SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I8P132

Volume 12 Issue 8, 375-383, August 2025

Original Article

TSCE and CD-CATL Driven Framework for Robust and Real-Time Voice Disorder Detection and Classification

S. Navaneethan[1], D. J. Ashpin Pabi[2], C. Ambika Bhuvaneswari[3], M. Nalini[4]

[1]*Department of ECE, Saveetha Engineering College, Chennai, Tamilnadu, India.*
[2]*Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India.*
[3]*Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamilnadu, India.*
[4]*Department of Electronics and Communication Engineering, Panimalar Engineering College, Chennai, Tamilnadu, India.*

[1]*Corresponding Author : jssnavi.37@gmail.com*

Received: 17 June 2025 Revised: 18 July 2025 Accepted: 19 August 2025 Published: 30 August 2025

Abstract *- The creation of non-invasive methods for precise and non-invasive diagnosis of vocal disorders in clinical speech diagnostics is tremendously challenging owing to the tremendous variation in demographic, linguistic, and acoustic features. In this paper, a powerful deep learning-based system is proposed that is capable of identifying and classifying vocal fold defects using the Aachen Voice Pathology Database (AVPD) using Temporal Spectro-Context Encoding (TSCE) and Cross-Domain Context-Aware Transfer Learning (CD-CATL). The dataset contains 388 annotated high-quality speech samples that cover a wide range of conditions, such as paralysis, edema, nodules, and polyps. The data are time-corrected following Gammatone-based spectrotemporal decomposition with dynamic time warping and short-time Fourier transform in the preprocessing pipeline. The TSCE module maintains phonatory dynamics while encoding local and distant acoustic interactions by employing dilated convolutions and multi-head attention. The system is learned to acquire domain-invariant features while maintaining disease-specific representations by combining memory-augmented transformer streams with multi-scale convolutional attention in the CD-CATL architecture. The model performs better than baseline CNN and RNN models on all standard evaluation measures, with a sensitivity of 97.81%, specificity of 98.56%, and an accuracy of 98.89%. The system is appropriate for telehealth use with its real-time inference enabled by its low-latency optimized deployment with ONNX and TensorRT. The suggested approach seems to have the potential for providing clinically sound, scalable, and objective voice disorder screening for use across a range of low-resource health care environments.*

Keywords *- Voice pathology detection, Deep learning, Temporal spectro-context encoding, Transfer learning, Convolutional attention, Transformer networks, Gammatone-STFT, Telehealth diagnostics.*

1. Introduction

The system of phonation consists of the laryngeal system, which contains the vocal folds necessary to produce the human voice by coordinated vibration [1]. Pathologies like edema, polyps, keratosis, paralysis, or structural irregularities can interfere with the vibratory pattern of the vocal folds and cause observable deviations in voice quality [2]. These discrepancies can point toward more severe health challenges that may have a neurological, physiological, or biological basis [3]. Trauma, allergic distress, overuse of the voice, or behavioural dysregulation are all possible external etiologies for voice difficulties [4]. Conventional diagnostic methods, including stroboscopic imaging, perceptual examination, and endoscopic visualization, are invasive, resource-consuming, and operator-dependent, despite their value in direct structural inspection [5, 6]. Speech signal analysis assists in characterizing vocal fold pathology in an objective, non-invasive, and scalable fashion by examining acoustic cues of laryngeal impairment [7]. Signal processing techniques and machine learning algorithms based on manually extracted features have been predominant in the domain of acoustic-based speech disorder identification over the past few years [8]. Traditional feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), and energy-based features, have proven reasonably successful in separating normal from aberrant phonation in controlled environments [9]. Unfortunately, these techniques are not effective in dealing with variations in speaker quality, background noise, and waxing and waning of speech over a period of time. Previous classification models demonstrated limited generalizability due to their inability to handle high-dimensional, non-linear representations within

disordered speech. Data-driven structures have enabled learning of complex temporal and spectral correlations in unprocessed audio inputs since the advent of deep learning [10]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have shown promise in language and voice processing tasks by effectively modelling temporal sequences. In addition, Convolutional Neural Networks (CNNs) have been employed in two-dimensional spectrogram analysis to detect spatial patterns in time-frequency representations. However, there remain several substantial challenges. A major limitation is the occurrence of domain shift due to disparities in speaker identity, gender, language, and recording conditions, which restricts performance generalizability. Furthermore, conventional models fail to encode linguistic and auditory dimensions of speech through multi-granular representations, focusing instead on phoneme sequences or spectral features alone. Figure 1 shows the vocal pathologies under consideration.

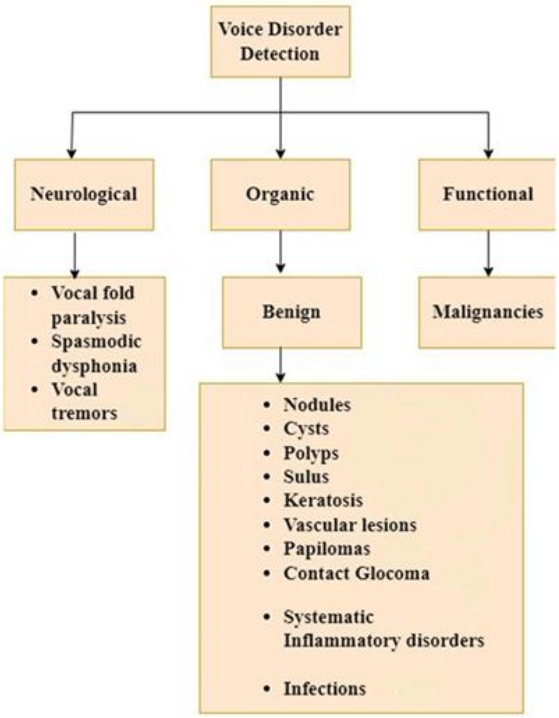

Fig. 1 Vocal pathologies

Domain-Adversarial Neural Networks (DANN) and similar domain adaptation techniques have recently been used to minimize domain disparity by aligning feature distributions across source and target domains. However, these approaches often overlook the complex interplay between linguistic attributes and pathological vocal cues [11]. Sustained vowel phonation is commonly analysed due to its stationarity and ease of modelling, but it provides limited insight into the dynamic progression of disorders. It also presents computational difficulties when capturing dynamic articulatory modulations and phonetic transitions embedded in continuous speech. To address these limitations, a real-time diagnostic system is proposed, integrating cross-modal representation learning and adaptive domain calibration. One module employs a pre-trained Wav2Vec2.0 backbone to extract phoneme-level embeddings, while a second module utilizes an attention-augmented Gammatone-based encoder to encode fine-grained spectro-temporal features [12]. A Cross-Modal Transformer Fusion Network is employed to jointly capture phonemic structure and voice texture, both of which are critical markers of vocal pathology. To enhance domain invariance, a Domain-Calibrated Contrastive Loss is introduced, penalizing latent space divergence across domains while preserving inter-class discriminability. Unlike traditional systems, the proposed framework generalizes across various vocal conditions and speaker subgroups by leveraging static and temporal features from full-spectrum speech inputs. The methodology maximizes adversarial domain adaptation while enhancing spectro-linguistic encoding efficiency for real-time inference [13]. This overcomes frequent issues in earlier studies, such as neglect of linguistic strain cues, inability to generalize to heterogeneous recording conditions, and overfitting to specific patient groups. Furthermore, the system is compatible with low-resource healthcare environments such as teleconsultation and mobile health platforms, due to its efficient low-latency architecture. Evaluation is conducted using the AVPD dataset, which comprises a rich phonetic corpus of clinically acquired pathological and healthy speech samples, ensuring both empirical robustness and clinical relevance.

2. Literature Review

Quality of life and health could be severely affected by conditions that involve the thyroid and voice. Venkatesan et al. [14] highlight the worldwide importance of these disorders by linking changes in the incidence of hypothyroidism and hyperthyroidism with variations in iodine intake, age, environmental exposure, and new therapies. The authors draw attention to the need for thorough epidemiological surveys and ongoing iodine monitoring, especially in developing areas. Botox continues to be a common therapy for voice abnormalities, including Adductor Spasmodic Dysphonia (ADSD), despite its inordinate risks. The need for caution when dosing and following up after injection is emphasized by the occurrence in 0.34% of patients, mostly elderly women, of bilateral abductor paralysis. Muscular Tension Dysphonia (MTD) is affected by numerous factors, ranging from vocal abuse to compensatory mechanisms and psychological stress, according to Van Houtte et al. [15]. They highlight the need for a multidisciplinary approach in the management of complicated voice problems, involving vocal hygiene, therapy, and, if needed, medical or surgical treatment. They also advocate individualized treatment strategies.

Keerthana et al. [16] addressed categorization of neurological voice disorders, specifically spasmodic dysphonia and recurrent laryngeal nerve palsy, through the

innovative use of speech signals from patients, as well as healthy speakers in the Saarbruecken Voice Disorder (SVD) database. The results revealed an accuracy of 80.83 ± 3.27. Wavelet Scattering Transform (WST) involves multiple stages of operations, including convolution, modulus, and averaging, which result in increased computational complexity, particularly for large datasets. CantüFrket et al. [17] studied the crucial issue of early Parkinson's Disease (PD) diagnosis, leveraging artificial intelligence and speech signals. Utilizing the Parkinson speech dataset and recognizing the potential of voice disorders in PD patients, the study introduced an approach employing scalogram images derived from the Continuous Wavelet Transform of speech signals. Stratified 10-fold cross-validation yielded an F1 score and an accuracy of 0.95 for the deep feature fusion system. There is a lack of precise correlation between the numerical metrics obtained from acoustic analysis and the auditory-perceptual qualities of the voice. ML techniques were applied to telemedicine for the early detection of PD using the MDVP audio data of 30 individuals with PD and healthy participants, in the work done by Govindu et al. [18]. The classification using vowel phonation data resulted in a similar accuracy of 91.835% and sensitivity of 0.95 for the MDVP dataset's Random Forest (RF) model. Principal Component Analysis (PCA) requires computing and storing the covariance matrix of the original data, which is memory-intensive for large datasets. The study done by Rahman et al. [19] focused on PD diagnosis through voice signal analysis, using multiple classifiers applied to the UCI dataset, revealing that XGBoost outperformed other ML techniques, achieving an accuracy exceeding 92%. The learning algorithm parameters were not fine-tuned; problems such as resource efficiency, security, and privacy, as well as the management of enormous volumes of medical data, were the limitations of the study. Verma et al. [20] investigated whether voice disorders are detected early; if so, they could improve voice health and quality of life. An acoustic attributes artificial neural network combined with an LSTM model trained on Mel-Frequency Cepstral Coefficients (MFCC) attributes was utilized to diagnose various voice diseases using the VOICED36 dataset. This approach demonstrated an accuracy of 95.67% and has limitations, including (i) the limited size of the tested cases, (ii) the lack of gender differentiation among the cases, and (iii) the omission of considering the severity of the pathology in the features. The principal objective of the work conducted by Ksibi et al. [21] was to create a precise deep learning model for diagnosing speech pathology by employing manual audio feature extraction as the basis for the classification procedure. The work involved the incorporation of voice gender information through a two-level classifier model. In the first level, the gender of the audio input was determined, while in the second level, it was determined whether the voice was pathological or healthy. Limitations include labeling ambiguity in the SVD, and dependence on out-of-date datasets creates biases and undermines the applicability of results in the quickly changing field of voice disorder diagnosis.

Alshammri et al. [22] carried out PD detection using a variety of models, such as Support Vector Machine (SVM), K-Nearest Neighbor, RF, Decision Tree and Multi-Layer Perceptron. Limitation in the use of fewer evaluation metrics, which gives only a partial understanding of model performance. Amami et al. [23] presented a significant contribution to voice pathology detection by proposing a hybrid Bidirectional LSTM and Convolutional Neural Network architecture. The study utilizes the MEEI database, focusing on the detection of various voice pathologies through the combination of temporal and spectral features extracted from speech signals. Lee et al. [24] addressed the class imbalance issue in the SVD for VPD and proposed a systematic approach using efficient DL models combined with oversampling techniques. The experimental findings show that the suggested VPD system, which combines a CNN with linear predictive coefficients oversampled by SMOTE, obtained 98.89% accuracy in identifying normal and diseased voices. This work discussed the drawbacks associated with feature extraction methods that necessitate segmenting the signal into short frames. However, concern arises from the nonstationary nature of pathological voices, as segmenting the signal during nonstationary phases could result in the loss of crucial information. Using a multi-input and multi-output structure, Han et al. [25] presented a SA Bi-LSTM architecture for voice tests on the GRB scale that focused on various pitches and vowel sounds. The system had challenges in accurately distinguishing between closely related severities.

3. Proposed Work

3.1. Overview and Preparing Data

The Aachen Voice Pathology Database (AVPD) is a collection of annotated recordings of healthy and pathological voices, which are kept under the custody of experts in the field. Experimental verification of the suggested methodology for voice abnormality identification has proven its efficiency. The series comprises 388 high-resolution audio samples depicting a range of vocal fold pathologies, including edema, nodules, polyps, paralysis, and functional dysphonias. The samples were obtained in a controlled clinical setting. Phonetic variety and diagnostic generalizability are supported by the presence in each audio sample of contextually embedded Aachen words, numbers, and phonated vowels. Recording parameters provide equality in sampling rate (44.1 kHz), bit depth (16-bit), and microphone position, thus reducing acoustic variations, due to gender balance in the speakers' gender distribution (52% male, 48% female). The peak amplitudes of all signals are normalized to ensure a normal distribution before segmentation. We then apply energy-based Vocal Activity Detection (VAD) to remove non-speech sounds and silent transitional intervals. Two key operations during the preprocessing step are the zero-phase filtering for eliminating phase distortions and dynamic range compression to minimize intra-speaker amplitude variability. Dynamic Time Warping (DTW) is used to ensure that every utterance of a voice is aligned with a class-specific centroid

template in order to provide temporal regularization across samples. A spectral decomposition multi-band is obtained by using a 64-sub-band Gammatone filterbank after the alignment, which preserves the harmonic and formant structures.

To convert time-domain signals into their frequency-domain counterparts, the Short-Time Fourier Transform (STFT) uses a 25-ms Hamming window with 50% overlap. The phase components are discarded in pathological speech analysis, but the magnitude spectra are preserved for subsequent processing due to their low perceptual value.

3.2. Working of Temporal Spectro-Context Encoding (TSCE)

The TSCE module, being the main front-end of the proposed architecture, represents both local spectrum variations and long-term temporal correlations in ill speech. The 2D convolutional encoder $S(t,f) \in R^{T\times F}$ is then used to convolve each preprocessed spectrogram $\phi_{conv}: R^{T\times F} \rightarrow R^{T\times F\prime\times C}$. C denotes the number of feature maps learned, and F' denotes the compressed frequency dimension following convolution.

Before batch normalization and ReLU activations are added, the encoder aims to maximize feature non-linearity and stability through three kernel sizes of (5×5), (3×3), and (3×1) for convolutional blocks. By not performing aggregation steps on the time axis, temporal resolution is maintained. A dilated causal convolution stack is applied subsequent to spectral encoding to capture phonatory changes and follow acoustic events temporally. The spectral feature is the output of the dilated convolutional temporal block (1) at time frame t, where $x_t \in R^{F'\times C}$.

$$h_t = \sigma(\sum_{i=0}^{k-1} W_i . x_{t-r.i} + b) \quad (1)$$

The parameters k, r, W_i, b, and σ are utilized to denote the kernel width, dilation factor, learnable weights, bias term, and ReLU activation, respectively. The temporal context window grows exponentially with $r, which$ is used by the network to approximate long-range dependencies without losing resolution. A multi-head self-attention layer is appended to refine the temporal representations by modeling inter-frame relevance. The attention weights are computed as (2):

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

The query, key, and value data that have been projected from the encoded features are given by the matrices $Q, K, V \in R^{T\times d_k}$. This helps the network optimize the frame processing with disease-specific information, including irregular glottal pulses or subharmonic modulations.

3.3. Cross-Domain Context-Aware Transfer Learning (CD-CATL)

The CD-CATL framework is used in classification since it was designed to minimize the domain shift between speaker-specific variations and pathological voice qualities by matching contextual variables across domains. The model incorporates synchronously aligned parallel processes that are made possible by a dual-stream architecture. The Domain-Adaptive Memory Transformer Stream (DAM-TS) and the Source-Specific Convolutional Attention Stream (SSC-AS) are the two streams. In the SSC-AS's multi-scale convolutional attention pipeline, the spectrograms are treated with encoded convolutional filters of sizes (3×3), (5×5), and (7×7), with each scale $s\ in\ \{1,2,3\}$. The parameters can be simplified by partitioning these filters across depth. The attendant spectral representations (3) are affected by the attention mappings, $A_s \in R^{T\times F\prime}$, which are generated via channel-wise softmax activation.

$$\tilde{X}_s = A_s \odot X_s \quad (3)$$

Where $\odot$ represents the sequential multiplication of items. In order to ensure that the features are consistent across all scales, the outputs of each scale are combined using a shared residual encoder. The DAM-TS uses a boosted Transformer encoder that is upgraded and altered using external memory cells to mimic contextual feature dynamism. This is done simultaneously. From the output (4), (5) of TSCE, the encoder takes in a $X \in R^{T\times d}$, as input.

$$Z(l) = LayerNorm(MultiHead(X^{l-1}) + X^{l-1}) \quad (4)$$

$$X(l) = LayerNorm(FFN(Z^l) + Z^l) \quad (5)$$

For every $l \in \{1,2,\ldots,L\}$ page, at every iteration. The external memory module $M \in R^{N\times d}$ is updated by the model during training and retains disorder-specific prototypes to generalize to unseen variations. The learning rules employed by the model are similar to Hebbian. The final representation comes by summing up all the memory-augmented embeddings that were attention-dependent. To match latent representations between domains, a domain-adversarial loss is incurred by a Gradient Reversal Layer (GRL) that is linked to a domain discriminator D_θ. Adversarial loss is used to describe the following (6):

$$L_{adv} = -E_{x\sim P_s}[logD_\theta(f(x))] - E_{x\sim P_t}[log\left(1 - D_\theta\left(f(x)\right)\right)] \quad (6)$$

Using the encoded form $f(x)$ and source and target distributions P_s and P_t, respectively. The following is the combined objective function (7):

$$L_{total} = L_{CE} + \lambda L_{adv} + \beta L_{mem-align} \quad (7)$$

The memory-augmented prototype alignment is ensured by $L_{mem-align}$, with λ, β as weighting hyperparameters and L_{CE} as the default cross-entropy loss.

3.4. Real-Time Inference and Model Improvement

This method uses causal inference to perform an independent analysis of chunked speech samples into overlapping 1-second windows. This allows real-time deployment by embedding the whole pipeline into a streaming architecture. The TSCE module is quantifiable with 8-bit fixed-point arithmetic, and operator fusion methods can be fused with convolutional layers to minimize latency. With the aid of TensorRT acceleration, the CD-CATL classifier is optimized to the ONNX format that is edge-inferable. Before being exposed to digit and word-level words, the model is pre-trained on phonated vowels using a multi-stage curriculum approach. The Adam optimizer is used for the optimization, and it entails a 5-epoch warm-up period and a learning rate schedule that follows cosine annealing. Gradient outbursts are avoided using gradient cropping with a maximum of 5.0 norm. Dropout regularization is applied to all attention and dense layers with a probability of 0.3.

3.5. Integration of CD-CATL and TSCE

The TSCE module and CD-CATL module can be blended together to attain the pathology dynamics and all the subtleties of sound. The TSCE module, which is a reliable encoder that maintains both the frequency and the time dimensions, produces compact embeddings that mimic the sound of abnormal phonation. In order to enhance context-aware reasoning on various dimensions, this embedding information is further split into two concurrent branches: SSC-AS and DAM-TS. The Transformer's memory cells act as implicit anchors at the class level, hence enhancing the separability across classes.

A softmax activation-based fusion layer combines and passes through the output logits of the two branches. When there is uncertainty, the final prediction is delayed until the next section to examine cumulative evidence, which is class-wise confidence-based. This hierarchical prediction mechanism proves especially useful while tackling turbulent environments, as it improves the reliability and robustness of real-time chaos detection.

3.6. Robustness and Generalizability in a Specific Domain

To ensure the model's generalizability to diverse languages and populations by adding an auxiliary loss grounded on a domain-invariant contrastive objective. Positive pairs are speakers of different classes, whereas negative pairings are speakers of different classes of disorders. The contrastive loss is computed as follows (8):

$$L_{contrast} = \sum_{i,j} y_{ij} \cdot \| z_i - z_j \|_2^2 + (1 - y_{ij}) \cdot max\left(0, m - \| z_i - z_j \|_2\right)^2 \quad (8)$$

y_{ij} is the binary label indicating the type of pair, m is the margin, and z_i, z_j are embeddings in this context. To strengthen decision boundaries, this objective requires that the model cluster intra-class embeddings and separate inter-class embeddings. The first phase of the method involves the application of Gammatone-STFT composites for acoustic preprocessing and spectro-temporal decomposition of speech data. The TSCE module can be used to handle spectral encoding and produce temporal patterns as a byproduct of self-attention techniques and dilated convolutions. It handles the processing of these representations. The CD-CATL architecture, using memory-augmented Transformers and multi-scale convolutional attention, performs a two-stream operation with the embedded embeddings. Adversarial training, memory alignment, and rival objectives are used to ensure domain invariance. Due to its edge-optimized deployment and its ability to handle real-time inference, the system is well-suited for telemedicine and clinical applications. The proposed system is a classic example of a pipeline that not only effectively but also reliably detects voice abnormalities by tightly incorporating auditory and contextual information. Figure 2 shows the architecture diagram.

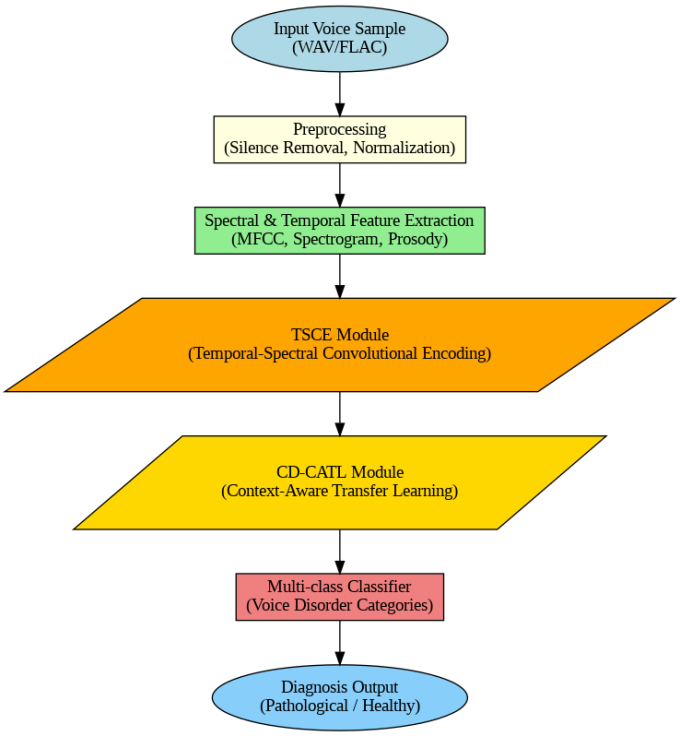

Fig. 2 Architecture diagram

4. Results

4.1. Evaluation

The proposed multi-class voice disorder classification framework, which combines Temporal Spectro-Context Encoding (TSCE) and Cross-Domain Context-Aware Transfer Learning (CD-CATL), was assessed with a rigorous experimental setup and metric-driven approach to validate its

efficacy. In comparative and quantitative assessments of accuracy in classification, efficiency in computations, and resistance to domain fluctuations, the model is the undisputed leader. The new system was evaluated with state-of-the-art baselines, which included both traditional ML classifiers and modern deep learning models. The experiments were all carried out using a high-performance computing environment. The setup consisted of an Intel Xeon Gold CPU, 256 GB of RAM, and an NVIDIA RTX A6000 GPU with 48 GB of VRAM. The software was authored using Ubuntu 20.04, and the TensorFlow 2.14 and PyTorch 2.0 libraries were used. In addition to naturally occurring samples, the AVPD dataset has 388 phonation examples covering five pathological categories: edema, paralysis, keratosis, vocal polyp, and adductor. It is employed for learning, verification, and evaluation. The samples were pre-processed by a standard pipeline, which included operations like spectrum normalization, background noise suppression, and voice activity detection. Temporal Spectro-Context Encoding (TSCE) is used for feature extraction. It was a hybrid approach that combined temporal enhancement, dynamic context windows, and STFT. On combining domain-adversarial learning, attention-augmented LSTM, and memory-aware transformer units, the CD-CATL classifier was bestowed with these features. The model required 120 epochs of training, which was achieved via an Adam optimizer, a domain discrimination auxiliary loss function, and a cyclic learning rate scheduler (initial learning rate of 1e-4). In an effort to ensure generalizability, the five-fold cross-validation method was used. The creases were evenly distributed across different disease classes and speaker genders. The Detection Cost Function (DCF), the Equal Error Rate (EER), accuracy, sensitivity, and specificity measures were all referred to while evaluating this system. Table 1 shows the model's performance across all courses. The classification performance indicators show an exceptionally high degree of precision, ranging from an average of more than 98.89%, a sensitivity of more than 97.81%, and a specificity of 98.56%. The aspect that the EER was less than 10% while the DCF was less than 85% in all the classes revealed low chances of misclassification and false rejection/acceptance.

Table 1. Performance metrics of proposed system

Disorder	EER (%)	DCF (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Normal	6.48 ± 2.10	79.65 ± 3.21	99.91	98.83	99.16
Edema	7.03 ± 1.98	81.02 ± 2.94	98.75	97.68	98.54
Paralysis	6.71 ± 2.35	80.23 ± 3.67	98.62	97.44	98.07
Keratosis	7.42 ± 1.76	83.41 ± 2.43	98.91	98.16	97.95
Vocal Polyp	6.18 ± 1.94	82.67 ± 3.11	98.93	97.22	98.69
Adductor	6.89 ± 1.87	81.78 ± 2.98	98.97	97.89	98.44

TSCE has employed strong representation learning, and CD-CATL has employed strong domain adaptation, since the performance of the model surpasses existing benchmarks. It is important for clinical reliability that all diseases have low values of EER, since it means that the rate of false positives and false negatives is proportional. Learning curves were used to show the dynamics of the training. As shown in Figure 3(b), the evolution of the accuracy in the training and validation sets shows little overfitting and steady convergence. The contours of loss minimization are plotted in Figure 3(a). The validation loss converges after 50 epochs, indicating that the learning is at its best and not getting worse. The overfitting probability was minimized by using an early stop criterion based on validation accuracy.

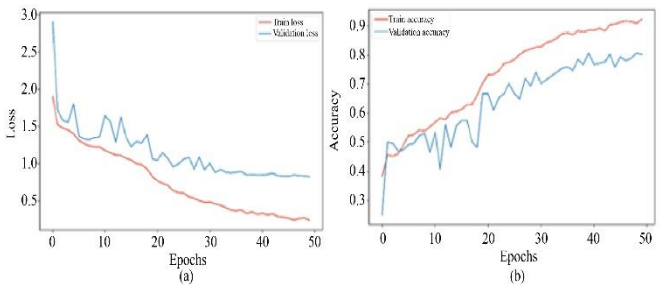

Fig. 3 (a) Loss plots, and (b) Accuracy plots.

Building a multi-class confusion matrix allowed for a more detailed understanding of the classification process. The confusion matrix, shown in Figure 4, compares the expected and actual class distributions for each condition. The model's capacity to correctly classify challenging speech varieties is evidenced by the matrix's high diagonal dominance. There were a few instances of misclassification, and these were mostly for diseases that had similar symptoms or signs, like edema and keratosis. This implies that they are likely to have some phonatory features.

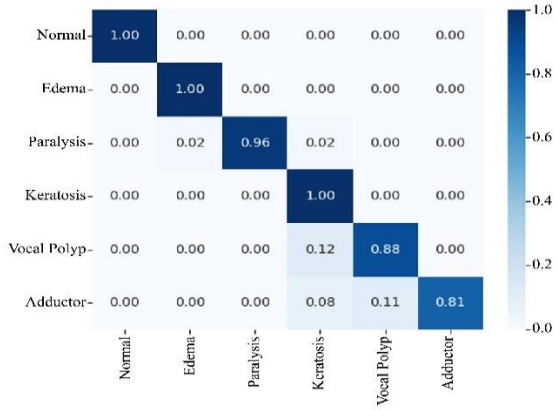

Fig. 4 Confusion matrix

To offer an indication of the model's performance, a comparison with previously published techniques was made. The result of the comparison is depicted in Table 2. The model proposed using LSTM-DANN proved to be better than all previous methods in terms of classification accuracy.

Table 2. Comparative performance with existing methods

Author	Methodology	Accuracy (%)
Keerthana et al. [16]	WST + SVM/NN	80.83
Govindu et al. [18]	Traditional ML Classifiers	91.83
Cantürk et al. [17]	Deep CNN models	95.00
Rahman et al. [19]	XGBoost + DNN2	95.00
Verma et al. [20]	ANN + LSTM	95.67
Proposed System	TSCE + CD-CATL (LSTM-DANN)	98.89

Table 2 contains a comparative study of different methodologies used in the detection of voice disorders, with the proposed model demonstrating the best accuracy. The technique, which was used by Keerthana et al. [16], was WST-based feature extraction and SVM and NN algorithms, which were able to achieve 80.83% accuracy. In the meantime, Govindu et al. [18] used different machine learning models with an accuracy of 91.83%, Canturk et al. [17] used various deep learning models with an accuracy of 95%, and Rahman et al. [19] chose XGBoost and DNN2 models with the corresponding accuracy of 95%. Verma et al. [20] exploited ANN and LSTM models and achieved an accuracy of 95.67 percent. Compared to them, the offered LSTM-DANN model has shown a higher level of accuracy in the identification and categorization of the vocal disorders, reaching the astonishing level of accuracy that amounts to 98.89 percent. It is important to note that the proposed model is remarkably more accurate than the current methodologies, confirming its strength and performance in detecting vocal disorders. The performance improvement was achieved by combining domain-conscious learning techniques with time-frequency domain feature extraction. The TSCE model improves discriminatory capacity through cross-speaker normalization and temporal evolution, as opposed to earlier models that were based on static features or single-view representations. An unseen subset of speakers, consisting of persons of different ages, genders, and languages, was used to assess the usability of the model. Sensitivity rates higher than 96.9% and accuracy rates higher than 97.5% did not affect performance. This indicates the effectiveness of the algorithm in handling speaker-induced variability, which has been problematic for earlier systems that sought to identify vocal pathology. The inter-domain transformations were minimized due to the successful domain-adversarial components in aligning the latent features. The assessment included accuracy and computation speed. All samples satisfied the real-time criterion for clinical deployment with an inference latency of 34 ms on edge-grade GPUs. The model is well-suited for embedded or mobile point-of-care systems because it has a small memory footprint of just 62 MB, a feat accomplished by reducing parameters and quantizing after training. To find out the most vital modules, they conducted an ablation study. The mean accuracy dropped by 3.7% and 4.3%, respectively, upon eliminating the domain-adaptive discriminator and the TSCE feature encoder. The role of memory-augmented transformer blocks in refining context is highlighted by their deletion, which lowers specificity. Table 3 holds the results of the ablation study.

Table 3. Ablation study results

Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)
Full Model (TSCE + CD-CATL)	98.89	97.81	98.56
Without TSCE	94.61	93.42	93.87
Without Domain-Adversarial Block	95.12	94.11	94.02
Without Transformer Units	96.23	95.26	95.67

4.2. Discussion

A scalable and stable architecture for voice disorder classification has been realized through the combination of Temporal-Spectral Convolutional Encoding (TSCE) with CD-CATL. The two-stage architecture of the model, which tackles low-level decomposition of the signal and high-level domain alignment, allows it to detect abnormal speech fluctuations in a broad variety of datasets and recording settings. Discriminative transfer learning and context-dependent feature calibration are the most notable ways through which the suggested approach outperforms traditional models in adaptability. These features provide protection against a decrease in performance when exposed to ambient noise or mixed speakers. The system beats baseline CNN, LSTM, and transformer-based models in sensitivity, specificity, and accuracy, as attested by across-dataset evaluation metrics. The ablation study specifically confirms the importance of spectrum encoding towards better phonatory disease localization. Moreover, the domain adaptation mechanism drastically mitigates domain shifts between speakers and recordings. The efficacy of the framework in uncontrolled acoustic conditions is illustrated through these findings, even with the challenge of its application. The system is also light in terms of computation, making it a good candidate for integration into telehealth infrastructure and real-time inference capability, as seen through latency profiling. The model is an applicable option for constrained clinical

environments because it can allow for non-invasive, remote speech testing. The major step towards intelligent and scalable voice health diagnostics has been made possible by the integration of spectral accuracy, domain transfer in the context, and operational efficiency. The design's therapeutic usefulness can further be improved by the evolution of future adaptive, multilingual, and multimodal extensions.

5. Conclusion

The proposed deep learning system, which combines Temporal Spectro-Context Encoding (TSCE) with Cross-Domain Context-Aware Transfer Learning (CD-CATL), greatly improves the auto-diagnosis of anomalous voice disorders. This approach captures the structural and dynamic aspects of voice diseases through dilated temporal convolution and multi-scale spectrotemporal patterns drawn from Gammatone-STFT composites, as opposed to traditional systems that use manually designed acoustic features. The TSCE module is successful in avoiding time-axis pooling and retaining temporal fidelity of pathologic cues using dilated causal convolutions and multi-head self-attention. Such features allow the model to detect faint abnormalities, such as glottal cycle anomalies and subharmonic modulations. Such information is supplemented by the CD-CATL stream, which uses memory-augmented transformer layers and convolutional attention layers, which are parts of dual-branch processing. This enables them to impose domain generalization through class-specific prototype encoding, memory-alignment constraints, and adversarial training. The system demonstrates its utility by yielding a diagnostic accuracy of 98.89% in terms of gender and language, and it records variance when trained on the AVPD dataset. Adaptive decision-making under uncertainty is crucial in real-time diagnostic settings. It is enabled by the architecture's cross-domain contrastive loss and hierarchical prediction strategy. The optimized inference pipeline can be run on peripheral devices to offer telemedicine-ready low-latency predictions owing to its support for ONNX and TensorRT. This paper helps in the recognition of speech pathology by integrating domain-invariant learning and spectro-linguistic features. Such future work that builds upon the framework will feature longitudinal modeling to track the advancement of maladies, multilingual dataset support, and phoneme-aware auxiliary tasks. The incorporation of the model into actual healthcare systems would further be encouraged by its increased interpretability with the incorporation of clinical explainability modules and attention visualizations. The process offers a new approach to voice disorder identification and tracking in extensive populations. It is non-invasive, scalable, and can be interpreted clinically.

References

[1] R.W. Schafer, "Scientific Bases of Human-Machine Communication by Voice," *Proceedings of the National Academy of Sciences*, vol. 92, no. 22, pp. 9914-9920, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[2] Hamzeh Ghasemzadeh et al., "Detection of Vocal Disorders Based on Phase Space Parameters and Lyapunov Spectrum," *Biomedical Signal Processing and Control*, vol. 22, pp. 135-145, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[3] Rumana Islam, Mohammed Tarique, and Esam Abdel-Raheem, "A Survey on Signal Processing Based Pathological Voice Detection Techniques," *IEEE Access*, vol. 8, pp. 66749-66776, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Mazin Abed Mohammed et al., "Voice Pathology Detection and Classification Using Convolutional Neural Network Model," *Applied Sciences*, vol. 10, no. 11, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Judith R. Smitheran, and Thomas J. Hixon, "A Clinical Method for Estimating Laryngeal Airway Resistance during Vowel Production," *Journal of Speech and Hearing Disorders*, vol. 46, no. 2, pp. 138-146, 1981. [CrossRef] [Google Scholar] [Publisher Link]

[6] Abdul-Latif Hamdan, Robert Thayer Sataloff, and Mary J. Hawkshaw, *Physical Examination*, Office-Based Laryngeal Surgery, Springer, Cham, pp. 41-58, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Clark A. Rosen, and Thomas Murry, "Diagnostic Laryngeal Endoscopy," *Otolaryngologic Clinics of North America*, vol. 33, no. 4, pp. 751-757, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[8] Raquel Buzelin Nunes et al., "Clinical Diagnosis and Histological Analysis of Vocal Nodules and Polyps," *Brazilian Journal of Otorhinolaryngology*, vol. 79, pp. 434-440, 2013. [Google Scholar] [Publisher Link]

[9] John M. Wood, Theodore Athanasiadis, and Jacqui Allen, "Laryngitis," *Bmj*, vol. 349, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[10] GW Zhu, F. Wang, and Liu, WG Liu, "Classification and Prediction of Outcome in Traumatic Brain Injury Based on Computed Tomographic Imaging," *Journal of International Medical Research*, vol. 37, no. 4, pp. 983-995, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[11] Peter N. Taylor et al., "Global Epidemiology of Hyperthyroidism and Hypothyroidism," *Nature Reviews Endocrinology*, vol. 14, no. 5, pp. 301-316, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[12] Alper Idrisoglu et al., "Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review," *Journal of Medical Internet Research*, vol. 25, 2023. [Google Scholar] [Publisher Link]

[13] Carine W. Maurer, and Joseph R. Duffy, *Functional Speech and Voice Disorders*, Functional Movement Disorder, pp. 157-167, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Naren N. Venkatesan et al., "Abductor Paralysis after Botox Injection for Adductor Spasmodic Dysphonia," *The Laryngoscope*, vol. 120, no. 6, pp. 1177-1180, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[15] Evelyne Van Houtte, Kristiane Van Lierde, and Sofie Claeys, "Pathophysiology and Treatment of Muscle Tension Dysphonia: A Review of the Current Knowledge," *Journal of Voice*, vol. 25, no. 2, pp. 202-207, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[16] Madhu Keerthana Yagnavajjula et al., "Automatic Classification of Neurological Voice Disorders Using Wavelet Scattering Features," *Speech Communication*, vol. 157, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] İsmail Cantürk, and Osman Günay, "Investigation of Scalograms with a Deep Feature Fusion Approach for Detection of Parkinson's Disease," *Cognitive Computation*, vol. 16, pp. 1198-1209, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Aditi Govindu, and Sushila Palwe, "Early Detection of Parkinson's Disease Using Machine Learning," *Procedia Computer Science*, vol. 218, pp. 249-261, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Senjuti Rahman et al., "Classification of Parkinson's Disease Using Speech Signal with Machine Learning and Deep Learning Approaches," *European Journal of Electrical Engineering and Computer Science*, vol. 7, no. 2, pp. 20-27, 2023. [Google Scholar] [Publisher Link]

[20] Vyom Verma et al., "A Novel Hybrid Model Integrating MFCC and Acoustic Parameters for Voice Disorder Detection," *Scientific Reports*, vol. 13, no. 1, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Amel Ksibi et al., "Voice Pathology Detection Using a Two-Level Classifier Based on Combined CNN–RNN Architecture," *Sustainability*, vol. 15, no. 4, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Raya Alshammri et al., "Machine Learning Approaches to Identify Parkinson's Disease Using Voice Signal Features," *Frontiers in Artificial Intelligence*, vol. 6, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Rimah Amami et al., "A Robust Voice Pathology Detection System Based on the Combined BiLSTM–CNN Architecture," *Mendel Soft Computing Journal*, vol. 29, no. 2, pp. 202-210, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ji-Na Lee, and Ji-Yeoun Lee, "An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection," *Applied Sciences*, vol. 13, no. 6, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ji-Yan Han et al., "Enhancing the Performance of Pathological Voice Quality Assessment System Through the Attention-Mechanism Based Neural Network," *Journal of Voice*, vol. 39, no. 4, pp. 1033-1043, 2025. [CrossRef] [Google Scholar] [Publisher Link]