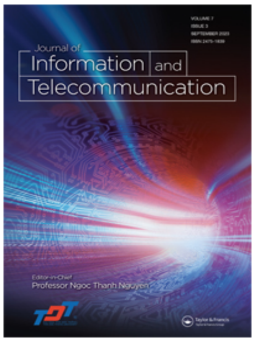


Taylor & Francis
Taylor & Francis Group

Journal of Information and Telecommunication

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tjit20

Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network

Nhat Truong Pham, Sy Dzung Nguyen, Vu Song Thuy Nguyen, Bich Ngoc Hong Pham & Duc Ngoc Minh Dang

To cite this article: Nhat Truong Pham, Sy Dzung Nguyen, Vu Song Thuy Nguyen, Bich Ngoc Hong Pham & Duc Ngoc Minh Dang (2023) Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network, Journal of Information and Telecommunication, 7:3, 317-335, DOI: 10.1080/24751839.2023.2187278

To link to this article: https://doi.org/10.1080/24751839.2023.2187278

Published online: 17 Mar 2023.

JOURNAL OF INFORMATION AND TELECOMMUNICATION
2023, VOL. 7, NO. 3, 317–335
https://doi.org/10.1080/24751839.2023.2187278

OPEN ACCESS Check for updates

Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network

Nhat Truong Pham [a,b], Sy Dzung Nguyen [c,d], Vu Song Thuy Nguyen[e], Bich Ngoc Hong Pham [f] and Duc Ngoc Minh Dang [g]

[a]Division of Computational Mechatronics, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam; [b]Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam; [c]Laboratory for Computational Mechatronics, Institute for Computational Science and Artificial Intelligence, Van Lang University, Ho Chi Minh City, Vietnam; [d]Faculty of Mechanical – Electrical and Computer Engineering, School of Technology, Van Lang University, Ho Chi Minh City, Vietnam; [e]Department of Computer Science and Engineering, Michigan State University, Michigan, MI, USA; [f]Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam; [g]Computing Fundamental Department, FPT University, Ho Chi Minh City, Vietnam

ABSTRACT
Speech emotion recognition (SER) has several applications, such as e-learning, human-computer interaction, customer service, and healthcare systems. Although researchers have investigated lots of techniques to improve the accuracy of SER, it has been challenging with feature extraction, classifier schemes, and computational costs. To address the aforementioned problems, we propose a new set of 1D features extracted by using an overlapping sliding window (OSW) technique for SER in this study. In addition, a deep neural network-based classifier scheme called the deep Pattern Recognition Network (PRN) is designed to categorize emotional states from the new set of 1D features. We evaluate the proposed method on the Emo-DB and the AESSD datasets that contain several different emotional states. The experimental results show that the proposed method achieves an accuracy of 98.5% and 87.1% on the Emo-DB and AESSD datasets, respectively. It is also more comparable with accuracy to and better than the state-of-the-art and current approaches that use 1D features on the same datasets for SER. Furthermore, the SHAP (SHapley Additive exPlanations) analysis is employed for interpreting the prediction model to assist system developers in selecting the optimal features to integrate into the desired system.

ARTICLE HISTORY
Received 13 October 2022
Accepted 24 February 2023

KEYWORDS
Feature extraction; overlapping sliding window; pattern recognition network; SHAP analysis; speech emotion recognition

1. Introduction

Since deep learning and neural networks have advanced significantly over the past ten years, speech emotion recognition (SER) has gained a lot of attention in the field of

CONTACT Duc Ngoc Minh Dang ✉ ducdnm2@fe.edu.vn

affective computing. On the basis of voice cues, it seeks to identify and understand human emotions. SER has recently been effectively incorporated into several applications in the real world (Badshah et al., 2019; Chatterjee et al., 2021; Chen et al., 2020; Gong & Luo, 2007; Yoon et al., 2007; Zhu & Luo, 2007). Yoon et al. (2007) trained a speech emotion recognition agent model and integrated it into a mobile communication service to monitor and analyze the mood of the person that consumers wanted to know via their smartphones. An SER system with different languages from diverse courses was proposed in an e-learning system to address the absence of emotional contact in Zhu and Luo (2007).

Researchers have worked to develop effective classifier schemes and determine the best feature sets to increase the accuracy of the SER systems. Energy, pitch, Mel frequency cepstral coefficients (MFCC) (Koduru et al., 2020; Kuchibhotla et al., 2016), linear predictive coding, zero-crossing rate (Lingampeta & Yalamanchili, 2020), Mel spectrograms (Chen et al., 2018; Meng et al., 2019; Pham et al., 2020), and wavelet features (Abdel-Hamid, 2020; Koduru et al., 2020) are just a few of the feature extraction techniques that have been presented throughout the years to extract the reliable and ideal characteristics. Additionally, many classifier schemes, including support vector machines (Swain et al., 2015), artificial neural networks (ANNs), adaptive-neuro-fuzzy inference systems (Giripunje & Bawane, 2007), linear discriminant analyses (LDAs), k-nearest neighbors (Lingampeta & Yalamanchili, 2020), and regularized discriminant analyses (RDA) (Kuchibhotla et al., 2016), have been developed to increase the recognition rate for the SER. However, these were handcrafted features and modeling methods that could not achieve the required precision and required a significant amount of time and money.

Deep learning has become increasingly important with the advancement of neural networks in a variety of fields, including image processing, text recognition, speech recognition, and speech emotion recognition. A 3-dimensional deep learning model that incorporated convolutional neural networks (CNN), a long short-term memory (LSTM) network and an attention mechanism for the SER was put out by Chen et al. (2018). Mustaqeem et al. (2020) employed a K-means clustering algorithm with the help of the radial basis function network (RBFN) to cluster the key segments from the input speech. The selected key segment sequence was transformed into spectrograms and then fed into ResNet-101 to extract features. Finally, these features were passed to a deep bidirectional LSTM to learn the temporal information and high-level representations for recognizing emotional states. Bao et al. (2019) introduced a new emotion style transfer as a data augmentation method to generate synthetic feature vectors for speech emotion recognition by using a cycle-consistent based generative adversarial network (CycleGAN) with a classification loss function. For multimodal emotion recognition, Nguyen et al. (2021) designed a deep learning framework including two branches for multimodal features: a 2D deep auto-encoder branch and a 1D deep auto-encoder branch to extract features from visual input and audio input, respectively. Then, these features are fused and fed into an LSTM network for emotion recognition. However, utilizing deep learning models has required a significant amount of data and computing power.

Recent researchers have also tried to reconfigure, combine, or propose new classification loss functions to improve the speech emotion recognition rate. For instance, Meng et al. (2019) created a new loss function for the SER that combined the center loss and the softmax loss for emotion classification. By merging contrastive-center loss and softmax loss, the loss function in Meng et al. (2019) was enhanced for the SER in Pham et al. (2020). To enhance emotion identification from speech, researchers have used an attention mechanism (Neumann & Vu, 2017) and a transformer for the SER (Siriwardhana et al., 2020). However, it is challenging to extract the proper characteristics from practically all SER systems, and consumers are unsure of which features to pick. Additionally, in SER systems, selecting the classifier schemes is crucial.

It is general knowledge that researchers have attempted to employ 2D feature representations and classifier systems based on deep learning with and without attention processes to enhance speech emotion identification. However, researchers have not yet determined which features should be employed to improve speech emotion recognition. Moreover, deep learning models, such as CNN, RNN, and their variant architectures, require a lot of data and computing resources. Therefore, this study developed a novel set of 1D features for speech emotion recognition using the overlapping sliding window (OSW) method. Time-domain and frequency-domain (TD-FD) characteristics serve as the foundation for these features. Then, utilizing the new set of 1D features, a deep Pattern Recognition Network (PRN) is created to train a classifier to improve the accuracy of the SER system. The suggested approach is also accessible as open source at https://github.com/nhattruongpham/osw-1d-prn-shap, which may be used to reproduce the experiments.

The following is a list of the study's key contributions:

- An OSW approach is used to make use of a new collection of 1D features. A simple deep pattern recognition network model is created to train a classifier from the new set of 1D features using a cross-entropy loss function, which drastically reduces the amount of computational work required compared to CNN, RNN, and their variant architectures since PRN is a simple deep neural network-based one;
- This OSW technique can also be used as a data augmentation method to enrich features for speech emotion recognition;
- The suggested technique is tested using the Emo-DB (Burkhardt et al., 2005) and AESSD (Vryzas et al., 2018; Vryzas, Matsiola et al., 2018) datasets, and SHAP analysis (Lundberg & Lee, 2017) is used to assess the contribution of each feature in the new collection of 1D features so that system developers may choose the best features for the desired system.

The rest of this paper is structured as follows. Related studies are summarized in Section 2. The proposed methodology is presented in Section 3. In Section 4, experimental results and discussion are depicted and analyzed. Finally, the paper is concluded in Section 5.

2. Related work

Speech emotion detection has faced significant difficulties with feature extraction. To extract reliable and ideal features for the SER, several researchers have looked at everything from low-level handmade features and conventional approaches to high-level representation and deep learning techniques. To identify emotional states from speech, Kuchibhotla et al. (2016) employed a total of 360 characteristics, including 324 MFCCs, 36 energy features, and 36 pitch features. Then, using sequential forward selection or sequential floating forward selection, these traits are fussed with or chosen (SFFS). When compared to other approaches, the experiment combining the RDA classifier scheme and SFFS provided competitive accuracy. In order to undertake a feature set for the SER, a lot of features were needed for this job. Two deep learning models were created by Zhao et al. (2019) that acquired local and global emotional characteristics from audio samples and log-Mel spectrograms to identify emotions in speech. The work of Zhao et al. (2019) did not, however, integrate several feature sets to create an algorithm that would mix the advantages of various deep features. Chen et al. (2020) proposed a two-layer fuzzy multiple random forest (TLFMRF) to recognize emotions from speech signals using a feature set of 16 basic features and 12 statistical values that belong to both personalized and non-personalized features.

In contrast, a three-step time-to-failure prognostic for rolling element bearings was developed by Wu et al. (2018) and covered feature extraction, feature reduction, and time-to-failure prediction. In the feature extraction process, TD-FD features were retrieved from raw vibration signals to create several statistical characteristics. Due to the rapid analysis of the speech signals in TD-FD, these characteristics are reliable and may be used for speech emotion identification. To evaluate audio content and extract useful information from it, Lerch (2012) identified several audio characteristics. Almost all aspects of speech processing, including speech emotion recognition, have made extensive use of some of them.

Most of the signal processing (Kusuma & Nuryani, 2019), mechanical systems (Clement et al., 2014; Li et al., 2018; Lin et al., 2021), and data streams (Domino & Gawron, 2019) have all used sliding window-based (SW) approaches. In addition, to minimize the noise in sound categorization, the OSW method was utilized in a short-time Fourier transform for assessing time origin from frame to frame (Tran et al., 2021). The length of the window (or frame) and the sliding step are crucial components of the OSW approach. These are created in accordance with the analysis's goal. Recently, the OSW for the 3-dimensional emotional system reacting to EEG (electroencephalography) input was proposed by Garg et al. (2021). It was also asserted in Liu et al. (2020) that the OSW with an ANN classifier technique might considerably increase the SER's accuracy. The EmoDB dataset's four emotional states are the only ones on which experimental findings are given.

As mentioned above, SW and OSW have been investigated in a lot of applications. The differences between a sliding window and an overlapping sliding window and the advantages of using the overlapping sliding window are summarized in some aspects. The sliding window method divides each input signal into several windows with a fixed interval size. There are two types of SW methods: OSWs and non-OSWs. In the OSW approach, there is an overlap between adjacent windows, whereas there is none between them in the non-OSW approach. For example, if the fixed interval size is 10 and the overlap is 5, the start and end of windows will be [1 10], [6 15], [11 20], etc. While in the non-overlapping window approach, they will be [1 10], [11 20], [21 30], and so on. In almost all popular windows, like Hanning and Hamming windows, choosing an overlap of 50% between adjacent windows works best in a lot of applications, such as signal processing, speech processing, mechanical systems, and control systems. In addition, because speech emotion recognition is a sequence process, each fixed interval size is not independent. If the OSW approach is employed, we can extract and capture as much information and features as possible. As a result, the suggested approach uses the OSW in various ways in this study. To extract the new set of 1D characteristics for SER, the OSW is first moved along with the voice signals. Second, the OSW is also regarded as a strategy for enhancing characteristics with incomplete data.

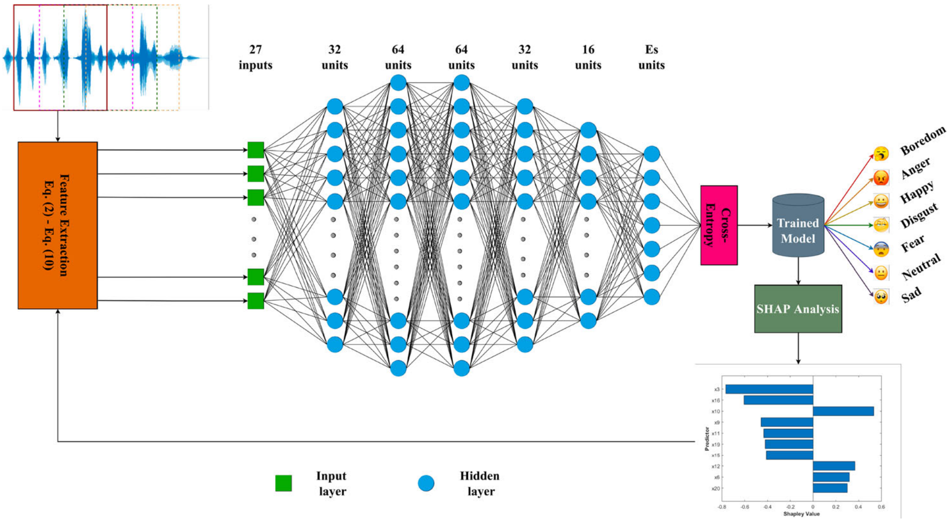


Figure 1. The proposed architecture.

3. Methodology

The proposed method is depicted in Figure 1 as an overview. It includes three main components: feature extraction, classification, and interpretation. For feature extraction, the OSW is created to calculate a new set of nine 1D features based on TD-FD characteristics (Lerch, 2012; Wu et al., 2018). Then these features are enriched by calculating the first and second derivatives. For classification, a simple neural network is designed to learn the representations from the extracted 1D features to discriminate emotions from speech signals. After that, SHAP analysis (Lundberg & Lee, 2017) is employed to interpret the contribution of each feature in the feature set to the prediction. Because all 27 extracted features are 1D features and the deep PRN is a shallow network, computing resources and training costs are drastically reduced. As a result, the SER system performs better. The details of using the OSW technique, designing a simple deep PRN model, and analyzing the model prediction are presented below.

3.1. A new set of 1D features using overlapping sliding window

In this work, a new set of 1D characteristics is suggested to improve the accuracy of the SER system. The new set of 1D features is built upon the TD-FD features of the mean value (*mean*), zero-crossing rate (*zrc*), root mean square value (*rmsv*), signal crest (*sc*), maximum absolute value (*mav*), kurtosis coefficient (*kc*) (Soualhi et al., 2014), square mean root value (*smrv*), root mean square logarithm (*rmsl*), clearance factor (*cf*), and root mean square frequency (*rmsf*).

Figure 2 is the first component in the proposed architecture that presents the OSW process to extract the new set of 1D features to train a classification model. As illustrated in Figure 2, a SW or frame $Fr(t_i)$ with a window size of Ws and a hop length of Hp is used to calculate these attributes at each time step t_i. There is a link between the hop length and the window size. The analysis time origin's hop length indicates how far it has advanced from frame to frame. This is heavily influenced by the analysis's goal. Increased overlap

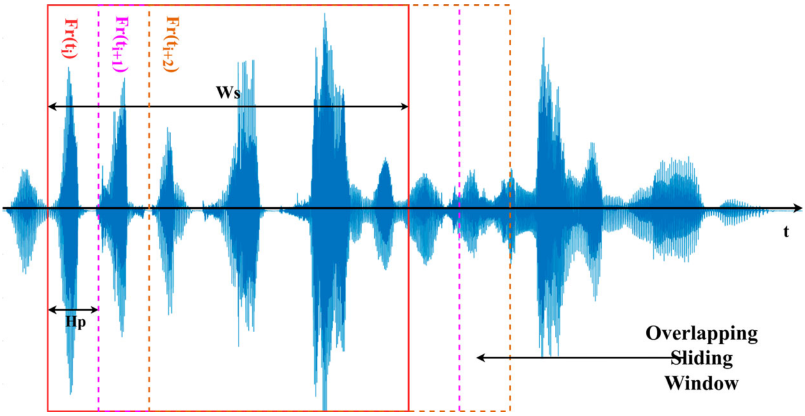


Figure 2. Process of extracting frame from speech.

leads to additional analysis points, which over time smooth out findings, but at a higher computational cost. The window size *Ws* of 25,000 samples is fixed in this study for both the training data and testing data sets, while the hop length *Hp* is fixed at 250 samples for the training data set and is randomly selected in a range of 250 to 1,000 samples for the testing data set. The window size is chosen based on empirical validity and experience. Furthermore, because the audio sampling rate is 16,000 kHz, each window size is approximately 1.5 s, providing a reasonable amount of time to speak the longest word. Then, a subset of the nine based features is obtained based on Equations (2) through (10), where $Ns \approx len(Ws)$ is the number of data samples in each frame $Fr(t_i)$. The mathematical models for nine 1D-based features extracted by using the OSW technique are described below.

$$fe_{mean} = \frac{1}{Ns}\sum_{i=1}^{Ns} Fr(t_i) \tag{1}$$

$$fe_{rmsv} = \sqrt{\frac{1}{Ns}\sum_{i=1}^{Ns} Fr^2(t_i)} \tag{2}$$

$$fe_{mav} = \max\left(Fr(t_i)\right) \tag{3}$$

$$fe_{smrv} = \left(\frac{1}{Ns}\sum_{i=1}^{Ns}\sqrt{\left|Fr(t_i)\right|}\right)^2 \tag{4}$$

$$fe_{kc} = \frac{1}{(fe_{rmsv})^3}\sum_{i=1}^{Ns}\left(Fr(t_i) - fe_{mean}\right)^3 \tag{5}$$

$$fe_{cf} = \frac{fe_{mav}}{fe_{rmsv}} \tag{6}$$

$$fe_{rmsf} = \sqrt{\frac{\sum_{i=2}^{Ns} \dot{Fr}^2(t_i)}{4\pi^2 \sum_{i=1}^{Ns} Fr^2(t_i)}} \tag{7}$$

$$fe_{rmsl} = 20\log\left(\sqrt{\frac{1}{Ns}\sum_{i=1}^{Ns} Fr(t_i)^2}\right) \tag{8}$$

$$fe_{sc} = \frac{fe_{mav}}{\sum_{i=1}^{Ns} Fr(t_i)} \tag{9}$$

$$fe_{zcr} = \frac{1}{2Ns}\sum_{i=1}^{Ns}\left| sign\left[Fr(t_i)\right] - sign\left[Fr(t_{i-1})\right]\right| \tag{10}$$

where

$$sign[Fr(t_i)] = \begin{cases} 1, & \text{if } Fr(t_i) > 0 \\ 0, & \text{if } Fr(t_i) = 0 \\ -1, & \text{if } Fr(t_i) < 0 \end{cases},$$

and $Fr(t_{i-1}) = 0$ is used as initialization if $Fr(t_{i-1})$ does not exist.

Since $n = 1, 2, 3$, let $FE^{(n)}_{feasub} = [fe_{rmsv}, fe_{mav}, fe_{smrv}, fe_{kc}, fe_{cf}, fe_{rmsf}, fe_{rmsl}, fe_{sc}, fe_{zcr}]$ be the subset feature. The first subset feature is calculated from the speech signals $Fr(t_i)$ when $n = 1$. To calculate the second subset feature for $n = 2$, first define $\dot{S}(t_i)$ as the derivative by time of $Fr(t_i)$, and then swap out every instance of $Fr(t_i)$ in Equations (2) through (10) with $\dot{Fr}(t_i)$. Replace all instances of $Fr(t_i)$ in Equations (2) through (10) with $\ddot{Fr}(t_i)$, which is obtained by dividing $\dot{Fr}(t_i)$ by time to compute the third subset feature with $n = 3$. The next step is to concatenate the three subset features, defining the set of features as $FE_{feature} = [FE^{(1)}_{feasub}, FE^{(2)}_{feasub}, FE^{(3)}_{feasub}]$.

3.2. A simple deep pattern recognition network architecture

To train a classifier model, a deep Pattern Recognition Network (PRN) is utilized. The deep PRN receives its inputs from the collection of 27 retrieved characteristics. As shown in Figure 1, the deep PRN's architectural layout includes five hidden layers, such as 32, 64, 64, 32, and 16 hidden units, respectively, and the final hidden layer includes Es hidden units that correspond to the number of emotional states of the dataset.

Finally, the weights are updated as training progresses using a cross-entropy (CE) loss function. The performance of classification tasks that compute the loss between predicted and ground truth values is frequently assessed using CE. The likelihood of emotional classes is calculated in this study using CE for multi-label classification, which is defined as follows:

$$CE = -\sum_{i=1}^{Es} y_i \log(\hat{y}_i), \tag{11}$$

where y_i represents the actual value and $\hat{y}_i$ represents the projected value of the emotion i.

3.3. Interpretability: query the optimal features using SHAP analysis

SHAP (SHapley Additive exPlanations) analysis, a machine learning method based on game theory, was introduced in Lundberg and Lee (2017). SHAP analysis may illustrate how much each attribute contributes to a prediction by measuring how much a forecast deviates from the average (response for regression or score of each class for classification). The LIME (Local Interpretable Model-Agnostic Explanations), which was initially proposed in Ribeiro et al. (2016), has been improved upon. A LIME strategy employs a simple interpretable model, like a decision tree or linear model, to estimate a complex model close to the desired prediction.

According to kernelSHAP in Lundberg and Lee (2017), the Shapley values are computed in this study to describe the contribution of each feature to prediction (response to the score for each class for classification). Let N be the total number of features and M be the total number of features in a set. The value function v defines the Shapley values of the i_{th} feature for the query point x as follows:

$$\Phi_i(v_x) = \frac{1}{N}\sum_{S\subseteq M\setminus\{i\}} \frac{v_x(S\cup\{i\}) - v_x(S)}{\frac{(N-1)!}{|S|!(N-|S|-1)!}}, \tag{12}$$

where $|S|$ is the cardinality of the set S or the total number of items in a set S, and $v_x(S)$ is the value function of the features for the query point x in a set S. The value of the function indicates how much the characteristics in S are probably going to influence the forecast for the input point x. The overall divergence of the forecast for the query point from the average is thus equal to the sum of the Shapley values for a query point across all characteristics, as shown below:

$$\sum_{i=1}^{N}\Phi_i(v_x) = f(x) - E\big[f(x)\big]. \tag{13}$$

The value function $v_x(S)$, in this case, must match the anticipated contribution of the features in S to the prediction f for the input point x.

The value function $v_x(S)$ of the kernelSHAP is obtained from Equations (12) and (13) as follows:

$$v_x(S) = E_D\big[f(x_S, X_{S^c})\big] \approx \frac{1}{O}\sum_{j=1}^{O} f\left(x_S, \left(X_{S^c}\right)_j\right), \tag{14}$$

where D stands for the interventional distribution, S^c for the joint distribution of the features, x_S for the query point value for the features in S, X_{S^c} for the features in S^c, O for the number of observations, and $(X_{S^c})_j$ for the values of the features in S^c of the j_{th} observation.

4. Experimental results and discussion

4.1. Dataset

4.1.1. Emo-DB

The German Emo-DB (Burkhardt et al., 2005) database of emotional speech is used in this study. Five male and five female speakers produced 535 samples at 44.1 kHz, which were further down-sampled to 16 kHz. There are various emotional states included in it, including anger, boredom, disgust, fear, happiness, neutral, and sadness. Figure 3 shows the distribution of various emotional states in the Emo-DB dataset.

4.1.2. AESSD

The AESSD (Acted Emotional Speech Dynamic Database), which includes 5 emotional states (anger, disgust, fear, happiness, and sadness), is recommended in Vryzas et al. (2018); Vryzas, Matsiola et al. (2018). AESSD was developed using SAVEE (Surrey Audio-Visual Expressed Emotion) in (Jackson & Haq, 2014), even though the utterances are in Greek rather than English. There are over 500 emotive spoken utterances in it, performed by five professional actors between the ages of 25 and 30. There are two men and three women in the ensemble. The AESSD dataset's distribution is shown in Figure 4.

Figures 3 and 4 make it abundantly evident that the AESSD dataset is regarded as balanced and the Emo-DB dataset as unbalanced. The imbalance dataset is therefore one of the most difficult problems in practically all SER systems.

4.2. Experimental setup

In this study, the Emo-DB and AESSD datasets were used to evaluate the proposed method. The OSW is used to extract 27 features from the raw audio signals of the Emo-DB and AESSD datasets. Then, the extracted features of each dataset were randomly divided into training and testing sets with a ratio of 70/30. As a result, the number of samples in the training set is 77,940, and in the testing set, it is 35,894 for the EmoDB

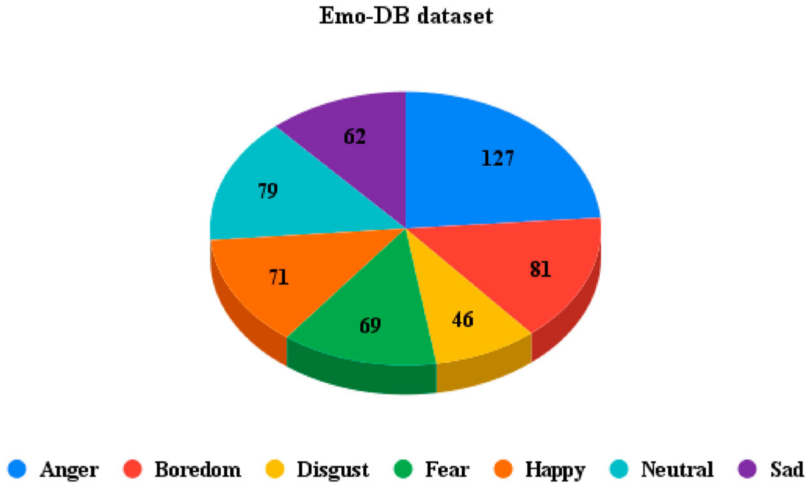


Figure 3. The distribution of emotional states in the Emo-DB dataset.

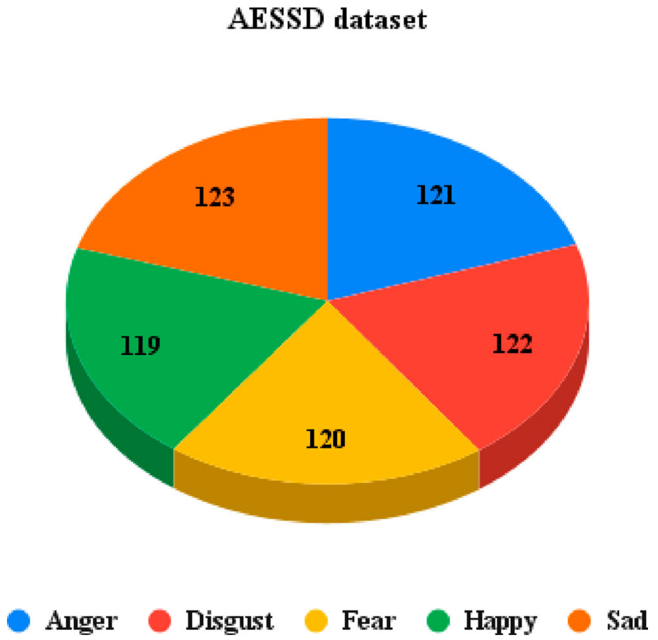
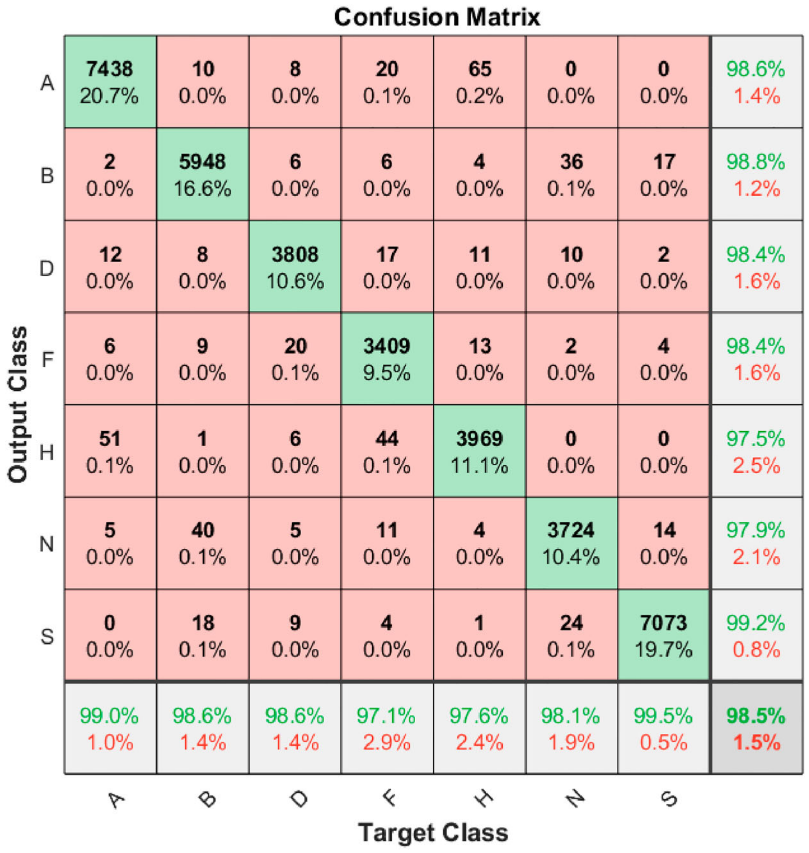


Figure 4. The distribution of emotional states in the AESSD dataset.


Confusion Matrix
Output Class
Target Class

	A	B	D	F	H	N	S	
A	7438 20.7%	10 0.0%	8 0.0%	20 0.1%	65 0.2%	0 0.0%	0 0.0%	98.6% 1.4%
B	2 0.0%	5948 16.6%	6 0.0%	6 0.0%	4 0.0%	36 0.1%	17 0.0%	98.8% 1.2%
D	12 0.0%	8 0.0%	3808 10.6%	17 0.0%	11 0.0%	10 0.0%	2 0.0%	98.4% 1.6%
F	6 0.0%	9 0.0%	20 0.1%	3409 9.5%	13 0.0%	2 0.0%	4 0.0%	98.4% 1.6%
H	51 0.1%	1 0.0%	6 0.0%	44 0.1%	3969 11.1%	0 0.0%	0 0.0%	97.5% 2.5%
N	5 0.0%	40 0.1%	5 0.0%	11 0.0%	4 0.0%	3724 10.4%	14 0.0%	97.9% 2.1%
S	0 0.0%	18 0.1%	9 0.0%	4 0.0%	1 0.0%	24 0.1%	7073 19.7%	99.2% 0.8%
	99.0% 1.0%	98.6% 1.4%	98.6% 1.4%	97.1% 2.9%	97.6% 2.4%	98.1% 1.9%	99.5% 0.5%	98.5% 1.5%

Figure 5. Confusion matrix using deep PRN with a new set of 1D features on the Emo-DB dataset.

dataset, while they are respectively 158,557 and 73,525 for the AESSD dataset. The training set is used to train and validate the proposed method, while the testing set is only used for prediction and evaluation.

We use MATLAB R2021a to implement our proposed method on an Intel Core i5 8th Gen computer without a graphics processing unit (GPU). The deep PRN is created using the *patternnet* MATLAB function. Since deep PRN is a deep shallow network, it does not support minibatch training. As a result, the entire dataset was applied for each epoch. The training function for the deep PRN is the scaled conjugate gradient backpropagation (*trainscg*) algorithm without an explicit learning rate. The maximum number of iterations is set to 5,000. In addition, early stopping is used to enhance the deep PRN model's generalization and prevent overfitting. Moreover, to explain the contribution of each feature to the prediction, *kernelSHAP* is calculated using the *shapley* function.

For performance evaluation, accuracy (*ACC*), precision (*PCS*), sensitivity (*Sn*), specificity (*Sp*), misclassification (*MC*), F1-score (*F*1), and Matthews correlation coefficient (*MCC*) evaluation metrics have been used to validate the effectiveness and robustness of the proposed method on both the Emo-DB and AESSD datasets. These metrics are defined below:

$$ACC = \frac{TP + TN}{TN + FP + FN + TP}; \tag{15}$$

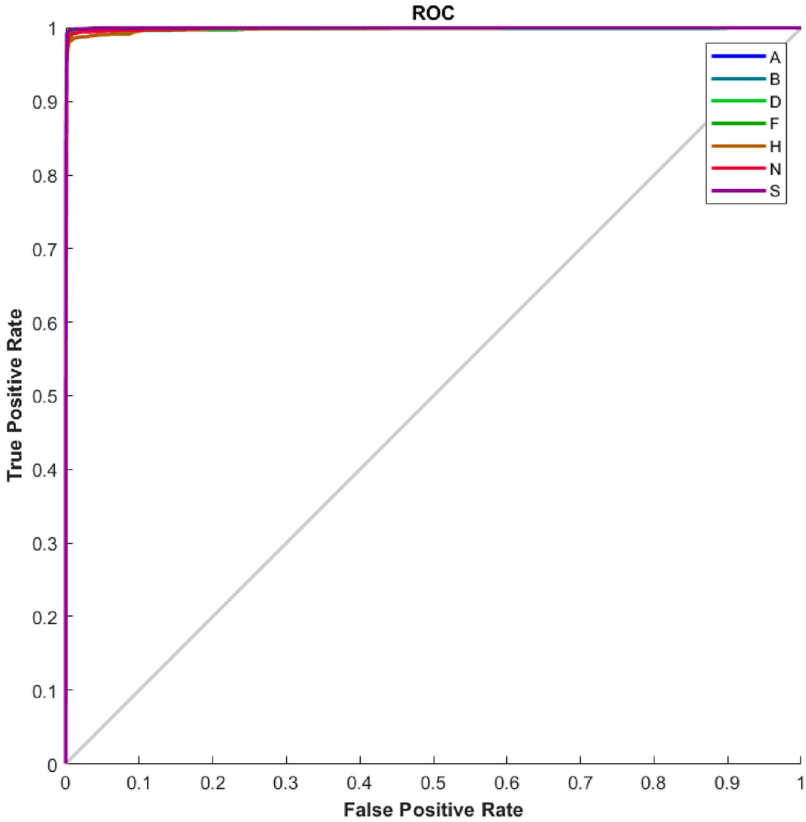


Figure 6. ROC curve using deep PRN with the new set of 1D features on the Emo-DB dataset.

$$PCS = \frac{TP}{TP + FP}; \tag{16}$$

$$Sn = \frac{TP}{TP + FN}; \tag{17}$$

$$Sp = \frac{TN}{TN + FP}; \tag{18}$$

$$MC = \frac{FP + FN}{TP + TN + FP + FN}; \tag{19}$$

$$F1 = 2 \times \frac{PCS \times Sn}{PCS + Sn}; \tag{20}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}; \tag{21}$$

where *TN*, *FP*, *FN*, and *TP* are true-negative, false-positive, false-negative, and true-positive,

Table 1. Comparison of the proposed method with the others on the Emo-DB dataset.

Solution	Dataset	Feature	Method	Result
Kuchibhotla et al. (2016)	Emo-DB	360 1D features	RDA + SFFS	92.6%
Zhao et al. (2019)	Emo-DB	Audio clip	1D & 2D CNN LSTM networks	92.3%
Chen et al. (2020)	Emo-DB	16 basic features + 12 statistics values	TLFMRF	85.6%
Ours	Emo-DB	27 1D features	PRN + OSW	98.5%

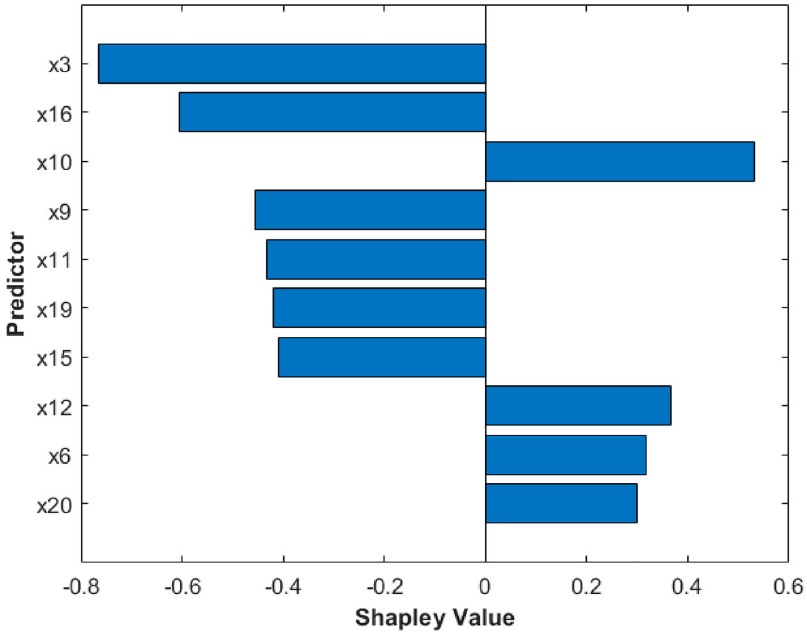


Figure 7. Shapley values impact on prediction of deep PRN with the new set of 1D features on the Emo-DB dataset.

respectively. Moreover, the receiver operating characteristic (ROC) curve has also been depicted to evaluate the quality of a multiclass classifier.

4.3. Results on the Emo-DB dataset

The deep PRN's confusion matrix, which employs a new set of 1D characteristics to confuse the predicted output class and the target class, is shown in Figure 5. The letters A, B, D, F, H, N, and S, respectively, stand for feelings of anger, boredom, disgust, fear, happy, neutral, and sad. According to the confusion matrix, the deep PRN's typical accuracy while employing the new set of 1D features is 98.5%. Additionally, Figure 6 shows the ROC curves of classifying seven emotions, where both the true-positive and true-negative rates of classifying all emotions are reaching 1. Besides, according to the ROC curves in Figure 6, the positive and negative rates of classifying all seven emotions are likely to be the same. As a result, the overall accuracy is higher.

This work performs better than several earlier attempts by employing 1D characteristics for the SER. As presented in Table 1, in comparison to the feature sets with classifier techniques in Chen et al. (2020), Kuchibhotla et al. (2016), and Zhao et al. (2019), the new

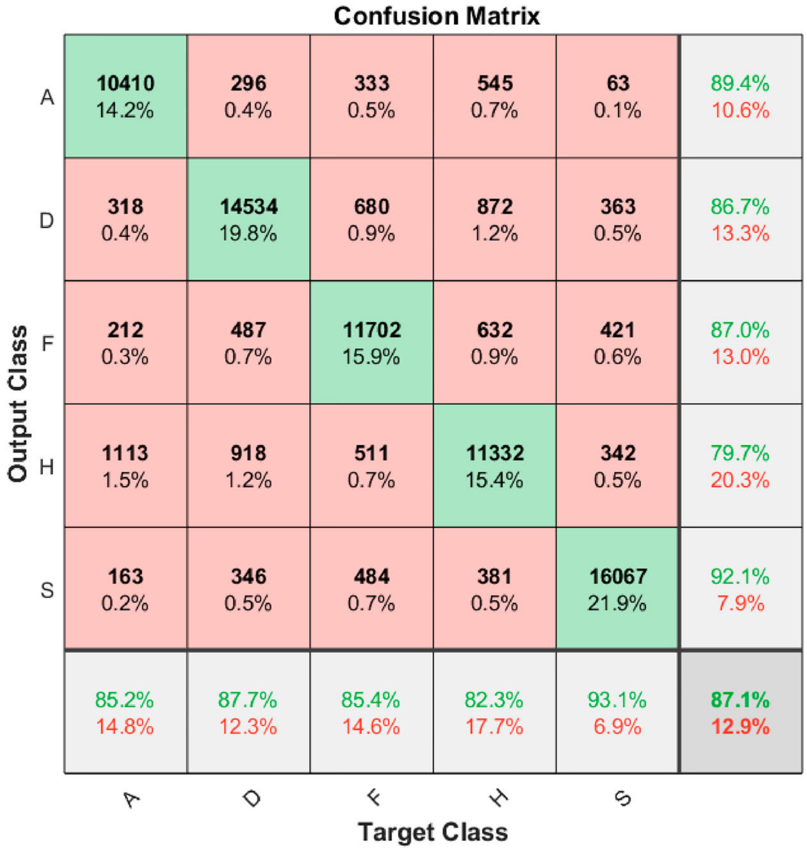


Figure 8. Confusion matrix using deep PRN with the new set of 1D features on the AESSD dataset.

set of 1D features with the deep PRN has greater accuracy, with the improvements on the Emo-DB dataset of 5.9%, 6.2%, and 12.9%, respectively.

Also shown in Figure 7 are the Shapley values, which use SHAP analysis to illustrate how each feature from the new collection of 1D features contributed to the deep PRN for the SER. As seen in Figure 7, $x3$, $x16$, and $x10$, respectively, have a significant effect on the target emotions' prediction, as evidenced by their values of fe_{smrv}, fe'_{rmsl}, and fe'_{rmsv}.

4.4. Results on the AESSD dataset

The deep PRN's confusion matrix, which employs a new set of 1D characteristics to confuse the predicted output class and the target class, is shown in Figure 8. The letters A, D, F, H, and S stand for anger, disgust, fear, happiness, and sadness, respectively. According to the confusion matrix, the deep PRN's typical accuracy while employing the new set of 1D features is 87.1%. Additionally, Figure 9 displays the ROC curves of classifying five emotions, where all the ROC curves are reaching for the upper left corner. As shown in Figure 9, the positive and negative rates of classifying all emotions are not the same. With significant improvements of 7.1% and 13.1% over the works in Vryzas et al. (2018) and (Vryzas, Matsiola et al., 2018), respectively, this study is also superior to those works, as summarized in Table 2.

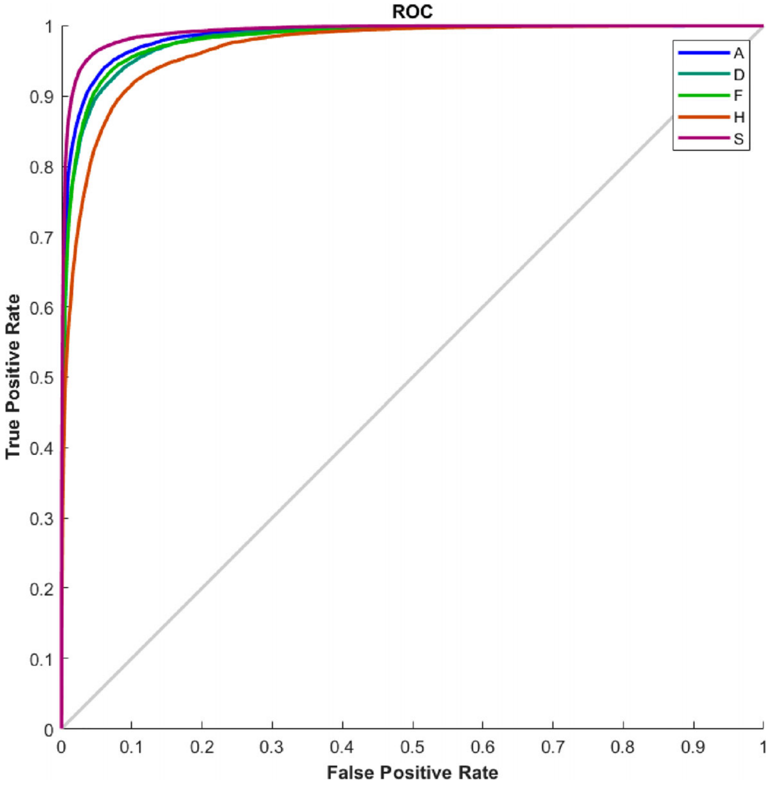


Figure 9. ROC curve using deep PRN with the new set of 1D features on the AESSD dataset.

Table 2. Comparison of the proposed method with the others on the AESSD dataset.

Solution	Dataset	Feature	Method	Result
Vryzas et al. (2018)	AESSD	40 features: TD-FD features, spectral statistics, and cepstral	Decision Tree (LMT)	~80%
Vryzas, Matsiola et al. (2018)	AESSD	Featurebased SER and performance is evaluated in Vryzas et al. (2018)		~74%
Ours	AESSD	27 1D features	PRN + OSW	87.1%

In order to demonstrate how each feature from the new collection of 1D features contributed to the deep PRN for the SER, Figure 10 shows the Shapley values using SHAP analysis. As seen in Figure 10, $x19$, $x15$, and $x12$, which stand for fe''_{rmsv}, fe'_{rmsf}, and fe'_{smrv}, respectively, have a significant impact on the prediction of the target emotions.

Moreover, Table 3 presents the performance evaluation of the proposed method on the Emo-DB and AESSD datasets with different evaluation metrics. As shown in Table 3, the *ACC*, *Sn*, and *Sp* scores are somehow the same on the Emo-DB dataset, but they are more different from each other on the AESSD dataset. As a result, the results of the proposed method on the Emo-DB dataset are better than those on the AESSD one in terms of effectiveness and robustness.

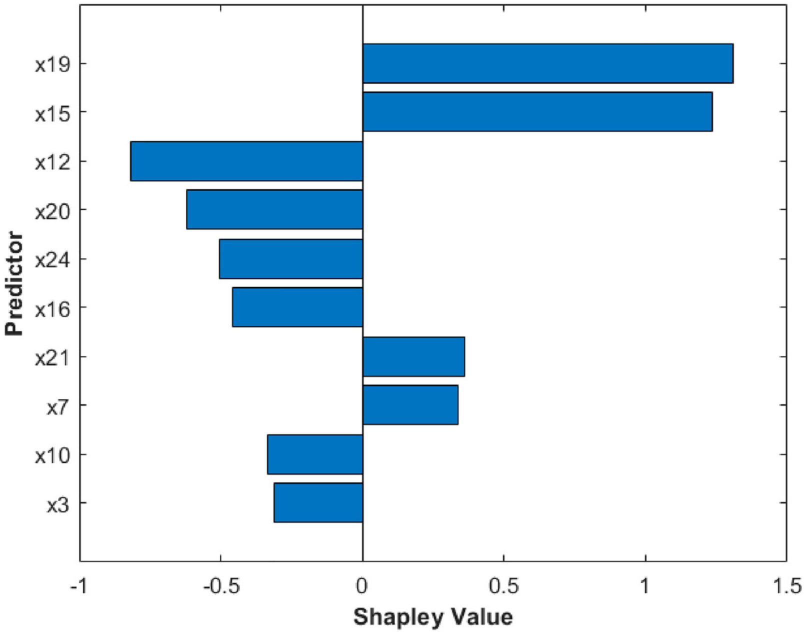


Figure 10. Shapley values impact on prediction of deep PRN with the new set of 1D features on the AESSD dataset.

Table 3. Performance evaluation on different metrics.

Dataset	*ACC*	*PCS*	*Sn*	*Sp*	*MC*	*F1*	*MCC*
Emo-DB	0.9958	0.9854	0.9854	0.9976	0.0042	0.9854	0.9829
AESSD	0.9484	0.8711	0.8711	0.9678	0.0516	0.8711	0.8388

4.5. Discussion

Our suggested strategy is superior and more comparable to recent research that employed 1D characteristics for voice emotion detection. Our suggested method is more comparable to earlier and current studies that use deep learning-based architectures like CNN and RNN and their pertinent architectures in terms of computing costs and computational resources because it uses a new set of features that includes all 1D features, and the deep PRN is a neural network-based architecture. As a result, the proposed method can be trained and deployed on a system without a GPU requirement.

As observed in the experiments, the OSW technique has played an augmentation role in this study to generate more and more features, so we can capture as much information as possible about the emotional states from each SW. Hence, using OSW also helps in improving the accuracy of the proposed method. Besides, although using OSW might lead to increased processing time, it will give more points of view to analyze and result in smoother performance over time.

Additionally, experimental findings demonstrate that our suggested approach is capable of handling the unbalanced data in the Emo-DB dataset. Moreover, system developers can choose the ideal characteristics for the intended system based on SHAP analysis.

5. Conclusion and future work

This paper suggests a brand-new set of 1D characteristics for voice emotion identification. Additionally, the deep Pattern Recognition Network is used to train a classifier system to identify emotions in speech. According to experiments, the new collection of 1D characteristics not only enhances accuracy but also the functionality of the SER system and requires less processing power. Another important consideration is the usage of the OSW approach, which may be applied as a data augmentation strategy to enhance speech emotion detection characteristics and prevent the deep PRN model from overfitting.

In order to allow users to trade-off between accuracy and performance and select an appropriate collection of features for the desired system, the SHAP analysis is also used to assess the contribution of each feature in the new set of 1D features. The new set of 1D characteristics may also be easily applied to other applications, such as fault identification and diagnosis, structural damage detection, and ambient sound categorization, as it is based on both time-domain and frequency-domain features.

Future research will examine wavelet features (Shegokar & Sircar, 2016) to create a feature set for the SER that is both optimal and resilient and contains a variety of characteristics. Prior to generating the deep learning model, it is also thought to apply an optimal clustering approach (Nguyen et al., 2022) to separate and choose the relevant characteristics. Additionally, a previously developed adaptive-neuro-fuzzy inference system (ANFIS) (Nguyen et al., 2017) is also considered while designing a new classifier system, called deep-ANFIS, to identify emotions in speech. Finally, a hybrid data augmentation approach (Pham et al., 2021) is required to produce and synthesize an increasing amount of data to address the imbalance and shortage of data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Nhat Truong Pham and Sy Dzung Nguyen would like to gratefully thank the Vietnam National Foundation for Science and Technology Development (NAFOSTED) for the support provided under grant number 107.01-2019.328.

ORCID

Nhat Truong Pham http://orcid.org/0000-0002-8086-6722
Sy Dzung Nguyen http://orcid.org/0000-0002-0145-7219
Bich Ngoc Hong Pham http://orcid.org/0000-0003-3182-1576
Duc Ngoc Minh Dang http://orcid.org/0000-0001-9302-3129

References

Abdel-Hamid, L. (2020). Egyptian arabic speech emotion recognition using prosodic, spectral and wavelet features. *Speech Communication*, *122*, 19–30. https://doi.org/10.1016/j.specom.2020.04.005

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S., & Wook Baik, S. (2019). Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, *78*(5), 5571–5589. https://doi.org/10.1007/s11042-017-5292-7

Bao, F., Neumann, M., & Vu, N. T. (2019). Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In G. Kubin, Z. Kacic (Eds.), *Interspeech 2019, 20th Annual conference of the international speech communication association* (pp. 2828–2832). ISCA.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. In *INTERSPEECH 2005 – Eurospeech, 9th european conference on speech communication and technology* (pp. 1517–1520). ISCA.

Chatterjee, R., Mazumdar, S., Sherratt, R. S., Halder, R., Maitra, T., & Giri, D. (2021). Real-time speech emotion analysis for smart home assistants. *IEEE Transactions on Consumer Electronics*, *67*(1), 68–76. https://doi.org/10.1109/TCE.30

Chen, L., Su, W., Feng, Y., Wu, M., She, J., & Hirota, K. (2020). Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences*, *509*, 150–163. https://doi.org/10.1016/j.ins.2019.09.005

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*(10), 1440–1444. https://doi.org/10.1109/LSP.97

Clement, S., Bellizzi, S., Cochelin, B., & Ricciardi, G. (2014). Sliding window proper orthogonal decomposition: application to linear and nonlinear modal identification. *Journal of Sound and Vibration*, *333*(21), 5312–5323. https://doi.org/10.1016/j.jsv.2014.05.035

Domino, K., & Gawron, P. (2019). An algorithm for arbitrary-order cumulant tensor calculation in a sliding window of data streams. *International Journal of Applied Mathematics and Computer Science*, *29*(1), 195–206. https://doi.org/10.2478/amcs-2019-0015

Garg, S., Patro, R. K., Behera, S., Tigga, N. P., & Pandey, R. (2021). An overlapping sliding window and combined features based emotion recognition system for eeg signals. *Applied Computing and Informatics*. https://doi.org/10.1108/ACI-05-2021-0130

Giripunje, S., & Bawane, N. (2007). ANFIS based emotions recognision in speech. In *International conference on knowledge-based and intelligent information and engineering systems* (pp. 77–84). Springer.

Gong, M., & Luo, Q. (2007). Speech emotion recognition in web based education. In *2007 IEEE international conference on grey systems and intelligent services* (pp. 1082–1086). IEEE.

Jackson, P., & Haq, S. (2014). *Surrey audio-visual expressed emotion (savee) database*. University of Surrey.

Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*, *23*(1), 45–55. https://doi.org/10.1007/s10772-020-09672-4

Kuchibhotla, S., Vankayalapati, H. D., & Anne, K. R. (2016). An optimal two stage feature selection for speech emotion recognition using acoustic features. *International Journal of Speech Technology*, *19*(4), 657–667. https://doi.org/10.1007/s10772-016-9358-0

Kusuma, B. A., & Nuryani, N. (2019). Heart sounds determination based on sliding window maximum method. In *Journal of physics: conference series* (p. 012075). IOP Publishing.

Lerch, A. (2012). *An introduction to audio content analysis: applications in signal processing and music informatics*. Wiley-IEEE Press.

Li, J., Li, M., Yao, X., & Wang, H. (2018). An adaptive randomized orthogonal matching pursuit algorithm with sliding window for rolling bearing fault diagnosis. *IEEE Access*, *6*, 41107–41117. https://doi.org/10.1109/Access.6287639

Lin, R., Liu, Z., & Jin, Y. (2021). Instantaneous frequency estimation for wheelset bearings weak fault signals using second-order synchrosqueezing s-transform with optimally weighted sliding window. *ISA Transactions*, *115*, 218–233. https://doi.org/10.1016/j.isatra.2021.01.010

Lingampeta, D., & Yalamanchili, B. (2020). Human emotion recognition using acoustic features with optimized feature selection and fusion techniques. In *2020 International conference on inventive computation technologies (ICICT)* (p. 221–225). IEEE.

Liu, X., Mou, Y., Ma, Y., Liu, C., & Dai, Z. (2020). Speech emotion detection using sliding window feature extraction and Ann. In *2020 IEEE 5th international conference on signal and image processing (ICSIP)* (pp. 746–750). IEEE.

Lundberg, S. M., & Lee, S. (2017, December 4–9). *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, California, USA

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access*, *7*, 125868–125881. https://doi.org/10.1109/Access.6287639

Mustaqeem, , Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, *8*, 79861–79875. https://doi.org/10.1109/Access.6287639

Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech. In: F. Lacerda (Ed.), *Interspeech 2017, 18th annual conference of the international speech communication association* (pp. 1263–1267). ISCA.

Nguyen, D., Nguyen, D. T., Zeng, R., Nguyen, T. T., Tran, S. N., Nguyen, T., Sridharan, S., & Fookes, C. (2021). Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition. *IEEE Transactions on Multimedia*, *24*, 1313–1324. https://doi.org/10.1109/TMM.2021.3063612

Nguyen, S. D., Choi, S. B., & Seo, T. I. (2017). Recurrent mechanism and impulse noise filter for establishing anfis. *IEEE Transactions on Fuzzy Systems*, *26*(2), 985–997. https://doi.org/10.1109/TFUZZ.2017.2701313

Nguyen, S. D., Nguyen, V. S. T., & Pham, N. T. (2022). Determination of the optimal number of clusters: a fuzzy-set based method. *IEEE Transactions on Fuzzy Systems*, *30*(9), 3514–3526. https://doi.org/10.1109/TFUZZ.2021.3118113

Pham, N. T., Dang, D. N. M., & Nguyen, S. D. (2020). A method upon deep learning for speech emotion recognition. *Journal of Advanced Engineering and Computation*, *4*(4), 273–285. https://doi.org/10.25073/jaec.202044.311

Pham, N. T., Dang, D. N. M., & Nguyen, S. D. (2021). Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. preprint arXiv:210909026.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. J. Smola, A. Smola, C. Aggarwal, D. Shen, and R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.

Shegokar, P., & Sircar, P. (2016). Continuous wavelet transform based speech emotion recognition. In T. A. Wysocki, B. J. Wysocki (Eds.), *10th International conference on signal processing and communication systems, ICSPCS* (pp. 1–8). IEEE.

Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., & Nanayakkara, S. (2020). Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, *8*, 176274–176285. https://doi.org/10.1109/Access.6287639

Soualhi, A., Medjaher, K., & Zerhouni, N. (2014). Bearing health monitoring based on hilbert–huang transform, support vector machine, and regression. *IEEE Transactions on Instrumentation and Measurement*, *64*(1), 52–62. https://doi.org/10.1109/TIM.2014.2330494

Swain, M., Sahoo, S., Routray, A., Kabisatpathy, P., & Kundu, J. N. (2015). Study of feature combination using HMM and SVM for multilingual odiya speech emotion recognition. *International Journal of Speech Technology*, *18*(3), 387–393. https://doi.org/10.1007/s10772-015-9275-7

Tran, T., Huy, K. B., Pham, N. T., Carratù, M., Liguori, C., & Lundgren, J. (2021). Separate sound into STFT frames to eliminate sound noise frames in sound classification. In *IEEE symposium series on computational intelligence, SSCI 2021* (pp. 1–7). IEEE.

Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C., & Kalliris, G. (2018). Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, *66*(6), 457–467. https://doi.org/10.17743/jaes.2018.0036

Vryzas, N., Matsiola, M., Kotsakis, R., Dimoulas, C., & Kalliris, G. (2018). Subjective evaluation of a speech emotion recognition interaction framework. In S. Cunningham, R. Picking (Eds.), *Proceedings of the audio mostly 2018 on sound in immersion and emotion* (pp. 341–347). ACM.

Wu, J., Wu, C., Cao, S., Wing Or, S., Deng, C., & Shao, X. (2018). Degradation data-driven time-to-failure prognostics approach for rolling element bearings in electrical machines. *IEEE Transactions on Industrial Electronics*, *66*(1), 529–539. https://doi.org/10.1109/TIE.41

Yoon, W., Cho, Y., & Park, K. (2007). A study of speech emotion recognition and its application to mobile services. In J. Indulska, J. Ma, L. T. Yang, T. Ungerer, J. Cao, (Eds.), *4th International conference on ubiquitous intelligence and computing UIC 2007, Proceedings, lecture notes in computer science* (Vol. 4611. pp. 758–766). Springer.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1d & 2d CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035

Zhu, A., & Luo, Q. (2007). Study on speech emotion recognition system in e-learning. In J. A. Jacko (Ed.), *12th International conference on human-computer interaction. HCI intelligent multimodal interaction Environments, HCI international 2007, Proceedings, Part III, Lecture Notes in Computer Science* (Vol. 4552, pp. 544–552). Springer.