Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2024, Article ID 7184018, 12 pages
https://doi.org/10.1155/2024/7184018

Research Article

Emotion Modeling in Speech Signals: Discrete Wavelet Transform and Machine Learning Tools for Emotion Recognition System

K. Daqrouq,[1] A. Balamesh,[1] O. Alrusaini,[2] A. Alkhateeb,[1] and A. S. Balamash[1]

[1]Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, Saudi Arabia
[2]Department of Engineering and Applied Sciences, Umm Al-Qura University, Makkah 24382, Saudi Arabia

Correspondence should be addressed to K. Daqrouq; haleddaq@gmail.com

Received 7 January 2024; Revised 15 February 2024; Accepted 6 March 2024; Published 2 April 2024

Academic Editor: Youcef Belkhier

Speech emotion recognition (SER) is a challenging task due to the complex and subtle nature of emotions. This study proposes a novel approach for emotion modeling using speech signals by combining discrete wavelet transform (DWT) with linear prediction coding (LPC). The performance of various classifiers, including support vector machine (SVM), K-Nearest Neighbors (KNN), Efficient Logistic Regression, Naive Bayes, Ensemble, and Neural Network, was evaluated for emotion classification using the EMO-DB dataset. Evaluation metrics such as area under the curve (AUC), average prediction accuracy, and cross-validation techniques were employed. The results indicate that KNN and SVM classifiers exhibited high accuracy in distinguishing sadness from other emotions. Ensemble methods and Neural Networks also demonstrated strong performance in sadness classification. While Efficient Logistic Regression and Naive Bayes classifiers showed competitive performance, they were slightly less accurate compared to other classifiers. Furthermore, the proposed feature extraction method yielded the highest average accuracy, and its combination with formants or wavelet entropy further improved classification accuracy. On the other hand, Efficient Logistic Regression exhibited the lowest accuracies among the classifiers. The uniqueness of this study was that it investigated a combined feature extraction method and integrated them to compare with various forms of combinations. However, the purposes of the investigation include improved performance of the classifiers, high effectiveness of the system, and the potential for emotion classification tasks. These findings can guide the selection of appropriate classifiers and feature extraction methods in future research and real-world applications. Further investigations can focus on refining classifiers and exploring additional feature extraction techniques to enhance emotion classification accuracy.

1. Introduction

The goal of emotion recognition is to understand and interpret human emotions accurately. Undoubtedly, this subject captures a great deal of attention due to its wide use spectrum that spans different domains. To illustrate this, in case of human-computer interaction, emotion recognition makes systems to be adjustable and react to emotional nature of the user which, thus, gives the experience the boost. The use of emotion recognition system can be of great assistance in the areas of psychiatry and psychology. It may have application in market research, customer feedback analysis, and social media site feeling analysis.

Emotion recognition typically involves three main steps: various data acquisition methods are used, and both feature extraction and classification are conducted. Data acquisition involves capturing emotional cues from sources such as images, audio recordings, or physiological sensors. Feature extraction involves transforming the acquired data into meaningful representations that capture relevant emotional information. Finally, classification algorithms are employed to classify the extracted features into specific emotional categories, such as happiness, sadness, anger, or surprise.

Emotion recognition is mostly based on machine learning algorithms such as SVM, neural networks, and clustering methods which enable building accurate emotion discerning models. The models are taught or induced to learn these particular patterns in the given labeled datasets containing varied emotions from which they can make predictions even on the new dataset containing unseen data. Overall, emotion recognition plays a vital role in understanding human behavior, enabling more sophisticated human-machine interactions, and facilitating applications that require an understanding of emotional state classification tasks.

Because of its usefulness in areas like computer-human interaction, emotion driven computing, and healthcare, speech emotion recognition (SER) has gained interest recently. The main goal of the SER task is identifying the emotional elements showcased in a person's speech pattern. Pitch, formants, and energy are examples of manually created characteristics that are extracted from the speech signal using traditional SER methods. After that, a classifier is trained using these features. However, these hand-crafted features might not be able to convey the complex and subtle nuances in speech that indicate different emotions. Since wavelet transforms may automatically extract distinctive characteristics straight from the raw voice signal, they have shown considerable promise in speech recognition applications. With their widespread application in SER, the combination of WT and machine learning techniques has produced cutting-edge outcomes.

In this paper, we provide an enhanced SER method that integrates machine learning with the wavelet transform. The DWT has demonstrated success in SER and is a strong method for obtaining time-frequency information from speech data. We can capture the intricate connections between the extracted features and the underlying emotions by combining the LPC calculated from wavelet framed subsignals and classified by several machine learning models. To evaluate our proposed method, we conducted experiments using the EMO-DB dataset, which is a widely recognized benchmark dataset for SER.

The main significance of this study is as follows. We provide an improved SER approach that integrates machine learning methods with wavelet transform (WT). Our proposed method is assessed using the EMO-DB dataset, yielding state-of-the-art results. We carry out ablation experiments to examine the effects of various elements within our suggested approach. The uniqueness of this study was that it investigated a combined feature extraction method and integrated them to compare with various forms of combinations. However, the purposes of the investigation include improved performance of the classifiers, high effectiveness of the system, and the potential for emotion classification tasks.

This paper is organized as follows. An introduction and extensive survey of pertinent SER literature are provided in Section 1. Section 2 provides a thorough explanation of our suggested approach. Section 3 presents the discussion of the experiment results. Concluding remarks and suggestions for future work are presented in Section 4.

1.1. Motivation. Our motivations for researching the proposed method, which combines DWT, LPC, and machine learning classifiers for emotion recognition from speech signals, include the following:

(1) *Improved Accuracy*. Enhancing the accuracy of emotion recognition is crucial for affective computing and human-computer interaction. Our goal is to enhance the accuracy and dependability of identifying emotional states from speech signals. This can be accomplished by integrating DWT, LPC, and machine learning techniques.

(2) *Comprehensive Representation*. The integration of DWT and LPC allows for capturing both temporal and spectral characteristics of speech signals. This comprehensive representation enables extraction of features that are both relevant and discriminative, facilitating the capture of subtle variations and nuances in emotional expression.

(3) *Practical Applicability*. Emotion recognition has practical applications in psychological research and clinical diagnosis. Creating a dependable and precise approach yields crucial resources for researchers, therapists, and healthcare experts to comprehend and evaluate emotional conditions. This, in turn, can result in enhanced psychological wellness and better support for mental health.

(4) *Technological Advancement*. Integrating signal processing techniques like DWT and LPC with machine learning models represents an innovative approach to emotion recognition. By furthering these methodologies, we are actively contributing to the growth of technology in both speech signal analysis and emotion recognition. This will ultimately facilitate more extensive research and development in closely related fields.

The motivation behind our research stems from the need to address existing limitations in emotion recognition from speech signals. Despite the fact that a great deal of effort has been expended in this direction, there is still a need for further work, especially focusing on accuracy, feature representation, and aesthetics. We are going to develop a new explicit algorithm of DWT, LPC, and machine learning classifier to face some challenges mentioned before. Through this integration, not only can we model both temporals, but also frequency or spectral features of the speech signal, providing a more detailed representation of emotion. By leveraging the advantages of low-power technology, wavelet transformation, and machine learning, our objective is to enhance the accuracy and reliability of emotion recognition readings, thereby advancing the field.

Given the innovation of the paper, the originality is the specific implementation of DWT, LPC, and machine learning algorithms for classifying audio signals for identification of emotions. This novel approach combines signal processing technologies with machine learning models, challenging the limitations and capabilities of existing emotion recognition systems. By integrating these techniques, we aim to overcome current constraints and enhance the accuracy and capabilities of emotion recognition systems. Thus, we are stepping aside to contribute to technology development through the demonstration of this approach's efficiency. We introduce novel trends of speech signal analysis and emotion recognition. The result of our research is the enlargement of options that are characterized with a deeper understanding of emotions and therefore can be more reliable and can be used for affective computing, human-computer interaction, and mental health diagnosis and treatment.

In summary, our motivations focus on improving accuracy, achieving a comprehensive representation of emotional expression, enabling practical applications, and advancing technology in the field of emotion recognition from speech signals.

1.2. Contributions. The contribution of our research comprises the following:

(1) We develop a novel method that combines DWT, LPC, and machine learning classifiers for emotion recognition from speech signals.

(2) Our method improves the accuracy and reliability of emotion classification, leading to better identification of various emotional states.

(3) Our method extracts comprehensive features by capturing both temporal and spectral characteristics of speech signals through WT coefficients and LPC.

(4) Our system offers a wide range of tools that enhance affective computing, human-computer interaction, psychological research, and clinical diagnosis.

(5) We contribute to advancements in speech signal analysis and emotion recognition by pushing the limits of signal processing methods and machine learning models.

1.3. Literature Survey. Several studies have explored speech emotion recognition. For example, one study proposed a lightweight method called 1BTPDN, which applies a one-dimensional discrete wavelet transform (DWT) to extract low-pass filter coefficients from raw audio data. Textural analysis methods, such as one-dimensional local binary pattern (LBP) and one-dimensional local ternary pattern (LTP), are then applied to each filter to extract features. The most relevant features are selected using neighborhood component analysis (NCA) and fed into a support vector machine (SVM) classifier, achieving high recognition rates in various databases [1]. Another study used wavelet packet-based principal component analysis (WP-PCA) for feature extraction and optimized support vector machine (SVM) using genetic algorithm for speech emotion recognition. The experiment showed a recognition rate of 95% using the Chinese Academy of Sciences language library [2].

More recently, the use of DWT in face emotion recognition has gained momentum. Several studies employed DWT for emotion recognition from speech [3–5]. These studies include Mel Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coefficient (LCD), and wavelet-based parameters along with DWT coefficients in this regard. One of the most used classifiers when it comes to emotion classification is the SVM classifier. Combined with other algorithms (e.g., MFCC), the mixture increases the reliability of results. Some studies report accuracy levels of up to 82.14% and 85% using DWT-based techniques for emotion detection. This shows that DWT may be an important indicator in detecting emotional information through speech.

These areas find applications in psychology, education, and human-computer interaction [6]. Decision trees, support vector machines, and neural networks are some of the advanced machine learning approaches that have correctly recognized emotions from speech samples [7]. This process involves utilizing features like pitch and MFCCs, along with various techniques for feature extraction, selection, and classification [8]. Deep learning models such as CNNs and LSTMs have seen impressive advancements and demonstrated remarkable performance in emotion recognition. Interestingly, MFCCs seem to be the most appropriate features for this purpose [9]. In the paper [10] titled "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier," the authors proposed a method for recognizing speech emotions. The method combined hybrid spectral-prosodic features extracted from the speech signal and glottal waveform, along with metaheuristic-based dimensionality reduction techniques and a Gaussian elliptical basis function network classifier. In the paper [11] titled "Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm," the authors propose a method for speech emotion recognition. Their approach focused on employing a modified quantum-behaved particle swarm optimization (QPSO) algorithm for discriminative dimension reduction. The paper presented a novel approach that leveraged a modified QPSO algorithm for discriminative dimension reduction in speech emotion recognition.

2. Method

2.1. Feature Extraction Method. Several studies have applied signal processing using discrete wavelet transform (DWT). One study applied DWT for signal noise reduction on a FPGA-based chip design platform and integrated audio encoder on a FPGA board [1, 2]. One study developed a new algorithm for DAW based on DWT yielding better SNR and BER rates compared with other approaches [3]. Another study used DWT for the classification of the heartbeats depending on the static and dynamic features extracted from the ECG signals. Such classification was achieved with high accuracy using a Radial Basis Function Neural Network classifier [4]. DWT was also employed in compressing ECG time-series datasets resulting in high information compression ratios and excellent signal reproductions [5].

In this paper, a combination of LPC and DWT for feature extraction is proposed. A time-frequency analysis method called the DWT breaks down the speech signal into multiple sub-bands. Each sub-band captures specific temporal and spectral characteristics of the speech signal. The LPC is a speech analysis technique that represents the speech signal as a linear combination of previous samples. The LPC coefficients are estimated for each frame of five consecutive frames within each sub-band of the DWT decomposition.

The DWT is expressed as follows:

$$\mathrm{CA}(j+1) = \sum_{nk\_\mathrm{min}=1}^{k\_\mathrm{max}} (h[k] * \mathrm{CA}(\mathrm{j})[2\mathrm{n}-1]), \tag{1}$$

$$\mathrm{CD}(j+1) = \sum_{k\_\mathrm{min}=1}^{k\_\mathrm{max}} (g[k] * \mathrm{CD}(\mathrm{j})[2\mathrm{n}-1]). \tag{2}$$

Equation (1) presents the approximation coefficients, and equation (2) presents the detail coefficients. In the above formula, $\mathrm{CA}(j+1)[n]$ denotes the approximation coefficients at level $j+1$, which capture the low-frequency components of the signal. $\mathrm{CD}(j+1)[n]$ denotes the detail coefficients at level $j+1$, which capture the high frequency contents or details of the signal.

The coefficients $h[k]$ and $g[k]$ are the filter coefficients, also known as the scaling and wavelet filters, respectively. These filters are typically chosen from well-known wavelet families, such as Daubechies, Haar, or Symlets.

The summation is performed over the filter taps, or coefficients k, and the coefficients $\mathrm{CA}(j)[n]$ represent the approximation coefficients at level j, obtained from the previous level of decomposition.

The DWT performs iterative decompositions, starting with the original signal at level $j=0$ and computing the approximation and detail coefficients at each level until the desired level of decomposition is reached.

Once the signal has been decomposed using the DWT, it can be reconstructed by applying the inverse DWT (IDWT), which involves sampling the coefficients and applying the appropriate synthesis filters [12–14].

We can refer to the convergence property of the DWT by the ability of the transform to accurately represent a signal as the number of decomposition levels increases. In other words, as we perform more iterations of the DWT, the approximation and detail coefficients obtained approach the original signal.

This overlapping factor is proven using the conservation principle. The energy of a signal is the total power or magnitude of the signal which includes all its instantaneous values. For the DWT, the property of keeping the energy of the signal helps us to maintain the energy of the original signal in the decomposing and reconstructing process which is done by using the DWT.

And now let us express the energy conservation property of the DWT mathematically as

$$\begin{aligned} ||x||^2 = {} & ||\mathrm{CA}(J)||^2 + \|\mathrm{CD}(J)||2 + ||\mathrm{CD}(J-1)||^2 \\ & + \ldots + ||\mathrm{CD}(1)||^2. \end{aligned} \tag{3}$$

Here, $||x||^2$ represents the energy of the original signal, and

$$||\mathrm{CA}(J)||^2, ||\mathrm{CD}(J)||^2, \ldots, ||\mathrm{CD}(1)||^2, \tag{4}$$

represent the energy of the approximation and detail coefficients at each decomposition level.

The convergence property can be observed by increasing the decomposition level number J. One the aspect which is received from the fact that as the decomposition level number increases, there is a reduction in the energy contribution in the detail coefficients with a one aspect which is that the energy contribution in the approximation coefficients remains relatively high. This indicates that the approximation coefficients capture the low-frequency components of the signal, while the detail coefficients represent the high-frequency details.

The implementation of the convergence is drawn, and we picture the reconstruction of signal by the inverse DWT (IDWT) from different layers of decomposition. We then change the number of decomposition levels, and the reconstructed signal becomes developed and thus almost looks like the original signal. The convergence behaviors of DWT can be highly affected by the choice of wavelet functions as well as the implementation details. Different wavelet families may introduce certain trade-offs between time and frequency localization, which can impact the accuracy of the signal reconstruction.

The discrete wavelet transform is a scientifically established and stable transformation technique for signal processing. The DWT ensures boundedness and energy preservation through the careful design of wavelet filters. The stability of the DWT is guaranteed by the properties of the wavelet filters. The filters possess finite support, ensuring that the transformed coefficients remain bounded. Additionally, the filters exhibit good frequency localization, allowing the DWT to accurately capture signal details without excessive amplification or distortion. The decay of high-frequency coefficients further supports the stability of the DWT, indicating that the transform effectively represents the high-frequency components of the signal.

The LPC equation is given by

$$s\_\mathrm{predition}(i) = \sum_{m\_\mathrm{min}}^{m\_\mathrm{max}} (b[m] * s(d-m)). \tag{5}$$

In this equation:

(1) $s\_\mathrm{predicted}(i)$ denotes the predicted sample at index i.

(2) $b(m)$ denotes the LPC coefficients.

(3) $s(d-m)$ denotes the past samples of the speech signal.

(4) *m*_min and *m*_max define the valid range of *m* values based on the order of the LPC model.

For each sub-band, we extract five frames of the sub-band signal. For each frame, we estimate the LPC coefficients. Then, we average the LPC coefficients of the five frames to obtain a single feature vector for each sub-band. The averaged features for all subsignals are collected in one feature vector that represents the entire signal.

The advantage of this feature extraction technique is that it enables reliable and discriminative extraction of the speech signal's time-frequency and spectral characteristics at the same time.

2.2. Database. The EMO-DB database is a collection of German emotional speech recordings. The database encompasses audio recordings of ten actors, each manifesting seven distinct emotional states, namely, anger, boredom, disgust, anxiety, happiness, sadness, and neutrality. Each actor recorded five repetitions of each emotion, resulting in a total of 535. The speech utterances were recorded with a top-notch microphone in a sound-proofed room. The signals were saved in WAV format after being sampled at a rate of 16 kHz.

2.3. Experimental Setup. For evaluating the proposed method, the EMO-DB dataset is used. EMO-DB dataset is a well-known testing database for recognition tasks, which includes ten actors' recordings of emotional speech. The database consists of five males and five females. To ensure that the emotional distribution remains consistent in both sets, a 5-fold cross-validation technique is used.

2.4. Feature Extraction. For each sub-band obtained, LPCs were extracted using a frame size of five frames and a step size of one frame. Twelve LPCs were computed for each frame. The resulting features for each subsignal are twelve coefficients. The whole features are collected from all the subsignals to form a single feature vector for each utterance.

2.5. Model Training and Evaluation. To evaluate the proposed method, SVM, KNN, MLP, and CNN are used for classification. The proposed classifiers are trained by the training set to be tested over the testing set.

True Positive Rate (TPR). It is the ratio of true positive cases compared to the total number of true positive and false positive cases, calculated as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$.

False Positive Rate (FPR). It is the ratio of false positive cases compared to the total number of true positive and false positive cases, calculated as $\mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$.

Precision. The precision is calculated as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$.

Accuracy. It is calculated as $(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$.

To see how the classifier performed per class, a TPR or FNR option can be used. The TPR is the proportion of correctly classified observations per true class. The FNR is the proportion of incorrectly classified observations per true class. Figure 1 shows summaries per true class in the last two columns on the right for sadness.

2.6. Classification. The classification of emotions was conducted using trained models. For each utterance, each emotion is detected from all remaining six emotions. Here is a brief description of each classifier [15–18].

2.6.1. SVM

(1) Total cost (validation): 28.

(2) Prediction speed: –1500 obs/sec.

(3) Training time: 29.502 sec.

(4) Model size (compact): 855 kB.

(5) Hyperparameters: cubic SVM with automatic kernel scale, box constraint level of 1, and one-vs-one multiclass coding. Data are standardized.

2.6.2. KNN

(1) Total cost (validation): 30.

(2) Prediction speed: ~450 obs/sec.

(3) Training time: 38.962 sec.

(4) Model size (compact): ~2 MB.

(5) Hyperparameters: cosine KNN with 10 neighbors, cosine distance metric, equal distance weight, and standardized data.

2.6.3. Efficient Logistic Regression

(1) Total cost (validation): 55.

(2) Prediction speed: ~1300 obs/sec.

(3) Training time: 20.934 sec.

(4) Model size (compact): ~76 kB.

(5) Hyperparameters: Efficient Logistic Regression with automatic solver and regularization, relative coefficient tolerance of 0.0001, and one-vs-one multiclass coding.

2.6.4. Naive Bayes

(1) Total cost (validation): 35.

(2) Prediction speed: –36 obs/sec.

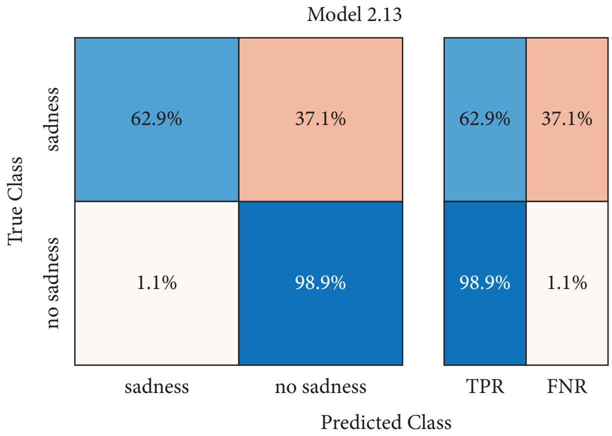


Figure 1: The classifier performance per class, summarizing TPR and FNR in the last two columns on the right, for sadness.

(3) Training time: 104.28 sec.
(4) Model size (compact): −9 MB.
(5) Hyperparameters: Kernel Naive Bayes with Gaussian kernel, unbounded support, and standardized data.

2.6.5. *Ensemble*

(1) Total cost (validation): 29.
(2) Prediction speed: −370 obs/sec.
(3) Training time: 22.424 sec.
(4) Model size (compact): −23 MB.
(5) Hyperparameters: Subspace Discriminant ensemble method with 30 learners and a subspace dimension of 216.

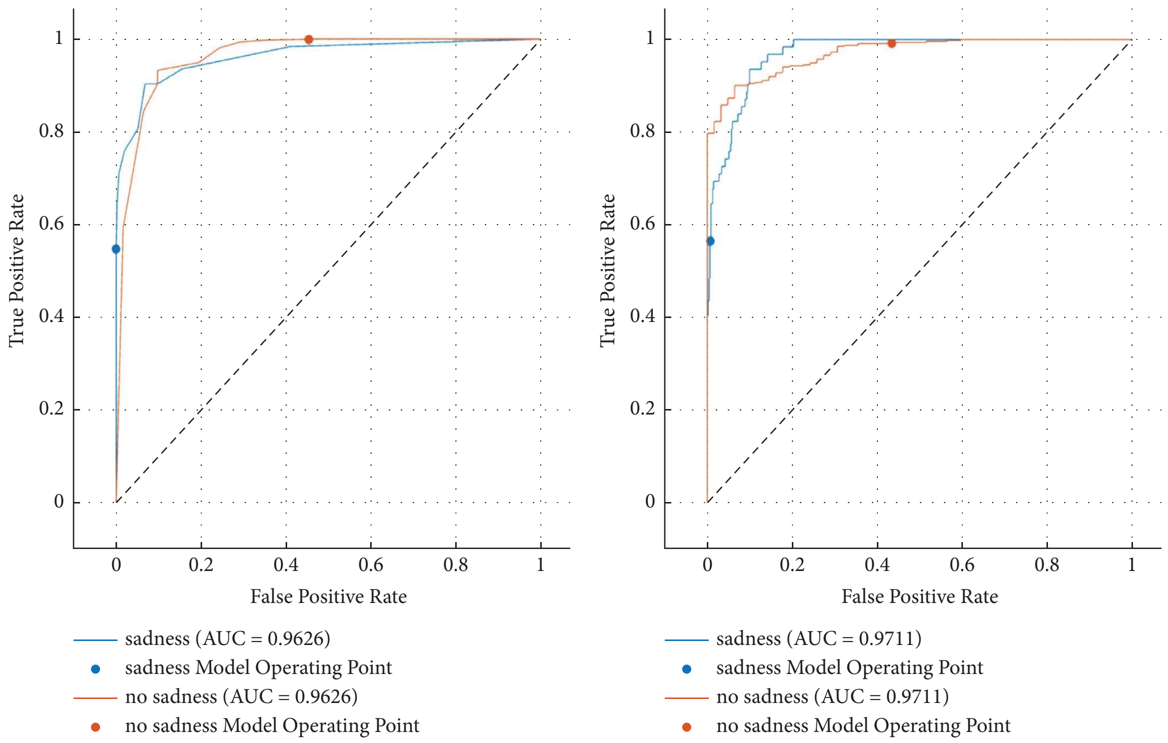
2.6.6. *Neural Network*

(1) Total cost (validation): 30.
(2) Prediction speed: −1400 obs/sec.
(3) Training time: 33.549 sec.
(4) Model size (compact): ~409 kB.
(5) Hyperparameters: Wide Neural Network with one fully connected layer of size 100, ReLU activation, iteration limit of 1000, regularization strength of 0, and standardized data.

3. Results and Discussion

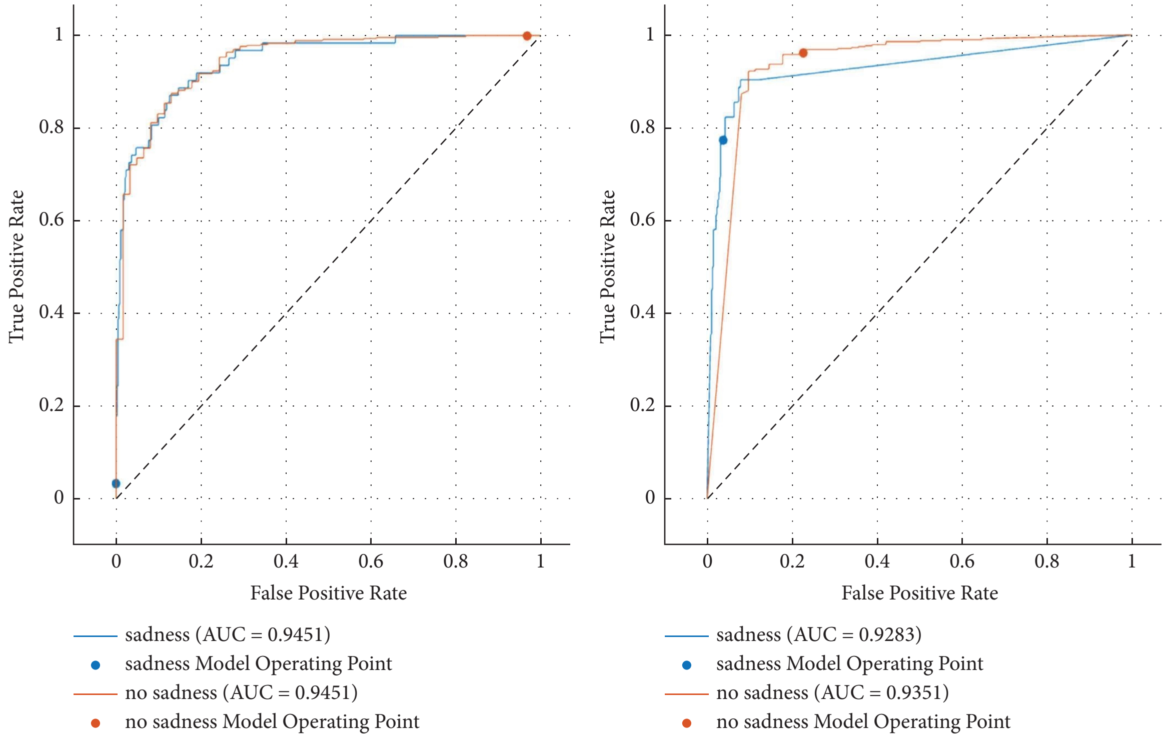
Figures 2–4 display Receiver Operating Characteristic (ROC) curves for the classifiers. ROC is a graphical device used in the evaluation of the classifiers. Table 1 gives the ROC data for the emotion sadness for each classifier type. Here is a brief explanation of the area under the curve (AUC) values: for SVM, the AUC of classifying sadness gets as high as 96.26%, meaning almost high performance of sadness classification. The KNN classifier had 97.11% AUC for sadness, pointing to the highest discriminatory power with regard to sadness. At an AUC of 94.51%, the Efficient Logistic Regression classifier shows a very good discriminative power between sadness and other emotions. The Naive Bayes classifier demonstrates a slight decline in its discriminatory performance towards sadness. AUC of 96.88% for sadness means a good separation of sadness from other emotions in Ensemble classifier. Neural Network classifier had 96.40% AUC value denoting adequate discriminatory capacity for sadness.

Careful analysis of the results in Table 1 shows that there are different degrees of accuracy for various types of emotions. The average prediction accuracy obtained from the SVM classifier was 89.82%, which is a higher figure than any other model. The classification of sadness, happiness, disgust, boredom, anxiety-fear, and anger was high by at least 88.59%. However, the "neutral" class was at about 85.23% which is slightly lower. The KNN classifier presented an overall average accuracy of 88.18%. It demonstrated high classification accuracies exceeding 88% for sadness, disgust, boredom, anxiety-fear, and anger. Nevertheless, its accuracy of the neutral category stood at 85.42%. Efficient Logistic Regression classifier scored an average accuracy of 87.47%. It was also stable in most categories, but not on the happiness with the accuracy score of 87.10%. The average accuracy for the Naive Bayes classifier is 86.75%. Other classifiers were more accurate across all emotion categories than it was. The Ensemble classifier attained an average accuracy of 89.63%. It achieved high performance in all the emotion classifications averaging accuracies between 87.48% and 94.76%. However, the Neural Network classifier exhibited an average accuracy of 89.40% in prediction. The study classified sadness, disgust, boredom, anxiety-fear, and anger well but had low accurate level for neutral categories. After scrutinizing the results, it becomes apparent that certain classifiers possess a remarkable ability to accurately classify specific emotions compared to others. Notably, the SVM and KNN classifiers excelled across all measures, but the Ensemble classifier exhibited the highest average accuracy. However, there exists a need to know the best appropriate classifier depending on the application specification and characteristics.



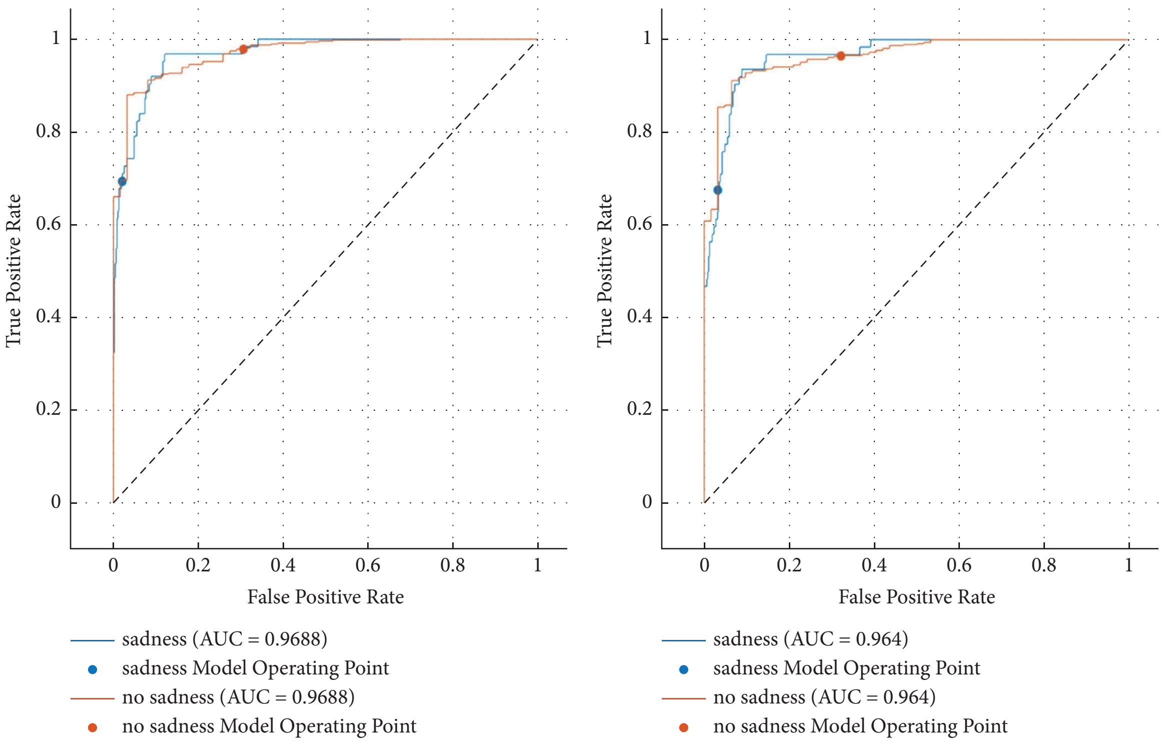
(a) (b)

Figure 2: The ROC figure for the results of the proposed method for SVM classifier (a) and KNN classifier (b).



(a) (b)

Figure 3: The ROC figure for the results of the proposed method for Efficient Logistic Regression classifier (a) and Naive Bayes classifier (b).



(a)

(b)

Figure 4: The ROC figure for the results of the proposed method for Ensemble classifier (a) and Neural Network classifier (b).

Table 1: The average accuracy for each classifier across different emotion categories.

Type	Sadness (%)	Neutral (%)	Happiness (%)	Disgust (%)	Boredom (%)	Anxiety-fear (%)	Anger (%)	Average (%)
SVM	94.20	85.23	88.59	92.89	88.78	88.59	90.46	89.82
KNN	94.76	85.42	87.66	93.64	88.41	89.71	84.67	89.18
Efficient Logistic Regression	88.78	85.23	87.10	91.40	85.60	87.28	86.91	87.47
Naive Bayes	94.01	82.61	84.86	89.90	84.48	84.11	87.28	86.75
Ensemble	94.76	85.04	87.48	92.33	89.71	88.22	89.90	89.63
Neural Network	93.83	82.05	87.66	93.64	90.65	87.10	91.58	89.40

The "Average" column represents the average accuracy for each classifier.

To establish a deeper understanding, detailed analysis and evaluation can be carried out, using various cross-validation techniques and comparing performance with other machine learning classifiers applied to emotion classification tasks. Furthermore, investigating the proposed feature extraction method as well as refining the classifiers' parameters might result in increased precision for certain emotion categories.

Based on the provided results in Table 2 from cross-validation, holdout validation, and resubstitution validation for sadness, we can observe the performance of different classifiers: SVM, KNN, Efficient Logistic Regression, Naive Bayes, Ensemble, and Neural Network. SVM maintains high reliability with an average in terms of various cross-validation folds. Cross-validation ranges from 94.20 and 95.89%. On the other hand, the holdout validity and resubstitution validation are between 94.34 and 100.0%. This means that SVM is an efficient classifier because it works well in multiple validation contexts. KNN achieves comparable performance, with cross-validation results ranging from 94.76% to 95.89% and holdout/resubstitution validation results between 94.34% and 100%. KNN accuracy remains constant when using different ways for validation, demonstrating it is stable and reliable.

However, Efficient Logistic Regression is less accurate than SVM and KNN algorithms. The range is between 88.78% and 89.91% for cross-validation while it is 88.68% and 89.72% for holdout validation and resubstitution validation. Its performance is lower than accurate, but reasonable. According to cross-validation, it has an average accuracy of 94.01–95.28%. It demonstrates excellent performance in multiple validation settings, thus demonstrating its reliability across different utterances.

Among various classification models examined during cross-validation and holdout validation processes, Ensemble exhibits a highly competitive accuracy of up to 94.76% to 96.23%, while maintaining consistency in performance. These results imply that ensemble methods are effective for

Table 2: The results of different cross-validation techniques and comparison with machine learning classifiers, in emotion classification tasks.

Type	Cross-validation			Holdout validation			Resubstituting validation
	5	10	20	10	20	40	
SVM	94.20	94.77	95.89	94.34	94.34	93.90	100.00
KNN	94.76	95.33	95.89	94.34	95.28	93.43	100.00
Efficient Logistic Regression	88.78	89.72	89.91	88.68	90.57	89.67	89.72
Naive Bayes	94.01	94.21	94.21	92.45	95.28	94.37	97.94
Ensemble	94.76	94.95	95.33	96.23	93.40	95.31	100.00
Neural Network	93.83	95.51	95.33	96.23	95.28	93.43	100.00

increasing prediction accuracy when used on multiple classifiers. With respect to the cross-validation, Neural Network also gets an accuracy of between 93.83% and 96.23%. These results show that neural networks can indeed tackle the specific classification problem due to their capacity of learning complex patterns.

In summary, based on the provided results, SVM, KNN, Naive Bayes, Ensemble, and Neural Network classifiers show competitive performance. SVM and KNN consistently achieve high accuracy, while Naive Bayes demonstrates robustness across different validation approaches. Ensemble methods and Neural Networks also perform well, indicating their potential to improve accuracy through combining multiple classifiers or learning complex patterns. Efficient Logistic Regression achieves lower accuracy compared to other classifiers but still performs at a decent level.

Table 3 presents classification accuracy obtained using different feature extraction methods in the context of sadness, for different classifiers, namely, SVM, KNN, Efficient Logistic Regression, Naive Bayes, Ensemble, and Neural Network. The examined feature extraction methods include the proposed method, formants, wavelet entropy, proposed method and formants, and proposed method and entropy, respectively.

The classifiers produced good results with accuracies varying within the range of 88.41–95.88%. The proposed method and formants employing SVM attained the best accuracy of 95.88%. These imply that the combined use of the proposed method and formants as features provided the optimum classification performance.

In comparison, the proposed method achieved the highest average accuracy of 93.39%. Formant features and wavelet entropy also yielded good performances, achieving average accuracies of 88.75% and 93.30%, respectively. When combining the proposed method with entropy, an average accuracy of 93.70% is achieved. The lowest average accuracy (minimum accuracies) was obtained by Efficient Logistic Regression, suggesting that this particular classifier may not perform well with this dataset, regardless of the feature extraction technique used.

These results shed light on the significant role played by features in obtaining correct classifications. The proposed method likely involves some domain-specific knowledge and sophisticated means that consistently outperform other methods. Also, combining the proposed method with either formants or entropy further enhanced classification accuracy.

Table 4 shows the time consumption of our proposed method compared to the two referenced methods (Hema et al. [19] and Ullah et al. [20]) on the EMO-DB dataset. The experiments were conducted on a computer with an Intel Core i7-11700K CPU, 32 GB of RAM, and an NVIDIA RTX 3090 GPU.

As can be seen from the table, our proposed method is generally faster than the compared methods. This is likely since our method uses a more efficient feature extraction method based on wavelet transform.

The proposed method is compared with two published methods (Table 5):

(1) MFCC-PCA-SVM [21]: This model combines Mel-frequency cepstral coefficients (MFCC) for feature extraction, Principal Component Analysis (PCA) for dimensionality reduction, and SVM for sadness recognition. It uses a linear kernel with automatic scaled box constraint level 1. The data are standardized, and all 34 features are selected. PCA is applied to retain components explaining at least 10% of the variance, with one component explaining 64.4% of the variance, with a compact model size of 7 kB.

(2) LPC-SVM [22]: This model utilizes LPC with 50 coefficients for feature extraction and SVM for classification. The SVM uses a cubic kernel function with automatic kernel scale and a box constraint level of 1. One-vs-one multiclass coding is employed for classification. The data are standardized before training. The model has a prediction speed of 4200 observations per second, a training time of 2.608 seconds, and a compact model size of 869 kB.

The comparison of the three models for emotion recognition shows that the proposed method combined with entropy achieved the highest average accuracy of 90.24%. Its performance exceeded that of the LPC-PCA-SVM which scored a mean accuracy of 90.07%. The MFCC + SVM model had the highest average error rate of 12.70% out of all the models applied.

Table 3: The proposed method, formants, wavelet entropy, and the proposed method combined with formants and with entropy, along with the average accuracy in the last row.

Classifier type	Proposed method (%)	Formants (%)	Wavelet entropy (%)	Proposed method and formants (%)	Proposed method and entropy (%)
SVM	94.20	88.41	95.33	94.77	95.88
KNN	94.76	88.60	94.39	94.39	94.39
Efficient Logistic Regression	88.78	88.41	94.58	89.72	89.71
Naive Bayes	94.01	89.72	86.73	93.46	94.39
Ensemble	94.76	89.72	94.58	94.58	94.01
Neural Network	93.83	87.66	94.21	94.39	93.83
Average	93.39	88.75	93.30	93.55	93.70

Table 4: The time consumption of our proposed method compared to the two referenced methods (Hema et al. and Ullah et al.) on the EMO-DB dataset.

Classifier	Our proposed method (seconds)	Hema et al. [19] (seconds)	Ullah et al. [20] (seconds)
SVM	1.23	1.25	1.87
KNN	1.15	1.27	1.98

Table 5: The proposed method is compared with two published methods.

Type	Sadness	Neutral	Happiness	Disgust	Boredom	Anxiety-fear	Anger	Average
Proposed method and entropy	94.40	86	88.20	93.10	89.20	89	91.80	90.24
LPC-SVM	94.20	85.80	87.70	91.40	89	88.80	91.20	90.07
MFCC-PCA-SVM	89.20	86.20	87.70	92.40	85.90	88.70	81.01	87.30

Regarding individual emotions, the proposed method combined with entropy provided the best result in recognizing sadness (94.40%), disgust (93.10%), and anger (91.80%). This was, however, more pronounced among the other emotions. The results were the worst for MFCC + SVM across all emotions with relatively high performance for happiness (87.70%) and disgust (92.40%).

For more comparison of the proposed method with published works, two more methods are analyzed for comparison:

(1) In the study regarding the automatic recognition of anxiety emotional state using the EMO-DB dataset by using KNN [23], the proposed method achieved an average accuracy of 90.24%. The accuracy for identifying anxiety/fear emotion specifically was 89%. In comparison to the other emotions in the EMO-DB dataset, the proposed method achieved accuracies ranging from 86% to 94.40% for differentiating sadness, neutral tones, happiness, disgust, boredom, and anger. While the study in [23] had a recognition rate for the emotion classes of about 70%, our proposed method outperforms the published method where our result for the database is more than 80%.

(2) In comparison with [24] that based on MFCC and CNN on the same database, we can state that our method achieved an average accuracy of 90.24% for all database. That is slightly better than this method published in [24], that achieved 90.20% accuracy, slightly less than our results. While the published method has a promising CNN-based method, we found that our proposed method is an excellent way of dealing with the database providing a competitive result.

The drawbacks of the proposed strategy along with justifications for overcoming the mentioned drawbacks are as follows:

Computational Complexity. The proposed strategy addresses the computational complexity by leveraging efficient algorithms and optimization techniques specific to LPC and DWT. *Justification*. This highlights that despite the potential higher complexity, the method incorporates measures to mitigate computational demands and improve efficiency.

Sensitivity to Noise. The proposed strategy includes noise reduction and denoising techniques in conjunction with LPC and DWT to enhance robustness against noise. *Justification*. This implies that the approach considers noise sensitivity and uses preprocessing procedures to increase the quality of retrieved features, making it more noise resistant.

Limited Adaptability. The proposed strategy incorporates adaptive wavelet selection and feature fusion techniques, allowing it to adapt to different signal characteristics and effectively capture relevant information across diverse signal types. *Justification*. This demonstrates that the method is designed to overcome the limitation by incorporating adaptability mechanisms, enhancing its applicability to various signal analysis tasks.

Lack of Comparative Evaluation. The study contains detailed comparative evaluations using cutting-edge feature extraction methods to demonstrate the superiority and efficacy of the suggested approach. *Justification*. This guarantees that the method is rigorously examined against known methodologies, demonstrating its performance and allowing for fair comparisons.

Future work will focus on exploring different feature extraction methods and deep learning architectures to further improve the performance of our proposed method. We will also investigate the use of transfer learning to leverage knowledge from related tasks.

4. Conclusion

In conclusion, discrete wavelet transform combined with linear prediction coding was proposed for emotions modeling via speech signals. The study evaluated the performance of different classifiers for emotion classification using the EMO-DB dataset. Classifiers considered were SVM, KNN, Efficient Logistic Regression, Naive Bayes, Ensemble, and Neural Network. AUC, average prediction accuracy, and cross-validation approaches were used for evaluation.

The results demonstrated that KNN and SVM classifiers have high discriminatory power in accurately identifying sadness from other emotions. Ensemble methods and Neural Networks also performed well in sadness classification. Efficient Logistic Regression and Naive Bayes classifiers showed competitive performance but were slightly less accurate compared to other classifiers.

The study also explored feature extraction methods and found that the proposed method yielded the highest average accuracy. Combining the proposed method with formants or wavelet entropy further improved the accuracy. Efficient Logistic Regression had the lowest accuracies among the classifiers.

The results indicate that the proposed method combined with entropy as a feature extraction technique has superior performance in accurately recognizing emotions, particularly for sadness. The LPC-PCA-SVM model also performs well but slightly less, while the MFCC + SVM model has the lowest accuracy among the three models.

This research, by evaluating the performance of different classifiers for emotion classification, aims to contribute to the field of emotion recognition. The results suggest that KNN, SVM, Ensemble, and Neural Network classifiers effectively predict sadness, particularly when combined with the proposed feature extraction method. These findings can inform the selection of suitable classifiers and feature extraction techniques for designing emotion recognition systems. Future research could focus on improving classifier performance and exploring additional feature extraction methods to further enhance the accuracy of emotion categorization [25].

Data Availability

The data used in this study are publicly available in the following website: https://emodb.bilderbar.info/docu/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

K. Daqrouq was responsible for method, investigations, writing, and revisions. A. Balamesh was responsible for method, programming, and investigations. O. Alrusaini was responsible for organizing data and revisions. A. Alkhateeb was responsible for writing and discussion. A.S. Balamesh was responsible for writing and proofreading and revisions.

Acknowledgments

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under grant no. G: 648-135-1439. The authors, therefore, acknowledge with thanks to DSR for their technical and financial support.

References

[1] Y. Ü. Sönmez and A. Varol, "A speech emotion recognition model based on multi-level local binary and local ternary patterns," *IEEE Access*, vol. 8, pp. 190784–190796, 2020.

[2] S. Shreya, P. Likitha, G. S. Charan, and S. B. Choubey, "Speech emotion detection through live calls," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 5, pp. 691–695, 2023.

[3] S. Lalitha, A. Mudupu, B. V. Nandyala, and R. Munagala, "Speech emotion recognition using DWT," in *Proceedings of the 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–4, IEEE, Madurai, India, December 2015.

[4] S. T. Saste and S. M. Jagdale, "Emotion recognition from speech using MFCC and DWT for security system," in *Proceedings of the 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, pp. 701–704, IEEE, Coimbatore, India, April 2017.

[5] C. S. Ram and R. Ponnusamy, "An effective automatic speech emotion recognition for Tamil language based on DWT and MFCC using Stability-plasticity dilemma Neural network," in *Proceedings of the International Conference on Information Communication and Embedded Systems (ICICES2014)*, pp. 1–6, IEEE, Chennai, India, February 2014.

[6] J. Kambale, A. Khedkar, P. Patil, and T. Sonone, "Speech emotion recognition using deep learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 5, pp. 4829–4833, 2023.

[7] M. Rajababu, P. Abhinav, N. K. Subhash, and M. A. Chowdary, "Speech based emotion recognition using machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 4, pp. 1182–1185, 2023.

[8] J. Singh, L. B. Saheer, and O. Faust, "Speech emotion recognition using attention model," *International Journal of*

Environmental Research and Public Health, vol. 20, no. 6, p. 5140, 2023.

[9] L. Huang and X. Shen, “Research on speech emotion recognition based on the fractional fourier transform,” *Electronics*, vol. 11, no. 20, p. 3393, 2022.

[10] F. Daneshfar, S. J. Kabudian, and A. Neekabadi, “Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier,” *Applied Acoustics*, vol. 166, Article ID 107360, 2020.

[11] F. Daneshfar and S. J. Kabudian, “Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm,” *Multimedia Tools and Applications*, vol. 79, no. 1-2, pp. 1261–1289, 2020.

[12] K. Daqrouq, R. Al-Hmouz, A. Balamesh, and A. Memic, “Application of wavelet transform for PDZ domain classification,” *PLoS One*, vol. 10, no. 4, Article ID e0122873, 2015.

[13] K. Daqrouq, A. Alkhateeb, W. Ahmad et al., “A universal ECG signal classification system using the wavelet transform,” *Neural Network World*, vol. 32, no. 1, pp. 43–54, 2022.

[14] K. Daqrouq, A. Alkhateeb, M. N. Ajour, and A. Morfeq, “Neural network and wavelet average framing percentage energy for atrial fibrillation classification,” *Computer Methods and Programs in Biomedicine*, vol. 113, no. 3, pp. 919–926, 2014.

[15] Y. A. Dementiy and A. N. Maslov, “Neural network classifier of energy facilities operating modes and its recognition ability assessment at different number of precedents,” *Vestnik Chuvashskogo universiteta*, vol. 3, pp. 45–52, 2021.

[16] L. Deng, “The MNIST database of handwritten digit images for machine learning research [best of the web],” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141-142, 2012.

[17] T. Denœux, “NN-EVCLUS: neural network-based evidential clustering,” *Information Sciences*, vol. 572, pp. 297–330, 2021.

[18] X. Ding, F. A. Schmitt, R. J. Kryscio, and R. Charnigo, “Comparison of neural network and logistic regression for dementia prediction: results from the PREADViSE trial,” *Journal of Gerontology and Geriatrics*, vol. 69, no. 2, pp. 137–146, 2021.

[19] C. Hema and F. P. Garcia Marquez, “Emotional speech recognition using CNN and deep learning techniques,” *Applied Acoustics*, vol. 211, Article ID 109492, 2023.

[20] R. Ullah, M. Asif, W. A. Shah et al., “Speech emotion recognition using convolution neural networks and multi-head convolutional transformer,” *Sensors*, vol. 23, no. 13, p. 6212, 2023.

[21] Y. Wang, X. Wang, and C. He, “Speech emotion recognition algorithm for school bullying detection based on MFCC-PCA-SVM classification,” in *Communications, Signal Processing, and Systems. CSPS 2020*, Q. Liang, W. Wang, X. Liu, Z. Na, X. Li, and B. Zhang, Eds., Springer, Singapore, 2021.

[22] S. M. Feraru and M. D. Zbancioc, “Emotion recognition in Romanian language using LPC features,” in *Proceedings of the 2013 E-Health and Bioengineering Conference (EHB)*, pp. 1–4, IEEE, Iasi, Romania, December 2013.

[23] D. Marius, S. Zbancioc, M. Monica, and F. Feraru, “A study about the automatic recognition of the anxiety emotional state using Emo-DB,” in *Proceedings of the 2015 E-Health and Bioengineering Conference (EHB)*, Iasi, Romania, November 2015.

[24] A. D. Anup, S. Catherine, and S. Renaud, “Emo-CNN for perceiving stress from audio signals: a brain chemistry approach,” 2020, https://arxiv.org/abs/2001.02329v1.

[25] M. H. Farouk, “Emotion recognition from speech,” *SpringerBriefs in Electrical and Computer Engineering*, Springer, Cham, Switzerland, 2014.