Ingénierie des Systèmes d'Information
Vol. 30, No. 9, September, 2025, pp. 2473-2486
Journal homepage: http://iieta.org/journals/isi

An Efficient Speaker Identification System with High-Level Feature Extraction and Database Dimensionality Reduction

Ahmed Hussein Shatti[1*], Haider A. Mohamed-Kazim[1], Rusul Noori Saraj[2], Ahmed Aldhahab[1]

[1] Department of Electrical Engineering, University of Babylon, Babylon 51001, Iraq
[2] Department of Electrical Engineering, Technical Institute of Babylon, Al-Furat Al-Awsat Technical University, Babylon 51001, Iraq

Corresponding Author Email: eng.ahmed.hussein@uobabylon.edu.iq

https://doi.org/10.18280/isi.300921

Received: 19 July 2025
Revised: 17 September 2025
Accepted: 25 September 2025
Available online: 30 September 2025

Keywords:
speaker Identification, 2D-DMWT-CS, PCA, CNN, dimensionality reduction, feature extraction, RAVDESS

ABSTRACT

Speaker identification is a biometric technology that leverages distinct characteristics obtained from vocal utterances to verify users' identities. Later advancements in multiple fields have raised the importance of speaker identification systems, particularly in security applications. The challenging task in speaker identification systems is how accurately to extract discriminative features from the speech signal. This paper presents a novel approach method that integrates the two-dimensional discrete multi-wavelet analysis-based critical sampling scheme (2D-DMWT-CS) with the principal component analysis (PCA) to employ a reliable and efficient speaker identification system. The proposed method incorporates four phases: preprocessing, feature extraction, dimensionality reduction, and training and classification. During the preprocessing phase, successive refinement techniques such as duration division, silence removal, resampling, and dimension reshaping are applied to the databases. All databases speech samples are then analyzed using the 2D-DMWT-CS. The resultant discriminative features of the wavelet analysis are further processed by the PCA during the supplementary dimensionality reduction phase. The latter provides high-level, hierarchically ordered features that come with a substantial benefit for enhancing the classification accuracy of the convolutional neural network (CNN). The suggested approach was validated by testing and evaluating the framework over many individuals using their speech identities in four online datasets: RAVDESS, TIMIT, ELSDSR and SALU-AC. The achievement results, in terms of the recognition rate, were 97.19% for the TIMIT database, 97.96% for the RAVDESS database, and 98.91% for the ELSDR database, which are higher results than those in the state-of-the-art literature. The reliable and efficient identification rates with high accuracy and fast learning with reducing dimensionality, are the main contributions of this work.

1. INTRODUCTION

Biometrics has seen an increase in popularity in line with the growing curiosity in security. Voice is an accurate and secure biometric that reveals behavioral information about personality characteristics, including nationality, age, sex, and emotional state [1]. In addition to using voice as a biometric, additional distinctive traits, such as the iris, retina, and face, help to distinguish people from each other. The iris, retina, fingerprint, and face are classified as psychological biometrics, while the voice, signature, and keystroke are classified as behavioral biometrics [2]. The performance of each biometric technology is classified in Table 1 in terms of cost, ease of implementation, simplicity of use, and accuracy [3]. According to the information contained in Table 1, it is clear that the voice shows superiority over other biometrics in terms of the aforementioned parameters. Due to the fact that the voice is the most intuitive means of human communication, it expresses the identity of the speaker, including feelings, gender, age, and race. In addition, due to the varying shapes and sizes of human organs such as the larynx and vocal tract, the voice produced is unique for everyone [4]. As such, voice identity is used in speaker identification systems as a robust biometric modality. Over the past sixty years, ongoing research on speaker (voice) recognition has grown substantially, thanks to developments in hardware, architecture, algorithms, and signal processing techniques [5]. Distinguishing between speaker recognition and speech recognition is essential to understanding the key difference between their respective roles. The former is used to identify the persons (speakers), while the latter is the words (speech) rather than speaker identification [3]. Speaker identification (SI) and speaker verification (SV) form the two main categories of speaker recognition. The term "Speaker verification" refers to the process of confirming a speaker's identity by analyzing the details within their speech signal. This process aims to ensure that the client is indeed who they claim to be, resulting in a one-to-one confirmation. In contrast, speaker identification involves determining the identities of individuals who speak anonymously, representing a 1: N

classification scenario [5].

Table 1. Comparison among several characteristics of various physical biometrics

Physical Biometric	Cost	Simplicity in Implemen-tation	Simplicity in Use	Accuracy
Face	Low	Medium	Low	Low
Iris	High	Medium	Medium	Medium
Retina	Medium	Low	Low	High
Fingerprint	Medium	High	Medium	High
Voice	Low	High	High	Medium

The speaker identification system plays a crucial role in ensuring security and authentication, offering a multitude of advantages that extend across various domains and applications, such as enhancing individuals' accessibility and inclusivity, personalizing user experiences, and contributing significantly to operational efficiency by automating processes. According to this rationale, the speaker recognition system has garnered the attention of several researchers, prompting numerous publications with an extensive investigation, as follows:

The authors in reference [6] proposed an approach that is based on optimization to enhance speaker recognition. The study integrates optimization techniques to enhance the achievement of the identification system, which poses a priority importance in various applications such as security and authentication. During the feature extraction phase, the study used the "Multiple Kernel Weighted Mel Frequency Cepstral Coefficient (MKMFCC)", while during the classification phase, the "Support Vector Neural Networks (SVNN)" was employed to classify the extracted features and identify the speaker. In the optimization phase, the weights and biases of the SVNN are optimally tuned using an "Adaptive Fractional Bat (AFB)" algorithm. This algorithm enhances the convergence rate of the standard algorithm. The "English Language Speech Database for Speaker Recognition (ELSDSR)" is used to validate the work. The ELSDSR database comprises voice messages from 22 speakers (12 male, 10 female) with a range of ages between 24 and 63 years, recorded as '.wav' files at a 16 kHz sampling rate. The outcomes of this method in terms of accuracy were 0.95% for 90% of the training data. The authors in reference [7] introduced a new construction called "SECNN (Squeeze-and-Excitation Convolutional Neural Network)", which combines squeeze-and-excitation (SE) elements with the basic "residual convolutional neural network (ResNet)". During the preprocessing and feature extraction phases, the model processes the time-frequency spectrograms as its input. It then measures the similarity between the utterances of each speaker with the models' speaker using cosine similarity. Speaker models are produced by averaging the utterance-level features of each input speaker. The system was evaluated using the TIMIT database (an acoustic-phonetic continuous speech corpus with 630 speakers) and the Librispeech database (a large-scale ASR corpus based on public domain audiobooks, comprising 1000 hours of speech sampled at 16 kHz). The achievable accuracy was 95.83% for TIMIT and 93.92% for Librispeech, respectively. The authors in reference [8] presented an emotional speaker identification system using machine and deep learning models. In the feature extraction phase, the study used the "Mel-frequency cepstral coefficients (MFCCs)", which capture the spectral properties of speech like pitch and loudness. In the training and classification phase, the study compared five machine learning models ("Support Vector Machine (SVM)", "Logistic Regression, Random Forest, XGBoost, and k-Nearest Neighbor (k-NN)"), and three deep learning models ("Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM) network, and Convolutional Neural Network (CNN)"). The performance of these models was evaluated utilizing the database in "RAVDESS", that consists of eight expressed emotions uttered by 24 speakers. The results showed the superiority in performance of the deep learning models with an accuracy of 92% over machine learning models, which attained an accuracy of 88%. The authors in reference [9] proposed a method for performance enhancement of the speaker identification system, particularly under extremely high-pitched conditions. The study used the modified SVM classifier with various speech datasets such as: "Arabic Emirati-accented database" (a corpus includes 50 speakers (30 males, 20 females) with eight utterances each), "Speech Under Simulated and Actual Stress (SUSAS)" (an English database comprises 32 speakers (19 males, 13 females) with 70 utterances each), and "Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)" (an English database includes 24 professional speakers (12 males, 12 females). For feature extraction, the study applied the MFCCs, including MFCCs-delta and MFCCs delta-delta, to extract distinctive features from each audio input and then treated them as input for the modified SVM classifier. The performance of the system achieved an accuracy of 93.95% based on the Arabic Emirati database, 93.31% based on the SUSAS database, and 93.01% based on the RAVDESS database. The authors in reference [10] suggested a comprehensive approach to speaker identification. The method proposed a system that relies on comparing two distinct feature extraction techniques ("the Reconstructed Phase Space (RPS)" and the MFCC) and applying the Random Forest as a classification algorithm. Based on 38 speakers from the TIMIT datasets, the study found that the MFCC, when combined with a Random Forest classifier, yields highly considerable outcomes for speaker identification as compared with the RPS method in terms of accuracy. The authors in reference [11] presented an advanced speaker identification system by combining the "Automatic Spokesperson Recognition (ASR)" with a hybrid machine learning approach. The work combines spectral features, resulting from applying (MFCCs, spectral kurtosis, skewness, NPF, formants), with a "Random Forest-Support Vector Machine (RF-SVM)" classifier. The system had achieved a highly desirable accuracy of 98% for speaker identification based on the database of ELSDSR. In reference [12], the core of the speaker system identification is a convolutional neural network (CNN), where the spectrogram method was used during the feature extraction phase. During the classification phase, the convolutional neural network was employed. To judge the system, the authors used 5 speakers, and each speaker uttered 4 voice samples. The system had achieved an identification rate of about 96.54%. In reference [13], the authors develop an

identification system framework that compares the effectiveness of the MFCC and MSE during the feature extraction phase. In the classification phase, the authors used the Gaussian mixture model (GMM) and support vector machine (SVM) techniques. The ELSDSR database was used to evaluate their system. Based on their findings, it can be concluded that the accuracy of MFCC features is superior to that of MSE features in both classifiers. Additionally, the GMM classifier has better performance than the SVM classifier. The authors in reference [14] proposed the use of the multiresolution analysis (MRA) based on the 2D-DMWT for the feature extraction phase, and the CNN for the training and classification phase. Their outcomes, based on the database used, were 96.30% for SALU-AC, 97.31% for ELSDSR, 96.05% for RAVDESS, and 93.59% for the TIMIT database. The authors in reference [15] explored a multi-level procedure that includes feature-level techniques, dimensionality reduction, and feature optimization strategies. The feature-level fusion (FLV) approach was employed during the feature extraction phase to combine different features to create a more robust representation for the speaker identification task. This is followed by the dimensionality reduction strategies, the principal component analysis (PCA) and the independent component analysis (ICA) techniques, to simplify the extracted data and improve the efficiency. Genetic algorithm (GA) and marine predator algorithm (MPA) were used for further feature enhancement. The proposed method was evaluated across various speech datasets under different noise levels and speaker counts. It achieved 92.7% accuracy on the TIMIT babble noise dataset (120 speakers) and 95.2% accuracy on the VoxCeleb1 dataset based on the PCA-MPA optimization. The authors in reference [16] proposed the use of CNN for feature extraction and the LSTM for classification in speaker identification, achieving an accuracy of 96.52%. This combination effectively captures both spatial and temporal information from audio data. The study demonstrated superior performance in speaker identification by effectively integrating spatial and temporal feature learning, validated against Gaussian Mixture Model (GMM), CNN, and SVM models on the RAVDESS database.

In this paper, a robust approach that integrates the 2D-DMWT-CS (Two-Dimensional Discrete Multi-Wavelet Transform-based Critical Sampling scheme) with the PCA (Principal Component Analysis) and CNN (Convolutional Neural Network) is developed to produce an efficient and reliable speaker identification system. On one hand, the DMWT technique was utilized to extract discriminative features, while PCA was employed to reduce the dimensionality of the extracted features and to produce high-level, descending ordered features. On the other hand, the CNN was used for the classification task, ensuring a high recognition rate. The system was evaluated and tested through four online speech databases, nominated "Salford university anechoic chamber (SALU-AC)", "English language speech database for speaker recognition (ELSDSR)", "Ryerson audio-visual database of emotional speech and song (RAVDESS)", and "TIMIT". The proposed algorithm showed superiority in performance in terms of accuracy as compared with the state-of-the-art literature.

The paper is structured as follows: in Section 2, the main structure of the speaker recognition system is introduced. The essential concepts of the 2D-DMWT and the PCA analysis are discussed in Section 3. The proposed system of speaker recognition is presented in Section 4. The results and their analysis are discussed in Section 5. Finally, Section 6 concludes the paper.

2. MAIN STRUCTURE OF SPEAKER RECOGNITION

The speaker recognition refers to the method of identifying an unknown speaker by matching their voice to others in a recorded database, resulting in a one-to-many comparison [3]. The core framework structure of the speaker recognition system is shown in Figure 1, which consists of two phases: the training phase, also known as the enrolment phase, and the classification phase, also known as the recognition or testing phase. The individual blocks of the main framework of the speaker recognition system are described in the following subsections.

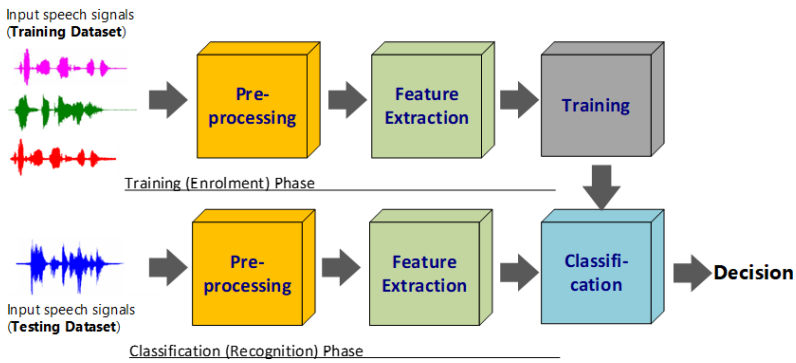

Figure 1. Speaker recognition system- main block diagram

2.1 Preprocessing

Pre-processing is the initial step in speech signal processing, which entails transforming an analogue input into a digital signal. It is a very crucial step when there is noise happening in the speech signal during recording, as incorrect pre-processing of such signals will reduce classification performance [3]. Some of the widely used preprocessing techniques, including signal cropping, resampling, framing, normalization, and reshaping, are used to transform the input speech signal into a suitable form for analysis. The main goal of the preprocessing step is to make the speech signal acceptable for the next step of the feature extraction.

2.2 Feature extraction

The key idea of this process is to extract a sequence of attributes (features) from every speech segment, which is, for

proper modelling, assumed to be short-term and stationary [3]. In this sense, this process preserves relevant and useful information about the speech signal and eliminates unnecessary and duplicate information. Multiple strategies exist in the literature to obtain characteristics from the speech signal as coefficients, including Linear Prediction Cepstral Coefficients (LPCC), Linear Prediction Coding (LPC), and Mel-Frequency Cepstral Coefficients (MFCC) [3].

2.3 Training and classification models

These models are broadly classified into two categories, which are discriminative and generative models [17]. Generative models depict the distribution of distinct classes, whereas discriminative models ascertain the borders that separate those groups. Training and classification models are not just dependent on the training and classification tasks, but also on the available features. Numerous aspects, including speech type, training simplicity, as well as storage and computational demands, must be evaluated prior to selecting the training and classification models [17].

3. ESSENTIAL CONCEPTS

3.1 Discrete multi-wavelet transform (DMWT)

The discrete multiwavelet transform (DMWT) is a mathematical technique for decomposing signals into distinct frequency components. It has developed as the cornerstone in most modern image and signal processing. The conventional discrete wavelet transform (DWT) uses a single scaling function (usually called, father wavelet, denoted as φ) and a single wavelet function (usually called, mother wavelet, denoted as ψ), to visualize the frequency components of the signals at different scales. However, the conventional scalar wavelet system faces inherent limitations and is unable to simultaneously investigate highly desirable properties like orthogonality, symmetry, compact support, and high-order approximation [18]. To address these limitations, the idea of multi-wavelets was presented [19]. The discrete multi-wavelet transform (DMWT) with multiple scaling and wavelet functions is a direct generalization of the conventional DWT and is based on the principle of multiresolution analysis (MRA). The main advantage of the MRA is to represent the signal at various levels of resolution. This is achieved using a sequence of nested subspaces. The central idea behind DMWT is to increase the number of basis functions at each level of multiresolution analysis. Instead of using one scaling and one wavelet functions, the DMWT system uses r scaling functions $\{\varphi_1(t), \varphi_2(t), \dots, \varphi_r(t)\}$ and r wavelet functions $\{\psi_1(t)., \psi_2(t)., \dots, \psi_r(t).\}$, where r is known as the "multiplicity" of the system [19]. For multiwavelets system, $r > 1$, while the standard DWT corresponds to the case of $r = 1$ .For notational convenience, the multiple scaling and wavelet functions can be written using the vector notation in Eqs. (1) and (2), as follows:

$$\Phi(t) = [\varphi_1(t), \varphi_2(t), \dots, \varphi_r(t)]^{\mathrm{T}} \quad (1)$$

$$\Psi(t) = [\psi_1(t)., \psi_2(t)., \dots, \psi_r(t)]^{T} \quad (2)$$

where, $\Phi(t)$ and $\Psi(t)$, denote the vectors of the multi-scaling function and the multi-wavelet function, respectively. In theory, the variable "r" has the potential to possess any size. The commonly used value of r in the literature of the multi-wavelets are often taken with a value of *2*. The multi wavelet and multi scaling functions with multiplicity $r = 2$ can be mathematically written in Eqs. (3) and (4), as in reference [14]:

$$\Phi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} H_k \,.\, \varphi(2t - k) \quad (3)$$

$$\Psi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} G_k \,.\, \psi(2t - k) \quad (4)$$

where, G_k and H_k are practical filters used to decompose the signal at multiple scales. Each filter represents a matrix with $r \times r$ dimensions instead of scalar filters in conventional DWT. The matrix components of these filters provide a number of degrees of freedom more than the traditional scalar wavelet. These extra degrees of freedom can be used to incorporate into the multiwavelet filters with great advantageous properties, such as orthogonality, symmetry, and high order of approximation [20]. The popular and commonly practical used filter in multiresolution analysis is the GHM filter [18]. The former is developed by Geronimo, Hardian, and Massopust, and it comes with unique advantages that incorporate orthogonality, symmetry, and compact support. The GHM system has a multiplicity of $r = 2$, meaning it employs two scaling functions $\{\varphi_1(t), \varphi_2(t)\}$ and two wavelet functions $\{\psi_1(t), \psi_2(t)\}$, which can be written in Eqs. (5) and (6), as in reference [14].

$$\begin{bmatrix}\varphi_1(t)\\ \varphi_{2(t)}\end{bmatrix} = \sqrt{2} \sum\nolimits_k H_k \begin{bmatrix}\varphi_1(2t-k)\\ \varphi_2(2t-k)\end{bmatrix} \quad (5)$$

$$\begin{bmatrix}\psi_1(t)\\ \psi_2(t)\end{bmatrix} = \sqrt{2} \sum\nolimits_k G_k \begin{bmatrix}\psi_1(2t-k)\\ \psi_2(2t-k)\end{bmatrix} \quad (6)$$

The H_k and G_k coefficient matrices in the GHM system consist of four scaling matrices and four wavelet matrices, denoted as H_0 , H_0 , H_0 , H_0 and $G_0, G_1, G_2,$ and G_3 , respectively, and can be written in Eqs. (7), (7') and (8), (8') as:

$$H_0 = \begin{bmatrix}\frac{3}{5\sqrt{2}} & \frac{4}{5}\\ -\frac{1}{20} & -\frac{3}{10\sqrt{2}}\end{bmatrix}, H_1 = \begin{bmatrix}\frac{3}{5\sqrt{2}} & 0\\ \frac{9}{20} & \frac{1}{\sqrt{2}}\end{bmatrix}, \quad (7)$$

$$H_2 = \begin{bmatrix}0 & 0\\ \frac{9}{20} & -\frac{3}{10\sqrt{2}}\end{bmatrix}, \; H_3 = \begin{bmatrix}0 & 0\\ -\frac{1}{20} & 0\end{bmatrix} \quad (7')$$

$$G_0 = \begin{bmatrix}-\frac{1}{20} & -\frac{3}{10\sqrt{2}}\\ \frac{1}{10\sqrt{2}} & \frac{3}{10}\end{bmatrix}, G_1 = \begin{bmatrix}\frac{9}{20} & -\frac{1}{\sqrt{2}}\\ -\frac{9}{10\sqrt{2}} & 0\end{bmatrix}, \quad (8)$$

$$G_2 = \begin{bmatrix}\frac{9}{20} & -\frac{3}{10\sqrt{2}}\\ \frac{9}{10\sqrt{2}} & -\frac{3}{10}\end{bmatrix}, G_3 = \begin{bmatrix}-\frac{1}{20} & 0\\ -\frac{1}{10\sqrt{2}} & 0\end{bmatrix} \quad (8')$$

The use of multiple functions implies important considerations. The corresponding digital filter bank, H_k and G_k, that are responsible for decomposition and reconstruction, now incorporates vector processing. They are designated by sequences of $r \times r$ matrices, instead of using scalar filter coefficients in the conventional DWT. This yields that the input signals must generally be transformed (during a preprocessing phase) from scalar sequences into vector sequences of length r.

In the literature, two popular preprocessing methodologies have been developed for preprocessing the input before applying the wavelet transformation: The *oversampling* and *critical* sampling schemes. These schemes differ in their methods of transforming scalar input into vector form, resulting in different consequences for computational efficiency, redundancy, and approximation characteristics. They can be defined in the following subsections.

3.1.1 Oversampling preprocessing scheme

Oversampling is a preprocessing technique that plans a scalar input sequence of length *N* to a vector sequence of the same length *N,* but each vector element has r components [14]. This method will efficiently increase the number of DMWT's coefficients by a factor of r. For multiple wavelets with multiplicity $r = 2$ and input length N, this oversampling scheme will result in a preprocessing input vector with $2N$ length. The most popular method to implement the oversampling scheme for a given signal is to repeat the signal. This procedure, denoted as "repeated row preprocessing," results in the repetition of the input data by a factor of two.

3.1.2 Critical sampling preprocessing scheme

Critical sampling is a preprocessing method that plans a scalar input sequence of length N to a vector sequence of length N/r, where each vector element has r components [14]. The most popular method to implement the critical sampling scheme for a given signal is called "the first-order approximation preprocessing". This approach preserves the same total number of DMWT's coefficients before and after preprocessing.

3.2 Discrete Multi-Wavelet Transform for 2D signals

By using the critical sampling scheme, the DMWT coefficients' matrix will be in the same dimensions as the input matrix, which should be a square matrix of $N \times N$ dimensions and N must be a power of two, as 2^a, where a represents an integer value. Before examining how the DMWT is constructed under a critical sampling framework, the first-order approximation technique will be used for preprocessing the rows of the input matrix. The operation of the first-order approximation-based row preprocessing is summarized as follows [14]: Let the input matrix be denoted by X, the following preprocessing will be applied at the odd and even numbers of rows:

$$R_{odd}^{new} = (0.373615)\, R_{odd}^{same} + (0.11086198)\, R_{even}^{next} + (0.11086198)\, R_{even}^{previous} \quad (9)$$

$$R_{even}^{new} = \left(\sqrt{2} - 1\right) R_{even}^{same} \quad (10)$$

When calculating for odd rows using Eq. (9), it's important to regard the number preceding the first row as an even-numbered row with a zero value. In the same way, when calculating for the ultimate odd row, the subsequent even row will take a zero value. The computation of the 2D-DMWT using the critical sampling scheme is completely discussed in [14] and can be concluded here by the following steps as follows:

1. Input checking-dimensions: Make sure that dimensions for each applied matrix be equal to $N \times N$, where N must equal to 2^a, and a is an integer number. A padding procedure can be applied to individual rows or columns of the non-square input matrix.

2. Transformation matrix preparation: The transformation matrix is constructed according to the size of the input matrix, but with $(N/r \times N/r)$ dimensions and take the form as in Eq. (11).

$$W = \begin{bmatrix} H_0 & H_1 & H_2 & H_3 & 0 & 0 & \dots \\ G_0 & G_1 & G_2 & G_3 & 0 & 0 & \dots \\ 0 & 0 & H_0 & H_1 & H_2 & H_3 & \dots \\ 0 & 0 & G_0 & G_1 & G_2 & G_3 & \dots \\ H_2 & H_3 & 0 & 0 & H_0 & H_1 & \dots \\ G_2 & G_3 & 0 & 0 & G_0 & G_1 & \dots \end{bmatrix} \quad (11)$$

where H_i and G_i, are the coefficient matrices defined in Eqs. (7) and (8), respectively. After substituting the GHM coefficients, the size of the transformation matrix, W, will becomes $N \times N$ (the same size as the input matrix), since each coefficient matrix has $r \times r$ dimension at $r = 2$.

3. Rows preprocessing: For preprocessing rows, the first-order approximation technique defined in Eqs. (9) and (10) are applied to the $N \times N$ input matrix rows that remarked as even and odd orders, respectively. The size of the matrix after preprocessing remains constant with $N \times N$ dimension.

4. Rows transformation: Can be done as follows:

a-Apply matrix multiplication between the row-preprocessed matrix of size $N \times N$ and the transformation matrix W of size $N \times N$.

b-Permute rows of $N \times N$ obtained matrix by arranging rows according to sequence "1, 2, and $5,6\dots, N-3, N-2$" consecutively, regarding the upper portion of the rows in the resultant matrix, after that, place the pairs of rows such that the sequence "3,4 and $7, 8, \dots, N-1, N$" becomes as the bottom rows in the resultant matrix.

5. Columns preprocessing: To repeat the same procedure used in preprocessing rows.

a-Transpose the resultant $N \times N$ matrix from step 4.

b-Repeat step 3 on the last transposed matrix to obtain an $N \times N$ columns preprocessed matrix.

6. Columns transformation: The column preprocessed $N \times N$ matrix undergoes the following transformation:

a-Multiply the columns preprocessed matrix by the transformation matrix W.

b-Permute the latest resultant $N \times N$ matrix through sequentially arranging rows combinations like "$1, 2,$ and $5, 6\dots N-3, N-2$" to be in the upper half, followed by inserting the row combinations $3, 4,$ and $7, 8\dots N-1, N$ in the lower portion.

7. The DMWT matrix: To obtain the final wavelet transformed matrix, it is necessary to do the following processes:

a-Transpose the resulting matrix in step 6 (the column transformation).

b-Apply the coefficients permutation to the resulting transposed matrix in step 7-(a).

The block diagram of the two-dimensional Discrete Multi Wavelet Transform using critical sampling scheme (2D-DMWT-CS) with first-order approximation-based row preprocessing technique is shown in Figure 2. As seen from the figure, the output matrix of the wavelet transform remains the same as the original input matrix with $N \times N$ dimensions. The resultant matrix behind applying the 2D-DMWT-CS for the two-dimensional preprocessed speech signal with dimensions of 256×256 will be subdivided into four primary sub-bands. Each sub-band possesses dimensions of 128×128. Additionally, each primary sub-band undergoes subdivision into another four secondary sub-sub-bands, where each has dimensions of 64×64. Consequently, the utilization of the 2D-DMWT-CS discovers the ability to use the high-level features existing in the speech signal when considering the LL sub-band from the output matrix.

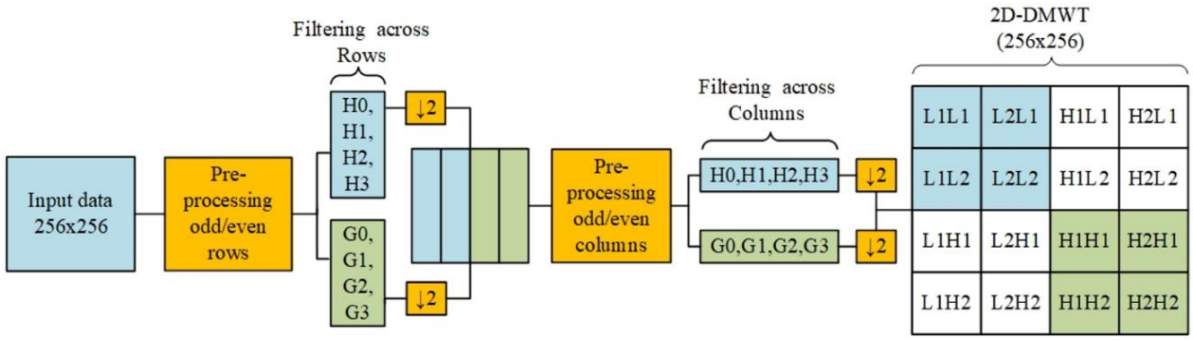

Figure 2. First-level DMWT decomposition for (256\times256\ matrix) using critical sampling scheme with the first-order row preprocessing-based method

3.3 Principal component analysis (PCA)

PCA is a statistical tool that uses an orthogonal transformation to display the dataset in another form. It converts a group of correlated variables into a group of uncorrelated variables by identifying the most prominent paths of data variations [21]. In other words, the PCA transforms the original dataset into a new coordinate system in which axes are arranged in a certain order based on the variations seen in the data. PCA is widely used as a data analysis technique that removes noise, redundancy, and correlation from the dataset [22]. Hence, it can be used as a dimensionality reduction technique as follows:

Let's consider a dataset matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ where n refers to the number of samples (rows) and m denotes the number of features (columns) in each sample. The dataset $\mathbf{X}$ can be viewed as a linear combination of n data rows (samples) given as $\mathbf{x^1}, \mathbf{x^2}, \dots, \mathbf{x^n}$. To reduce the number of features per sample from m into k, where $k \ll m$, the following PCA procedure can be described [23]:

Step 1: Standardize the data rows (samples) by centering and normalizing the data columns in $\mathbf{X}$ as follows:

$$\mathbf{x}_j^i = \frac{\mathbf{x}_j^i - \overline{x}_j}{\sigma_j} \quad j = 1,2,\dots,m, \quad \text{and} \quad i = 1,2,\dots,n \tag{12}$$

where, $\overline{x}_j, \sigma_j$ are the mean and the variance of each column ($\mathbf{x}_{jth}$ feature) in the dataset, respectively. This yields a centered-normalized matrix, which can be denoted as $\mathrm{X_{cn}}$.

Step 2: Calculate the covariance matrix of $\mathrm{X_{cn}}$: using the sample covariance matrix

$$\mathbf{C} = \frac{1}{m}\mathbf{X}_{\mathrm{cn}}^{\mathrm{T}}\mathbf{X}_{\mathrm{cn}} = \frac{1}{m}\sum_i^m \mathbf{x}_j\mathbf{x}_j^T \ , \in \mathbb{R}^{n \times n} \tag{13}$$

Step 3: Calculate the eigenstructure of the covariance matrix $\mathbf{C}$ using the Eq. (14):

$$\mathbf{C}\mathbf{v}_i = \lambda_{\mathrm{i}}\mathbf{v}_i \tag{14}$$

where, $\mathbf{v}_i$, and λ_i , denote the eigenvectors and eigenvalues in $\mathbf{C}$, respectively.

Step 4: Rearrange the eigenvalues and their corresponding eigenvectors of $\mathbf{C}$ in descending order and finding the proportions of each eigenvalue (variance) that correspond to the total variance. The eigenvectors denote the directions of maximal variance (the principal components (*PC*)), whilst the eigenvalues indicate the extent of variation described by each direction. The proportion of each principal component (variance) tells us how the data are spread along those principal components and is given by:

$$\text{proportion of } PC \ = \frac{\lambda_i}{\sum_{j=1}^{f}\lambda_j} \times 100\% \tag{15}$$

Step 5: Project the centered-normalized dataset $\mathbf{X}_{\mathrm{cn}}$ onto the first top k eigendirections to get a matrix with a k-dimensional subspace. The top k eigenvectors will form the new basis for the data and can be written in a matrix denoted by $\mathbf{V}_{\mathrm{k}}$. Hence, the projected matrix will represent the resultant matrix of the PCA, and is given by:

$$\mathrm{P} = \mathrm{X_{cn}^{T}V_k} \tag{16}$$

3.4 The convolutional neural network (CNN)

Convolutional neural networks (CNNs) are among the most advanced deep learning architectures in machine learning. Their ability to handle data using a grid-like structure has pushed progress in computer vision, speech identification, and further fields. It uses a set of kernels or filters in the convolution process, which is mathematically performed by the element-wise multiplications and additions of the kernel parameters with a segment of the input data. The proposed

system in this paper uses a discriminative CNN for training and classification. The biggest advantage of employing CNNs is their ability to mitigate the overfitting issue seen in traditional neural networks. The reason for this is that the CNN possesses a fixed number of parameters within each kernel that are uniformly shared across all input data. In addition to that, the parameters are independent of the number of features in each sample of the dataset [24]. In contrast to the fully-connected layers used in the conventional neural network, the CNN comprises sparse-connected layers, which means that each output value in each layer only depends on a small number of inputs rather than requiring all inputs. The typical architecture for CNN often consists of three major layers, which come in order as convolution, pooling, and fully connected layers. The convolutional layer has a number of discrete trainable kernels (filters), which are employed to extract number of varieties for feature mapping. The individual features are linked to the previous layer's receptive field. The new feature map is generated initially by convolving the input with the kernels. Following that, a function with a nonlinear activation is applied to the convolving procedure's output. The pooling layer takes a small portion of the convolutional layer's output as input and down samples it to obtain a single result. The purpose of the pooling layer is to reduce the computational complexity and dimensionality [24]. The topmost layer of a CNN contains one or more fully connected layers, which are typical of feedforward neural networks, that accept the input from the previous pooling or convolutional layer.

3.5 Online speech databases

The four online speech datasets, namely, RAVDESS, TIMIT, ELSDSR, and SALU-AC, were used in this work to train and test the CNN. These datasets come with attributes which can be briefly described as follows:

3.5.1 RAVDESS dataset

RAVDESS is a group of 24 emotional speeches of public practice speakers (12 males and 12 females). Each speaker has 60 spoken utterances (features) with durations 3 to 4 seconds. The dataset contains eight diverse emotions: joy, sadness, astonishment, calmness, disgust, indifference, fear, and anger. See the link to the web "https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio" to reach the RAVDESS speech dataset.

3.5.2 TIMIT dataset

TIMIT is a collection of 630 speakers of eight American with English acoustic-phonetic accent. There are ten utterances per speaker, each lasting 2 to 3 seconds. The TIMIT dataset's speech sampling frequency was 16 kHz, utilizing 16 bits per sample for quantization. To access the "TIMIT" speech dataset, see the web link "https://www.kaggle.com/datasets/tommyngx/timit-corpus".

3.5.3 ELSDSR dataset

ELSDSR is a group of 22 English speakers (twenty as Danes, one as an Icelander, and one as a Canadian). Twelve males and ten females. Each speaker has 9 spoken utterances. The speech in the ELSDSR dataset was sampled at 16 kHz. To download the ELSDSR dataset, use the web link "http://www2.imm.dtu.dk/~lfen/elsdsr/".

3.5.4 SALU-AC dataset

SALU-AC is a collection of 110 English speakers. Each speaker shares three samples in the speech dataset. The first sample has a duration of 60 seconds, while the other two samples have a duration of 40 seconds. The samples come from reading passages from different resources such as books, newspapers, and others. The SALU-AC dataset is available online on "https://salford.figshare.com/". This work uses 104 speakers from the SALU-AC dataset, with 48 males and 56 females.

4. PROPOSED SYSTEM AND METHODOLOGY

This section presents and discusses the proposed speaker identification system followed in this paper. As depicted in Figure 3, our proposed system consists of four stages: preprocessing, feature extraction, dimensionality reduction, and the training-classification stage. Although the traditional identification systems come with three stages, our proposed system uses an additional inner step after the feature extraction, which is based on the PCA analysis for obtaining high-level features. Four online databases, namely TIMIT, ELSDSR, SALU-AC, and RAVDESS, are employed to assess the suggested identification model. These aforementioned databases handle multiple speech changes involving gender, ambient noise, and age. The first preprocessing stage includes various processing techniques such as duration segmentation, resampling, removing silence, and 1D-to-2D reshaping. The second stage of the proposed system applies the 2D-DMWT-CS for the 2D preprocessed speech signal to obtain the important features from the desired signal. The third stage is concerned with applying the PCA to the extracted features localized in the main LL sub-band of the output matrix of the wavelet transform. Although the main advantage of the PCA lies in dimensionality reduction, it provides another advantage of getting high-level features in descending order of importance and relative information. The latest advantage will contribute to improving the learning process (the training and classification phases) that is done by the CNN at the fourth stage. Our proposed speaker identification system was used to extract the features (utterances) from each speech sample within the databases and confidently recognize the identity of each speaker with reliable accuracy. The four proposed stages are elaborated in the following subsections.

4.1 Preprocessing stage

The preprocessing stage is the initial stage of any speaker identification system that provides convenient speech signals for further processing. Since our databases contain speech signals with different utterances belonging to different persons, it is essential to format the samples within the speech databases correctly. In other words, the preprocessing stage is needed to put the speech samples of each database in an acceptable form for the next step of the speaker identification system, which is, in our case, to put the samples in 2D form for applying the wavelet transform. First, each database is segmented into other databases with different time durations of (0.5, 1, 2, 3, and 5) seconds. Then, each speech sample within the segmented database experiences silence removal, 16KHz resampling, and 1D-to-2D reshaping. Since every speech signal contains repetitive vocal pauses, which are redundant and irrelevant information in the speech signal,

hence, removing such silence pauses will contribute significant improvements in the performance of the proposed system. On the other hand, the resampling at a 16 KHz sampling rate will contribute to making each sample within the database possible to be converted from 1D into a 2D signal with a dimension of 256×256 (power of 2), for the capability of applying the 2D-DMWT-CS.

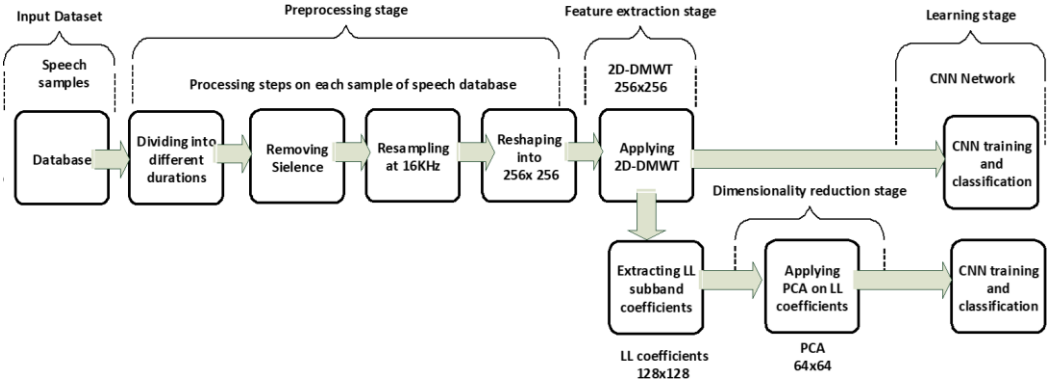

Figure 3. The speaker identifier proposed system

4.2 Feature extraction and dimensionality reduction stages

The common issue in any speaker identification system is that of feature extraction. Feature extraction refers to a technique that transforms a dataset space into a feature set space. This transformation is structured to represent the dataset with a reduced number of effective features while preserving most of the critical information content of the data. This means that the data experience dimensionality reduction. The dimensionality reduction property offers numerous benefits, including data compression and a decrease in computational time, which ultimately enhance the model's performance. Our proposed system employs the two-dimensional discrete wavelet transform-based critical sampling scheme (2D-DMWT-CS) technique for feature extraction and dimensionality reduction purposes. It is one of the most powerful tools that can extract discriminative features from large datasets, and it coherently performs dimensionality reduction.

On the other hand, the principal component analysis (PCA) is used to get high-level features from the speech dataset. It provides a deep feature extraction process at the same time as doing further dimensionality reduction. The majority of discriminative features in the desired signal are concentrated in the main low-low $(L-L)$ sub-band, as shown in Figure 2; therefore, it is advisable to retain the main $(L-L)$ sub-band while excluding the remaining sub-bands. However, according to wavelet analysis, the L_1L_1 sub-sub band represents the matrix obtained from the average of four sub-sub-bands concentrated in the main $(L-L)$ sub-band. In other words, it represents the matrix of the high-level discriminative features, which in our case comes with dimensions 64×64. Hence, for obtaining high-level features from the speech signal with more dimensionality reduction, the main $(L-L)$ sub-band in Figure 2 is further processed by the PCA. The latest will find the main directions of the relevant information by keeping only the L_1L_1, and rearranging them in descending order for training and classification convenience. This in-role will improve the accuracy of the recognition rate while preserving the execution time of the training CNN as low as possible.

4.3 Training-Classification stage

The learning stage in this work contains two main phases: phase #1: the training and classification without using the PCA, and phase #2: the training and classification with the PCA, as shown in Figure 3. Table 2 shows the configuration parameters of the CNN layers employed in this work. The CNN comprises sixteen layers. During phase #1, without using the PCA, the input of the CNN will be the main LL sub-band matrix with a dimension of 128×128. On the other hand, during phase #2: using the PCA for dimensionality reduction, the input of the CNN will be the L_1L_1 sub-sub-band matrix with a dimension of 64×64. To begin the training and classification processes, the CNN receives the input image, which is a 4-dimensional matrix ($N \times N \times 1 \times$number of samples) where $N = 128$ for phase #1, and $N = 64$ for phase#2, when using the PCA. During the training phase of the CNN, a value of 0.01 learning rate was employed with epochs staff equal to 500. For CNN optimization, the $sgdm$ function was used and tested as a best optimizer for executing the optimization operation.

Table 2. CNN configuration layers

Layer, Number: Name	Layer Description
Layer1: input	128×128×1×number of samples, for phase #1. 64×64×1×number of samples, for phase #2.
Layer2: convolution	3×3,24 filters, padding' same'
Layer3: batch normalization	
Layer4: relu	
Layer5: max pooling	
Layer6: convolution	Pool size = 2×2
Layer7: batch normalization	3×3,36 filters, padding' same'
Layer8: relu	
Layer9: max pooling	
Layer10: convolution	
Layer11: batch normalization	Pool size = 2×2
Layer12: relu	3×3,48 filters,padding'same'
Layer13: max pooling	
Layer14: fully Connected Layer	
Layer15: softmax	
Layer16: classification output	Pool size = 2×2

4.4 Algorithm of the proposed system

The overall algorithm steps of our proposed speaker identification system, including database preprocessing and applying 2D-DMWT-CS for feature extraction and the PCA for dimensionality reduction, and the CNN for classification, can be concluded in Algorithm 1, as follows:

Algorithm 1. Preprocessing phase

1. **Input**: Database directory containing (folders of speakers, and files of samples(.wave) per speaker in each speaker folder)
2. **Initialization**: No. of speakers, No. of samples, and a cell array named 'audioDS' for storing audio data of samples
3. **Reading audio data**: Call the database from a path to the (.wav) files
4. For loop:(1: No. of speakers)
5. For loop:(1: No. of samples)
6. Create an 'audioDataStore' object for the current (.wav) file with a 'labelSource' set to the folder name
7. Store each audioDataStore object in the audioDS cell array
8. End for loop of No. of samples
9. End for loop of No. of speakers
10. **Preprocessing audio data**: Resampling at 16 KHz
11. Initialize an empty cell array audioDS_resampled to store resampled (.wav) files
12. Initialize an empty cell array audioDS_new_label to store labels of resampled (.wav) files
13. Set sampling rate to fs_new=16000 Hz
14. For loop:(1: length of audioDS)
15. Read the audio .wav file and its sampling frequency fs_original
16. Remove silence from the .wav file
17. Resample the audio .wav file by new fs_new at 16KHz
18. Overwrite the resampled .wav file on the original .wav file
19. Store the .wav file to the cell array audioDS_resampled
20. Store the label to the cell array AudioDS_new_label
21. End for loop of audioDS

Feature Extraction Phase

22. Initialize a 3D matrix with dimensions (N x N x length of audioDS) for storing the extracted features (set N=256)
23. Initialize a 3D matrix with dimensions (N/4 x N/4 x length of audioDS) for storing the extracted features
24. For loop: (1: length of audioDS)
25. Read the resampled .wav file and its fs_new
26. Set a constant named windowsize with size N x N (256x256)
27. Create a speaker vector
28. If the length of the resampled .wav file $\leqslant$ windowsize
29. Pad resampled .wav file with zeros to N=256 (windowsize length)
30. else
31. Truncate resampled .wav file to N=256 (windowsize length)
32. Reshape the speaker vector into a speaker matrix with NxN (256x256)
33. Apply the 2D- DMWT based critical sampling (2D-DMWT-CS) for the speaker matrix 256x256
34. Store the resultant 2D-DMWT-CS in the 3D matrix (NxN x length of audioDS)

Dimensionality Reduction Phase

35. Input: the 3D matrix of the LL sub band coefficients matrix (N/2xN/2x length of audioDS)
36. Extract the LL sub-band coefficients matrix (N/2xN/2x length of audioDS)
37. Apply the PCA for the extracted features for the LL coefficients matrix (N/2xN/2x length of audioDS), keeping only the first components of size (N/4xN/4 x length of audioDS)
38. Store the resultant 2D-DMWT-CS-PCA in the 3D matrix (N/4xN/4 x length of audioDS) for training
39. End For loop: (1: length of audioDS)

Training and Classification: with and without Using PCA

40. Split the resultant matrix (preprocessed database from phase#1 and phase#2 into training and testing
41. Apply CNN for the 3D matrix in phase #1 (the training and classification without PCA)
42. Apply CNN for the 3D matrix in phase #2 (the training and classification with PCA)

5. FINDING AND DISCUSSION

During this section, the simulation outcomes of the proposed system are discussed, analyzed, and compared with the state-of-the-art works in the literature [6-11, 14-16]. The obtained results from training the CNN with the four databases, nominated SALU-AC, ELSDSR, RAVDESS, and TIMIT, are recorded in Table 3, under different durations of time, and with the number of speakers, 104 for SALU-AC, 22 for ELSDSR, 24 for RAVDESS, and 630 for TIMIT. As seen in Table 3, the SALU-AC and ELSDSR databases are divided into five databases with shorter durations of time "0.5 sec., 1 sec., 2 sec., 3 sec, and 5 sec". The database of RAVDESS comes with 3 seconds in length and is divided into databases with durations "0.5 sec., 1 sec., 2 sec., and 3 sec". The database of TIMIT comes with a duration of 2 seconds and is divided into databases with durations "0.5 sec., 1 sec., and 2 sec". The deep learning model of the proposed system was trained with these databases separately. The accuracy metric describes how accurately the deep learning model identifies the speakers during the classification phase; hence, it is used to evaluate the performance of the proposed system. The recognition rate of the proposed system in terms of the accuracy metric for each database with different time durations is shown in Table 3.

Table 3. Results in terms of accuracy

Database Name	PCA?	Accuracy in (%)				
		0.5 sec.	1 sec.	2 sec.	3 sec.	5 sec.
SALU-AC	No	99.55	98.45	97.48	97.15	95.37
	Yes	99.67	98.63	97.84	97.96	95.74
ELSDSR	No	98.60	98.01	95.97	95.01	94.30
	Yes	98.91	98.51	96.12	95.99	95.10
RAVDESS	No	95.76	95.26	92.60	90.58	------
	Yes	97.96	96.64	95.71	93.75	------
TIMIT	No	95.90	89.67	89.29	------	------
	Yes	97.91	89.99	89.59	------	------

To compare the proposed system with the state-of-the-art literature, some of the related training and classification figures are chosen as those shown in Figures 4-10 with 0.5 and 1 second sample durations. As seen in Table 3, when the sample's duration increased, the accuracy of the classification decreased. In other words, the length of the database is another factor in the classification results of the recognition system. This is due to the fact that when there are many features available in each sample, the learning model will suffer from the high bias problem, which leads to a decrease in the accuracy of the system. To assess the work's performance, the results of the suggested model are evaluated in comparison with [6-11, 14-16] in Table 4 in terms of database used, feature extraction method, classification method, and resultant recognition rate, as shown in Table 4. The recognition rates in

Table 4 are expressed by the training and classification accuracies shown in Figures 4-10. When the accuracy is high, this indicates a high recognition rate and vice versa. As shown in Table 4, the recognition rates of the proposed model present superiority in performance over other state-of-the-art works. The reason behind that comes with the use of the 2D-DMWT-CS-PCA combination that provides high-level feature extraction and dimensionality reduction.

As seen in Table 4, in the case of the RAVDESS database, the authors in reference [8] have employed hybrid techniques combining the time-frequency and cepstral domains to obtain the crucial features from the desired signal and the multi-layer perceptron (MLP) method for classification. It is seen that in Table 4, the proposed model investigates a recognition rate of 97.96%, which is higher than the work in reference [8], see Figure 5. Also, as noticed from Figures 4 and 5, the recognition rate with the PCA is higher than the system without using the PCA. The reason behind that is the use of the 2D-DMWT-CS in conjunction with the PCA to enable more distinctive feature extraction.

Table 4. Comparison results

Data-base	Work in Literature	Feature Extraction Method	Classification Method	Recognition Rate in (%)
RAVDESS	(**Proposed**)	2D-DMWT-CS + PCA	CNN	**97.96**
	[8]	Hybrid techniques	MLP	92
	[9]	Modified-MFCC	SVM	93.01
	[14]	2D-DMWT	CNN	96.05
	[16]	CNN	LSTM	96.52
TIMIT	(**Proposed**)	2D-DMWT-CS + PCA	CNN	**97.91**
	[10]	MFCC	random forest	97
	[7]	MFCC	SECNN	95.83
	[14]	2D-DMWT	CNN	93.59
	[15]	FLV+PCA+ICA	MPA	92.7
ELSDSR	(**Proposed**)	2D-DMWT-CS + PCA	CNN	**98.51**
	[6]	AFB	SVNN	95
	[11]	Hybrid techniques	RF-SVM	98.16
	[14]	2D-DMWT	CNN	97.31

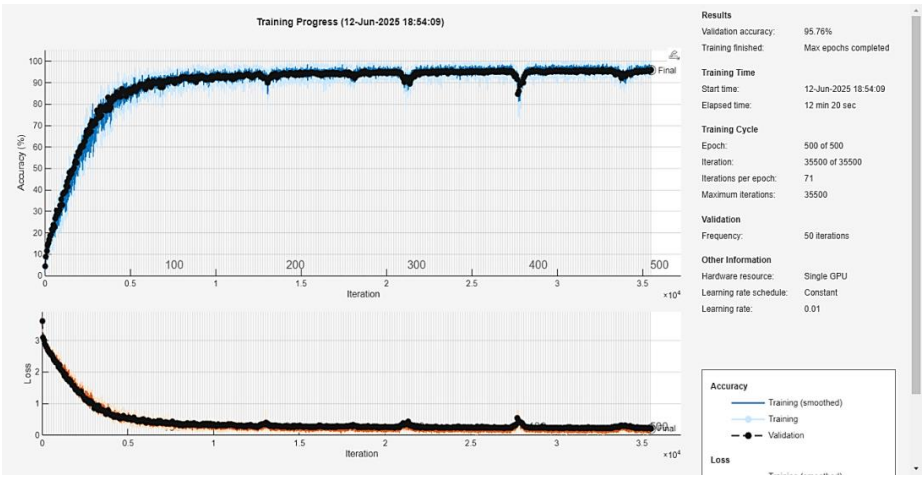

Figure 4. The training process using a 0.5-second length RAVDESS database without PCA

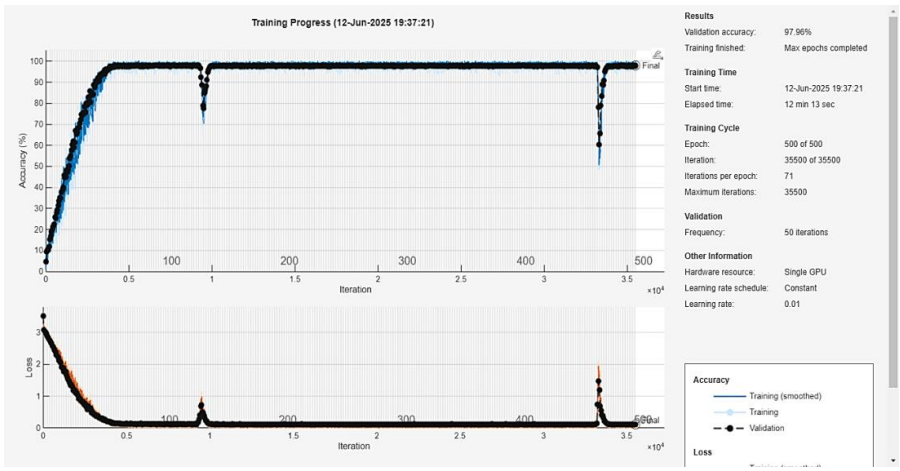

Figure 5. The training process using a 0.5-second length RAVDESS database using PCA (proposed)

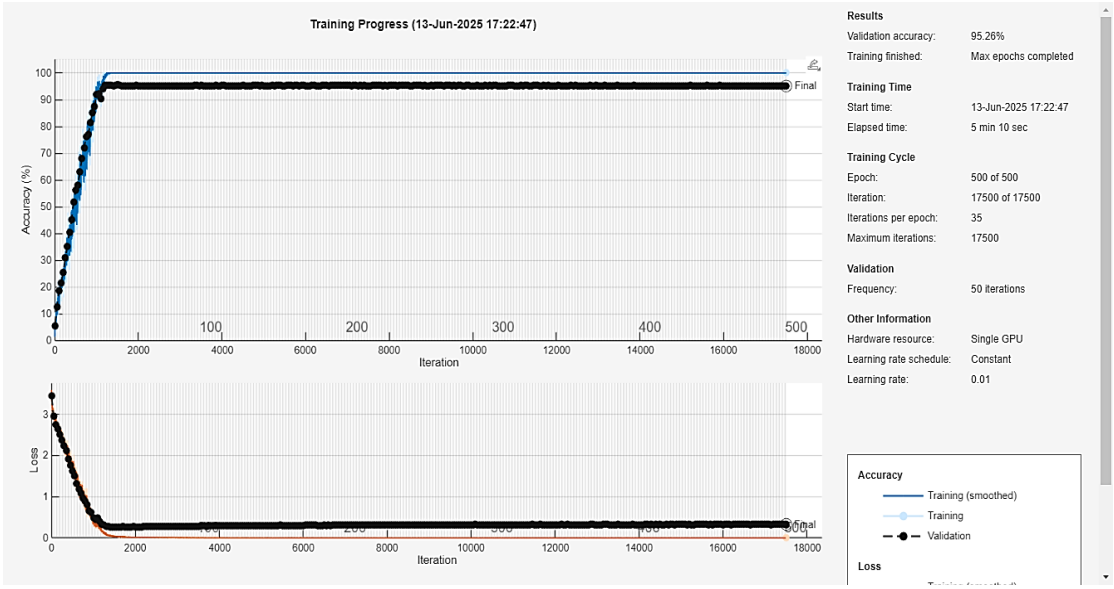

Figure 6. The training process using a 1-second length RAVDESS database without PCA

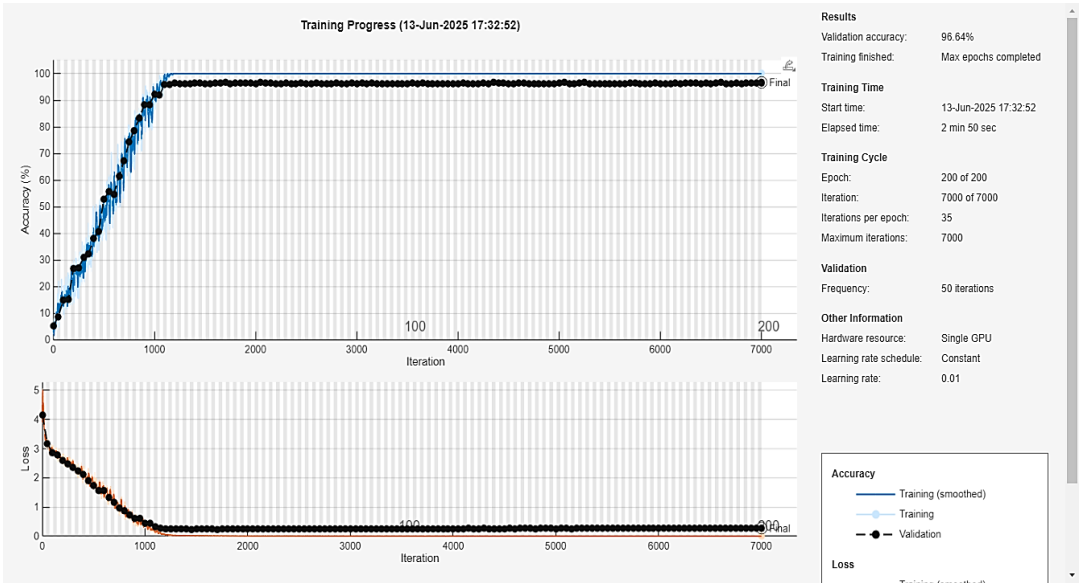

Figure 7. The training process using a 1-second length RAVDESS database with PCA (proposed)

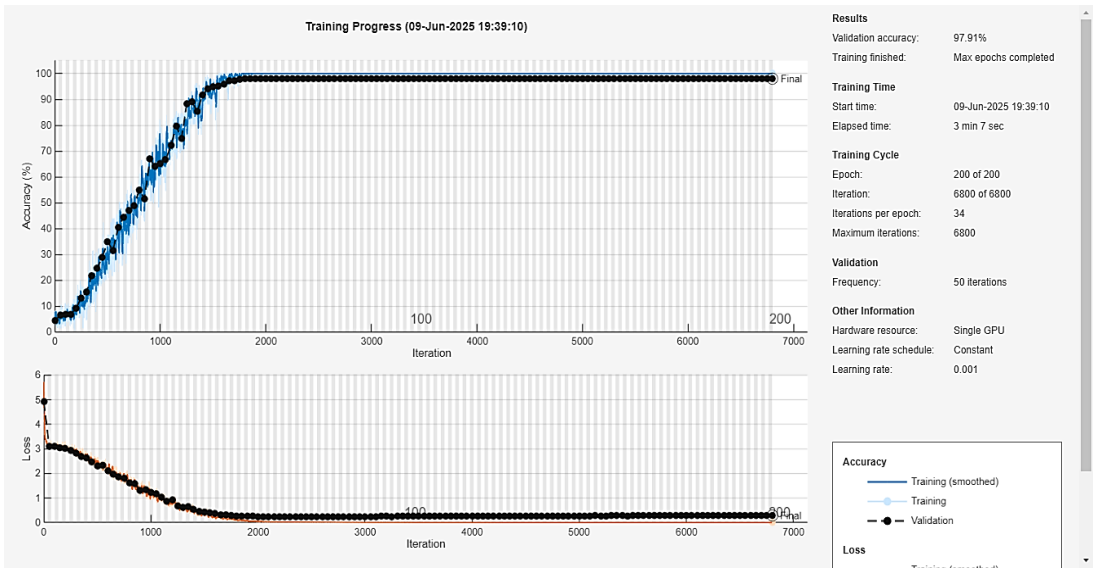

Figure 8. The training process using a 0.5-second length TIMIT database with PCA (proposed)

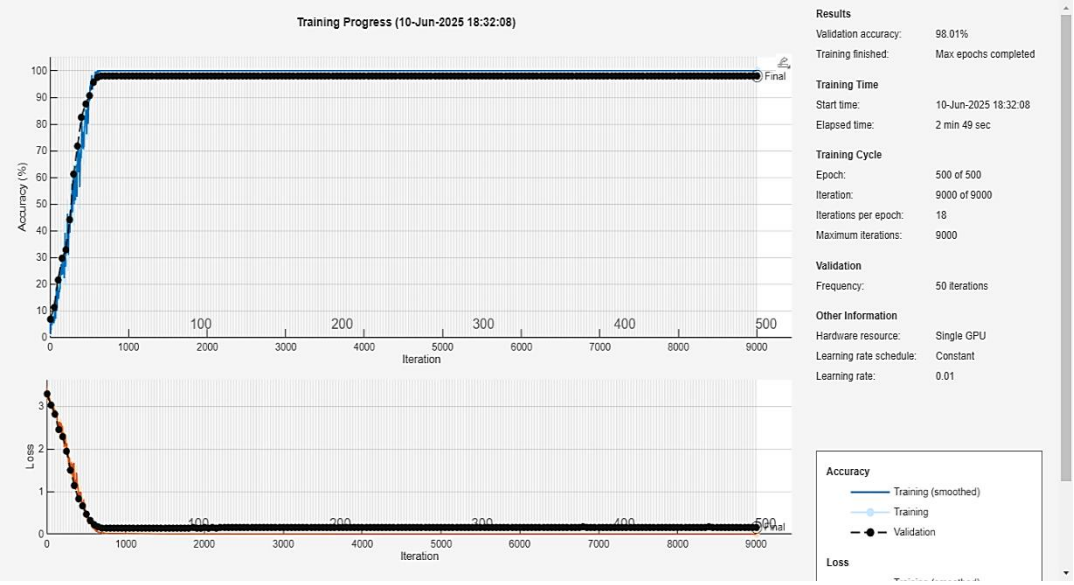

Figure 9. The training process using a 1-second length ELSDR database without PCA

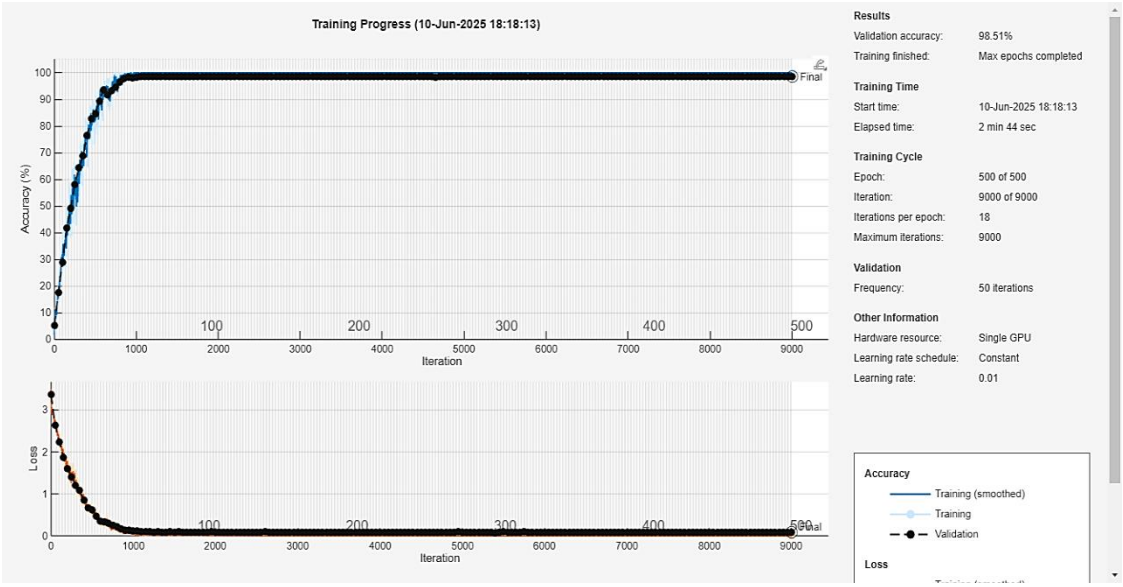

Figure 10. The training process using a 1-second length ELSDR database using PCA (proposed)

Our proposed method supports high-resolution properties that make feature extraction feasible with high-level features from the input speech signal. The authors in reference [9] have employed the modified method of Mel Frequency Cepstral Coefficients (MFCC) for feature extraction, while the Support-Vector-Machine (SVM) algorithm was used for classification purposes. To ensure impartial comparison, the study in reference [9] used only 12 samples for every speaker in the database, while in our work, different samples per speaker were used depending on the length of the RAVDESS database. For example, in the case of 0.5 sec. RAVDESS length, the number of samples used is 380, and in the case of 1 second. RAVDESS length, the number of samples used is 190. In other words, the number of samples per speaker is inversely proportional to the length of duration of the database and plays as another factor that affect the results of the model's accuracy, see Figures 6 and 7. As seen in Table 4, the proposed system outperforms the method applied in references [14, 16].

In the case of using the TIMIT database, the authors in reference [10] have employed the method of Mel frequency cepstral coefficients (MFCC) for feature extraction, supported by the random forest method for classification. The approach in reference [10] followed a fair comparison among speakers, where only 38 speakers from the TIMIT database were used. In our proposed system, 630 speakers with different segments of duration are used. As seen in Table 4, the proposed system outperforms the method applied in references [7, 14, 15], while it shows a comparable recognition rate (up to 97.91) as compared with the work in reference [10], see Figure 8.

In the case of using the ELSDSR database, the authors in reference [6] employed a combination of techniques, including multiple kernel weighted MFCC (MKMFCC), spectral skewness, spectral kurtosis, and autocorrelation, for the feature extraction stage of the speaker identification system. Although a hybrid method was followed in the literature [6], our proposed system, which uses the 2D-DMWT in conjunction with the PCA and CNN, has achieved a higher recognition rate (up to 98.51 in case of 1-second duration) as seen in Table 4 and Figure 10. On the other hand, Figures 9 and 10 shows that the recognition rate with using PCA is

higher than those without using the PCA The proposed method also attains a higher recognition rate as compared with the work in references [11, 14].

Regarding the results in Table 4, it is clearly seen that the conjunction between the 2D-DMWT and PCA techniques with the CNN leads to higher results in recognition rates as compared with the works in literature [6-11, 14-16]. This corresponds to the powerful properties of the DMWT described by the orthogonality, symmetry, and compact support. In addition to that, the DMWT offers many advantages: perfect reconstruction while preserving orthogonality, linear phase symmetry, and higher order approximation. Due to the highly desirable features offered by the DMWT, this technique will make the opportunity of improving the performance of the system very high. On the other hand, the resilience techniques in the preprocessing stage of the speech signal, such as speech database length splitting, silence removal, and speech resampling were had a significant and direct impact on the learning process and the performance of the proposed system. Also, using the PCA as a technique for dimensionality reduction and for getting high-level and descending ordered features will improve the performance of the system in terms of accuracy and consequently recognition rate.

5.1 Complexity and dimensionality reduction

The complexity of computations is crucially related to the number of additions and multiplications during the training and classification phase. Hence, the dimensions of the database will play a major role in this venue. As the dimensions of the database increase, the complexity will increase. On the other hand, the training time will also increase. The overall dimensionality reduction of the proposed system can be measured by the following Eq. (17):

$$D_{Reduction} = 1 - \frac{M_{resultant}}{M_{input}} \times 100\% \quad (17)$$

where, $M_{resultant}$ is of 64×64 matrix that represents the resultant matrix after applying the PCA, and M_{input} denotes the input matrix, in our case, it represents the reshaped speech sample of any speaker from the processed database, and it has a size of 256×256. After substitutions, the dimensionality reduction in Eq. (6) will be $D_{Reduction} = 93.75\%$. These results will add another advantage to our work, which is preserving the storage space of the extracted features, as compared with the state-of-the-art literature in [14-19, 22-24], where the dimensions are the same before and after feature extraction methods. Due to these significant reductions in dimensions of each speech sample, the complexity of computations and the training time will decrease significantly. As shown in the Figures 4-10, it is shown that the training time is approximately reduced to half, see Figures 6 and 7 in the case of RAVDESS with and without using the PCA.

6. CONCLUSION

This paper demonstrates a robust methodology for employing an efficient identification system constructed on using the 2D-DMWT-CS for feature extraction, the PCA for dimensionality reduction, and the CNN for training and classification. Refinement preprocessing techniques, such as duration division, silence removal, resampling, and dimension reshaping, were applied to the database before the feature extraction phase. The proposed model harnesses the highly desirable properties of the DMWT, such as orthogonality, symmetry, compact support, and dimensionality reduction, to extract high-level discriminative features from the speech signal. On the other hand, uses the PCA to provide another dimensionality reduction percentage. This has resulted in an enhancement in the CNN training and classification process, yielding high recognition rates. The reason behind that is the descending order of discriminative features generated by the PCA, which contributes to improving the CNN learning.

The proposed system has been assessed with the well-known online speech databases nominated SALU-AC, ELSDSR, RAVDESS, and TIMIT, and exhibited significant recognition rates as compared with the other literature in Table 4. The proposed model has investigated up to 93.75% dimensionality reduction, which in turn contributed to reducing the time of training and the classification process. The potential limitations such as the sensitivity to speech length and noise environments may be overcome in this work, since a decimation preprocessing step for database speech samples have used to divide the sample's duration into small durations, which have proven in the results sections in providing high accuracy and fast learning, in addition to that, using of the 2D-DMWT have verified its immunity to noise. For future work, the researchers can propose a hybrid model that incorporates the DMWT with other methods, such as MFCC to extract the features, and then use the PCA for dimensionality reduction, and the temporal prediction model using CNN-LSTM for the learning process.

REFERENCES

[1] Shetty, P., Rodricks, R., Malgundkar, S., Pamnani, H., Katke, S. (2023). Speech biometrics: A comprehensive deep learning-based speaker identification system. In 2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, pp. 1-5. https://doi.org/10.1109/MIUCC58832.2023.10278329

[2] Sarkar, A., Singh, B.K. (2020). A review on performance, security and various biometric template protection schemes for biometric authentication systems. Multimedia Tools and Applications, 79(37): 27721-27776. https://doi.org/10.1007/s11042-020-09197-7

[3] Hanifa, R.M., Isa, K., Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. Computers Electrical Engineering, 90: 107005. https://doi.org/10.1016/j.compeleceng.2021.107005

[4] Zhang, Z. (2016). Mechanics of human voice production and control. The Journal of the Acoustical Society of America, 140(4): 2614-2635. https://doi.org/10.1121/1.4964509

[5] Yuan, X., Li, G., Han, J., Wang, D., Tiankai, Z. (2021). Overview of the development of speaker recognition. Journal of Physics: Conference Series, 1827(1): 012125. https://doi.org/10.1088/1742-6596/1827/1/012125

[6] Srinivas, V., Santhirani, C. (2020). Optimization-based support vector neural network for speaker recognition. The Computer Journal, 63(1): 151-167. https://doi.org/10.1093/comjnl/bxz012

[7] Qi, M., Yu, Y., Tang, Y., Deng, Q., Mai, F., Zhaxi, N. (2020). Deep CNN with se block for speaker recognition.

In 2020 Information Communication Technologies Conference (ICTC), Nanjing, China, pp. 240-244. https://doi.org/10.1109/ICTC49638.2020.9123307

[8] Sefara, T.J., Mokgonyane, T.B. (2020). Emotional speaker recognition based on machine and deep learning. In 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Kimberley, South Africa, pp. 1-8. https://doi.org/10.1109/IMITEC50163.2020.9334138

[9] Al Hindawi, N.A., Shahin, I., Nassif, A.B. (2021). Speaker identification for disguised voices based on modified SVM classifier. In 2021 18th International Multi-Conference on Systems, Signals Devices (SSD), Monastir, Tunisia, pp. 687-691. https://doi.org/10.1109/SSD52085.2021.9429403

[10] Nawas, K.K., Barik, M.K., Khan, A.N. (2021). Speaker recognition using random forest. ITM Web of Conferences, 37: 01022. https://doi.org/10.1051/itmconf/20213701022

[11] Karthikeyan, V., Suja Priyadharsini, S. (2022). Hybrid machine learning classification scheme for speaker identification. Journal of Forensic Sciences, 46(3): 1033-1048. https://doi.org/10.1111/1556-4029.15006

[12] Nugroho, K., Noersasongko, E. (2022). Enhanced Indonesian ethnic speaker recognition using data augmentation deep neural network. Journal of King Saud University-Computer and Information Sciences, 34(7): 4375-4384. https://doi.org/10.1016/j.jksuci.2021.04.002

[13] Abakarim, F., Abenaou, A. (2022). Comparative study to realize an automatic speaker recognition system. International Journal of Electrical and Computer Engineering, 12(1): 376-382. https://doi.org/10.11591/ijece.v12i1

[14] Al-Dulaimi, H.W., Aldhahab, A., Al Abboodi, H.M. (2023). Speaker identification system employing multi-resolution analysis in conjunction with CNN. International Journal of Intelligent Engineering Systems, 16(5): 350-363. https://doi.org/10.22266/ijies2023.1031.30

[15] Chauhan, N., Isshiki, T., Li, D. (2024). Enhancing speaker recognition models with noise-resilient feature optimization strategies. In Acoustics, 6(2): 439-469. https://doi.org/10.3390/acoustics6020024

[16] Suryamritha, M., Balaji, V., Kannan, S., Murali, K. (2024). Speaker identification using CNN-LSTM model on RAVDESS dataset: A deep learning approach. In 2023 4th International Conference on Intelligent Technologies (CONIT), Bangalore, India, pp. 1-6. https://doi.org/10.1109/CONIT61985.2024.10626802

[17] Yan, H., Lei, Z., Liu, C., Zhou, Y. (2024). GMM-ResNeXt: Combining generative and discriminative models for speaker verification. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, pp. 11706-11710. https://doi.org/10.1109/ICASSP48485.2024.10447141

[18] Lowe, B., Salman, H., Zhan, J. (2022). Ghm wavelet transform for deep image super resolution. arXiv preprint arXiv:2204.07862. https://doi.org/10.48550/arXiv.2204.07862

[19] Strela, V. (1996). Multiwavelets--theory and applications (Doctoral dissertation, Massachusetts Institute of Technology). http://hdl.handle.net/1721.1/10631.

[20] Cui, L.H. (2005). Some properties and construction of multiwavelets related to different symmetric centers. Mathematics and Computers in Simulation, 70(2): 69-89. https://doi.org/10.1016/j.matcom.2005.04.001

[21] Greenacre, M., Groenen, P.J., Hastie, T., d'Enza, A.I., Markos, A., Tuzhilina, E. (2022). Principal component analysis. Nature Reviews Methods Primers, 2(1): 100. https://doi.org/10.1038/s43586-022-00184-w

[22] Hasan, B.M.S., Abdulazeez, A.M. (2021). A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1): 20-30. https://doi.org/10.30880/jscdm.2021.02.01.003

[23] Reddy, G.T., Reddy, M.P.K., Lakshmanna, K., Kaluri, R., Rajput, D.S., Srivastava, G., Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. IEEE Access, 8: 54776-54788. https://doi.org/10.1109/ACCESS.2020.2980942

[24] Bačanin Džakula, N. (2019). Convolutional neural network layers and architectures. In Sinteza 2019-International Scientific Conference on Information Technology and Data Related Research, pp. 445-451. https://doi.org/10.15308/Sinteza-2019-445-451