



SpecAug-GAN-Based Synthetic Speech Augmentation for Voice Disorder Classification: Evaluation of Machine Learning Models

Kavitha Krishnan^{1*} , Rajkumar Nalliah² 

¹ Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore 641014, India

² Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai 625009, India

Corresponding Author Email: kavi.skcet@gmail.com

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430411>

ABSTRACT

Received: 1 June 2026

Revised: 27 July 2026

Accepted: 5 August 2026

Available online: 31 August 2026

Keywords:

generative adversarial networks, machine learning, pathological voice detection, SpecAug-GAN, synthetic speech augmentation, voice disorder classification

To enable timely diagnosis and effective intervention, abnormal speech needs to be characterized precisely. The purpose of this research is to investigate the influence of different synthetic data generation (SDG) approaches on deep learning (DL)-based voice abnormality classification. Standard evaluation criteria, such as accuracy, sensitivity, specificity, F1-score, and area under the curve (AUC), are used to evaluate the model performance. A comparative evaluation of multiple machine learning (ML) models, including support vector machines (SVM), Multi-Layer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), Categorical Boosting model (CatBoost), and Multi-Level Classification Framework (MLCF), is performed to assess their effectiveness in classification tasks. This research investigates numerous SDG approaches, such as variational autoencoders (VAEs), long short-term memory (LSTM) networks, Generative Adversarial Networks (GANs), and Spectrogram Augmentation-based Generative Adversarial Network (SpecAug-GAN), to determine their potential to improve model performance. The findings show that MLCF performs better than the other classifiers, with the greatest accuracy (0.89) and F1-score (0.88). SpecAug-GAN showed notably higher accuracy and F1-score metrics compared to the other SDG methods assessed under the experimental conditions. The Saarbrücken Voice Database (SVD), including 1037 speech samples for the glottal dataset, was considered. The results obtained highlight the potential classification capabilities under the considered experimental conditions. More validation on larger and clinically more diverse datasets is needed prior to the comprehensive assessment of its prospective clinical use.

1. INTRODUCTION

Millions of people around the world struggle to communicate or to live their lives normally each day due to abnormalities of their voice. The rapid and precise detection of aberrant speech is critical in efficiently diagnosing and coordinating therapy for the patient. Conventional diagnostic approaches are susceptible to bias because they depend on speech-language pathologists' subjective evaluations throughout the diagnosis process. In recent years, several breakthroughs have been achieved in machine learning (ML) and deep learning (DL), making automated speech disorder classification a potentially useful tool to enhance the precision and efficiency of the classification process [1].

However, the lack of quality labelled datasets is a common limitation for classifiers of voice disorders, such as ML and DL models. The difficulties in collecting a large clinical voice disorder database have made SDG techniques increasingly popular to provide additional training data and improve the generalizability of the model. Spectrogram Augmentation-based Generative Adversarial Network (SpecAug-GAN), variational autoencoders (VAEs), long short-term memory

(LSTM) networks, and Generative Adversarial Networks (GANs) are some of the most successful methods for producing genuine voice representations for classification model training. These methods are used to create accurate speech representations. The purpose of this study is to look at how synthetic data production affects the effectiveness of several ML and DL models, such as support vector machines (SVM), Multi-Layer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), Categorical Boosting model (CatBoost), and Multi-Level Classification Framework (MLCF). The research evaluated these models using common performance criteria including accuracy, precision, recall, F1-score, and area under the curve (AUC).

Current synthetic data augmentation techniques do not maintain pathological speech characteristics, thus compromising the quality of the data samples and the generalization of the models. The proposed SpecAug-GAN framework is used for synthetic speech generation to overcome the above limitations, yielding more representative synthetic speech for experimental voice disorder classification.

Over the years, the most common route to diagnosis of voice

disorders has been through perception testing by speech-language pathologists and by voice specialists. While these methods are very beneficial, they are also very individualistic [1]. This means that there may be variations in outcomes and bias in therapy. With recent advances in AI, specifically ML and DL, it is now possible to analyse speech automatically in a totally new way that was previously imaginable. These technologies can be used to render diagnostic procedures objective, consistent, and very energy-efficient. The devices can even decipher speech communications that contain complex acoustic patterns that are inaudible to the human ear. Automated systems can help medical workers, make screening more inclusive for others, and improve the accuracy of diagnostic processes through the use of algorithms that can accurately distinguish whether voices are healthy or ill. Initial studies have yielded significant results, and some models have proven competent in voice disorder detection.

This study examines the effectiveness of synthetic data in improving the classification of vocal abnormalities. The research was conducted to examine this inquiry. This paper presents a comprehensive analysis of several data generation techniques, including VAE, LSTM networks, traditional GANs, and a unique diffusion-based approach known as SpecAug-GAN. Voice disorder classification was performed using a set of robust machine learning (ML) classifiers, including SVM, MLP, XGBoost, CatBoost, and the proposed MLCF, to assess the efficacy of these synthetic data models. Existing diffusion-based speech generation methods have limited capability to preserve pathological speech characteristics. The proposed SpecAug-GAN enhances data diversity and representation and addresses class imbalance in experimental voice disorder classification.

Medical speech generation remains a challenge due to a small number of datasets, class imbalance, and pathological speech variation. The purpose of this study is to determine if SpecAug-GAN can improve the quality of the synthetic samples and improve the classification performance under the experimental conditions.

There are three main contributions to this paper:

- In this paper, the authors extensively review the application of DL approaches for generating synthetic speech for voice disorder detection.
- Research comparing ML classifiers for multiple vocal disorders diagnosis.
- Empirical results show that SpecAug-GAN significantly enhances the classification accuracy.

Proposed work aims to address the main limitation of the dataset availability of automated speech disorder diagnostic systems, to make them more reliable and accurate. These findings underscore the critical role advanced data augmentation plays in constructing precise and dependable classification models, which can lead to the creation of user-friendly and effective tools for clinical diagnosis and patient care. To conclude, SpecAug-GAN showed better results, and the results show that using synthetic data has a great positive impact on classification accuracy. The results of this research highlight the need to use more sophisticated data augmentation methods to create accurate voice disorder classification algorithms. The enhancement of automated speech issue detection systems' reliability and the mitigation of dataset limitations are achieved via the use of synthetic data. Proposes an experimental assessment of the proposed framework based on the publicly accessible Saarbrücken Voice Database (SVD) repository of experimental data. The goal is to explore the

ability of synthetic data augmentation for the voice disorder classification task and not to create a clinically validated diagnostic system.

2. LITERATURE REVIEW

Voice disorders are significant health problems that necessitate advanced diagnostic methods. Recent years have seen an examination of ML and DL methodologies to enhance the accuracy and efficiency of vocal pathology classification. The aim of this literature review is to identify and discuss significant literature within the subject, concentrating on methods and data, as well as the main results.

This study by Rehman et al. [2] shows that the SVM and PCA (SVM-PCA) feature selection method yields an accuracy of 99.97% in voice disorder detection. However, their work does point to the limitations of the datasets and the effect that different acoustic transducers have on categorisation ability. The authors believe that the hybrid classifiers and DL models should be combined with ensemble classifiers, fuzzy logic, and novel features like the Enhanced Teager Energy Operator (ETEO) to improve the performance of mobile health applications.

In the context of Smartphone-Based Voice Pathology Detection, Di Cesare et al. [3] explored edge computing for smart healthcare by examining the feasibility of using it to analyse vocal samples recorded by smartphones from the publicly available VOICED dataset for voice pathology detection. They use Mel Frequency Cepstral Coefficients (MFCCs) to extract features, and the optimized k-Nearest Neighbour (KNN) is used to test various classifiers and achieves an accuracy of 98.3%. The key strength of their work is the ability of the smartphone to be used as an early detection tool, a continuous monitoring tool, and an accessible tool in healthcare.

In the context of cross database evaluation for model generalization, Latiff et al. [4] characterised cross database assessment for model generalisation as a technique to enhance model resilience using cross-database assessment methodologies. Subsequently, they examined ML methodologies, specifically Online Sequential Extreme Learning Machine (OSLM), SVM, Decision Tree (DT), and Naïve Bayes (NB), utilising two datasets: SVD and Malaysian Voice Pathology Database (MVPD). The data indicate that OSLM may attain the highest level of accuracy, underscoring its use in many practical applications.

Chaiani et al. [5] investigated the importance of feature selection in monaural speech segregation and the challenges associated with using a single feature approach. They tested eight speech characteristics with a Deep Neural Network (DNN) and found that adding feature clusters significantly enhances speech segregation. They are doing research that helps with voice augmentation, noise reduction, and speech recognition in conditions with lots of noise.

Ding et al. [6] conducted a comprehensive study on the use of AI-driven speech analysis for the diagnosis of Alzheimer's disease (AD). It presents the AI methodologies and benchmark datasets that were used for the research, such as ADReSS and ADReSSo, highlighting the importance of DL models, particularly those that combine auditory and linguistic features. Despite achieving high accuracy rates, challenges like limited data availability and model interpretability remain substantial difficulties.

In the context of voice pathology detection using IoT, bin Sham et al. [7] proposed a voice pathology detection system based on IoT using MFCC and SVM. The proposed web-based system provides for remote diagnostics in real time based on vowel /a/ samples of the German SVD. This study sheds light on the potential of IoT for diagnosis of intelligent voice disorders.

In studies on data augmentation in predictive modelling, Yao et al. [8] studied how well Tabular Generative Adversarial Networks (TGAN) can be used in data augmentation for predictive modelling. They show that models trained on augmented data outperform the models trained on the original data, and that neural networks have an accuracy improvement of 15.2%.

Investigates the development of alternative ML systems using vocal data recorded by cell phones to differentiate pathological and normal voices. The research used the publicly available VOICED dataset containing 58 healthy voice samples and 150 samples of voices with pathological conditions, and ML methods for the classification of healthy patients and patients with pathological voice from the employment of Mel-frequency cepstral coefficients [9].

For time series classification for Parkinson's disease detection, Rey-Paredes et al. [10] presented a time series classification approach for Parkinson's disease detection. For data augmentation, they used the Big Vocoder Slicing Adversarial Network (BigVSAN), thereby improving DL models, including LSTM-FCN. Their results highlight how synthetic data could help to enhance diagnosis performance.

Moëll et al. [11] carried out a systematic overview of the ML approaches for speech-based diagnosis of neurological disorders, analysing 91 studies. They found that models of speech categorization work well for detecting neurological diseases, with high accuracy for Parkinson's disease, laryngeal issues, and dysarthria. However, there is a need for further substantiation for other issues like autism and OCD. A literature study comparison of different datasets, techniques, and accuracy is done.

A general overview of recent studies of voice disorder classification is presented in Table 1. The comparison of the works reported is not intended as a performance benchmarking, as the different works are based on different data, preprocessing, and evaluation methods.

Table 1. Literature review on voice disorder classification

Author	Year	Dataset	Technique	Accuracy
Rehman et al. [2]	2024	Synthetic Dataset	SVM with PCA-based feature selection	99.97%
Di Cesare et al. [3]	2024	VOICED Dataset (58 healthy, 150 pathological)	KNN with MFCC features	98.3%
Latiff et al. [4]	2025	Saarbrücken Voice Database (SVD) and Mel Frequency Cepstral Coefficient	OSELM, SVM, DT, NB with cross-dataset evaluation	85.71% (SVD) / 80.77% (MVPD)
Chaiani et al. [5]	2023	Glottal dataset	DNN for feature selection	91.00%
Ding et al. [6]	2024	ADReSS and ADReSSo	BERT-based deep learning (DL) models with acoustic and linguistic features	85%
bin Sham et al. [7]	2023	Saarbrücken Voice Database (SVD)	IoT-based voice pathology detection using SVM and MFCC	89.82%
Yao et al. [8]	2024	Geopolymer Concrete Dataset (930 samples)	TGAN for data augmentation with ML models	+15.2% (NN), +9.8% (Tree Models), +7.5% (SVM)
Parasyris et al. [9]	2023	Synthetic Seismic Dataset	PIX2PIX GAN with ResNet-9 for velocity model reconstruction	SSIM: 0.92, MAE: 0.025 km/s
Rey-Paredes et al. [10]	2025	PC-GITA (Sustained vowel /a/)	BigVSAN GAN for synthetic data + CNN-based classifiers	+15.87% (CDIL-CNN vs. Baseline)
Moëll et al. [11]	2025	91 Studies (Meta-Analysis)	ML methods for speech-based neurological disorder diagnosis	High for PD, Laryngeal Disorders, Dysarthria

Note: SVM = Support Vector Machines; KNN = k-Nearest Neighbour; MFCC = Mel Frequency Cepstral Coefficient; DT = Decision Tree; NB = Naïve Bayes; MVPD = Malaysian Voice Pathology Database; OSELM = Online Sequential Extreme Learning Machine; GAN = Generative Adversarial Networks.

The reviewed literature has pointed out the revolutionary role of ML and DL in the classification of voice disorders. While feature selection, data augmentation, and model optimization have improved the accuracy, problems of data accessibility, generalizability, and model interpretability persist. There is a need to focus on cross database evaluations, hybrid classifiers, and multimodal feature fusion to enhance identification of speech pathology in future studies.

3. MATERIALS AND METHODS

The first step in the process is to load the data and prepare it for analysis, then extract the essential features that capture the patterns in the data. Synthetic samples are created to tackle the data imbalance and enhance the robustness of the models. Then, various models are trained on the improved dataset, and their performance is meticulously assessed on the basis of

accuracy and privacy metrics. Lastly, results are interpreted and conclusions drawn, which lead to good decisions and sound judgment.

3.1 Dataset description

The SVD Glottal database is a benchmark database publicly available for speech processing, voice pathology analysis, and computer-aided diagnosis of vocal disorders research. Standardised digital speech recordings from healthy subjects and patients with different laryngeal disorders were recorded under controlled conditions and stored in the database. The audio recordings are used for the extraction of acoustic, spectral, cepstral, and glottal features to be used in automatic voice disorder detection and classification. A subset of the SVD, corresponding to speech data from the Glottal class, was used in this study, containing 1037 speech recordings: 560 Healthy, 392 Vox Senilis, and 85 Laryngoceles. To prepare the

recordings for the proposed framework, SpecAug-GAN, noise reduction, removal of silence, resampling, and normalization were performed.

Data Splitting Methodology: In order to obtain reliable

evaluation results, prevent data leakage, and improve the reliability and reproducibility of the devised framework, the SVD was split into independent sets for training, validation and testing purposes.

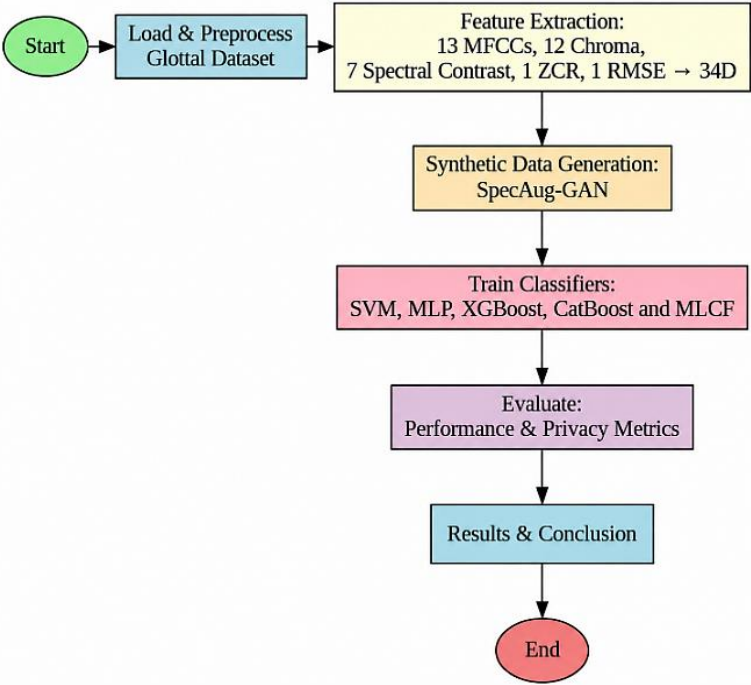


Figure 1. Overall framework of the proposed Spectrogram Augmentation-based Generative Adversarial Network (SpecAug-GAN) voice disorder classification system

Table 2. Overview of glottal dataset classes

Classes	Disease Type	Count of Files	Description
Laryngozele	Laryngocele	85	Includes vocal recordings from persons diagnosed with laryngocele, defined by the presence of an air-filled sac within the larynx.
Vox Senilis	Vox Senilis	392	Includes speech signals from individuals with Vox Senilis, a disease linked to age-related alterations in vocal quality.
Normal	Healthy	560	Comprises voice signals from healthy persons, functioning as a control group for comparative analysis.

As proposed, the work classification framework is shown in Figure 1. The proposed approach begins with speech preprocessing of signals, followed by conducting SpecAug-GAN-based data augmentation to diversify the training dataset. Both real and synthetic speech samples are used during the feature extraction phase, in which various characteristics such as time-domain, acoustic, pitch, and glottal features are derived. After the feature extraction phase, various features are normalized and combined into a single feature vector, which is given as input to the Dynamic MLCF framework, allowing prediction of whether a patient suffers from a voice disorder or not with help of the developed classifier. The glottal dataset comprises many categories of vocal disorders, as shown in Table 2.

3.2 Data preprocessing

Several vital preprocessing steps are carried out to ensure high-quality data generation of synthetic data for voice disorder classification and improve the quality of the model. Many essential preprocessing processes are applied to ensure

high-quality data generation of the synthetic data generation for the classification of voice disorders and to improve the quality of the model by Korba et al. [12].

Speech preprocessing: Sounds that were taken from the glottal collection of SVD were subjected to extensive processing in order to enhance the quality and achieve uniformity of signals before extracting the features. Firstly, the spectral subtraction technique was implemented to eliminate the background noise and artifacts while keeping the primary acoustic features that are connected with pathological voices intact. Then, the voice activity detection (VAD) was utilized in order to delete soundless and non-speech portions of the speech signal and keep only useful regions of the speech signal for further processing. After that, the speech records underwent a new sampling procedure and were retrieved with the same sampling frequency of sixteen kHz to guarantee overall uniformity. Finally, the users conducted peak and/or RMS normalization in order to prevent the amplitude variations caused by the surroundings of the speakers record in. Hence, the speech preprocessing was done successfully, creating the uniform, high-quality speech signal.

Noise Mitigation: Background noise is reduced by calculating the noise spectrum over non-speech segments and subtracting it from the spoken signal; this is called "Spectral Subtraction. This will increase the signal to noise ratio (SNR) and improve the clarity of the data. The clean speech spectrum $\hat{S}(f)$ is computed as:

$$\hat{S}(f) = |Y(f)| - |N(f)| \quad (1)$$

where, $|Y(f)|$ – noisy speech spectrum and $|N(f)|$ – estimated noise spectrum, f – frequency.

Silence Removal: Extraneous silence is removed by using VAD. This approach uses characteristics like short term energy and zero crossing rates (ZCR) to distinguish between speech and non-speech parts and thus targets speech data.

Resampling: For uniformity, all audio samples are resampled to 16 kHz, using the Resample Method of Librosa. This ensures uniformity and compatibility with DL models.

Normalization: Amplitude levels in all voice samples are standardized using Peak Normalization or Root Mean Square (RMS) Normalization to achieve consistent loudness. RMS normalization is defined as:

$$x' = \frac{x}{RMS(x)} \quad (2)$$

Let's assume x' – normalized speech signal, x – original speech signal, $RMS(x)$ – root mean square value of the speech signal [13].

3.3 Generation of synthetic data

In DL, synthetic data production is vital for generating synthetic data in the event that true data is unavailable, expensive, or poses privacy concerns. In the case of GAN, the data quality is improved through adversarial training, and for VAE, the purpose is to gain latent representations to create a variety of more authentic samples. The LSTM model has proven to be useful in speech processing to handle the sequential relationships, mitigating the disappearing gradients with gating mechanisms. An improved voice synthesis model is created by using SpecAug-GAN, a diffusion-based model, to enhance mel-spectrogram representations. Wang et al. [14] showed that these methods complement DL by addressing the challenges associated with data scarcity and improving model performance in specific applications like medical imaging, speech disorder detection, and network security.

3.3.1 Generative Adversarial Networks

GAN generates synthetic data through a generator (G) that produces samples and a discriminator (D) that distinguishes between genuine and fraudulent data. The adversarial training method incrementally enhances data quality. The discriminator minimizes binary cross-entropy loss.

$$L_D = -\sum_{x \sim p_{data}} [\log D(x)] - \sum_{x \sim p_z} [\log (1 - D(G(z)))] \quad (3)$$

where, L_D – discriminator loss, x – genuine sample from the data distribution p_{data} , and z – random noise sourced from the distribution p_z . $G(z)$ – output generated by the generator, $D(x)$ – likelihood that x is authentic, and $D(G(z))$ – likelihood that $G(z)$ is authentic. The generator

improves its ability to mimic genuine data distributions by reducing.

$$L_G = -\sum_{x \sim p_z} [\log (1 - D(G(z)))] \quad (4)$$

where, L_G – generator loss, p_z – random noise vector drawn, $D(G(z))$ – discriminator output for the generated sample. while GANs are employed in image generation, tabular data augmentation, text and speech synthesis, 0 and anomaly detection. Challenges include mode collapse, training instability, and evaluation metrics. GANs are essential for SDG in contexts with little real data, as ongoing improvements in training and assessment enhance their effectiveness.

Algorithm 1: SpecAug-GAN for synthetic data generation (SDG)

Input: Phoneme sequence y , Speaker embedding s , Diffusion step t

Output: Generated synthetic mel-spectrogram x_0

Steps:

1. Utilize a Transformer encoder to encode the phoneme sequence y $h = \text{TransformerEncoder}(y)$
2. Enhance the phoneme characteristics by forecasting duration, pitch, and energy $h' = \text{VarianceAdaptor}(h)$
3. Initialize the hidden state $H = h'$ and the noisy spectrogram x_t for the diffusion process.
4. For each diffusion step t from T to 1
Introduce Gaussian noise to the spectrogram
 $q(x_t|x_{t-1}) = N(\sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I)$
Estimate the mean μ_θ using the diffusion decoder $\mu_\theta = \text{Decoder}(x_t, y, s, t, H)$
Estimate the clean spectrogram $x_{t-1} = \text{Sample}(x_t, \mu_\theta, \sigma_t)$
Update the hidden state H .
5. Integrate the speaker embedding s into the decoder to produce varied speech samples.
6. Use sinusoidal positional encoding to encode the diffusion step t .
7. Generate the final synthetic mel-spectrogram $x_0 = \text{FinalConv}(x_1)$

Model training involves

MSE Loss between the generated spectrogram x_0 and the ground truth spectrogram. $L_{MSE} = ||x_0 - x_{gt}||^2$

The Kullback-Leibler Divergence Loss is employed to regularize the diffusion process.

$$L_{adv} = \mathbb{E} [\log (D(x_{gt}))] + \mathbb{E} [\log 1 - (D(G(y, s)))]$$

Adjust model parameters by back propagation. $L_{total} = L_{MSE} + \lambda_1 L_{KL} + \lambda_2 L_{adv}$

Use the trained SpecAug-GAN to produce synthetic speech samples for data augmentation.

Algorithm 1 presents the SpecAug-GAN model, the authors utilized the Adam optimization method, implementing a learning rate of 0.0001 with a batch size of 32 over a period of 100 epochs. The diffusion process used in the model consisted of 200 steps, which made it possible to produce high-quality mel-spectrograms. In order to enhance reconstruction quality and latent feature learning, the model optimized both Mean Squared Error (MSE) and Kullback-Leibler divergence losses. The authors generated synthetic speech samples for the

purpose of addressing the class imbalance within the Glottal corpus as well as combining them with the original samples used for classifier training. The implementation choices make the obtained results reproducible and provide consistency in the experimental evaluation process. Let's assume y – phoneme sequence, s – speaker embedding, t – diffusion step, T – total number of diffusion steps, x_t – noisy spectrogram at diffusion step, x_0 – ground-truth clean mel-spectrogram, H – hidden state, $\mu_\theta = \text{Decoder}(x_t, y, s, t, H)$ – estimated mean produced by the diffusion model.

3.3.2 Variational autoencoders

VAE produces synthetic speech data by acquiring a concise latent representation of vocal sounds. The encoder transforms input speech x into a latent space z , whereas the decoder reconstructs speech from z , maintaining acoustic properties. The VAE loss function comprises:

- Reconstruction Loss: Guarantees that the synthesized speech closely resembles authentic samples.
- KL Divergence Loss: regularizes the latent space by reducing the disparity between $q(z|x)$ and $p(z)$.

$$L_{VAE} = \sum_{z \sim p(z|x)} [\log p(x|z)] - D_{KL}(q(z|x)||p(z)) \quad (5)$$

Let's assume D_{KL} – Kullback–Leibler divergence, x – input speech signal, z – latent variable, $q(z|x)$ – learned latent distribution parameterized, $p(z)$ – prior distribution of the latent variable. The VAE produces varied, authentic speech, reflecting differences in voice disorders, and is beneficial for enhancing datasets in voice disorder classification.

3.3.3 Long short-term memory

LSTM networks are crucial in speech processing for modelling sequential dependencies and capturing long-term trends in audio. In contrast to RNNs, LSTMs employ gating mechanisms forget, input, and output gates to control information flow, hence mitigating the vanishing gradient problem. The cell state C_t is updated as:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (6)$$

$$h_t = o_t \odot \tanh(C_t) \quad (7)$$

where, h_t is used to transmit speech information in time, the information can be seamlessly and coherently created. The LSTM is widely used in voice synthesis, recognition and enhancement, ensuring accurate and natural speech modelling Liu et al. [15]. Let's assume $\tilde{C}_t$ – cell state at time step, C_{t-1} – cell state at the previous time step, f_t – forget gate, i_t – input gate, C_t – candidate cell state, O_t – output gate, h_t – hidden state at time step, t – time step, $\odot$ – element-wise multiplication.

3.3.4 Spectrogram Augmentation-based Generative Adversarial Network

SpecAug-GAN is a diffusion-based model for synthetic speech synthesis, improving data augmentation for voice disorder categorization. It comprises transformer encoders, a variance adaptor, and a diffusion decoder.

SpecAug-GAN data augmentation: After completing the speech preprocessing, the standardized recordings were

utilized as input to the SpecAug-GAN framework, enabling the generation of synthetic representations of pathological speech. Initially, preliminary signals obtained from the preprocessing stage were converted to a mel-spectrogram, which retains appropriate spectral and temporal features for pathological voice analysis. Learning from the original distributions of the speech recordings by the generator and constant assessment of its performance by the discriminator, the process continued until the desired representations of synthetic speech were achieved. The generated synthetic samples were used only in the training data, as it was necessary to improve data diversity and solve the class imbalance problem for the under-sampled Laryngocele class in comparison to other classes. Validation and testing subsets of the speech dataset were not influenced by the procedure so as to avoid the data leakage problem. Both types of recordings produced in the project were delivered to the feature extraction stage, ensuring that the proposed specification allows for learning of better models of pathological speech.

The diffusion process systematically enhances the impaired mel-spectrogram (x_t), represented as:

$$p(x_{t-1}|x_t, y, s) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y, s, t), \sigma_t^2 I) \quad (8)$$

Let y represent the phoneme sequence, s denote the speaker embedding, and t indicate the diffusion step index. The model utilizes Deep Speaker embedding for speaker-specific speech production and sinusoidal positional encoding for enhanced time step modelling. The ultimate mel-spectrogram is produced as:

$$x_0 = G(x_t, y, s, t) \quad (9)$$

SpecAug-GAN enhances voice synthesis by guaranteeing natural prosody, speaker conditioning, and superior data augmentation for DL models.

Data Augmentation Strategy: Once the process of dividing data into separate folders was completed, SpecAug-GAN was limited to generating synthetic speech samples for the training dataset only.

The class distributions of the Glottal subset of SVD have been presented in Table 3, showing the results of the data augmentation procedure and synthetic sample generation for the training dataset in order to achieve balanced results.

Table 3. Class distribution before and after Spectrogram Augmentation-based Generative Adversarial Network (SpecAug-GAN) augmentation

Voice Disorder Class	Original Dataset	Training Dataset (Before Augmentation)	Training Dataset (After Augmentation)
Healthy	560	448	448
Vox Senilis	392	314	448
Laryngocele	85	68	448
Total	1,037	830	1,344

3.4 Feature extraction

Following data augmentation, the original and synthetic speech signals were subjected to feature extraction to provide a complete description of the pathological characteristics of the voices. The acoustic features were extracted to quantify the speech signal in terms of spectral characteristics, energy level, and frequency components to detect variations in speech due

to the occurrence of speech disorders. Next, the time-domain characteristics of the waveforms were calculated to describe the features of the signal in terms of energy profile and dynamic fluctuations of the signal during speech. Types of quantifying pitch properties were extracted to gain insights into the F0 and its variations, which provide information about the movements of vocal cords, the state of voice, and the occurrence of disorders concerning voice production. Glottal feature extraction was performed to provide insights regarding the behaviour of vocal folds. Each category of features provides complementary information describing pathological speech. The process of analysing speech in several ways allows obtaining data about both spectral and physiological characteristics of speech.

To ensure that all the features are expressed on a scale that allows for comparison among them, the features that could be extracted have been normalized before proceeding to Integration. Normalization has the task of reducing the influence of variations in the value of features, so that features with larger values will not dominate the process of learning and also improve the stability of the computations used in the process of training the classifier. By changing the range of extracted features, we make it possible to make learning more balanced with respect to different categories of features, and at the same time keeping their discriminative capabilities.

Feature Integration: Integration of Features: After the features were normalized, acoustic, temporal, pitch-related, and glottal features were combined together to create the final feature vector that represents each speech sample. This integration enabled the collection of useful information from various feature types and the use of all of it at the same time in the proposed system because it involved the use of unique properties of different categories of the derived characteristics. Furthermore, this final feature vector can be very useful to make sure that the positive discriminative properties, which are specific to both kinds of speech, healthy and pathological, are preserved while minimizing the redundancy of the

components of this vector. The final integrated feature vector has made it possible to get a compact and valuable representation of each speech record, thereby facilitating proper learning by the classification system. Finally, this integrated feature vector served as the input for the model trained for voice disorder classification.

Feature Processing Strategy: The described framework consists of a proper feature processing strategy so as to implement extracted speech features in the process of voice disorder classification. The process includes preprocessing and augmentation based on SpecAug-GAN. In the process of feature extraction, unique acoustic, temporal, pitch, and glottal features were extracted, normalized, and processed in the form of an integrated feature vector, which is a common feature representation. It was entered as input information into the Dynamic MLCF model. No further feature selection or dimensionality reduction option was used, which allowed utilizing all extracted features in the final feature representation. Hence, it is guaranteed that the entire pathological speech information is preserved without losing potentially distinguishing features related to the minority voice disorder classes. As a result, the proposed framework presents a solution where comprehensive speech information is utilized to enable higher classification accuracy, robustness, and generalization for Healthy, Vox Senilis, and Laryngocele categories.

The feature extraction procedure is essential for preparing audio data for analysis and classification, as it converts raw speech signals into significant numerical representations. The collected features encapsulate diverse acoustic attributes crucial for speech processing de Brito Santos and Gupta [16, 17]. Table 4 shows the acoustic features for the proposed model. Table 5 shows the time-domain features. The voice quality and pitch-related features are illustrated in Table 6. The glottal features were shown in Table 7.

Table 4. Acoustic features

Feature Name	Description	Equation
Mel Frequency Cepstral Coefficients (MFCCs)	Captures short-term spectral envelope characteristics	$MFCC_k = \sum_{n=1}^N X(n) \cdot \cos \left[k \left(n - \frac{1}{2} \right) \frac{\pi}{N} \right]$
Delta and Delta-Delta MFCCs	Represents dynamic changes in speech	$\Delta MFCC_t = \frac{MFCC_{t+1} - MFCC_{t-1}}{2}$
Gammatone Cepstral Coefficients (GTCC)	Simulates auditory sense of glottal vibrations	$GTCC_k = \sum X(n) \cdot g(n)$
Spectral Centroid	Determines the "luminance" of the speech signal	$SC = \frac{\sum fS(f)}{\sum S(f)}$
Spectral Bandwidth	Quantifies the dispersion of spectral frequencies	$SB = \sqrt{\sum (f - SC)^2 S(f)}$
Spectral Contrast	Highlights differences between spectral peaks and valleys	$SC_k = \frac{E_{max} - E_{min}}{E_{max} \mp E_{min}}$
Spectral Roll-off	Determines the frequency boundary for most of the energy distribution	$SR = \min_f (\sum_0^{SR} S(f) \geq 0.85 \sum S(f))$
Mel Spectrogram	Provides a time-frequency representation of the signal	$m(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right)$

Table 5. Time domain features

Feature Name	Description	Equation
Zero-Crossing Rate (ZCR)	Differentiates between voiced and unvoiced portions	$ZCR = \frac{1}{N} \sum_{n=1}^{N-1} 1(x_n \cdot x_{n-1} < 0)$
Root Mean Square Energy (RMSE)	Estimates vocal intensity and breathiness	$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^N x_n^2}$
Entropy of Energy	Assesses unpredictability in energy fluctuations, beneficial for identifying abnormal glottal cycles	$H = - \sum p_i \log p_i$

Table 6. Voice quality and pitch-related features

Feature Name	Description	Equation
Pitch Function	Records fundamental frequency fluctuations, essential for identifying vocal problems	$F_0 = \arg \max_f S(f)$
Harmonic Ratio	Distinguishes between harmonic and noisy elements in speech	$HR = \frac{E_{harmonic}}{E_{Total}}$
Tonnetz (Tonal Centroid Features)	Illustrates harmonic relationships in vocalized speech	$T = \sum_n C_n \cdot \cos(2\pi f n)$
Chroma STFT	Beneficial for monitoring harmonic structures in phonation	$C(f) = \sum S(f) \cdot W(f)$

Table 7. Glottal-specific features

Feature Name	Description	Equation
Linear Predictive Cepstral Coefficients (LPCC)	Models the glottal source and vocal tract dynamics	$LPCC_k = a_k + \sum a_n LPCC_{k-n}$
Deviation of ZCR and Energy	Facilitates the evaluation of phonation stability	$\sigma_{ZCR} = \sqrt{\sum (ZCR - \bar{ZCR})^2}$ $\sigma_E = \sqrt{\sum (E - \bar{E})^2}$
Haar Transform and Fourier Transform	Useful for analysing periodicity in glottal waveforms	$H(n) = x(2n) - x(2n + 1)$ $X(f) = \sum x(n) e^{-j2\pi f n}$

3.5 Machine learning classification

In this research, the SVM, MLP, XGBoost, CatBoost and a Dynamic MLCF were considered efficient classifiers. For SVM was used for high-dimensional data, while MLP was used for intricate non-linear patterns. CatBoost was good at dealing with the categorical features, and XGBoost was excellent with feature interactions via gradient boosting. In the case of Dynamic MLCF, these models were applied in adaptive layers, selecting classification routes based on the level of confidence. This framework enhanced precision and resilience, particularly with heterogeneous and imbalanced information.

3.5.1 Support vector machines

Trijatmiko and Rahman [18] demonstrated that SVM can provide synthetic speech data for the categorization of voice disorders by interpolating between support vectors and utilizing perturbation techniques. Employing the polynomial kernel:

$$K(x, y) = (x \cdot y + c)^d \quad (10)$$

where, (x, y) represent feature vectors, with SVM models encompassing both linear and nonlinear speech fluctuations. Synthetic data is generated by sampling near support vectors, thereby augmenting feature variety in MFCCs, spectral centroid, and chroma. This enhances model generalization and classification precision in CNNs, DNNs, and perceptron-based networks.

3.5.2 Multi-Layer Perceptron

MLP is a neural network with input, hidden, and output layers, enabling it to represent nonlinear functions. Every neuron calculates a weighted aggregation of inputs Peng et al. [19]:

$$z_j^{(l)} = \sum w_{ji}^{(l-1)} a_i^{(l-1)} + b_j^{(l)} \quad (11)$$

This is subsequently processed by an activation function such as $ReLU(f(z) = \max(0, z))$ or $\text{sigmoid}(f(z) =$

$\frac{1}{1+e^{-z}})$ weights are modified through gradient descent.

$$w_{ji}^{(l)} = w_{ji}^{(l)} - \eta \frac{\partial L}{\partial w_{ji}^{(l)}} \quad (12)$$

where, L denotes the loss function, exemplified by Mean Squared Error (MSE): $L = \frac{1}{N} \sum (y_i - \hat{y}_i)^2$ or Cross Entropy Loss for classification: $L = - \sum y_i \log(\hat{y}_i)$.

3.5.3 Extreme Gradient Boosting

XGBoost is an ML algorithm that employs boosting, wherein numerous weak learners are amalgamated through additive training to create a robust predictor. XGBoost is widely used thanks to its efficiency, scalability, and robustness. The model minimizes a regularized objective function that balances prediction accuracy with model complexity (represented by the regularization term), which is given as:

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (13)$$

where, $l(y_i, \hat{y}_i)$ is a loss function (e.g., least squares for regression, log loss for classification), and $\Omega(f_k)$ is a regularization term that prevents overfitting. Gradient boosting is used in XGBoost to reduce residual error through an iterative process, and the sub-sampling technique (column and row sub-sampling) is used to reduce the variation in the prediction. In real-world scenarios, XGBoost proves to be a top choice for classification, regression, and ranking tasks, particularly when dealing with large-scale datasets, missing values, and feature selection [20].

3.5.4 Categorical Boosting model

CatBoost is a gradient boosting decision tree (GBDT) method that is optimized for categorical features. It reduces overfitting with ordered boosting and target statistics (TS) where categorical values are transformed during training.

$$\hat{x}_{ik} \approx E(y|x_i = x_{ik}) \quad (14)$$

Uses random permutations to reduce bias and categorical

transformations as follows:

$$x_{\sigma_{p,k}} = \frac{\sum_{j=1}^{p-1} 1[x_{\sigma_{j,k}}=x_{\sigma_{j,k}}]Y_{\sigma_j} + \beta P}{\sum_{j=1}^{p-1} 1[x_{\sigma_{j,k}}=x_{\sigma_{j,k}}] + \beta} \quad (15)$$

Additionally, ordered boosting improves the predictions by iterative refinement:

$$M_i = M_i + \Delta M \quad (16)$$

where, ΔM represents acquired modifications. CatBoost uses stochastic permutations to improve generalization and reduce the complexity of the model from $O(s \times n^2)$ to $O(s \times n)$ and is therefore an appealing choice for modelling categorical data [21].

3.5.5 Multi-Level Classification Framework

The MLCF is an iterative ensemble learning method, which boosts the classification ability through multiple layers feature fusion. After feature extraction and feature integration, the unified feature vector is fed into the Dynamic MLCF framework, which helps in the classification of voice disorders. At the first layer, a group of ML classifiers, including SVM, MLP, XGBoost, and CatBoost, is used to learn independent discriminative features extracted from the voice signal. Every classifier examines the representation of various aspects of the voice signals, such as acoustic features, time-domain features, pitch features, and glottal features, and outputs individual classification results regarding Healthy, Vox Senilis, and Laryngocele classes. At the second layer, the outputs of each classifier are fused together using a dynamic model fusion process to generate combined outputs. This model fusion technique is very useful as it overcomes the drawbacks of single-model usage. To finish with, at the last layer, the fused classifier predictions are utilized in order to generate the final prediction of the model by choosing one of the three classes. As a result of the usage of different classifiers in the system, the proposed Dynamic MLCF framework improves classification accuracy and reduces errors in minority class types.

The method starts with model prediction $M_i = 0$ for an ordered dataset $\{(X_k, Y_k)\}_{k=1}^n$, and recalculates them iteratively. Residual errors are computed as $r_i = y_i - M_{\sigma(i)-1}(X_i)$, and the model is updated during the boosting rounds $t = 1$ to I as the following model update $\Delta M = \text{LearnModel}[X_i, r_j]; \sigma(j) \leq i]$. Then the predictions are updated in each instance $M_i = M_i + \Delta M$. In the multi-layer fusion phase ($l = 1$ to L), the following feature fusion is performed:

$$M^{(l)} = M^{(l-1)} + \sum_{i=1}^N w_i g_i \quad (17)$$

where, w_i are the weights and g_i are the modification of the features in the instance where the modification is to be applied. Boosting and hierarchical feature fusion are integrated to improve classification accuracy, without compromising computational efficiency, making MLCF suitable for complex learning tasks.

Classification Algorithm:	Dynamic Multi-Layer Classifier Fusion (Dynamic MLCF)
Data	$\{(X_k, Y_k)\}_{k=1}^n$ Ordered according to σ .

•Number of trees I .

Hyper parameters: Learning rate η , regularization λ , number of layers L .

Steps

1. Initialize Generate random permutation σ of $[1, n]$

2. Initialize model predictions $M_i = 0$ for all $i = 1, \dots, n$.

3. Boosting Iterations $t = 1$ to I :

For $i = 1$ to n do:

Compute residual error: $r_i = y_i - M_{\sigma(i)-1}(X_i)$.

End for

For $i = 1$ to n :

Train Model update: $\Delta M = \text{LearnModel}[X_i, r_j]; \sigma(j) \leq i]$.

Update Model: $M_i = M_i + \Delta M$

End for

4. For Multi-Layer Fusion for each layer $l = 1$ to L :

Fuse features: $M^{(l)} = M^{(l-1)} + \eta \sum_{i=1}^N w_i g_i$.

End for.

4. RESULTS AND DISCUSSION

This section focuses on the effectiveness of ML algorithms in voice problem classification and the effect of DL based SDA methods in enhancing the accuracy of voice problem classification. The performance of the models is evaluated based on standard metrics like accuracy, precision, recall, F1-score, and AUC with ML models like SVM, MLP, XGBoost, CatBoost, and MLCF. The impact of DL algorithms (VAE, LSTM, GAN, and SpecAug-GAN) on the creation of high-quality synthetic data on the improvement of classification performance is explored. The results highlight the most suitable model-synthetic data combination for the classification of voice pathologies, emphasizing the benefits of data augmentation and advanced learning approaches for speech analysis.

The evaluation of the statistical reliability of the proposed SpecAug-GAN and Dynamic MLCF framework is given in Table 8. The table gives the average, standard deviation, and 95% confidence interval of the metrics calculated through repeated experiments, showing the stability, consistency, and reliability of the voice disorder classification system being proposed.

Table 8. Statistical reliability analysis of the proposed SpecAug-GAN with Dynamic MLCF framework

Performance Metric	Mean (%)	Standard Deviation (SD)	Confidence Interval
Accuracy	94.02	0.78	93.34 – 94.70
Precision	93.61	0.83	92.89 – 94.33
Recall	93.27	0.91	92.48 – 94.06
F1-Score	93.48	0.86	92.73 – 94.23

Note: SpecAug-GAN = Spectrogram Augmentation-based Generative Adversarial Network; Dynamic MLCF = Dynamic Multi-Layer Classifier Fusion.

4.1 Metrics for performance analysis

To test the effectiveness of SDG for DL classification of voice disorders, traditional evaluation metrics are used, such as accuracy, sensitivity, specificity, and F1-score, as used by Rahman and Direkoglu [22]. The equations for these measures are given by.

Accuracy: Assesses the comprehensive precision of the

classifier.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

Sensitivity (Recall): Assesses the model's capacity to accurately recognize aberrant speech.

$$Sensitivity = TP/(TP + FN)$$

Specificity: The measure indicates the model's proficiency in accurately classifying healthy speech

$$Specificity = \frac{TN}{TN + FP}$$

F1-score: Harmonizes precision and recall to enhance overall performance assessment.

$$F1 - score = (2 \cdot TP)/(2 \cdot TP + FP + FN)$$

where, TP = True Positives (accurately identified abnormal speech), TN = True Negatives (accurately identified healthy speech), FP = False Positives (incorrectly assessed healthy speech as abnormal), and FN = False Negatives (misclassification of disordered speech as healthy).

Regarding the earlier studies applied in the literature review, the use of the SpecAug-GAN framework allows obtaining better classification results, as it incorporates high-quality synthetic speech with beneficial feature extraction. While previously proposed approaches were generally aimed at increasing the amount of training data used, the framework proposed today is focused on keeping the pathology of speech during the process of augmentation. Consequently, the success rate is increased when tested on the Glottal dataset. Yet, additional testing on the corresponding datasets is required to gain more validation of the suggested approach.

4.1.1 Comparative analysis of machine learning model performance on benchmark dataset

The classification of Laryngocele, Vox Senilis, and Normal speech is compared among different ML models such as SVM, MLP, XGBoost, CatBoost, and MLCF in this section. The models are evaluated with the key classification metrics, such as accuracy, precision, recall, and F1-score, with the aim of determining their effectiveness in distinguishing voice disease. The results highlight advantages and disadvantages of each model and provide insights into the most suitable model for precise classification. Five ML models (SVM, MLP, XGBoost, CatBoost, and MLCF) were used for three categories of voice disorder: Laryngocele (L), Normal (N),

and Vox Senilis, as demonstrated in Figure 2. The figure shows that all of the models have perfect accuracy (1.00) for the laryngocele class except for MLP, with an accuracy of 0.75. The average accuracy for Normal and Vox Senilis is 0.83-0.90 and 0.75-0.87, respectively, with MLCF outperforming. Ensemble methods improve the accuracy, with a possible trade-off in less common situations, such as in laryngocele. The recall structure assesses the model's ability to identify all relevant events. The laryngocele dataset is imbalanced with a low recall (0.12–0.53 across models), which may be a limitation in the detection of laryngocele. Normal and Vox Senilis recall is better (0.88–0.96 and 0.80–0.91), with MLCF again outperforming (0.89). These results demonstrate that recall is complementary to accuracy and data augmentation may enhance sensitivity, as shown by vocal pathology detection studies.

The proposed framework has improved the process of representation learning by subjecting classifiers to a large number of naturally given variations of speech obtained via SpecAug-GAN. This means the newly proposed system is different than standard augmentation techniques in that it is able to prepare realistic artificial samples that are more in correspondence with the acoustic features of the speech of people suffering from a speech disorder, thus eliminating the effect of overfitting and improving the ability of the model to generalize. Thanks to this improved representation, the classifiers are able to easily discriminate among the speech recordings of Healthy, Vox Senilis, and Laryngocele people in accordance with the experimental conditions under which the respective experiments were conducted.

Table 9 shows the comparative evaluation of the ML models for discriminating speech pathology based on precision, recall, and F1-score measures for the Laryngocele (L), Healthy (N), and Vox Senilis (V) classes. As follows from the results, proposed Dynamic MLCF framework has the overall best results with average precision of 0.89, an average recall of 0.89, and average F1-score of 0.88. This can be explained by the unique nature of the Dynamic MLCF framework that includes different layers of learning, providing the models with complete complementary information that allows them to differentiate various speech pathologies more efficiently. Moreover, the proposed framework shows the most remarkable recall (0.96) and F1-score (0.91) for the Healthy class which indicates successful classification of normal speech recordings. To some extent, the situation with the Laryngocele class is opposite because its recall is less than in any other category. Though this is due to the absence of numerous samples of this class, Dynamic MLCF reveals the best results (0.50 for recall and 0.66 for F1-score) thus improving recognition of the minority classes.

Table 9. Comparative performance of machine learning (ML) models

ML Model	Precision (L)	Precision (N)	Precision (V)	Average Precision	Recall (L)	Recall (N)	Recall (V)	Average Recall	F1-Score (L)	F1-Score (N)	F1-Score (V)	Average F1-Score
SVM	1.00	0.83	0.75	0.81	0.12	0.88	0.84	0.80	0.21	0.85	0.79	0.78
MLP	0.75	0.90	0.85	0.87	0.53	0.89	0.91	0.87	0.62	0.90	0.88	0.87
XGBoost	1.00	0.84	0.84	0.85	0.47	0.94	0.80	0.85	0.64	0.89	0.82	0.84
CatBoost	1.00	0.87	0.85	0.87	0.42	0.95	0.85	0.87	0.59	0.90	0.85	0.86
MLCF	1.00	0.89	0.87	0.89	0.50	0.96	0.88	0.89	0.66	0.91	0.87	0.88

Note: SVM = Support Vector Machines; MLP = Multi-Layer Perceptron; XGBoost = Extreme Gradient Boosting; CatBoost = Categorical Boosting model; MLCF = Multi-Level Classification Framework.

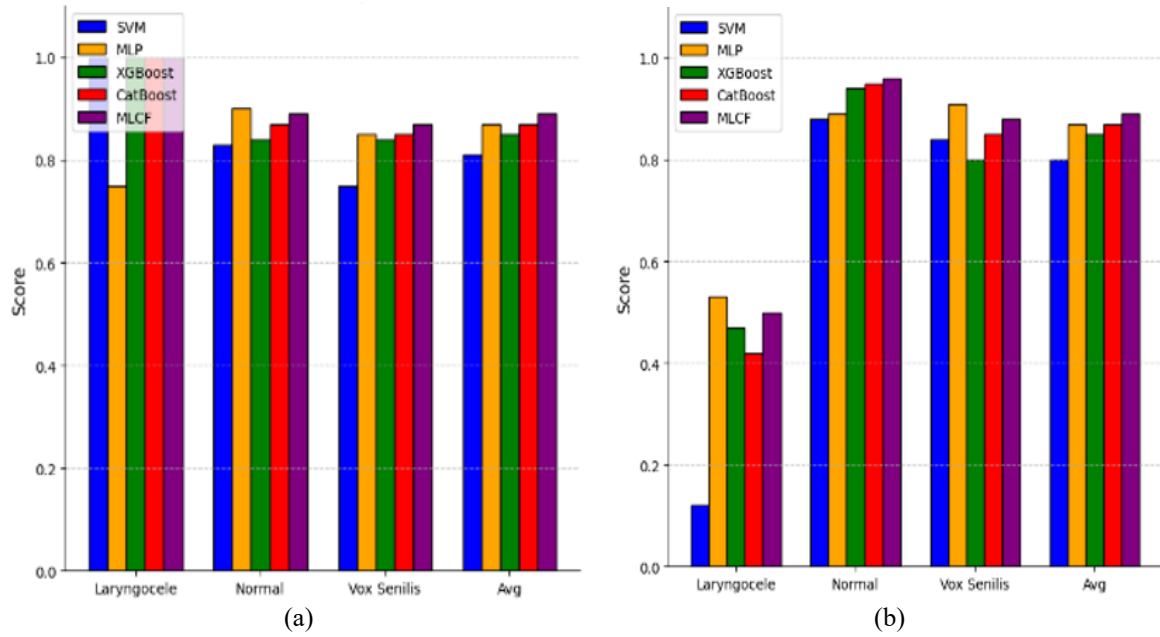


Figure 2. Comparative performance analysis, (a) precision comparison, (b) recall comparison

The classification results of the ML models are evaluated in terms of accuracy, precision, recall, and F1-score. The results have shown that the proposed framework called Dynamic Multi-Layer Classifier Fusion (Dynamic MLCF) has shown the highest overall performance (0.89, 0.89, 0.89, and 0.88 values for accuracy, precision, recall, and F1-score, respectively). The results indicate that the proposed approach is efficient in picking important traits of speech disorder by combining the results of different classifiers. The models MLP and CatBoost also demonstrated decent performance by achieving the corresponding accuracy of 0.87, while XGBoost managed to get 0.85 accuracy rate. Despite the results of SVM being satisfactory (0.80) its precision, recall, and F1-score rates show that the model is not very reliable in terms of generalization. The good results achieved by Dynamic MLCF show that the fusion of different techniques is effective and allows for better representation of features and good classification results.

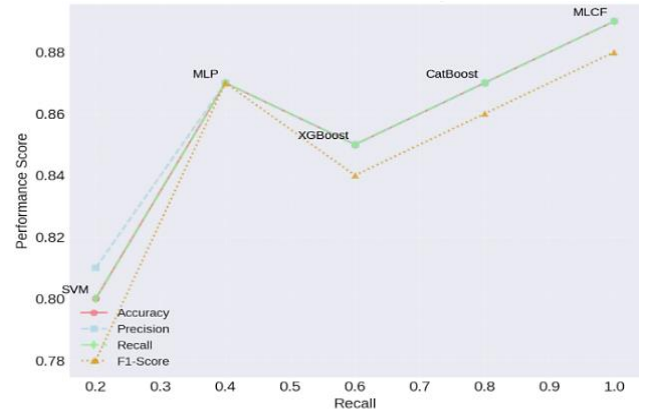


Figure 4. Comparison of machine learning (ML) model performance

Table 10. Comparative analysis of machine learning (ML) model performance model accuracy

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.80	0.81	0.80	0.78
MLP	0.87	0.87	0.87	0.87
XGBoost	0.85	0.85	0.85	0.84
CatBoost	0.87	0.87	0.87	0.86
MLCF	0.89	0.89	0.89	0.88

Note: SVM = Support Vector Machine; MLP = Multi-Layer Perceptron; XGBoost = Extreme Gradient Boosting; CatBoost= Categorical Boosting model; MLCF = Multi-Level Classification Framework.

Figure 3 compares the performance of the synthetic data models with ML classifiers. The comparison of MLCF model performance is shown in Figure 4. The comparative analysis of ML models. It includes the values of various methods. The comparison of the performance of ML models and the classification of the synthetic data is shown in Table 10. The MLCF model outperforms the other models with accuracy, precision, recall, and F1-score of 89%, making it the most efficient model. The performance of MLP and CatBoost is equal at 87% across all criteria, which shows that both models are stable and reliable. The accuracy of XGBoost is 85%,

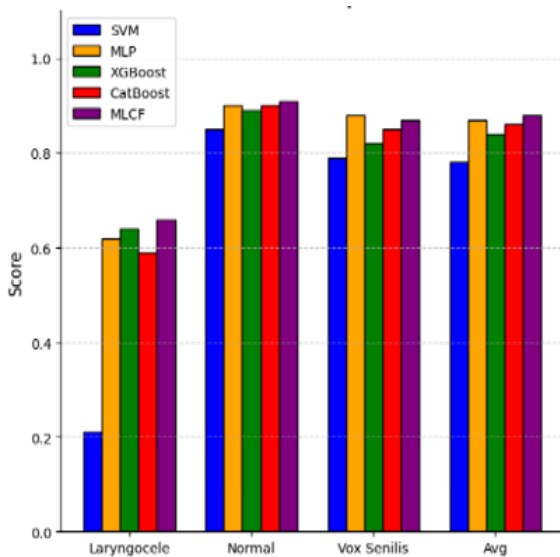


Figure 3. Comparative performance of machine learning (ML) models

which is competitive but slightly lower. The SVM model gives the lowest result in terms of accuracy (80%) and F1-score (78%). These are shown graphically in the accompanying plot, which shows an improvement in classification performance for the MLCF model.

Table 11 and Figure 5 illustrate the effectiveness of the SDG techniques (VAE, LSTM, GAN, and the proposed SpecAug-GAN) when analysed together with the methodology of four ML classifiers (SVM, MLP, XGBoost, and CatBoost). As a result, it has been concluded that the results of producing high-quality synthetic data plays an important role for the entire classification process. pecAug-GAN has shown the best results among the analyzed techniques, producing more useful samples of pathological speech.

Table 11. Comparative performance of synthetic data models across machine learning (ML) classifiers

SVM					MLP			
Synthetic Model	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
VAE	0.83	0.84	0.82	0.81	0.83	0.84	0.82	0.81
LSTM	0.84	0.85	0.83	0.82	0.84	0.85	0.83	0.82
GAN	0.86	0.86	0.85	0.84	0.86	0.86	0.85	0.84
SpecAug-GAN	0.88	0.88	0.87	0.86	0.88	0.88	0.87	0.86
XGBoost					CatBoost			
VAE	0.87	0.87	0.86	0.86	0.89	0.96	0.87	0.88
LSTM	0.88	0.88	0.87	0.87	0.90	0.96	0.88	0.89
GAN	0.89	0.89	0.88	0.88	0.91	0.97	0.89	0.90
SpecAug-GAN	0.90	0.90	0.89	0.89	0.92	0.97	0.90	0.91

Note: SVM = Support Vector Machines; MLP = Multi-Layer Perceptron; VAE = Variational Autoencoders; LSTM = Long Short-Term Memory; GAN = Generative Adversarial Networks; XGBoost = Extreme Gradient Boosting; CatBoost= Categorical Boosting model.

SWhile VAE and LSTM contribute to enhancing the classification results as compared to cases when limited original data is used, both of them are still less effective in terms of capturing complex characteristics of pathological speech. The technique of GAN significantly improves the results, as it provides more authentic synthetic data, but SpecAug-GAN still shows the best accuracy, precision, sensitivity, and F1-score. The biggest increase is observed when using the CatBoost algorithm (0.92 – accuracy, 0.97 – precision, 0.90 – sensitivity, and 0.91 – F1). A similar pattern is demonstrated by XGBoost, MLP, and SVM regarding efficacy, which supports the assumption that the methodology proposed in this study is beneficial for the increase of feature diversity, reduction of class imbalance, and better classifiers’ performance in the field of pathological speech.

The model performance comparison is shown in MLCF (Figure 6). Table 12 shows MLCF model performance, which is a comparative assessment of the different synthetic models by considering Accuracy, Precision, Recall, and F1-score. The SpecAug-GAN model demonstrates the best performance with an accuracy of 94%, a precision of 94%, a recall of 93%, and an F1-score of 93%, making it the most effective technique to enhance categorization. The GAN-based model shows 93% accuracy, precision, and recall, meaning quite good generalization performance. At the same time, the LSTM-based model gets 92% on all the measures, proving its effectiveness in learning sequential data. The VAE model achieves competitive but slightly lower accuracy, precision, and recall of 91%, 91%, and 90%, respectively, when it comes to generating synthetic data, though with slightly reduced classification performance in comparison to other models. The outcomes highlight how effective the GAN-based SDG can be, especially for improving the performance of the MLCF model, where SpecAug-GAN was found to be the most robust technique. The findings demonstrate the significant effect on

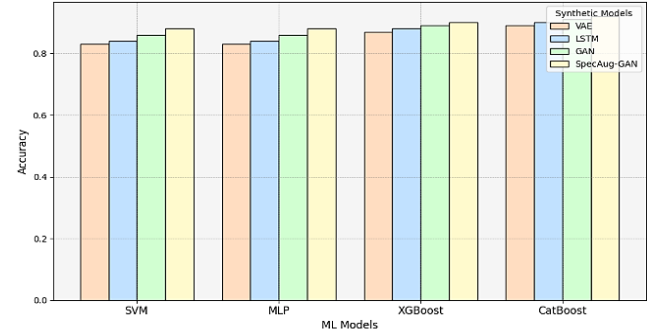


Figure 5. Performance comparison of synthetic data models across machine learning (ML) classifiers

classification performance when using advanced GAN architectures in the augmentation process. The data preprocessing and comparison of the model performance.

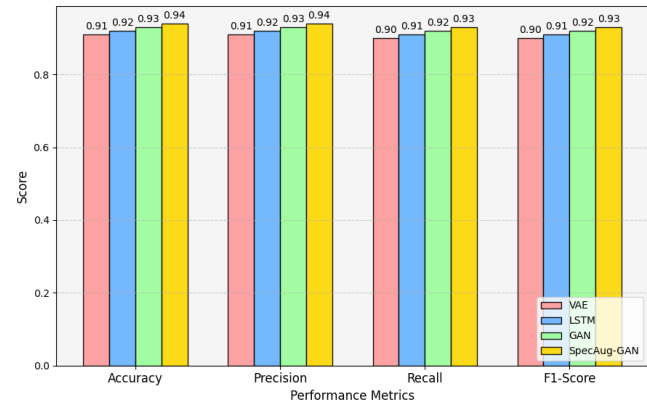


Figure 6. Multi-Level Classification Framework (MLCF) model performance comparison

Table 12. Multi-Level Classification Framework (MLCF) model performance

Synthetic Model	Accuracy	Precision	Recall	F1-Score
VAE	0.91	0.91	0.90	0.90
LSTM	0.92	0.92	0.91	0.91
GAN	0.93	0.93	0.92	0.92
SpecAug-GAN	0.94	0.94	0.93	0.93

The results of preprocessing on the model performance are presented in Figure 7. The accuracy of the raw data is 0.85, and significant improvements are seen after preprocessing the data. Noise is eliminated, improving accuracy up to 0.87. By removing the silence, speech segmentation accuracy is raised

to 0.89. Resampling with 16 kHz standardizes the dataset and increases the accuracy to 0.91. The normalization process gives better uniformity of the amplitude and achieves 0.93 accuracy. Finally, the accuracy of the voice disorder classification task is enhanced by the SpecAug-GAN augmentation, reaching 0.94, along with 0.94 precision, 0.93 recall, and 0.93 F1-score, showing the effectiveness of the preprocessing pipeline for voice disorder categorization.

The accuracy of this raw data is 0.85, and improvements are noticeable after the preprocessing processes are used. Noise was removed, achieving an accuracy of 0.87. The elimination of silence greatly helps in speech segmentation, raising accuracy to 0.89. When resampling at 16 kHz, the dataset is standardized, which increases the accuracy to 0.91. The accuracy is improved to 0.93 using normalization. Finally, the accuracy, precision, recall, and F1-score of the augmented dataset with SpecAug-GAN were compared, yielding 0.94, 0.94, 0.93, and 0.93, respectively, and demonstrating the effectiveness of the preprocessing pipeline for voice disorder classification. The confusion matrix in Figure 8 illustrates the overall performance of the classification model in three classes: Laryngozele, Normal, and Vox senilis.

The accuracy of the raw data is 0.85, and significant improvements are seen after preprocessing the data (Table 13). The model was able to correctly classify 8 cases as Laryngozele and 6 as Normal, and 5 cases as Vox senilis. The 109 cases in the Normal class were correctly classified, and 6 cases were misclassified as Vox senilis; no cases were misclassified as Laryngozele. The identification of Vox senilis was correct in 63 cases; 11 were given the name of Normal and none were given the name of Laryngozele. For laryngozele, the number of true positive (TP) values was 8, false negative (FN) was 11, false positive (FP) was 0, and true negative (TN) was 189. Normal demonstrated 109 TP, 6 FN, 17 FP, and 76 TN. Vox senilis presented 63 TP, 11 FN, 11 FP, and 123 TN. This analysis aims to highlight the strengths and areas of improvement of the model to distinguish between these classes. The 3D plot of the overall performance of the ML model.

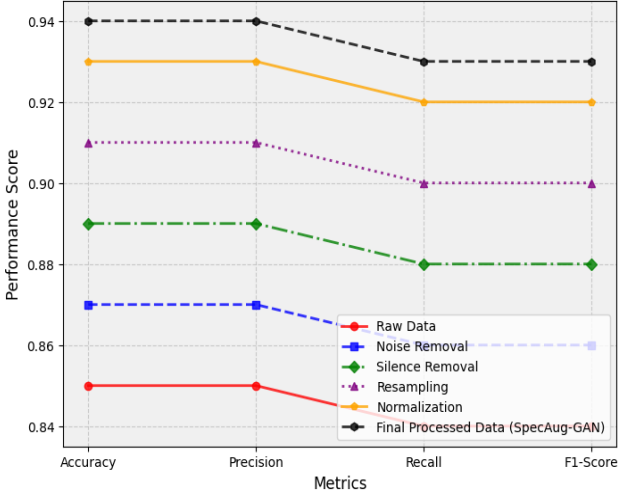


Figure 7. Impact of preprocessing on model performance

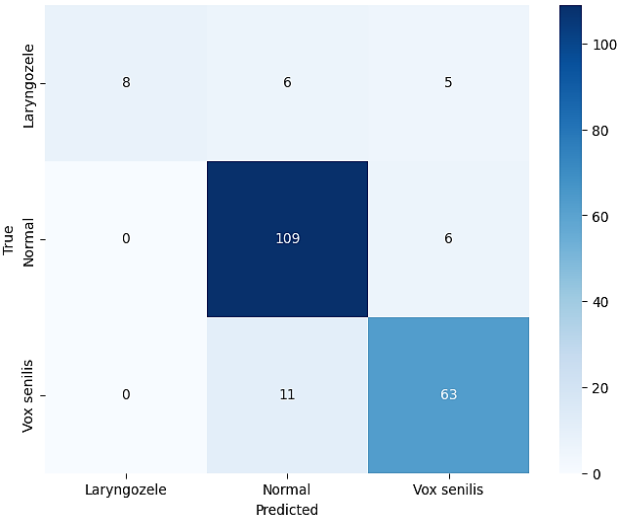


Figure 8. Confusion matrix

Table 13. Data preprocessing steps and impact on model performance

Processing Step	Description	Accuracy	Precision	Recall	F1-Score
Raw data	Original unprocessed voice recordings.	0.85	0.85	0.84	0.84
Noise removal	Spectral Subtraction applied to enhance clarity by removing background noise.	0.87	0.87	0.86	0.86
Silence removal	Voice Activity Detection (VAD) eliminates unnecessary silent regions.	0.89	0.89	0.88	0.88
Resampling	Standardizing all speech samples to 16 kHz for consistency.	0.91	0.91	0.90	0.90
Normalization	Peak or RMS Normalization applied to maintain uniform amplitude levels.	0.93	0.93	0.92	0.92
Final processed data (SpecAug-GAN)	Enhanced dataset with SpecAug-GAN for synthetic data augmentation.	0.94	0.94	0.93	0.93

Class-level performance analysis: The proposed method was assessed by examining the classification performance of each voice disorder class of the Glottal dataset. The Healthy and Vox Senilis classes were successfully recognized due to their relatively high number of training samples, while the Laryngozele class only included 85 recording samples so it is more difficult to classify correctly. The proposed method improves recognition of the minority class by generating synthetic samples, reducing issues of class imbalance. Although classification performance related to Laryngozele

was improved, the number of original samples of this class is very limited, so further evaluation using more extensive pathological speech datasets is necessary.

Figure 9 illustrates the surface plot that provides a three-dimensional comparison of the overall performance of the different ML models in terms of their performance metrics such as accuracy, precision, recall, and F1-score. The surface plot shows the performance trend from SVM to the proposed Dynamic MLCF. SVM does not perform well especially in terms of F1-score, which shows that SVM has lesser

effectiveness in differentiating voice categories in terms of pathology. The performance of MLP, XGBoost and CatBoost models improves gradually in terms of their performance metrics which indicates better learning of features of speech classification. The proposed Dynamic MLCF performs best with a value near 0.89 with respect to all metrics. The flat nature of the surface around the upper performance level confirms that the classifier is able to classify equally well without any major difference in terms of performance metrics. Hence the use of copasetic classifiers along with the synthetic speech generated by SpecAug-GAN improves feature capturing abilities, thereby overcoming the effects of class imbalance.

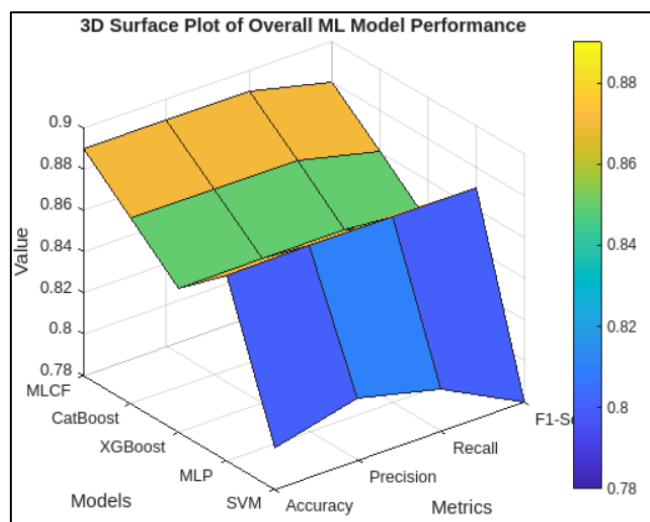


Figure 9. 3D surface plot of machine learning (ML) model performance

5. CONCLUSION

The findings of this research highlight the pivotal role of SDG and advanced ML models in enhancing voice disorder classification. By generating high-fidelity synthetic mel-spectrograms, data augmentation, and solving data scarcity problems, the incorporation of SpecAug-GAN improved classification accuracy significantly. Among all the classifiers evaluated, the Dynamic MLCF achieved the highest accuracy (0.94) and F1-score (0.93), outperforming the traditional classifiers including SVM, MLP, XGBoost, and CatBoost. The results demonstrate the synergistic effect of effective SDG and ensemble learning techniques in extracting the fine-grained audibility patterns of diseases like Laryngocele and Vox Senilis. The results deliver scalable techniques for early diagnosis and targeted intervention, particularly in resource-limited settings where there are limited, labelled datasets. So far, although SpecAug-GAN is effective, some problems still remain, such as the complexity of calculation and cross-dataset validation. Further studies are recommended to explore hybrid methods of incorporating both SpecAug-GAN and Transfer Learning, investigate real-time deployment on edge devices, and verify models across different linguistic and demographic communities. Furthermore, making the frameworks more interpretable can increase the confidence of the clinicians in AI-aided diagnoses. Addressing these challenges will enable the creation of reliable, easily accessible, and automated systems for the detection of vocal pathologies, which will

optimize patient care through timely and accurate diagnosis. The SpecAug-GAN and MLCF models have shown promising results in the Glottal data trials conducted. However, such results require further tests using bigger, well-balanced, and clinically diverse datasets in order to establish their applicability in practice. The class-wise analysis reveals that the suggested model improves the classification of the underprivileged pathological speech by diversifying the underrepresented class of the Laryngocele. Moreover, even though there has been an overall improvement in the classification results, the small number of original recordings of the Laryngocele might affect the recall of the class. Therefore, future work aimed at validating the model in bigger and well-balanced pathological speech datasets should be undertaken. As such, the suggested model has demonstrated its significance in the task of voice disorder classification on the Glottal subpart of the SVD corpus. However, the above results should be treated as preliminary validation of the conducted experiments rather than having significance for the healthcare sphere.

Limitations and future work: In this study, the only dataset used was the Glottal subset of SVD, which comprised of 1 037 speech recordings. Although the proposed method has shown remarkable performance for classification, the lack of external validation or cross-database assessment limits the applicability of the results obtained. Moreover, the rather small volume of recordings of laryngocele may play an important role in the quality of recognition of the minority class. Future studies will aim at validating the proposed approach based on larger multi-centre and clinically diverse pathological speech datasets.

REFERENCES

- [1] Jirsa, T., Verde, L., Marulli, F., Marrone, S., Vrba, J. (2024). Improving voice pathology classification using artificial data generation. *Procedia Computer Science*, 246: 5175-5184. <https://doi.org/10.1016/j.procs.2024.09.612>
- [2] Rehman, M.U., Shafique, A., Jamal, S.S., et al. (2024). Voice disorder detection using machine learning algorithms: An application in speech and language pathology. *Engineering Applications of Artificial Intelligence*, 133: 108047. <https://doi.org/10.1016/j.engappai.2024.108047>
- [3] Di Cesare, M.G., Perpetuini, D., Cardone, D., Merla, A. (2024). Assessment of voice disorders using machine learning and vocal analysis of voice samples recorded through smartphones. *BioMedInformatics*, 4(1): 549-565. <https://doi.org/10.3390/biomedinformatics4010031>
- [4] Latiff, N.M.A.A., Al-Dhief, F.T., Sazihan, N.F.S.M., et al. (2025). Voice pathology detection using machine learning algorithms based on different voice databases. *Results in Engineering*, 25: 103937. <https://doi.org/10.1016/j.rineng.2025.103937>
- [5] Chaiani, M., Selouani, S.A., Boudraa, M., Yakoub, M.S. (2022). Voice disorder classification using speech enhancement and deep learning models. *Biocybernetics and Biomedical Engineering*, 42(2): 463-480. <https://doi.org/10.1016/j.bbe.2022.03.002>
- [6] Ding, K., Chetty, M., Noori Hoshyar, A., Bhattacharya, T., Klein, B. (2024). Speech based detection of Alzheimer's disease: A survey of AI techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):

325. <https://doi.org/10.1007/s10462-024-10961-6>
- [7] bin Sham, A.H., Latiff, N.M.A.A., Al-Dhief, F.T., Sazihan, N.F.S.M., Rahman, S., Muhammad, N.A. (2023). Voice pathology detection system using machine learning based on internet of things. In 2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Kuala Lumpur, Malaysia, pp. 136-141. <https://doi.org/10.1109/SKIMA59232.2023.1038737>
- [8] Yao, D., Koivu, A., Simonyan, K. (2025). Applications of artificial intelligence in neurological voice disorders. *World Journal of Otorhinolaryngology-Head and Neck Surgery*, 11(4): 491-517. <https://doi.org/10.1002/wjo2.70017>
- [9] Parasyris, A., Stankovic, L., Stankovic, V. (2023). Synthetic data generation for deep learning-based inversion for velocity model building. *Remote Sensing*, 15(11): 2901. <https://doi.org/10.3390/rs15112901>
- [10] Rey-Paredes, M., Pérez, C.J., Mateos-Caballero, A. (2024). Time series classification of raw voice waveforms for Parkinson's disease detection using generative adversarial network-driven data augmentation. *IEEE Open Journal of the Computer Society*, 6: 72-84. <https://doi.org/10.1109/OJCS.2024.3504864>
- [11] Moëll, B., Aronsson, F.S., Östberg, P., Beskow, J. (2025). The order in speech disorder: A scoping review of state of the art machine learning methods for clinical speech classification. *arXiv preprint arXiv:2503.04802*. <https://doi.org/10.48550/arXiv.2503.04802>
- [12] Korba, M.C.A., Doghmane, H., Khelil, K., Messaoudi, K. (2024). Improved laryngeal pathology detection based on bottleneck convolutional networks and MFCC. *IEEE Access*, 12: 124801-124815. <https://doi.org/10.1109/ACCESS.2024.3454825>
- [13] Goyal, M., Mahmoud, Q.H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, 13(17): 3509. <https://doi.org/10.3390/electronics13173509>
- [14] Wang, J., Xu, H., Peng, X., Liu, J., He, C. (2023). Pathological voice classification based on multi-domain features and deep hierarchical extreme learning machine. *The Journal of the Acoustical Society of America*, 153(1): 423-435. <https://doi.org/10.1121/10.0016869>
- [15] Liu, Y., Yang, T., Tian, L., Huang, B., Yang, J., Zeng, Z. (2024). Ada-xg-CatBoost: A combined forecasting model for gross ecosystem product (GEP) prediction. *Sustainability*, 16(16): 7203. <https://doi.org/10.3390/su16167203>
- [16] de Brito Santos, M., de Moraes Calazan, R. (2024). Improved spectral dynamic features extracted from audio data for classification of marine vessels. *Intelligent Marine Technology and Systems*, 2(1): 18. <https://doi.org/10.1007/s44295-024-00029-0>
- [17] Gupta, N., Ahlawat, P. (2024). Audio classification using ensemble model. In 2024 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT), Gwalior, India, pp. 1-6. <https://doi.org/10.1109/INSPECT63485.2024.10896172>
- [18] Trijatmiko, D., Rahman, A.Y. (2023). Voice classification of children with speech impairment using MFCC kernel-based SVM. In 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), pp. 703-707. <https://doi.org/10.1109/ICCoSITE57641.2023.10127773>
- [19] Peng, X., Xu, H., Liu, J., Wang, J., He, C. (2023). Voice disorder classification using convolutional neural network based on deep transfer learning. *Scientific Reports*, 13(1): 7264. <https://doi.org/s41598-023-34461-9>
- [20] Kumar, S.P., Narayanan, N., Ramachandran, J., Thangavel, B. (2023). Convolutional neural network for voice disorders classification using kymograms. *Biomedical Signal Processing and Control*, 86: 105159. <https://doi.org/10.1016/j.bspc.2023.105159>
- [21] Gupta, R., Jin, C.T., Gunjawate, D.R., Nguyen, D.D., Stasak, B., Chacon, A.M., Madill, C. (2026). Voice disorders classification using machine learning: A scoping review. *Frontiers in Digital Health*, 8: 1800132. <https://doi.org/10.3389/fdgth.2026.1800132>
- [22] Rahman, M.U., Direkoglu, C. (2025). A hybrid approach for binary and multi-class classification of voice disorders using a pre-trained model and ensemble classifiers. *BMC Medical Informatics and Decision Making*, 25(1): 177. <https://doi.org/10.1186/s12911-025-02978-w>