold. This approach guarantees that the model's generalization capability is assessed on unseen data, maintaining a clinically relevant evaluation framework. In addition, the VOICED dataset is utilized to ensure the model's generalization capability on novel data. Table 4 presents the outcomes of the performance evaluation. It reveals the average accuracy, precision, recall, F1-score, specificity of classifying voice samples.

Table 5 outlines the performance of the SD classification models on the SVD datasets. The superior results reflect the model's exceptional ability to distinguish healthy and pathological speech. The proposed model outperformed the pre-trained CNNs and ViTs using its innovative architecture. The existing models face challenges in capturing complex temporal and spectral patterns. The outstanding performance can be credited to the extensive training and effective Mel-Spectrogram generation.

The outcomes presented in Table 6 demonstrate the effectiveness of the proposed model, achieving a generalization accuracy of 98.2%, precision of 98.8%, recall of 98.1%, F1-score of 98.4%, and specificity of 97.4%. By leveraging the hybrid feature extraction architectures, the proposed model captured detailed and hierarchical features. MobileNet V3

TABLE 3. Computational configurations.

Parameters	Values
MobileNet V3	
Depth Multiplier	1.0
Input Size	224 × 224
Activation Function	Swish
Batch Size	32
EfficientNet B7	
Dropout	0.5
Optimizer	Adam with a learning rate of 1e-4, decay set of 0.96 per epoch
Linformer	
Sequence Length	196
Low-rank Approximation Dimension	64
Attention Head Size	8 heads
Activation	GeLU
Performer	
Attention Kernel	FAVOR+
Number of attention heads	8
Hidden layer size	512
Dropout rate	0.2
Statistical and Handcrafted Feature Extraction	
Pitch	Window size=25ms and Step = 10 ms
Intensity	Framelength = 20ms and FFT size = 512
Mel-Frequency cepstral coefficients	13 coefficients and window = 25ms
Spectral Centroid	FFT size = 1024 and Hop Length = 256
Spectral Bandwidth	Order = 2 and FFT size = 512
Spectral Contrast	6 Frequency Bands
Jitter	Local Jitter Function
Shimmer	Local Shimmer Function
Harmonics –to-noise Ratio	Window = 30ms and Step = 10ms
Zero Crossing Rate	Frame Size = 25ms and Hop = 10ms
Voiced / Unvoiced Ratio	Extracted from Pitch Contour
Formant Frequencies	Linear Predictive Coding
SpecGAN	
Generator Input Latent Vector	100-dimensional normal distribution
Generator Number of Layers	5 Convolutional Layers
Generator Stride	2
Generator Activation function	LeakyReLU
Discriminator Network Architecture	Convolutional Layers With Downsampling
Discriminator Sride	2
Discriminator Activation Function	LeakyReLU
Loss Function	Wasserstein loss with gradient penalty
Batch Size	64
XGBoost	
Learning Rate	0.1, adjustable via BOHB
Max Depth	6
Subsample	0.8
Colsample by Tree	0.8
Number of Tree	1000, with early stopping based on validation loss

TABLE 4. Findings of K – fold cross validation - SVD dataset.

Fold	Accuracy	Precision	Recall	F1-Score	Specificity
1	98.6	99.1	98.7	98.9	96.9
2	97.9	98.3	97.9	98.1	97.6
3	99.5	99.4	98.8	99.1	98.8
4	98.8	98.5	97.9	98.2	98.3
5	98.1	97.8	97.3	97.6	97.0

show significantly lower performance due to their limited ability to identify the intricate SD patterns.

TABLE 5. Findings of performance evaluation – SVD dataset.

Models	Accuracy	Precision	Recall	F1-Score	Specificity
Proposed Model	98.1	97.8	97.3	97.6	97.0
MobileNet V3	93.2	93.0	91.5	92.2	90.1
EfficientNet B7	94.8	92.9	91.7	92.3	89.2
Linformer	94.1	93.1	92.0	92.5	88.4
Performer	93.6	90.4	89.5	89.9	89.2

TABLE 6. Findings of performance evaluation – VOICED dataset.

Models	Accuracy	Precision	Recall	F1-Score	Specificity
Proposed Model	98.2	98.8	98.1	98.4	97.4
MobileNet V3	90.1	88.4	87.9	88.1	86.2
EfficientNet B7	89.5	87.5	86.8	87.1	85.1
Linformer	90.5	87.4	85.4	86.3	85.9
Performer	88.7	85.7	86.7	86.2	83.9

Figure 4 illustrates the remarkable classification performance of the proposed model. It underscores the model's robustness and reliability in identifying healthy and pathological samples. Mel-Spectrogram is the primary contributor for the successful classification. It transforms raw speech signals into detailed time-frequency representations, enabling the model to capture frequency irregularities. In addition, the dynamic weighting approach of the proposed feature fusion facilitates the critical features to the fine-tuned XGBoost classifier, strengthening the classification performance.

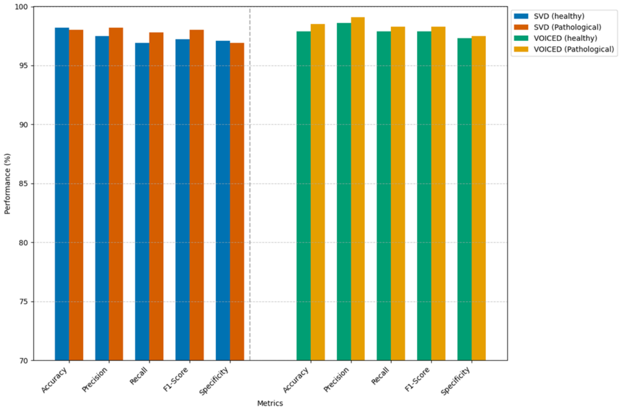

FIGURE 4. Findings of classification analysis – SVD and VOICED datasets.

Table 7 provides an insightful comparison in terms of computational configurations and statistical analysis. The proposed model achieves a better outcome with 12.3 Million parameters and 13.6 GFLOPs. This low computational footprint allows the deployment of the model in real-world settings. Standard deviation of 0.006 reflects the model's stability on novel data. The low loss value and confidence interval indicates the model's reliability. MobileNet V3 required minimum number of parameters. However, the

TABLE 7. Outcomes of statistical analysis and computational efficiencies –VOICED dataset.

Models	Parameters (in Millions	FLOPs (in Giga)	Standard Deviation	Confidence Interval	Loss
Proposed Model	12.3	13.6	0.006	[95.8-96.9]	0.13
MobileNet V3	5.4	27.2	0.008	[96.1-96.9]	0.37
EfficientNet B7	36.5	69.4	0.005	[97.1-98.2]	0.24
Linformer	14.8	15.6	0.007	[95.6-96.7]	0.27
Performer	18.5	17.9	0.009	[95.3-97.8]	0.35

higher loss of 0.37 represents lower classification accuracy. Generally, ensemble models are employed to have a higher parameter count than individual models due to the combination of multiple ML and DL architectures. However, the proposed model achieves a minimum number of parameter through a feature extraction-based ensemble strategy. It extracts compact feature representations from each model, preventing unnecessary duplication of learned parameters. It employs parameter sharing across feature extractors. As a result of efficient feature extraction, parameter sharing, and feature fusion strategies, an exceptional outcome is achieved with limited computational resources compared to individual CNNs and ViTs architectures.

Figure 5 displays AUROC curves for the SVD and VOICED datasets. It indicates the model's high discriminatory power. The proposed model achieved a high true positive rate while maintaining a low false positive rate. It identified pathological speech with minimal false alarms. The VOICED dataset AUROC curve highlights the ideal generalization performance of the proposed model. The advanced feature extraction and fusion mechanisms contributed to the outstanding performance.

Figure 6 illustrates the model's ability to maintain high precision and recall by achieving an exceptional AUPRC for the SVD and VOICED datasets. The slight drop in precision indicates the model's effort in detecting pathological samples. This trade-off is acceptable in clinical scenarios to avoid missing pathological cases. AUPRC reinforces the model's reliability, guaranteeing accurate identification of pathological cases.

Table 8 outlines the performance of SD classification models. Compared to the existing model, the proposed model demonstrated outstanding performance. The high performance is validated by low loss and standard deviation. The inclusion of hybrid feature extraction approaches enhanced the proposed model's performance, outperforming the existing models. In addition, the integration of SHAP values added an innovative layer to the proposed model. The unique feature extraction and fusion strategies enabled the proposed model to offer an outstanding performance.

V. DISCUSSIONS

The proposed study offers transformative implications for clinical practice and academic research. It establishes a new benchmark through the integration of statistical, handcrafted acoustic, and learned features with a fine-tuned XGBoost classification technique. Unlike traditional feature extraction techniques, it extracted localized features and global temporal dependencies from Mel-Spectrograms, contributing to the improved model's generalization. The proposed model identified the subtle SD patterns using the extracted features. In existing models, there is a lack of interpretability, a significant barrier to implementing these models in the healthcare sector. By including SHAP values, the proposed model provided an additional layer of interpretability. The SHAP values offer an in-depth knowledge on features, such as jitter, shimmer, and Mel-spectrogram patterns, contributing to the model's decision. SHAP values illustrate the model's decision-making process by presenting predictions into feature contributions. For instance, it can identify role of features in differentiating normal and abnormal speech signals. The proposed model

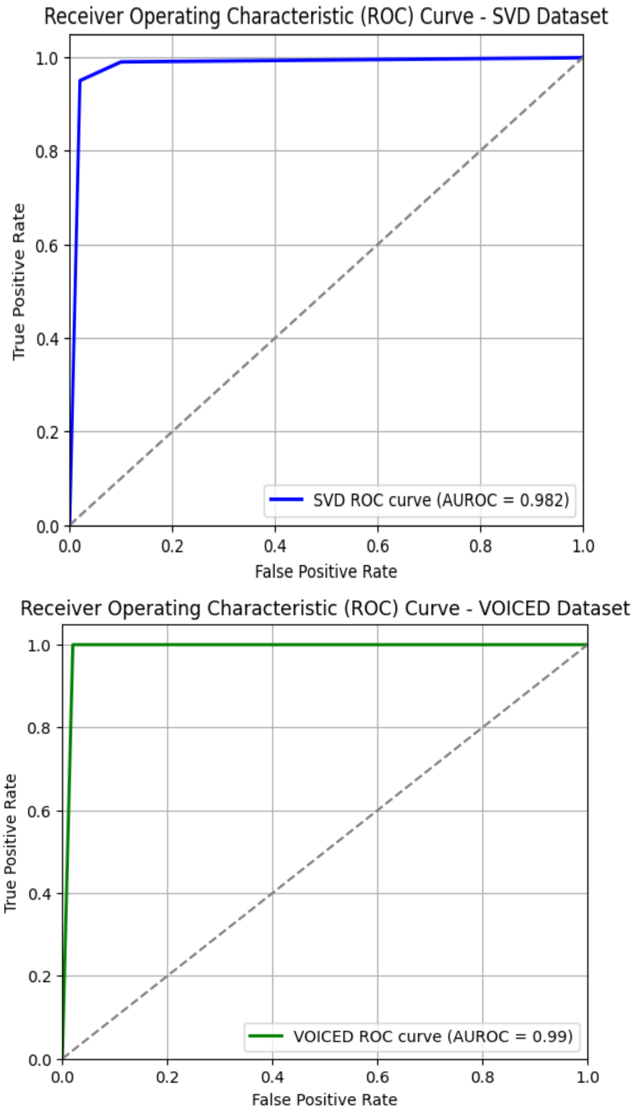

FIGURE 5. AUROC –SVD and VOICED datasets.

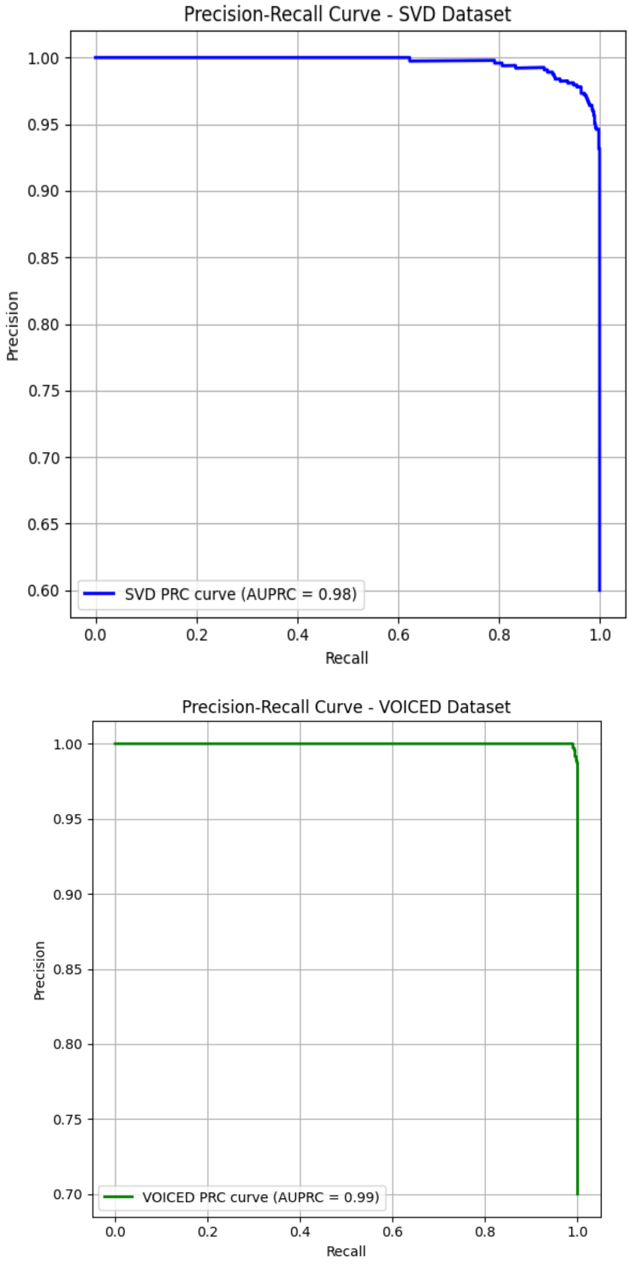

FIGURE 6. AUPRC – SVD and VOICED datasets.

TABLE 8. Findings of comparative analysis.

Studies	Feature Extraction	Dataset	Performance
Proposed SD classification model	MobileNetV3-EfficientNet B7 and Linformer-Performer	SVD	Accuracy: 98.1% Precision: 97.8% Recall: 97.3% F1-Score: 97.6% Specificity:97.0% Loss: 0.13 Standard deviation:0.006 FLOPs:13.6 G
Proposed SD classification model	MobileNetV3-EfficientNet B7 and Linformer-Performer	VOICED	Accuracy: 98.2% Precision: 98.8% Recall: 98.1% F1-Score: 98.4% Specificity:97.4% Loss: 0.15 Standard deviation:0.007 FLOPs: 13.6G
Chaiani *et al.* [5]	CNNs and Long-short term memory	SVD	Accuracy: 70.62%
Mohammed *et al.* [23]	ResNet24 based feature extraction	SVD	Precision: 94.9% Recall: 95.9% F1-Score: 93.9%
Mohaghegh *et al.* [24]	Audio spectrogram transformer	NewHandPD [53]	Average Accuracy: 80.0%
Verde *et al.* [25]	Customized CNN with SVM classifier	VOICED	F1-score:90.0%
Ribas *et al.*[27]	Customized CNNs with Random Forest classifier	SVD	Accuracy: 83.8% AUROC: 86.7
Abdullah *et al.* [46]	Feature optimization and K-nearest neighbor	SVD	Accuracy: 95.0% Precision: 98.0% AUROC: 0.90 Loss: 0.12
Ribas *et al.* [47]	Customized CNNs with Random Forest classifier	SVD	Accuracy: 93.9% AUROC: 97.6%
Hegde *et al.* [48]	ViTs and CNNs	SVD	Accuracy: 90.0% Precision: 90.3% Loss: 3.6% F1-score: 83.7%
Hegde *et al.* [49]	ViTs and CNNs	SVD	Average Accuracy: 97.6%
Irshad *et al.* [50]	Pixel's localized and spatial feature extraction	UA Speech [54] +TORGO [55]	Average Accuracy: 97.7%
Rehman *et al.* [51]	Attribute selection process and PCA with SVM	SVD	Average Accuracy: 90.23% Sensitivity: 89.99% Specificity: 89.65% F1-Score: 88.00% Recall: 90.00%

maintain the degree of transparency in order to establish confidence with clinicians and guarantee the decision align with medical knowledge. The model's interpretability have far-reaching implications for the development of assistive technologies. Through the identification of key features, healthcare centers can develop personalized therapy plans. Clinicians can analyze a patient's progress by visualizing the features like shimmer and pitch variability over time.

The lightweight architecture of the proposed SD classification, incorporating MobileNet V3, Linformer, and Performer, guarantees interoperability with resource-constrained devices. With the increasing accessibility, empowering educators and parents to identify SD in the initial stages. Timely speech therapy treatments can produce effective outcomes, supporting individuals to improve their quality of life. The proposed model demonstrated resilience in handling real-world challenges. The recommended feature fusion addressed the shortcomings of the existing models by delivering a remarkable generalization accuracy of 98.2% on the VOICED dataset. Table 4–8, revealed the exceptional performance of the proposed SD classification. The proposed model outperformed the recent models by achieving a remarkable outcome with limited computational capability.

Mohammed et al. [23] trained the residual networks (ResNet24) model to classify the voice samples using Mel-Spectrograms. They achieved precision of 94.9%, recall of 95.9%, and F1-score of 93.9%. However, the increased computational overhead reduced the model's applicability in resource-constrained environment. Mohaghegh et al. [24] use an audio spectrogram transformer. The shortcomings of direct voice classification reduced the model's ability in capturing temporal dependencies. Verde et al. [25] use lightweight CNNs to detect pathological speech. The vanishing gradients with skip connections caused challenges, leading to higher computational overhead and diminishing performance improvements. Ribas et al. [26] introduce an open-source voice disorder detection. The lack of interpretability limited the model's capability. Abdullah et al. [46] use a feature optimization and K-Nearest Neighbor techniques for classifying SD. The lack of feature fusion influenced the model's classification accuracy. Ribas et al. [47] achieve a classification accuracy of 93.9%. Deep neural networks lack the potential to generate temporal dependencies, causing challenges in capturing patterns over time. The limited memory mechanisms reduced the model's ability in differentiating subtle variations in speech patterns. Hegde et al. [48] optimize ViTs and CNNs to detect vocal cord paralysis. They achieved an average accuracy of 90.0% using the ViTs architecture. The lack of effective feature fusion strategies affected the model's performance. Hegde et al. [49] use ViTs architecture for detecting abnormalities in speech signals. Similarly, Irshad et al. [50] use pixel's localized and spatial feature extraction with ViTs architecture. The absence of model's interpretability and extensive computational overhead reduced the trustworthiness and applicability of the models. Rehman et al. [51] propose a feature selection technique using attribute selection process. SVM is used to classify the extracted features. The shortcomings of SVM impact the model's performance. Chaiani et al. [5] introduce a hybrid SD classifier using CNNs-LSTM. The one dimensional CNNs lacks in identifying subtle voice patterns, affecting the overall performance of the model. In contrast, the proposed model identified the SD patterns using its diverse feature extraction approaches.

The existing studies either focus on Mel-Spectrogram-based CNN classification or utilize acoustic features, lacking the representational depth required to capture the complex phonatory variations in pathological speech. In contrast, the proposed study introduces a hybrid feature fusion pipeline integrating Mel-Spectrograms, statistical acoustic features, and deep feature embedding. The major innovation of this study lies in its multi-dimensional feature integration and architectural synergy, allowing it to selectively emphasize salient feature across domains. This is evident in the model's performance, surpassing traditional and recent baseline models. The incorporation of SHAP values enhance the interpretability of the SD classification. The SHAP values help break down model decisions by highlighting the contribution score to each feature. It offers a clear rationale for classifying healthy and pathological speech signals. It reduces unnecessary computations through the identification of redundant or non-contributing factors. For instance, the SHAP values, including Harmonic-to-Noise ratio (HNR) = +0.47, jitter =-0.05, and spectral tilt = +0.33, indicates the healthy speech with high confidence. Similarly, jitter = +0.62, and formant F2 shift = +0.35, underscore the influence of positive jitter and negative HNR in classifying pathological speech. By explaining the feature contribution towards a specific classification, the SHAP values enhance model's transparency. This ensures that the proposed model aligns with clinical speech pathology insights, making the proposed model interpretable and clinically reliable for real-time applications. The proposed SD classification delivered remarkable outcomes, contributing a significant approach for classifying speech signals. However, it has few limitations. The preprocessing steps, including noise removal and segmentation is essential to maintain consistency in distinguishing healthy and pathological speech. Irrelevant noise or missing critical speech segments may degrade the quality of Mel-Spectrograms. The manual training and domain expertise are essential to tune the optimal parameters for generating Mel-Spectrograms. The hybrid feature extraction approaches may introduce significant computational resources in real-time settings. The attention-based fusion mechanism may overemphasize certain features while neglecting subtle patterns. However, dynamic weighting feature is used to overcome this limitation. The integration of SHAP values may cause challenges in deploying the model in low-resource settings. The feature engineering process can be improved using the ensemble and extreme gradient boosting techniques. The proposed model can be extended to sperm morphology analysis and organ sound identification using advanced DL techniques.

By integrating video (lip movements) or text (transcription), the model could provide a comprehensive understanding of SD. The use of multimodal techniques can capture pauses and patterns of articulation, assisting the model to provide a valuable outcome. Mel-Spectrogram generation can be automated with adaptive techniques. The integration of unsupervised and self-supervised learning methods can reduce the extensive preprocessing tasks. Exploring lightweight ViTs architecture can minimize the computational overhead. The multi-modal attention mechanisms could enhance the feature fusion strategies in integrating diverse features. Expanding the model to support multi-class SD classification could significantly enhance its clinical utility.

VI. CONCLUSION

The proposed study offers a robust and innovative SD classification model, achieving state-of-the-art performance on SVD and VOICED datasets. The proposed model addressed the complexities in classifying SD using an image-based classification approach, integrating advanced DL architectures. It combines statistical and handcrafted acoustic features with learned features extracted through MobileNet V3-EfficientNet B7 and Linformer- Performer architectures. The recommended attention-based feature fusion prioritized the critical features associated with pathological and healthy speech signals. The use of BOHB optimized the XGBoost classification, enhancing the model's generalizability across diverse datasets. With an exceptional accuracy of 98.1% on the SVD dataset and 98.2% on the VOICED dataset, the proposed model underscores its potential for real-world applications in healthcare and assistive technologies. The model's robustness was underscored through remarkable accuracy and low computational loss. However, certain limitations may affect the model's performance in real-time settings. The variations in recording conditions and speech accents may influence the Mel-Spectrogram generation. The model may demand substantial optimization in the resource-constrained environment. Incorporating multi-class classification technique may be beneficial to expand the proposed model for multiple SD classification. The real-time feedback and progress tracking can be integrated with the proposed model to facilitate interactive speech therapy tools. The development of multi-lingual datasets could enable the model deployment across the globe, supporting diverse languages and dialects.

REFERENCES

[1] Q. Yang, X. Li, X. Ding, F. Xu, and Z. Ling, "Deep learning-based speech analysis for Alzheimer's disease detection: A literature review," *Alzheimer's Res. Therapy*, vol. 14, no. 1, p. 186, Dec. 2022.

[2] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[3] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, Dec. 2018.

[4] A. A. Anthony, C. M. Patil, and J. Basavaiah, "A review on speech disorders and processing of disordered speech," *Wireless Pers. Commun.*, vol. 126, no. 2, pp. 1621–1631, Sep. 2022.

[5] M. Chaiani, S. A. Selouani, M. Boudraa, and M. Sidi Yakoub, "Voice disorder classification using speech enhancement and deep learning models," *Biocybernetics Biomed. Eng.*, vol. 42, no. 2, pp. 463–480, Apr. 2022.

[6] S. Yadav and D. Yadav, "Dysarthria voice disorder detection using mel frequency logarithmic spectrogram and deep convolution neural network," in *Proc. 1st Int. Conf. Electron., Commun. Signal Process. (ICECSP)*, Aug. 2024, pp. 1–6.

[7] M. Lesnichaia, V. Mikhailava, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin, "Classification of accented English using CNN model trained on amplitude mel-spectrograms," in *Proc. INTERSPEECH*, Sep. 2022, pp. 3669–3673.

[8] S. Mehra, V. Ranga, and R. Agarwal, "A deep learning approach to dysarthric utterance classification with BiLSTM-GRU, speech cue filtering, and log mel spectrograms," *J. Supercomput.*, vol. 80, no. 10, pp. 14520–14547, Jul. 2024.

[9] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Appl. Sci.*, vol. 13, no. 9, p. 5521, Apr. 2023.

[10] H.-Y. Lai, C.-C. Hu, C.-H. Wen, J.-X. Wu, N.-S. Pai, C.-Y. Yeh, and C.-H. Lin, "Mel-scale frequency extraction and classification of dialect-speech signals with 1D CNN based classifier for gender and region recognition," *IEEE Access*, vol. 12, pp. 102962–102976, 2024.

[11] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Aug. 2021, pp. 12116–12128.

[12] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10211–10221.

[13] S. Kim, J. Nam, and B. C. Ko, "ViT-Net: Interpretable vision transformers with neural tree decoder," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 11162–11172.

[14] D. Paul, A. Chowdhury, X. Xiong, F.-J. Chang, D. Carlyn, S. Stevens, K. L. Provost, A. Karpatne, B. Carstens, D. Rubenstein, C. Stewart, T. Berger-Wolf, Y. Su, and W.-L. Chao, "A simple interpretable transformer for fine-grained image classification and analysis," 2023, *arXiv:2311.04157*.

[15] M. Rigotti, C. Miksovic, I. Giurgiu, T. Gschwind, and P. Scotton, "Attention-based interpretability with concept transformers," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–16.

[16] J. B. Lee and H. G. Lee, "Quantitative analysis of automatic voice disorder detection studies for hybrid feature and classifier selection," *Biomed. Signal Process. Control*, vol. 91, May 2024, Art. no. 106014.

[17] W. Lambamo, R. Srinivasagan, and W. Jifara, "Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition," *Appl. Sci.*, vol. 13, no. 1, p. 569, Dec. 2022.

[18] K. Basak, N. Mishra, and H.-T. Chang, "TranStutter: A convolution-free transformer-based deep learning method to classify stuttered speech using 2D mel-spectrogram visualization and attention-based feature representation," *Sensors*, vol. 23, no. 19, p. 8033, Sep. 2023.

[19] L. Sheng, D.-Y. Huang, and E. N. Pavlovskiy, "High-quality speech synthesis using super-resolution mel-spectrogram," 2019, *arXiv:1912.01167*.

[20] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1147–1157, 2022.

[21] M. Shahin, U. Zafar, and B. Ahmed, "The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 400–412, Feb. 2020.

[22] I. Sindhu and M. S. Sainin, "Automatic speech and voice disorder detection using deep learning—A systematic literature review," *IEEE Access*, vol. 12, pp. 49667–49681, 2024.

[23] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. K. Abd Ghani, M. S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, and F. T. Al-Dhief, "Voice pathology detection and classification using convolutional neural network model," *Appl. Sci.*, vol. 10, no. 11, p. 3723, May 2020.

[24] M. Mohaghegh and J. Gascon, "Identifying Parkinson's disease using multimodal approach and deep learning," in *Proc. 6th Int. Conf. Innov. Technol. Intell. Syst. Ind. Appl. (CITISIA)*, Nov. 2021, pp. 1–6.

[25] L. Verde, N. Brancati, G. De Pietro, M. Frucci, and G. Sannino, "A deep learning approach for voice disorder detection for smart connected living environments," *ACM Trans. Internet Technol.*, vol. 22, no. 1, pp. 1–16, Feb. 2022.

[26] N. Q. Abdulmajeed, B. Al-Khateeb, and M. A. Mohammed, "Voice pathology identification system using a deep learning approach based on unique feature selection sets," *Expert Syst.*, vol. 42, no. 1, p. 13327, Jan. 2025.

[27] D. Ribas, M. A. P. Yoldi, A. Miguel, D. Martínez, A. Ortega, and E. Lleida, "S3prl-disorder: Open-source voice disorder detection system based in the framework of S3PRL-toolkit," in *Proc. IberSPEECH*, Nov. 2022, pp. 136–140.

[28] D. Park, Y. Yu, D. Katabi, and H. K. Kim, "Adversarial continual learning to transfer self-supervised speech representations for voice pathology detection," *IEEE Signal Process. Lett.*, vol. 30, pp. 932–936, 2023.

[29] D. Hemmerling, M. Wodzinski, J. R. Orozco-Arroyave, D. Sztaho, M. Daniol, P. Jemiolo, and M. Wojcik-Pedziwiatr, "Vision transformer for Parkinson's disease classification using multilingual sustained vowel recordings," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 1–4.

[30] R. Islam and M. Tarique, "Spectrogram and mel-spectrogram based dysphonic voice detection using convolutional neural network," in *Proc. Int. Conf. Electr., Comput. Energy Technol. (ICECET*, Jul. 2024, pp. 1–5.

[31] S. Rubio Felipo, D. Ribas González, E. Lleida Solano, A. Ortega Giménez, and A. M. Artiaga, "Assessing the impact and potential of TTS for pathological voice data augmentation on pathology detection systems," in *Proc. IberSPEECH*, Nov. 2024, pp. 41–45.

[32] H. S. Lau, M. Huntly, N. Morgan, A. Iyenoma, B. Zeng, and T. Bashford, "Interpreting pretrained speech models for automatic speech assessment of voice disorders," in *Proc. Int. Conf. AI Healthcare*. Cham, Switzerland: Springer, Jan. 2024, pp. 59–72.

[33] M. Sayadi, V. Varadarajan, M. Langarizadeh, G. Bayazian, and F. Torabinezhad, "A systematic review on machine learning techniques for early detection of mental, neurological and laryngeal disorders using patient's speech," *Electronics*, vol. 11, no. 24, p. 4235, Dec. 2022.

[34] N. Memari, S. Abdollahi, S. Khodabakhsh, S. Rezaei, and M. Moghbel, "Speech analysis with deep learning to determine speech therapy for learning difficulties," in *Proc. INFUS Conf. Intell. Fuzzy Techn., Smart Innov. Solutions*, Istanbul, Turkey. Cham, Switzerland: Springer, Jul. 2021, pp. 1164–1171.

[35] S. Bindas and E. Onuiri, "A deep learning approach to speech recognition for detection of mental disorders," *CURRENT TRENDS Inf. Commun. Technol. Res.*, vol. 2, no. 1, pp. 28–46, 2023.

[36] B. Koonce, *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, 1st ed., Berkeley, CA, USA: Apress, 2021, doi: 10.1007/978-1-4842-6168-2.

[37] V. Krishna Kancharla and P. Kumar Kaveti, "Exploring self-supervised learning with U-Net masked autoencoders and EfficientNet B7 for improved classification," 2024, *arXiv:2410.19899*.

[38] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," 2020, *arXiv:2006.04768*.

[39] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller, "Rethinking attention with performers," 2020, *arXiv:2009.14794*.

[40] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[41] *SVD Dataset*. Accessed: May 7, 2024. [Online]. Available: https://www.spsc.tugraz.at/databases-and-tools/saarbruecken-voice-database.html

[42] *VOICED Dataset*. Accessed: May 5, 2024. [Online]. Available: https://physionet.org/content/vvoiced/1.0.0/

[43] K. E. Smith and A. O. Smith, "A spectral enabled GAN for time series data generation," 2021, *arXiv:2103.01904*.

[44] R. Bey, R. Goussault, F. Grolleau, M. Benchoufi, and R. Porcher, "Fold-stratified cross-validation for unbiased and privacy-preserving federated learning," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 8, pp. 1244–1251, Aug. 2020.

[45] U. Imran, A. Waris, M. Nayab, and U. Shafiq, "Examining the impact of different k values on the performance of multiple algorithms in K-Fold cross-validation," in *Proc. 3rd Int. Conf. Digit. Futures Transformative Technol. (ICoDT2)*, Oct. 2023, pp. 1–4.

[46] S. M. Abdullah, T. Abbas, M. H. Bashir, I. A. Khaja, M. Ahmad, N. F. Soliman, and W. El-Shafai, "Deep transfer learning based Parkinson's disease detection using optimized feature selection," *IEEE Access*, vol. 11, pp. 3511–3524, 2023.

[47] D. Ribas, M. A. Pastor, A. Miguel, D. Martínez, A. Ortega, and E. Lleida, "Automatic voice disorder detection using self-supervised representations," *IEEE Access*, vol. 11, pp. 14915–14927, 2023.

[48] K. Jayashree Hegde, K. Manjula Shenoy, and K. Devaraja, "Performance evaluation of pre-trained models for classification of vocal cord paralysis over vowels," in *Proc. 2nd Int. Conf. Netw., Multimedia Inf. Technol. (NMITCON)*, Aug. 2024, pp. 1–7.

[49] J. Hegde, M. Shenoy, and K. Devaraja, "Analysis of frequencies and pitches for vocal cord paralysis classification through transfer learning," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2024, pp. 1–6.

[50] U. Irshad, R. Mahum, I. Ganiyu, F. S. Butt, L. Hidri, T. G. Ali, and A. M. El-Sherbeeny, "UTran-DSR: A novel transformer-based model using feature enhancement for dysarthric speech recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2024, no. 1, p. 54, Oct. 2024.

[51] M. Ur Rehman, A. Shafique, Q.-U.-A. Azhar, S. S. Jamal, Y. Gheraibia, and A. B. Usman, "Voice disorder detection using machine learning algorithms: An application in speech and language pathology," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108047.

[52] *NewHandPD Dataset*. Accessed: 2024. [Online]. Available: https://www.kaggle.com/datasets/claytonteybauru/spiral-handpd

[53] *UA Speech Dataset*. Accessed: 2024. [Online]. Available: https://ieee-dataport.org/documents/uaspeech

[54] *Torgo Dataset*. Accessed: 2024. [Online]. Available: https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html

ABDUL RAHAMAN WAHAB SAIT (Member, IEEE) received the Ph.D. degree from Alagappa University, India, in 2017. He is currently an Assistant Professor with the Department of Archives and Communication, King Faisal University, Saudi Arabia. He has published research articles in the reputed journals and conferences. His research interests include machine learning, artificial intelligence, web mining, data mining, and deep learning techniques.

SURESH SANKARANARAYANAN (Senior Member, IEEE) is currently a Full Professor in computer science with the College of Computer Sciences and Information Technology, King Faisal University, Al Hofuf, Saudi Arabia. He has authored more than 100 international publications in refereed journals and conferences. He is a holder of five Indian patents and one U.S. patent to his credit in the field of the IoT, edge/fog computing, and artificial intelligence. His current Google Scholar citation is 2408 with an H-index of 23 and Scopus citation of 1252 with an H-index of 16. He has graduated five Ph.D.'s and more than 30 graduate research thesis and projects in the area of wireless sensors, the IoT, fog computing, intelligent agents, and machine learning. He has also been involved in lot of international collaborative project pertaining to the IoT, fog computing, and healthcare. His current research interests include the IoT, wireless sensor networks, QoS, edge computing, and machine learning. He is a reviewer and a technical committee member of a number of IEEE conferences and journals.

P. GOUTHAMAN received the master's degree in computer and information systems from Federation University Australia, Australia, in 2010, and the Ph.D. degree in computer science from the SRM Institute of Science and Technology, Chennai, in 2022. He has professional experience working as a Network Support Engineer and an IT Analyst in industries in Australia and India. He has been passionate teaching professional with the SRM Institute of Science and Technology, since 2015. His research interests include the Internet of Things, software engineering, and project management, where he has published articles in SCIE journals and other internal conferences like IEEE and Springer.