

Original Research

Received 19 June 2026

Revised 05 Aug 2026

Accepted 20 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (16s): 1439–1453

KEYWORDS:

Pathological voice, Multi-Dimensional Voice Program (MDVP), Stacked Extreme Learning Machine (SELM), Kernel Extreme Learning Machine (KELM), Extreme Learning Machine (ELM).

Clinically Interpretable MDVP Feature Learning for Robust Pathological Voice Classification Using Stacked and Kernel Extreme Learning Machines

Anil Kumar V, R Venkata Siva Reddy

School of Electronics and Communication Engineering, REVA University, Bengaluru-560064, Karnataka, India
anilkvmr@gmail.com

ABSTRACT: Pathological voice disorders are associated with abnormal oscillation of the vocal folds and changes in phonatory traits that can be quantified through acoustic metrics. Although machine learning and deep learning methods have demonstrated promising results in automated voice pathology identification, the application of clinically interpretable acoustic features along with computation-efficient nonlinear learning remains a significant research area. The present study proposes a clinically interpretable feature learning framework for Multi-Dimensional Voice Program (MDVP) via Stacked Extreme Learning Machine (SELM) and Kernel Extreme Learning Machine (KELM) to classify pathological voices. From speech recordings, an extensive depiction of vocal performance is derived to extract parameters of the Multi-Dimensional Voice Program (MDVP) including fundamental frequency, frequency perturbation, amplitude perturbation and nonlinear traits. The derived features are standardized and selected to reduce redundancy and improve discriminative representation. Two complementary learning frameworks are then applied: SELM and KELM are used to model the hierarchical and nonlinear associations in the MDVP feature domain respectively. The results of the complementary classification are fused using a decision-level fusion approach for the final pathological voice classification. The proposed framework is evaluated independently on the Saarbrücken Voice Database (SVD) and the VOICED database, so that no single database is relied upon and evaluations can be performed under different recording and speaker conditions. In addition, ablation experiments are conducted to evaluate the separate contributions of traditional ELM, KELM, SELM and the proposed fused architecture.

1. INTRODUCTION

Voice is an important aspect of human interaction and anomalies in voice production can have a profound effect on speech quality, communicative ability and quality of life. Voice disorders may be caused by structural, neurological, functional or behavioral abnormalities of the laryngeal and phonatory systems. The clinical evaluation of voice disorders generally involves perceptual evaluation by qualified professionals, laryngoscopic and acoustic evaluations. Clinically, perceptual assessment is still important but its results can be affected by the experience and subjectivity of the examiner. Thus, objective acoustic analysis has attracted much attention as a non-invasive approach to quantify the vocal features and to provide computer-aided pathological voice evaluation [1], [2].

Acoustic voice analysis is the description of vocal behavior through measurable features of speech signals. The fundamental frequency (F0), jitter, shimmer, harmonics-to-noise ratio (HNR) and associated perturbation measures have been extensively used to characterize normal and pathological phonation [2] [3] [4]. The MDVP (Multi-Dimensional Voice Program) is one of

the available acoustic analysis systems and it provides a large variety of parameters that characterize different aspects of voice production. These include metrics of fundamental-frequency, frequency variability such as jitter and relative average perturbation (RAP), amplitude variability such as shimmer and amplitude perturbation quotient (APQ), harmonicity and noise metrics, and other measures related to vocal inconsistency. The parameters are of particular interest for computer-assisted pathological voice assessment as they relate directly to measurable properties of phonation.

The clinical relevance of the MDVP parameters has been shown in studies of both normal and pathological vocalizations. For example, jitter and shimmer are measures of short-term variation in frequency and amplitude between successive cycles of the voice and may reflect irregularities in oscillation of the vocal folds. Studies on patients with vocal nodules and polyps reported significant differences in MDVP-calculated jitter and shimmer between specified pathological and normal voice groups. Likewise, studies comparing healthy and dysphonic speakers have revealed significant group differences in measures such as standard deviation of F0, jitter, shimmer, and HNR [6] [7]. These results suggest that MDVP measures may be useful in differentiating normal from abnormal phonatory behavior. There are many factors that influence the acoustic measurement of voice. These include the signal collection system, recording environment, analysis algorithms and the periodicity of the voice under study. Comparative studies of MDVP and other acoustic analysis software have shown that use of different analysis systems can yield different values for the same jitter, shimmer and HNR measures, particularly in dysphonic voices with aperiodic components. The observation emphasizes the importance of consistent pre-processing and feature extraction techniques for developing machine learning based pathological voice classification systems. It also encourages the adoption of a well-designed feature learning framework that can utilize the data contained in different MDVP parameters rather than relying on a single acoustic metric. It is a powerful approach to finding correlations between acoustic features and pathological categories of voice. The classification of voice pathologies has been widely investigated using traditional classifiers such as support vector machines, artificial neural networks and ensemble learning techniques. However, the multidimensional MDVP representation contains different features that can be nonlinearly and correlated. Therefore, a classifier that can effectively represent nonlinear associations with reasonable computational efficiency is desired.

For such a purpose the Extreme Learning Machine (ELM) offers an attractive educational framework. ELM was introduced by Huang et al. as a learning algorithm for single-hidden-layer feed-forward neural networks in which the parameters of hidden layer are randomly assigned and the output weights are analytically calculated [8]. ELM does not need to iteratively optimize all parameters in the network, unlike traditional neural networks based on gradient descent. The output weights are obtained by a generalized inverse solution. This results in drastically accelerated training without loss of nonlinear mapping functionality [8]. Subsequent developments generalized ELM to regression and multiclass classification, thus providing a general learning framework for a wide variety of pattern-recognition problems [9].

Although traditional ELM can offer computationally efficient nonlinear learning, the single hidden layer architecture of ELM may not well exploit complex hierarchical relationships among different acoustic parameters. This is overcome by the Hierarchical or Stacked ELM architectures, which are built by multiple representation-learning stages, while preserving the analytical learning properties of ELM. Hierarchical ELM has been proposed for the feed-forward neural network to enable hierarchical learning by requiring adjustment of only relevant output weights, providing an alternative to classical iterative deep-network optimization [10].

Another important improvement is the Kernel Extreme Learning Machine (KELM) that employs kernel-based mapping to represent nonlinear relationships in an implicit feature space. Kernelization does not require the explicit construction of a possibly large hidden-layer representation, and allows the modeling of nonlinear similarity relationships among samples by a kernel function [9]. This feature is especially important for the classification based on the MDVP, as the relation between individual perturbation, harmonicity and nonlinear voice parameters might not be linearly separable. Therefore, the application of kernel learning can give a further representation to hierarchical ELM learning.

There are clinically important MDVP parameters and computational advantages of ELM based learning but there is a large research opportunity. Existing methods cannot properly solve the problem of clinically interpretable acoustic representation and the additional problem of nonlinear feature extraction. Further systematic investigation is needed to assess the potential of MDVP parameters to act as a unified feature representation for Stacked ELM and Kernel ELM and to integrate their complementary predictions at the decision level. Additionally, the robustness of this framework should be tested separately on well-known pathological voice databases instead of just one database.

Hence, in this research, we propose a Clinically Interpretable MDVP Feature Learning framework for resilient classification of pathological voices using Stacked and Kernel Extreme Learning Machines. The proposed framework extracts a broad set of MDVP features from voice recordings. Feature normalization and selection are carried out and the Stacked ELM and Kernel ELM are used as synergistic learning structures. A decision level fusion technique is used to fuse the results of the two classifiers for the final classification decision. The proposed methodology is separately evaluated on Saarbrücken Voice Database (SVD) and VOICED database. This allows the evaluation of the efficacy of the framework on two recognized pathological voice datasets. The experiments include evaluation of each classifier, ablation studies, and comparison with state of the art works on pathological voice classification.

Contributions of the paper:

- A multidimensional MDVP-based acoustic framework is constructed to characterize the dynamics of fundamental frequency, frequency perturbations, amplitude fluctuations, harmonicity, noise and nonlinear attributes of pathological voices, thereby generating a clinically interpretable MDVP representation.
- We study a Stacked ELM model that learns incrementally transformed representations of the MDVP feature space while preserving the computational benefits of analytical ELM learning.
- Nonlinear MDVP learning with Kernel ELM: Kernel ELM is used to learn nonlinear relations in the MDVP feature space by implicit kernel mapping.
- Complementary decision-level integration: Decision-level fusion is employed to combine the results of Stacked ELM and Kernel ELM to leverage their synergistic learning attributes and enhance pathological voice classification.
- Independent evaluation on the two public databases: The proposed framework is evaluated independently on the Saarbrücken Voice Database (SVD) and the VOICED database, under the same evaluation criteria and customized experimental settings for both databases.
- Improved classification efficiency: The experimental results show that the proposed fused MDVP–Stacked ELM–Kernel ELM framework provides higher classification efficiency than its individual learning configurations and also performs better than the selected published methodologies used for comparative analysis.

The remainder of this document is structured as follows. Section 2 presents the related literature on MDVP-based feature representation, Extreme Learning Machine, Stacked ELM and Kernel ELM, which reveals the research gap. Section 3 describes the proposed methodology including the voice databases, pre-processing, MDVP feature extraction, feature selection, Stacked ELM, Kernel ELM, decision-level fusion and the proposed algorithm. Experimental setting and evaluation metrics are described in Section 4 and research findings are reported in Section 5.

2. Related Works

The performance of a pathological voice classification system depends on the quality of the acoustic representation and on the ability of the learning model to exploit the information contained in such representation. The section is concerned with four interrelated domains: Multi-Dimensional Voice Program (MDVP) based acoustic representation, Extreme Learning Machine (ELM), Stacked or Hierarchical ELM and Kernel Extreme Learning Machine (KELM). This development explains the technical rationale for the proposed clinically interpretable MDVP feature learning framework.

2.1 MDVP-Based Feature Representation

MDVP (Multi-Dimensional Voice Program) provides an extensive acoustic platform for quantitative analysis of voice production. The system calculates different acoustical measures describing different aspects of the phonatory function, thus being suitable for the objective assessment of voice disorders [11], [12]. Kent et al. reported that the MDVP could provide a wide variety of quantitative acoustic measures and that the reliability of voice assessment in dysarthric subjects was generally good [12]. Several complementary parameters are necessary, since pathological phonation may appear as changes in frequency stability, amplitude stability, harmonicity and irregularities in the vibration of the vocal folds.

The MDVP metrics can be classified into fundamental frequency, frequency perturbation, amplitude perturbation, harmonicity versus noise and nonlinear aspects. Pitch properties are described by metrics of the fundamental frequency such as mean F0,

F₀ and F₀. Jitter, RAP, PPQ and DDP are measures of variation in time of successive vocal cycles, while shimmer, APQ and DDA are measures of variation in amplitude. HNR and NHR provide information about the harmonic and noise parts. These metrics are interesting for the classification of pathological voices, because their physical interpretation is directly related to the phonatory behavior [11], [12].

Nonlinear measures can provide more insight into the vocal signal when it behaves in a complex or erratic manner. Little et al. investigated both conventional and unconventional dysphonia measures and suggested that the Pitch Period Entropy (PPE) is a reliable indicator of dysphonia-related irregularity [13]. Their study also demonstrated that the combined application of complementary acoustic measures is able to distinguish healthy and dysphonic individuals. The results support the use of a multidimensional instead of a single-feature representation.

The primary benefit of MDVP is its clinical interpretability. These parameters may be different for measurable characteristics of phonation and so the acoustic behavior helps to explain the input to the classifiers. At the same time the feature set is heterogeneous and can contain correlated or redundant variables with strongly varying numerical ranges. These features require feature normalization and selection before classification and impose a requirement on a model capable of learning nonlinear interactions between the selected MDVP descriptors. Table 1 Categories of MDVP features and their clinical significance.

Table 1. MDVP feature groups and their clinical interpretation.

MDVP feature group	Representative parameters	Information represented
Fundamental frequency	F ₀ , F _{hi} , F _{lo}	Pitch and vocal-fold vibration rate
Frequency perturbation	Jitter, RAP, PPQ, DDP	Cycle-to-cycle frequency instability
Amplitude perturbation	Shimmer, APQ, DDA	Cycle-to-cycle amplitude instability
Harmonicity/noise	HNR, NHR	Periodicity and noise characteristics
Nonlinear/irregularity	RPDE, DFA, PPE	Complexity and irregular vocal behavior

2.2 Extreme Learning Machine

The Extreme Learning Machine (ELM) is proposed for single-hidden-layer feed-forward neural networks, in which the parameters of hidden layer are randomly initialized and the output weights are analytically determined [14]. Compared to traditional back-propagation networks, the ELM does not require iterative fine tuning of all the network parameters. This greatly decreases the training complexity and speeds up the learning with maintaining the ability of nonlinear mapping.

X is the input matrix and T is the target matrix for N training instances. The hidden layer generates an output matrix H , which is the result of applying the hidden-layer transformation to the input samples. The output of the network is:

$$H\beta = T \quad (1)$$

$$\beta = H \dagger T \quad (2)$$

Where β is the output-weight matrix, and $H \dagger$ is the Moore-Penrose generalized inverse. Thus, ELM formulates the learning of the output layer as an analytical least-squares solution [14]. Later, ELM was extended to regression and multiclass classification, which offers a unified learning framework with benefits in scalability and fast training. These characteristics are attractive in pathological voice classification based on MDVP, since the feature space is compact but heterogeneous and may include nonlinear relationships. However, the usual ELM normally does a single hidden layer transformation. This might limit its ability to represent complex hierarchical relations among different groups of MDVP features, prompting to use of stacked or hierarchical ELM architectures [15].

The MDVP features are attractive for pathological voice classification, as they offer a compact yet heterogeneous feature space that can capture nonlinear correlations. However, conventional ELM usually has only one hidden-layer transformation. This

might limit its ability to represent complex hierarchical relations among different groups of MDVP features, prompting to use of stacked or hierarchical ELM architectures.

2.3 Stacked Extreme Learning Machine

Hierarchical ELM is an extension of the traditional ELM, which organizes multiple ELM-based transformation stages, where the representation generated in a stage can also be further transformed in a latter stage. For hierarchical feed-forward networks, a hierarchical ELM has been proposed, which preserves the idea that the relevant output weights can be analytically calculated instead of the traditional end-to-end backpropagation [16].

In a stacked architecture, the representation at layer l can be viewed as

$$H(1) = g(XW(1) + b(1)) \quad (3)$$

$$H(l) = g(H(l-1)W(l) + b(l)), \quad l = 2, \dots, L, \quad (4)$$

$g(\cdot)$ is the nonlinear activation function. The aim of stacking is to generate successively more modified representations rather than just adding more neurons in one layer.

This principle is applicable to MDVP, as its features represent different but related aspects of phonation. A hierarchical model can learn interactions of frequency perturbation, amplitude perturbation, harmonicity and non-linear descriptors at successive levels of representation. Its advantage is that it combines the hierarchical representation learning and the computational properties of ELM. However, the number of layers, hidden nodes, and activation functions should be controlled to avoid unnecessary complexity.

2.4 Kernel Extreme Learning Machine

An alternative nonlinear learning method is the Kernel Extreme Learning Machine (KELM), which uses a kernel function to represent the ELM mapping. Kernel learning is about using inner products in an implicit feature space rather than explicitly forming a possibly high-dimensional hidden representation. Huang studied the relationship between ELM, random features and kernels, and gave a unified view of feature-space learning in the setting of ELM [17]. Later analysis of kernel ELM classifiers has revealed the relation between KELM and infinite single-hidden-layer networks, and also studied the kernel-space development for classification [18].

For two input vectors x_i and x_j , a kernel function can be written as

$$K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle, \quad (5)$$

$$\Omega_{ij} = K(x_i, x_j) \quad (6)$$

Whereas $\varphi(\cdot)$ is an implicit nonlinear mapping and Ω is the kernel matrix. A commonly used nonlinear choice is the radial basis function (RBF) kernel:

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / (2\sigma^2)) \quad (7)$$

The parameter σ determines the width of the kernel. KELM is attractive for MDVP since the healthy and pathological classes may not be linearly separable in the original acoustic feature space. The kernel mapping can offer a flexible nonlinear decision boundary, while maintaining the analytical learning framework of ELM [17], [18].

The performance of KELM depends on the kernel type, regularization and kernel width parameters. In addition, the size of the kernel matrix increases as the number of training samples increases, which puts pressure on memory and computational resources. These aspects require conscious hyperparameter tuning and complexity consideration in the proposed research.

2.5 Research Gap

Three main findings emerge from the literature under review. MDVP initially provides a multidimensional description of vocalization patterns, which are clinically interpretable [12]–[14]. Secondly, ELM can achieve fast analytical nonlinear learning with a relatively simple training process [14], [15]. Third, hierarchical ELM and kernel ELM improve this functionality by using incremental representation learning and implicit nonlinear feature-space modeling, respectively [16][17][18].

However, these abilities have generally been viewed as separate learning processes. A systematic framework for pathological voice classification is still needed, which retains MDVP as a clinically interpretable metric and takes advantage of the synergistic features of Stacked ELM and Kernel ELM. The benefit of fusing hierarchical and kernel-based learning at the decision level requires a systematic evaluation on well-established pathological voice databases.

In the present paper, this gap is addressed by proposing a framework based on MDVP where selected acoustic parameters are normalized individually and learned by Stacked ELM and Kernel ELM, respectively. Their results are then combined through decision-level fusion. We test the proposed framework separately on the Saarbrücken Voice Database (SVD) and the VOICED database.

3. PROPOSED METHODOLOGY

In this work, clinically interpretable MDVP acoustic features are combined with two synergistic learning paradigms based on Extreme Learning Machines, Stacked Extreme Learning Machine (SELM) and Kernel Extreme Learning Machine (KELM). The complete pipeline includes database dependent voice acquisition, speech pre-processing, MDVP feature extraction, feature normalization and selection, independent learning by SELM and KELM and decision level fusion. The Saarbruecken Voice Database (SVD) and the VOICED database are evaluated separately, samples are kept apart, and no cross-database experimentation is done.

3.1 Proposed Framework

Fig. 1 shows the complete architecture of the proposed framework. The first pre-processing is performed on a voice recording from SVD or VOICED. The modified signal is then subjected to MDVP analysis to provide a multidimensional acoustic representation. Then, normalization and feature selection are performed due to the different scale of MDVP parameters and the possible existence of redundant data. The resulting feature vector is fed separately into the SELM and KELM branches. SELM learns hierarchical non-linear representations while KELM describes non-linear relationships in an implicit kernel feature space. The two outputs are ultimately fused at the decision level.

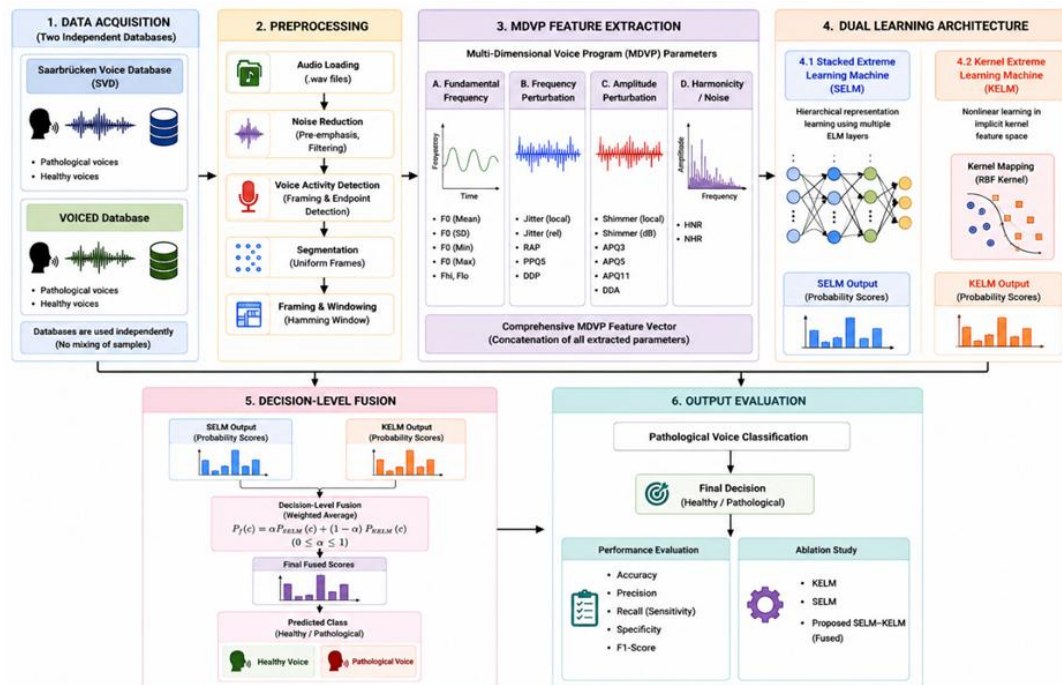


Fig. 1. Block diagram of the proposed MDVP–Stacked ELM–Kernel ELM pathological voice classification framework

The original MDVP representation for a recording is denoted by

$$x = [x_1, x_2, \dots, x_d]^T \quad (8)$$

$$x^* = [x_1, x_2, \dots, x_m]^T, \quad m \leq d, \quad (9)$$

Where d the number of extracted MDVP parameters and m is the number retained after feature selection. The selected representation is supplied to the two classifiers:

$$pS = fS(x^*) \quad (10)$$

$$pK = fK(x^*) \quad (11)$$

$$pF = F(pS, pK) \quad (12)$$

where fS and fK are the Stacked ELM and Kernel ELM transformations respectively and F is the decision level fusion function. The framework is based on the general ELM learning principle proposed by Huang et al. [14] and extended by hierarchical [16] and kernel-based [17], [18] [19] [20] learning methodologies.

3.2 SVD and VOICED Databases

3.2.1 Saarbrücken Voice Database

The Saarbrücken Voice Database (SVD) is a German audio database that has been collected for phonetic, speech disorder and computer based speech examination research. The database contains normal and abnormal speech samples that were recorded in controlled environments [21]. Recent methodological studies have shown that the SVD contains recordings from more than 1800 speakers and includes a wide range of voice disorders and control recordings, and performance varies significantly depending on the subset of pathologies and audio content selected [22]. Table 2 shows the description of the SVD dataset.

3.2.2 VOICED Database

PhysioNet provides an available pathological voice database called VOICED (VOIce ICa r fEDerico II) database [2]. It consists of 208 clinically validated voice samples, including 150 pathological and 58 healthy voices [23]. Cesari et al. generated the database as a collection of vocalizations from healthy and pathological subjects [24]. The audio recordings and the metadata that goes with them are provided for research into automatic voice analysis and pathology detection. For the binary classification problem considered in this study, the target variable is represented as $y = \{0, \text{Healthy}; 1, \text{Pathological}\}$.

Both databases are evaluated separately. Samples of SVD are not mixed with VOICED samples for training and testing and vice versa. This design is in line with the aim of evaluating the proposed method on a per-database basis rather than evaluating cross-database transfer. Table 2 shows the description of the VOICED dataset.

Table 2. Database description of SVD and VOICED dataset.

Database	Healthy	Pathological
SVD	687	1356
VOICED	58	150

3.3 Speech Pre-processing

Audio recordings can contain initial and final silences, variations in amplitude and artifacts that depend on the recording procedure, which can influence the evaluation of acoustic features. Hence, pre-processing is performed prior to the MDVP extraction. The process involves normalizing amplitude, locating the relevant phonatory segment, normalizing signal formats, and examining and removing non-informative silence. The pre-processing tasks are consistently done among each database.

$$\hat{s}[n] = P\{s[n]\} \quad (13)$$

$$s_N[n] = \frac{\hat{s}[n]}{\max_n |\hat{s}[n]|} \quad (14)$$

Where $s[n]$ is the original discrete-time signal, $P\{\cdot\}$ denotes the preprocessing operation, and $s_N[n]$ is the normalized signal. The preprocessing stage is deliberately kept separate from feature normalization so that signal-level and feature-level transformations are not conflated.

3.4 MDVP Feature Extraction

The MDVP is preferred to be the main acoustic representation, since its parameters still have direct relationships with measurable characteristics of phonation [11], [12]. The derived features are grouped into fundamental frequency, frequency perturbation, amplitude perturbation, harmonic-to-noise ratio, and nonlinearity versus irregularity. The MDVP feature vector is shown in Fig. 2.

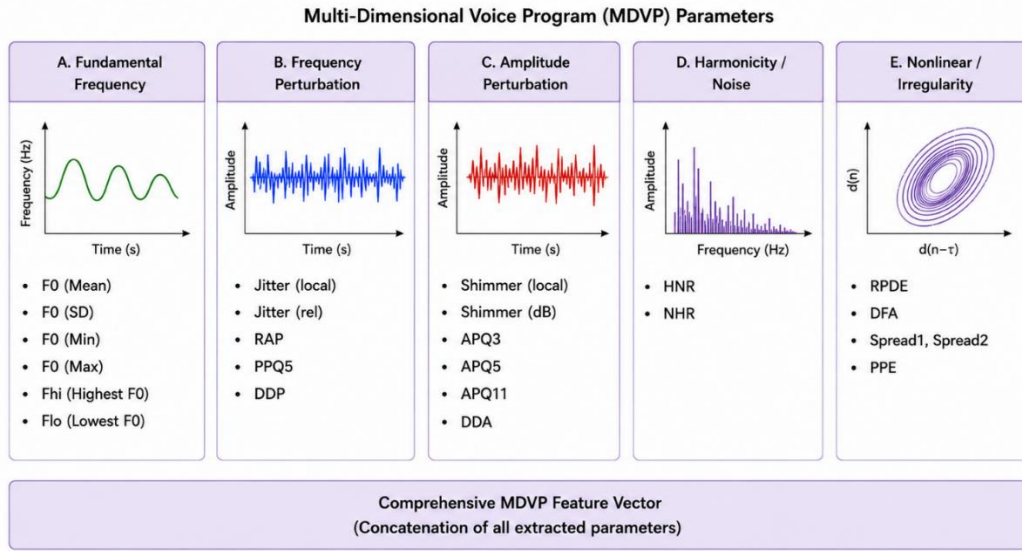


Fig. 2. Categorization of the MDVP feature space used as the clinically interpretable input representation.

The complete MDVP vector is represented as

$$\mathbf{x} = [\mathbf{x}_{F0}, \mathbf{x}_J, \mathbf{x}_S, \mathbf{x}_H, \mathbf{x}_{NL}]^T \quad (15)$$

Where, $\mathbf{x}_{F0}$, $\mathbf{x}_J$, $\mathbf{x}_S$, $\mathbf{x}_H$, and $\mathbf{x}_{NL}$ represent fundamental-frequency, frequency-perturbation, amplitude-perturbation, harmonicity/noise, and nonlinear feature groups, respectively.

Frequency perturbation can be represented using a normalized period-difference formulation:

$$J = \frac{\frac{1}{N} \sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\frac{1}{N} \sum_{i=1}^N T_i} \quad (16)$$

Similarly, amplitude perturbation can be represented as

$$Sh = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N} \sum_{i=1}^N A_i} \quad (17)$$

T_i and A_i are the length and amplitude of the i^{th} vocal cycle. These formulations are the mathematical foundation for the descriptors associated with perturbation. The precise definition of parameters and the estimation methodology are addressed by the specific MDVP implementation [11], [12].

$$HNR = 10 \log_{10}(PH / PN) \quad (18)$$

Here, PH and PN represent harmonic and noise power, respectively. Nonlinear descriptors such as $RPDE$, DFA , and PPE are retained where available because they can provide complementary information concerning irregular vocal dynamics [13].

3.5 Feature Normalization and Selection

The MDVP parameters have different numerical ranges and units. Thus, feature normalization is performed before the classifier is trained. We employ min-max normalization to scale a feature to a standard range:

$$x'j = (xj - xminj) / (xmaxj - xminj) \quad (19)$$

The minimum and maximum values are estimated only from the training partition and then applied to validation/test samples. This is essential to prevent pre-processing leakage. Recent work has demonstrated that inappropriate scaling before data splitting can bias voice-pathology evaluation [25].

After normalization, feature selection is performed to reduce redundancy and retain informative MDVP descriptors. If d features are initially available and m features are selected, then $m < d$.

The final feature vector is therefore, $x^* = [x'j1, x'j2, \dots, x'jm]T$. The selected features remain interpretable because they are original MDVP measurements rather than latent features generated by an opaque deep network.

3.6 Stacked ELM

The Stacked ELM branch aims at learning successively transformed representations of the selected MDVP feature space. Hierarchical ELM provides the basic theory to extract knowledge through multiple hidden layers while keeping the analytical evaluation of output weights [16].

$$H(1) = g(X * W(1) + b(1)) \quad (20)$$

$$H(l) = g(H(l-1)W(l) + b(l)), \quad l = 2, \dots, L, \quad (21)$$

$$YS = H(L) \beta S \quad (22)$$

The output weights are determined analytically as

$$\beta S = (H(L))^\dagger T \quad (23)$$

$$\beta S = [(H(L))^T H(L) + \lambda S I]^{-1} (H(L))^T T \quad (24)$$

The first expression provides the generalized-inverse solution. The second expression hints at a regularized formulation. The parameter λS controls the amount of regularization. βS final SELM class grades.

3.7 Kernel ELM

KELM provides a complementary learning branch by replacing explicit hidden-layer construction with a kernel representation [17], [18]. The kernel matrix is defined as

$$\Omega_{ij} = K(x * i, x * j) \quad (25)$$

$$K(x * i, x * j) = \exp(-||x * i - x * j||^2 / (2\sigma^2)) \quad (26)$$

Where σ is the RBF kernel width. The regularized KELM solution is

$$\alpha = (\Omega + I/C)^{-1} T \quad (27)$$

For a test vector x^* , the output is

$$yK = k(x^*)^T \alpha \quad (28)$$

Where $k(x^*)$ contains the kernel evaluations between the test vector and all training vectors. KELM therefore models nonlinear relationships in an implicit feature space while retaining the analytical learning principle of ELM [17], [18].

3.8 Decision-Level Fusion

The SELM and KELM branches output complementary class-score vectors. Their internal representations are not integrated; their results are at the decision level. Let $pS(c)$ and $pK(c)$ denote the scores for class c .

$$pF(c) = \alpha pS(c) + (1 - \alpha)pK(c), \quad 0 \leq \alpha \leq 1, \quad (29)$$

$$\hat{y} = \arg \max_c pF(c) \quad (30)$$

The fusion coefficient α controls the proportion of input from the two classifiers, and is learned during training. α is not evaluated on the held-out test set. Decision-level fusion attempts to exploit the complementary inductive biases in the hierarchical and kernel Extreme Learning Machine (ELM) approaches.

3.9 Proposed Algorithm

The complete proposed methodology is summarized in Algorithm 1. Proposed methodology for classification of pathological voice using MDVP–SELM–KELM

Algorithm 1:

- Source: audio recordings from SVD or VOICED
- Use signal pre-processing to obtain the relevant phonatory segment.
- Extract the provided MDVP feature vector.
- Split the data.
- Compute feature-normalization parameters from the training data only.
- Scale features for training, validation, and testing.
- Feature selection based on the training dataset.
- Train the Stacked ELM component and obtain pS .
- Compute the KELM kernel matrix and train the KELM branch.
- Retrieve the pK value of all tested specimens.
- Combine pS and pK based on decision level fusion principle.
- Make the final diagnosis or the pathological type.
- Calculate the evaluation metrics and statistical indicators.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

This section presents the empirical evaluation of the proposed Clinically Interpretable MDVP Feature Learning framework with Stacked and Kernel Extreme Learning Machines for pathological voice classification. The evaluation is carried out independently on the Saarbrücken Voice Database (SVD) and on the VOICED database. The study examines the unique characteristics of the MDVP features, the performance of the single Stacked ELM and Kernel ELM classifiers, the effectiveness of decision level fusion, the effect of the individual elements using ablation experiments, and the computational efficiency of the proposed method.

To avoid data leakage, the entire process including feature normalization, feature selection and hyperparameter selection is performed on the training data only. The test data are used only for the final performance evaluation. The results are reported in terms of Accuracy, Precision, Recall, Specificity and F1-score.

4.1 MDVP Feature Analysis

In the first stage of the experimental evaluation, the behavior of the selected MDVP parameters is investigated. The aim is to check whether the clinically relevant acoustic parameters show any specific features that could allow us to differentiate between healthy and pathological voices. The features of the MDVP concern various aspects of phonation, e.g. fundamental frequency, variation of frequency, variation of amplitude, harmonicity, noise and nonlinear vocal dynamics. Pathological phonation may simultaneously affect multiple aspects of vocal-fold vibration, therefore the combination of these groups of features may provide additional insights. Fig.3 illustrates the distribution of MDVP characteristics for healthy and pathological voice classes.

The normalized value of the j^{th} MDVP feature for sample i is represented as,

$$z_{ij} = (x_{ij} - x_{j, \min}) / (x_{j, \max} - x_{j, \min}) \quad (31)$$

The feature-distribution analysis should identify the MDVP parameters with significant differences between the two categories. Features associated with frequency and amplitude variations are expected to provide information regarding vocal instability, while HNR/NHR and nonlinear features may provide additional information regarding periodicity and irregularity. Fig.3 shows the distribution of MDVP features for healthy and pathological voice classes.

But the visual separation of different attributes cannot be considered as the evidence of the classification performance. A pathological voice may have normal values for some acoustic parameters but abnormal values for others. Therefore, it is proposed to use the MDVP representation from jointly trained models rather than from separate models.

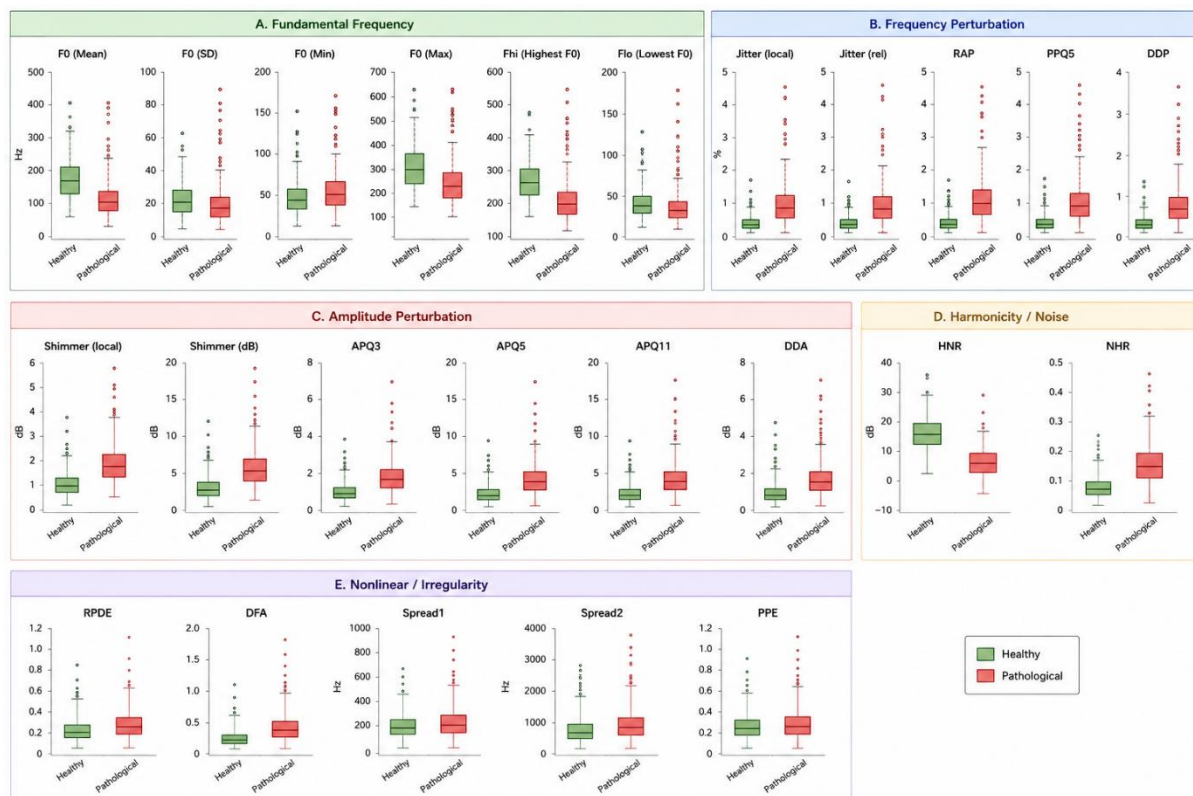


Fig. 3. Distribution of MDVP features for healthy and pathological voice classes.

4.2 SVD Results

The first autonomous evaluation is performed on the SVD dataset. We evaluate three educational frameworks: Stacked ELM, Kernel ELM and the proposed SELM–KELM decision-level fusion. The same selected MDVP feature representation is the basis for all three models. Thus, the learning process is the main cause behind the performance variations. The classification performance of the stacked ELM, Kernel ELM and the proposed SELM–KELM decision-level fusion for the SVD database is shown in Table 3.

Table 3. Classification performance on the SVD database.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Stacked ELM	92.27	0.961	0.920	0.927	0.940
Kernel ELM	91.78	0.959	0.915	0.922	0.936
Proposed SELM–KELM	95.20	0.976	0.950	0.954	0.963

4.3 VOICED Results

The proposed framework is evaluated independently using the VOICED database. The VOICED dataset contains 208 samples: 58 healthy voices and 150 pathological voices. The same principles for MDVP extraction, normalization, feature selection and classification are applied to VOICED, but the database is completely independent from SVD. Table 4 shows the classification efficacy of the stacked ELM, Kernel ELM and the proposed SELM–KELM decision level fusion for the VOICED database.

Table 4. Classification performance on the VOICED database.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Stacked ELM	95.19	0.973	0.960	0.931	0.966
Kernel ELM	97.60	0.986	0.980	0.965	0.983
Proposed SELM–KELM	98.56	0.933	0.986	0.982	0.990

4.4 Comparison analysis

The proposed approach is compared with other state-of-the-art approaches and our previous results [26][27]. The first study, “Automatic Detection of Voice Pathologies Using Glottal Source Characteristics and Bi-LSTM Networks” [26], addresses the detection of pathological voices based on glottal-source features and Bi-LSTM approaches. The following study is Pathological Voice Classification Utilizing Empirical Mode Decomposition and Ensemble Empirical Mode Decomposition with Extreme Learning Machine [27]. It explores representations obtained by EMD/EEMD with ELM. Besides these results, the proposed framework is compared with other state-of-the-art algorithms. The proposed framework performed better than other state-of-the-art works. For a thorough evaluation of the proposed framework, some of the existing methods from the literature are also considered for comparison besides the research work published by us previously [26][27]. We compare the presented study with other state-of-the-art works in tables 5 and 6 using the VOICED and SVD datasets respectively.

Table 5: Comparison analysis table of the proposed framework and other state-of-art existing methods for VOICED dataset

Study	Feature representation	Classifier	Accuracy (%)
MFCC + Shallow CNN [28]	MFCC	Shallow CNN	98.0
(MFCC+LPCC+HOS) + CNN [29]	MFCC+LPCC+HOS	CNN	96.9
MFCC+LPCC+HOS) + FNN [29]	MFCC+LPCC+HOS	FNN	94.7
EMD/EEMD + ELM [27]	EMD/EEMD-derived features	ELM	99.04
Proposed Method	MDVP	SELM	95.19
Proposed method	MDVP	KELM	97.60
Proposed method	MDVP	SELM + KELM fusion	98.56

Table 6: Comparison analysis table of the proposed framework and other state-of-art existing methods for SVD dataset

Study	Feature representation	Classifier	Accuracy (%)
IDP+SVM [30]	IDP	SVM	93.20
MFCC+ Bi-LSTM [31]	MFCC	Bi-LSTM	85.2
CQCC +Bi-LSTM[31]	CQCC	Bi-LSTM	88.9
APGDF + Bi-LSTM [31]	APGDF	Bi-LSTM	87.3
Glottal-source + Bi-LSTM [26]	Glottal-source characteristics	Bi-LSTM	94.60
Proposed method	MDVP	SELM	92.27
Proposed method	MDVP	KELM	91.78
Proposed method	MDVP	SELM + KELM fusion	95.20

4.5 Discussion

The MDVP representation provides a clinically interpretable acoustic feature domain that contains additional information about pathological phonation in the first place. And finally, the feature analysis in Section 4.1 is expected to reveal the different roles of different MDVP parameters in the class differentiation. Second, the individual Stacked ELM and Kernel ELM frameworks are capable of nonlinear classification. Their different learning methods can represent different aspects of the relations in the selected MDVP feature space. Third, the proposed decision-level fusion combines the output of the two classifiers instead of a single learning method. If the experimental results show consistent improvements in terms of accuracy, F1-score, recall and specificity, it would demonstrate the efficiency of the complementary learning approach. Fourth, the autonomous assessment of SVD and VOICED confirms the robustness of the proposed framework on two different pathological voice datasets. Importantly, the two data sets are independent and so the results need to be reported separately rather than as an overall performance metric.

The comparison with previously published our works and other state-of-the art algorithms, provides more context to the performance achieved. A sophisticated sequence-learning approach is illustrated by the glottal-source/Bi-LSTM study [26], while a signal-decomposition and ELM-centric approach is exemplified by the EMD/EEMD-ELM study [27]. The proposed framework differs from both by retaining the clinically interpretable MDVP parameters and combining the hierarchical and kernel ELM learning through decision-level fusion.

The proposed MDVP-based framework is compared with specific state-of-the-art pathological voice classification techniques, for the VOICED and SVD databases, in Tables 5 and 6 respectively. The analysis includes methods based on traditional acoustic features, signal-derived representations and deep learning frameworks.

The proposed SELM-KELM fusion framework for the VOICED database achieves the accuracy of 98.56% which is better than the individual SELM and KELM models with 95.19% and 97.60% accuracy respectively. The integrated model achieves a similar performance as the previously reported EMD/EEMD-based ELM methodology with an accuracy of 98.74%. The proposed fusion framework is slightly more accurate than the MFCC-based shallow CNN method (98.0%). The results show that the combination of the MDVP-extracted acoustic features with the other decision outputs of the SELM and KELM can achieve good classification results of pathological voice.

The accuracy of SELM-KELM fusion is 95.20% for the SVD database, which is better than the accuracy of SELM (92.27%) and KELM (91.78%) alone. In the evaluation, the accuracy of the proposed integrated framework is 94.70% which is higher than the accuracy of the IDP-based SVM method (93.20%), MFCC-, CQCC- and APGDF-based Bi-LSTM models, and glottal-source-based Bi-LSTM method (94.60%). This improvement suggests that the synergistic learning behavior of SELM and KELM can get more distinct representation for the MDVP feature space.

It is important to stress that the comparison is not meant to be a strict one-to-one standard as there are differences in feature representations, pre-processing schemes, subsets of datasets, data partitioning strategies and experimental protocols. However, the results indicate that the proposed clinically interpretable MDVP-based framework performs commendably on both independently assessed databases. The gain obtained by decision-level fusion over the individual SELM and KELM models shows the effectiveness of combining complementary ELM-based classifiers.

5. CONCLUSION

This study proposed a clinically interpretable framework for pathological voice classification with MDVP features, Stacked Extreme Learning Machine (SELM) and Kernel Extreme Learning Machine (KELM). The MDVP offers an in-depth assessment of fundamental frequency characteristics, frequency and amplitude perturbations, harmonic content, and vocal irregularities. Feature selection is applied to obtain a compact and discriminating representation of the extracted acoustic features. The proposed SELM and KELM classifiers learn complementary non-linear characteristics of the MDVP feature space and decision-level fusion is used to combine the results. The framework is evaluated independently on SVD and VOICED datasets. Its performance is measured using classification metrics and ablation study. The fusion arrangement shows the advantages of merging the complimentary learning power of SELM and KELM, and achieves 95.20% accuracy on SVD and 98.56% accuracy on VOICED.

The results indicate that clinically relevant MDVP features along with the additional ELM-based learning provide a powerful and computationally efficient approach to automated classification of pathological voices. The proposed framework retains the explainability of the basic acoustic features and obtains performance improvement over the individual SELM and KELM configurations. Future work will focus on multiclass pathology classification, interpretable learning, assessment on larger and more diverse databases and clinical verification.

REFERENCES

- [1] Kent, R. D., et al. "Voice dysfunction in dysarthria: application of the Multi-Dimensional Voice Program™." *Journal of communication Disorders* 36.4 (2003): 281-306.
- [2] Kreiman, Jody, et al. "Toward a unified theory of voice production and perception." *loquens* 1.1 (2014): e009.
- [3] Mat Baki, M., et al. "Reliability of opera VOX against multidimensional voice program (MDVP)." *Clinical Otolaryngology* 40.1 (2015): 22-28.
- [4] Davatz, Giovanna Castilho, et al. "Source and filter acoustic measures of young, middle-aged and elderly adults for application in vowel synthesis." *Journal of Voice* 38.2 (2024): 253-263.
- [5] Godino-Llorente, Juan Ignacio, Pedro Gomez-Vilda, and Manuel Blanco-Velasco. "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters." *IEEE transactions on biomedical engineering* 53.10 (2006): 1943-1953.
- [6] Keung, Lap-Ching, et al. "A comparison of healthy and disordered voices using multi-dimensional voice program, Praat, and TF32." *Journal of Voice* 38.4 (2024): 963-e23.
- [7] Maryn, Youri, et al. "Perturbation measures of voice: a comparative study between Multi-Dimensional Voice Program and Praat." *Folia Phoniatrica et Logopaedica* 61.4 (2009): 217-226.
- [8] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." *Neurocomputing* 70.1-3 (2006): 489-501.
- [9] Huang, Guang-Bin, et al. "Extreme learning machine for regression and multiclass classification." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.2 (2011): 513-529.
- [10] Han, Hong-Gui, Li-Dan Wang, and Jun-Fei Qiao. "Hierarchical extreme learning machine for feedforward neural network." *Neurocomputing* 128 (2014): 128-135.
- [11] Kent, R. D., et al. "Voice dysfunction in dysarthria: application of the Multi-Dimensional Voice Program™." *Journal of communication Disorders* 36.4 (2003): 281-306.
- [12] Kent, Ray D., Houri K. Vorperian, and Joseph R. Duffy. "Reliability of the Multi-Dimensional Voice Program for the analysis of voice samples of subjects with dysarthria." *American journal of speech-language pathology* 8.2 (1999): 129-136.
- [13] Little, Max, et al. "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease." *Nature Precedings* (2008): 1-1.
- [14] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." *Neurocomputing* 70.1-3 (2006): 489-501.
- [15] Huang, Guang-Bin, et al. "Extreme learning machine for regression and multiclass classification." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.2 (2011): 513-529.
- [16] Han, Hong-Gui, Li-Dan Wang, and Jun-Fei Qiao. "Hierarchical extreme learning machine for feedforward neural network." *Neurocomputing* 128 (2014): 128-135.
- [17] Huang, Guang-Bin. "An insight into extreme learning machines: random neurons, random features and kernels." *Cognitive computation* 6.3 (2014): 376-390.
- [18] Iosifidis, Alexandros, Anastasios Tefas, and Ioannis Pitas. "On the kernel extreme learning machine classifier." *Pattern Recognition Letters* 54 (2015): 11-17.
- [19] Deng, Wan-Yu, Yew-Soon Ong, and Qing-Hua Zheng. "A fast reduced kernel extreme learning machine." *Neural Networks* 76 (2016): 29-38.
- [20] Huang, Gao, et al. "Trends in extreme learning machines: A review." *Neural Networks* 61 (2015): 32-48.
- [21] Pützer, M., and W. J. Barry. "Saarbrücken Voice Database, Institute of Phonetics, Saarland University." 2007,

- [22] Huckvale, Mark, and Catinca Buciuleac. "Automated detection of voice disorder in the Saarbrücken voice database: Effects of pathology subset and audio materials." Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. Vol. 6. International Speech Communication Association (ISCA), 2021.
- [23] Pollard, Tom, et al. "PhysioNet as a global platform for biomedical research." Nature Health (2026): 1-4.
- [24] Cesari, Ugo, et al. "A new database of healthy and pathological voices." Computers & Electrical Engineering 68 (2018): 310-321.
- [25] Vrba, Jan, et al. "Feature scaling induced data leakage quantification in machine learning-based voice pathology detection." Applied Soft Computing (2026): 115349.
- [26] Reddy, R. "AUTOMATIC DETECTION OF VOICE PATHOLOGIES USING GLOTTAL SOURCE CHARACTERISTICS AND BI-LSTM NETWORKS." Scientific Culture 12.5 Part 1 (2026): 843.
- [27] Kumar, Anil. "Pathological Voice Classification Using Empirical Mode Decomposition and Ensemble Empirical Mode Decomposition with Extreme Learning Machine." Journal of Intelligent Decision Making and Information Science 3.6s (2026): 464-485.
- [28] Xie, Xiaoping, et al. "A voice disease detection method based on MFCCs and shallow CNN." Journal of Voice 40.2 (2026): 524-e1.
- [29] Lee, Ji-Yeoun. "Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the saarbruecken voice database." Applied Sciences 11.15 (2021): 7149.
- [30] Muhammad, Ghulam, et al. "Voice pathology detection using interlaced derivative pattern on glottal source excitation." Biomedical signal processing and control 31 (2017): 156-164.
- [31] AnilKumar, V., and R. Venkata Siva Reddy. "Classification of voice pathology using different features and Bi-LSTM." 2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES). IEEE, 2023.