Neural Computing and Applications (2022) 34:17581–17599
https://doi.org/10.1007/s00521-022-07410-2

ORIGINAL ARTICLE

Bidirectional parallel echo state network for speech emotion recognition

Hemin Ibrahim[1] · Chu Kiong Loo[1] · Fady Alnajjar[2]

Received: 6 December 2021 / Accepted: 9 May 2022 / Published online: 31 May 2022

Abstract

Speech is an effective way for communicating and exchanging complex information between humans. Speech signal has involved a great attention in human-computer interaction. Therefore, emotion recognition from speech has become a hot research topic in the field of interacting machines with humans. In this paper, we proposed a novel speech emotion recognition system by adopting multivariate time series handcrafted feature representation from speech signals. Bidirectional echo state network with two parallel reservoir layers has been applied to capture additional independent information. The parallel reservoirs produce multiple representations for each direction from the bidirectional data with two stages of concatenation. The sparse random projection approach has been adopted to reduce the high-dimensional sparse output for each direction separately from both reservoirs. Random over-sampling and random under-sampling methods are used to overcome the imbalanced nature of the used speech emotion datasets. The performance of the proposed parallel ESN model is evaluated from the speaker-independent experiments on EMO-DB, SAVEE, RAVDESS, and FAU Aibo datasets. The results show that the proposed SER model is superior to the single reservoir and the state-of-the-art studies.

Keywords Speech emotion recognition · Reservoir computing · Random resampling · Recurrent neural network

1 Introduction

Emotion in speech is considered a basic principle of human interaction and plays an important role in decision making, learning, and daily communications. Additionally, speech as a fast and effective method to communicate can be measured as a valuable mechanism for human-computer interaction (HCI). Identifying emotions from speech signals can have an effective role in several services, such as call center services for checking the customer's emotion during the call to provide better assistance [1]. Additionally, it can be useful for the in-car board system which can detect the driver's depressed status to provide more safety, because the driver's emotional state often affects the driving performance [2]. The emotion recognition system is also valuable for interactive educational systems, which can be able to truthfully identify a child's emotions that helps for positive evaluations [3]. Moreover, it is used to automatically classify the children's personalities through their speech when they are interacting with computers [4]. However, detecting emotions from speech is a big challenging task in the field of artificial intelligence and human-machine interface application [5]. Speech as a human physiological signal has several dependencies that are affected for recognizing emotions such as gender, culture, age, and health.

One of the most difficult problems in speech emotion recognition (SER) systems for researchers is to explore, catch and extract the most related and effective emotion features from the raw speech signal. Therefore, the performance of SER systems mainly depends on how the relevant emotion features are extracted [6]. There are two

✉ Chu Kiong Loo
ckloo.um@um.edu.my

Hemin Ibrahim
hemn@c4kurd.com

Fady Alnajjar
fady.alnajjar@uaeu.ac.ae

[1] Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

[2] College of Information Technology, UAE University, Al Ain, United Arab Emirates

main methods to extract features which are deep learned emotion features and handcrafted features. Each sample can be represented as one vector when global handcrafted features are adopted. Features can also be extracted locally from the speech signal frames when each sample will be represented as several time steps each as a vector of features.

Besides from choosing the most related emotional features from speech signals, developing an efficient model is another important step to have a better SER system. Hence, researchers in the SER area examine many approaches for better performance and an efficient system that recognizes emotion from speech. In the first decade in the current century, support vector machine (SVM) has been widely used in many works and gains a good performance. However, in the last decade, research works started focusing on deep learning models which have become a promising approach to gain better performance compared to classic models [7, 8]. Additionally, frame-based features and multivariate time series data have been adopted in some models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) [9, 10].

The training nature of the parameters for the majority of deep learning models, such as LSTM, increases the time complexity and requires a large amount of data and hardware resources. For example, authors in [11] reported that the train time duration for their experiments on GPU with deep neural network model was 2–14 days when they applied it on the IEMOCAP dataset. Additionally, the proposed deep learning model from [12] suffered from high computational costs and high memory required for their experiments. The computationally expensive, low convergence speed, and high memory requirements were the major drawbacks for using the deep learning model in [12].

To avoid the complexity problem of deep learning models, some researchers have used echo state network (ESN) as a special type of reservoir computing and RNN for SER systems. Researchers in [13] proposed the functional echo state network (FESN) model to adopt temporal dependency of time series data to reduce the time complexity. They validated the FESN model on UCR Time Series Data [14] as a common time series datasets and achieved a comparable or outstanding performance compared with other proposed models for temporal data. Additionally, [15] used multivariate time series emotion speech features with ESN for an efficient SER system. The paper conducted some experiments, which compared the time consuming of the LSTM and ESN models that adopted for SER and showed that ESN is much less time consuming than LSTM. The main reported advantage of ESN is that it has a simple architecture as it contains the input layer, a reservoir layer, and the output layer [16].

The reservoir layer has sparsely connected neurons that are randomly assigned without training. The input data contain multivariate time series data which will be multiplied by the reservoir input weights, the output of the reservoir will be processed inside the reservoir layer based on the nodes and their consequent reservoir sparse weights. Figure 1 shows the three main layers in ESN, which are input layer, reservoir layer, and output layer.

The small size of current speech emotion benchmark datasets makes roadblocks for some SER models. Therefore, some techniques such as data augmentation [17] including bidirectional signal are used to feed more information to the classifiers [15]. The trainless weights of the reservoir in the ESN, by assigning non-trainable random weights, lead to avoiding the time complexity of deep recurrent networks [18] and make ESN a candidate for real-time applications [19] such as time series forecasting.

Instability in ESN has been addressed in some works because in the reservoir layer, the weights are assigned randomly only once and fixed [16]. To overcome this issue some researchers used bidirectional input which feeds the data to the reservoir layer in both forward and backward directions to capture two distinct versions of information from the input layer [20], which will improve the memorization task [21]. Additionally, [22] proposed DeepESN to overcome the ESN instability problem by adopting multiple reservoir layers and obtained a valuable effect to improve the ESN model. Most of the speech emotion datasets such as EMO-DB, SAVEE, and FAU Aibo are imbalanced (see Fig. 3); hence, the data balancing approaches such as over-sampling or under-sampling are necessary to reduce the impact of the class imbalance on emotion recognition systems [3].

In this work, we proposed a novel bidirectional reservoir computing model by adopting two parallel reservoirs, when the same direction output from the different reservoirs is fused together lately. The proposed model used parallel

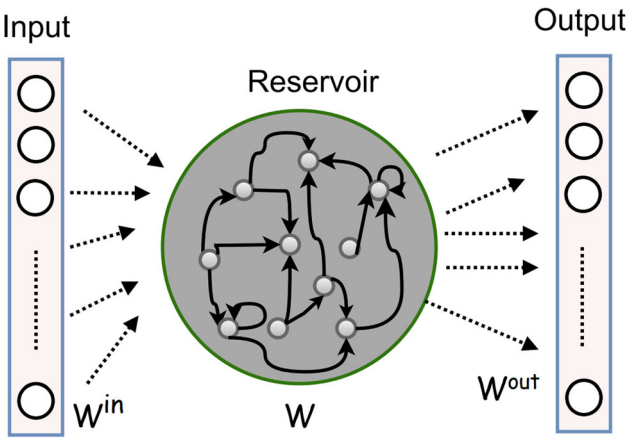


Fig. 1 The basic structure of the ESN model

reservoirs to create a more typical representation and capture independent information of the input data. To overcome the common imbalanced issue of the emotional datasets, we adopted the use of random over-sampling and random under-sampling approaches where samples are duplicated randomly or removed randomly for over-sampling and under-sampling approaches, respectively. Moreover, multivariate time series handcrafted features such as Mel-frequency cepstral coefficients (MFCCs) and Gammatone cepstral coefficients (GTCCs) are extracted to feed the reservoir layers. This novel proposed model to recognize emotion from speech with its trainless nature assists to improve the classification accuracy.

The remainder of this paper is divided as follows: Sect. 2 presents the literature review of the SER models, and Sect. 3 shows the methodology of the proposed model. Section 4 explains the involved datasets in this work, and the experimental setup is presented in Sect. 5. The detailed results for all four datasets and discussion of the performance of the proposed model are present in Sect. 6, and finally, the conclusion and future work come in Sect. 7.

2 Literature review

The speech signals have been widely used to recognize emotions which are preferred as a better interaction in the field of HCI. The SER design traditionally focuses on the extraction of robust emotion features from a speech in addition to the use of a proper classification model [23]. Feature extraction is a challenging task in SER models, for instance, some researchers are preferring deep learned features, while others are using handcrafted features.

One of the recent approaches for extracting features is learning from the deep learning models, and thus, many researchers have designed learned features using deep learning models [24]. Authors in [25] adopted LSTM to learn the extracted frame-based features, and the 3-D log mel-spectrogram features are learned by using a convolutional neural network (CNN). Researchers in [26] were focusing on learned emotion features directly from mel-spectrograms by adopting CNN deep learning model, which helped their model performance to improve by over 3%.

The handcrafted features for a speech sample can be globally represented in one vector, or it can also be represented locally when its extracted from frames. There are many toolkits for extracting emotional features from speech signals such as the openSMILE [27] toolkit which is an open-source toolkit for extracting global features and DeepSpectrum [28] toolkit for extracting global and local deep features. The openSMILE toolkit has been used in many studies for extracting non-temporal global features, such as [29] who used it for extracting global features from the raw signal, authors in [30] used openSMILE with SVM, and [9] adopted the low-level descriptions (LLDs) features that are extracted by openSMILE to feed it to bi-directional long-short term memory (bi-LSTM) with Directional Self-Attention deep learning model. The log-energy, pitch, TEO, ZCR, and MFCC have been identified as important emotion features with the RBF neural network model [31]. Authors in [32] extracted 113 acoustic handcrafted features and combined with textual features to input the Bimodal Deep Autoencoder (EBDA) model to recognize emotions from the large data from the Internet. However, some authors preferred time series frame-based features to recognize emotions from speech. Researchers in [19] used frame-based spectral features and fed them to the ESN model to detect emotions in real-time. Additionally, [33] used various handcrafted features with 512 frames for each speech sample with the use of CNN and bi-LSTM deep model. Therefore, the multivariate time series data representation requires a convenient and computationally intensive classifier such as RNN.

Another important aspect to building SER systems besides feature extraction from the speech is developing a robust mathematical emotion classification model. The strong classification has a vital role for better performance to detect emotions from speech signals [34]. Authors in [8] proposed a multi-learning trick deep learning model based on 1D dilated CNN architecture to recognize emotion from speech. [35] used a bidirectional long short-term memory (BLSTM) model with high-level features and combined it with maximum-likelihood-based learning process for emotion recognition. The artificial neural network (ANN) with one hidden layer and SVM classifiers have been used in [3] and reported that the ANN classifier has a better performance than SVM in a pairwise approach for SER. Due to the necessity of having a large amount of data for deep learning models, the data augmentation approach is adopted in [36] on Acted Emotional Speech Dynamic Database (AESDD) for continuous emotion recognition from speech with the use of the CNN model.

However, there are few works that reported the use of ESN in the field of detecting emotion from speech. Authors in [19] proposed the ESN model for real-time SER, although the work could not provide a successful real-time system. [37] investigates a preliminary system for automatic emotion recognition from a speech by inputting a time series features to an ESN model which is trained on a multi-classification task over different classes of emotion. Additionally, authors in [38] are using only neutral and anger emotional data with an ESN model by using the memristive circuits for real-time SER. However, in [15], the bidirectional multivariate time series features are used to feed a single reservoir layer and the proposed model has

come out with a high classification performance in the SER system.

A single reservoir suffers from the assigned random weights, which creates instability and yields high variance since the weights are assigned randomly only once and fixed in reservoir part [16]. The single reservoir captures a representation that may not detect all of the relationships between the information of various time steps, due to the sparse connection of the reservoir in the hidden layer. Therefore, some researchers have proposed an ESN model with more than one reservoir layer. Authors in [39] used multiple reservoir layers in a series configuration and optimized the number of reservoirs with hyper-parameters. Different experimental analysis architectures are proposed in [40] such as deepESN, deepESN-IA, and groupedESN. Additionally, researchers in [41] used more than one reservoir layer and proposed a functional deep echo state network (FDESN) for multivariate time series classification to introduce temporal and spatial aggregation. The proposed grouped multi-layer ESN model in [42] helped to improve linear separability on the readout layer.

The current benchmark speech emotion datasets are problematic due to the class imbalanced data which influences the classification techniques significantly. To overcome this issue, [3] followed the Synthetic Minority Oversampling TEchnique (SMOTE) to increase the number of samples in the minority class. Furthermore, authors in [43] used under-sampling by reducing the size of the majority emotion class to the same size of the minority class.

3 Methodology

In this section, the design of the ESN-based proposed model is presented. The proposed model in [15] gained a good performance by using late fusion of bidirectional ESN with only one reservoir layer. However, due to the instability of the ESN model, they adopted the use of optimization of many parameters such as internal units R, spectral radius (ρ), size of connectivity (β), input scaling (ω), and amount of leakage to control the timescale of its neurons inside reservoirs. The reason of randomness is basically due to the non-trained (fixed) random weights inside the reservoir layer. Logically more random representation of the data produces a more knowledgeable characterization of the space. Consequently, increasing the number of reservoirs that are parallelly fed by the input data may increase the performance of the SER model. Nevertheless, multiple reservoir layers may bias toward the classes with a high number of samples. To overcome this issue, one may apply a balancing technique on the involved datasets.

In this work, we have adopted two parallel reservoir layers and imbalanced data resampling using random oversampling and under-sampling in addition to the well-known SMOTE technique. The bidirectional time series data characterization has been reported to enhance the memorization capability [44]. However, having parallel reservoirs will produce multiple representations for each direction from the bidirectional data. The sparse random projection (SRP) [45] approach has been adopted to reduce the high-dimensional sparse output for each direction separately from both reservoirs. The outputs from dimension reduction stage are concatenated based on the same direction, such that each direction is represented separately from the other directions. Consequently, the outputs for the two forward directions after dimension reduction from both reservoirs will be concatenated separately from the concatenation of the backward direction outputs produced from the same parallel reservoirs. In this model, two fusion methods are applied; firstly, the fusion is applied after dimension reduction which is based on each direction separately, and secondly, the fusion of both forward and backward directions is applied after reservoir model space stage.

In the next stage, the reservoir model space method is applied on each direction, and their outputs will be fused lately to produce the final representation r_X of the speech signal. The ridge regression classifier takes the r_X vectors as an input to make the final decision on its emotion class label.

The structure of the proposed model is shown in Fig. 2. The next subsections present the details of the proposed model steps.

3.1 Feature extraction

The feature extraction is a primary part of any SER system that could mainly impact the model performance. Therefore, the first challenge in any SER model is to determine the most relevant emotion features that can be extracted from the raw speech signal. Moreover, selecting the right set of emotion features reflects the most available information about emotion characteristics from the speech signal. In this research, we extracted handcrafted multivariate time series features to feed the proposed model. Thirteen MFCCs are among the features that have been extracted and used in this work. MFCCs are the well-known features from speech that has a strong link to the human perception system. MFCCs are the most widely used features, most popular, and robust method for extracting handcrafted emotion features in SER systems and speech recognition in general. It was first introduced by [46] and explored that the outperformance of MFCC leads to providing a better accuracy of the short-term speech spectrum.

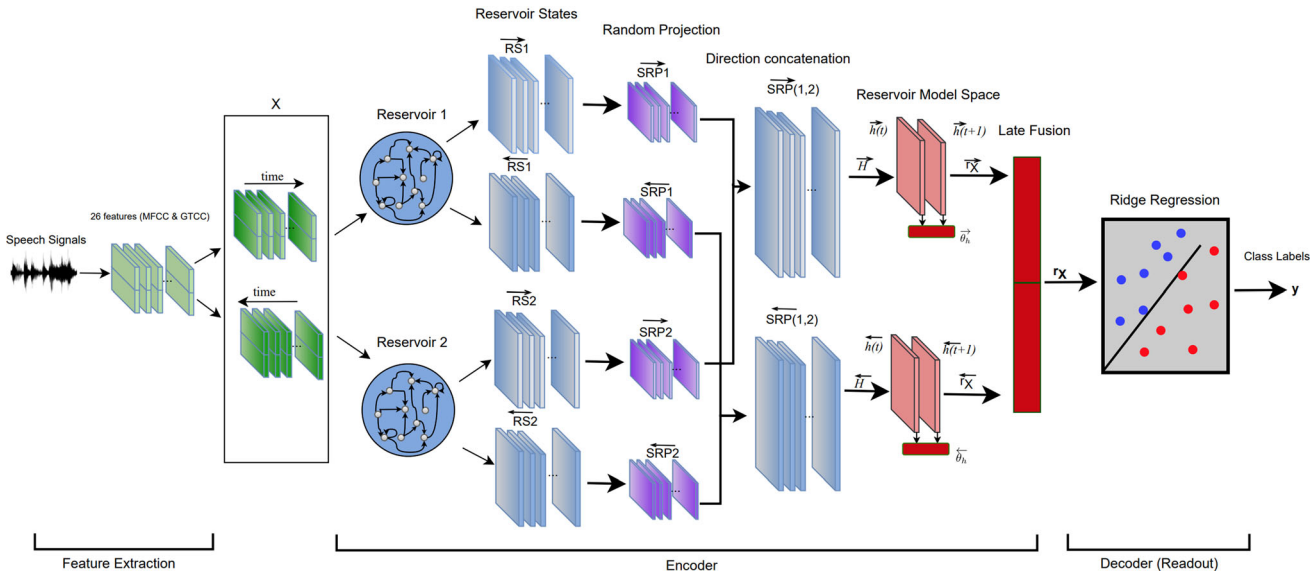


Fig. 2 The overview of the proposed model design, which presents the parallel ESN model with two reservoirs and direction concatenation with the late fusion

Nonetheless, MFC-based systems are deteriorated by not having a good performance under noisy conditions because the extracted MFCC features are biased by noise which triggers mismatched likelihood calculation [47]. Accordingly, the thirteen GTCCs extracted features help to overcome this issue which can have a better performance under noisy conditions. GTCCs can be described as a biologically inspired modulation of the MFCC by using the linear filter called Gamma-tone filters instead of the mel filter bank. Since MFCCs lack of robustness under noise conditions, we have adopted the use of GTCC features with MFCCs to produce more robust representation under noisy conditions. A total of 26 handcrafted features are used as an input to the proposed model.

For extracting both sets of features the MATLAB audio feature extractor (audioFeatureExtractor) has been adopted with windows of length 30 ms that are overlapped by 20 ms. Therefore, the length of the samples is different in all of the four datasets, and we have settled the lengths by pruning or padding with zeros at the start and the end of each sample. Thus, 500, 600, 400, and 300 frame sizes have been adopted for EMO-DB, SAVEE, RAVDESS, and FAU Aibo, respectively, based on the almost high length for each benchmark dataset.

3.2 Over-sampling and under-sampling

The class imbalanced dataset is a challenging problem in both deep learning models and traditional models for classification prediction. The unequal classes in the training set produce a classification problem because the minority classes, where they have a few samples, are hard to predict and learn the characteristics compared to the majority classes. Additionally, the imbalanced datasets may cause the models to overfit the classes that have a few samples and increase the generalization error [48]. This issue often affects the testing set performance or real-world application, however, the training set performance may have a good result. Most of the real-world datasets have an imbalanced nature, due to the challenges for collecting real data. Consequently, the data balancing and resampling approaches are necessary to reduce the impact of the class imbalance on classification models. Many research works have shown that in general, experiments on the balanced datasets are performing better than the imbalanced datasets [49]. There are two main strategies for data resampling to overcome the data imbalance issue. Firstly, over-sampling increases the minority samples, whereas the under-sampling method removes the samples from the majority classes. By increasing the minority class samples and reducing the majority class samples, the train set data can be balanced.

The adopted four datasets in this work are imbalanced in terms of sample sizes per class as shown in Fig. 3. For example, the EMO-DB dataset has 127 utterances of the anger class, but only 46 samples for disgust class and the rest of the samples per emotion class are: 81 boredom, 79 neutral, 69 fear, 71 happiness and 62 sadness utterances. Moreover, the SAVEE and RAVDESS datasets are less imbalanced, the neutral class in SAVEE participated by 120 samples and the rest are only 60 samples per class. However, the neutral class in RAVDESS has 96 utterances, but the rest are 192 samples for each class. The number of chunks per class in the FAU Aibo Emotion corpus is

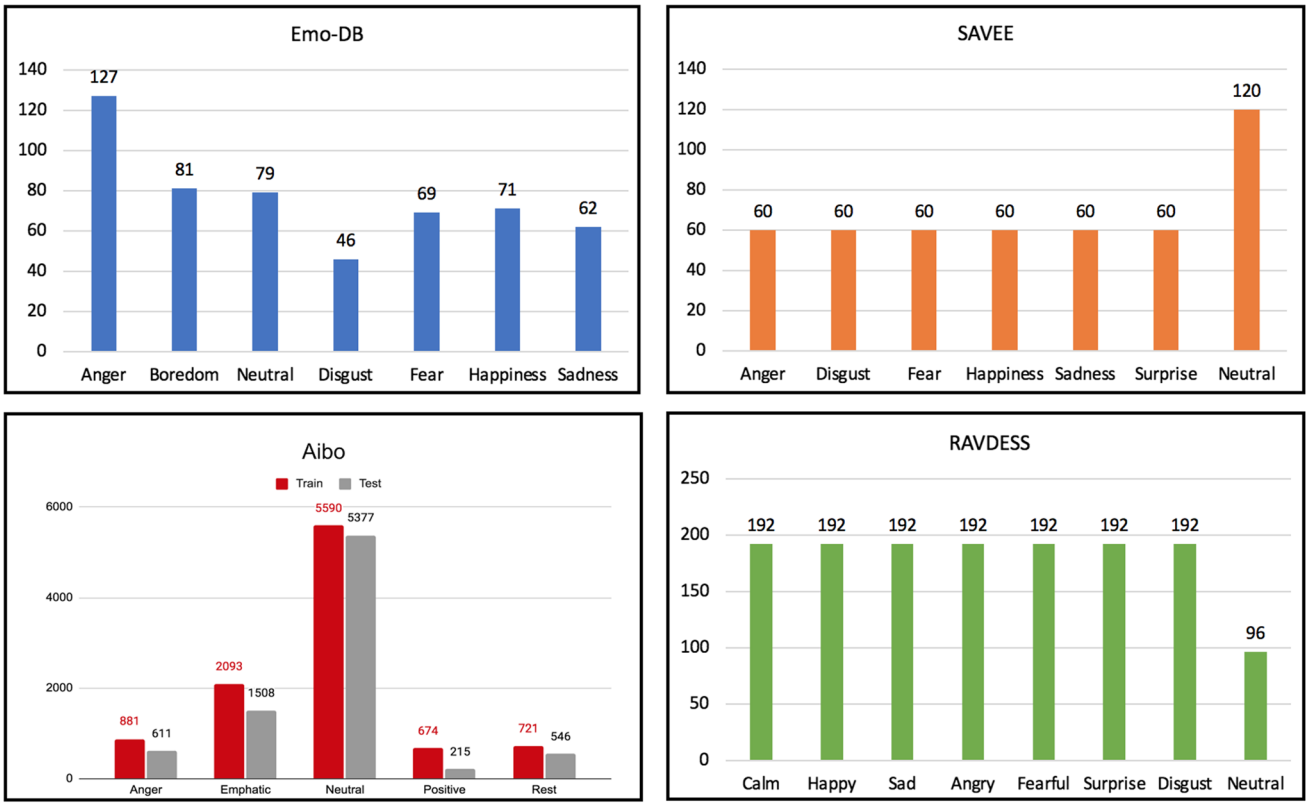


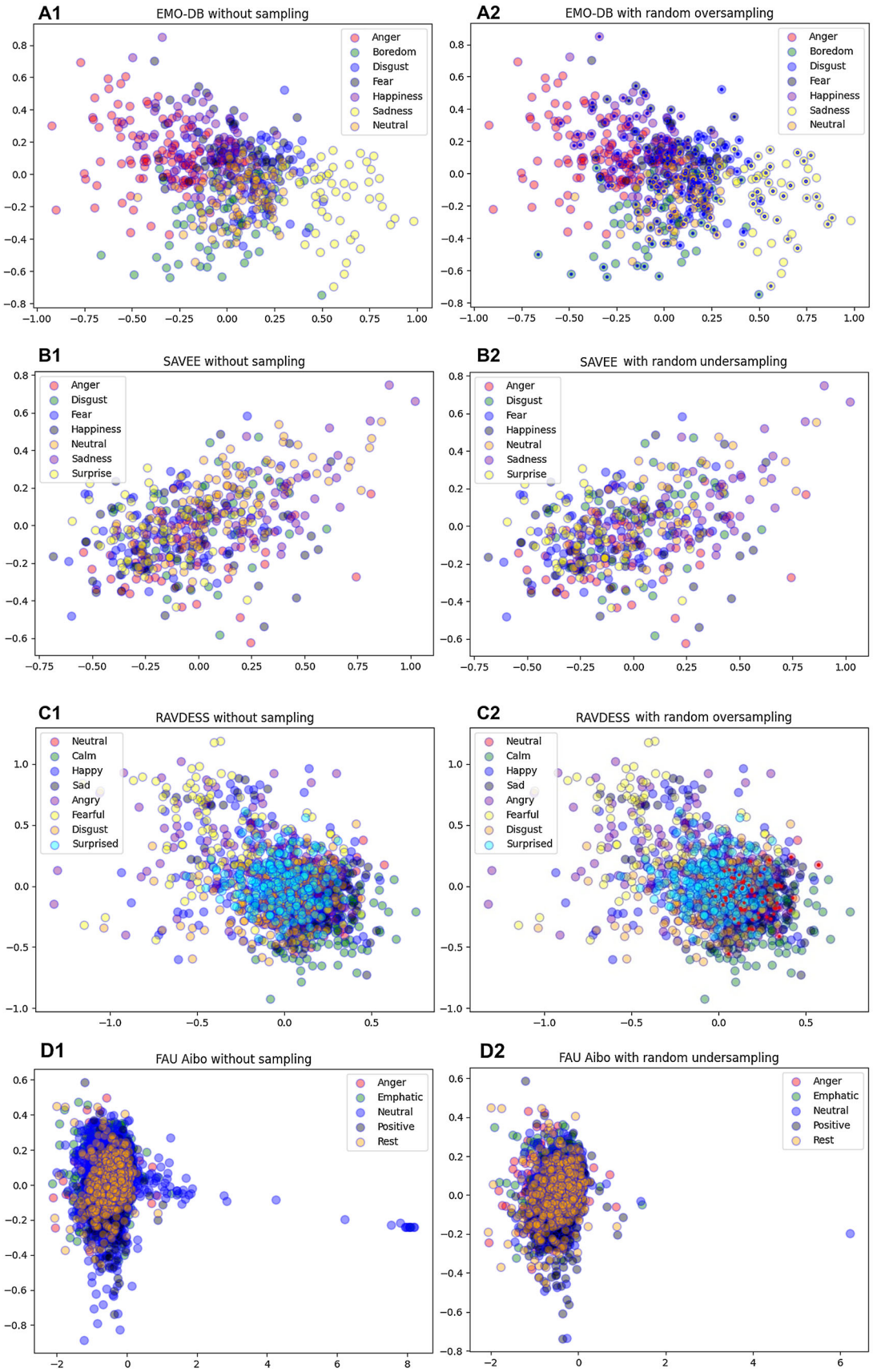
Fig. 3 The total number of emotion class samples for EMO-DB, SAVEE, RAVDESS, FAU Aibo datasets

extremely unbalanced as shown in Fig. 3, wherein the training set the 56.1% of the data are labeled as neutral, 21% are emphatic, 8.8% are angry, 6.8% are positive, and 7.2% are the rest of the other emotions.

To overcome this problem, the over-sampling approach is applied on EMO-DB and RAVDESS datasets, while the under-sampling is applied on SAVEE and FAU Aibo datasets. The reason behind adopting under-sampling on the SAVEE dataset is that all of the classes are balanced (60 samples for each class) with the exception of the neutral class which includes 120 samples. Consequently, the only class that needs balancing is the neutral class. After obtaining the extracted handcrafted features of speech signal samples, for EMO-DB and RAVDESS dataset, the random over-sampling method [50] has been adopted by selecting samples from the training set at random with replacement. Additionally, the random under-sampling method [50] has been applied on the SAVEE and FAU Aibo datasets to under sample the majority of emotion classes by randomly selecting samples with or without replacement. Overall, the random over-sampling technique is the simplest form by duplicating samples of the minority class; however, the random under-sampling method is taking away some data from the frequent class. Figure 4 shows how the over-sampling technique affects EMO-DB and RAVDESS dataset by adding more samples randomly, for example, some of the sadness emotion samples in EMO-DB dataset (as shown in A2), which is represented in yellow color, have been duplicated and denoted by a black dot inside it. On the other hand, the random under-sampling method reduces the number of samples randomly in the neutral class for SAVEE dataset from 120 samples to 60 samples. The random under-sampling reduced the number of neutral samples in FAU Aibo dataset (see Fig. 4 D1 and D2). The figures are presenting two features among 26, which have been selected randomly, and each feature represents the mean of the time steps values of that feature.

3.3 Bidirectional parallel echo state network

Echo state networks (ESNs) proposed by [51] as a powerful form of reservoir computing and a recent type of RNN for learning nonlinear systems. The reservoir computing (RC) as a computational framework is a kind of RNN model which does not have the training inside the layer and weights are initiated randomly [52]. In general, ESN has three main layers, the input layer, the hidden layer which is called the reservoir layer, and the output layer. The input data will be connected sparsely inside the reservoir layer, where nonlinear neurons are connected randomly. The main advantage of ESN is the untrained nature in the hidden layer which helps to avoid the vanishing gradient


A1
EMO-DB without sampling
Anger
Boredom
Disgust
Fear
Happiness
Sadness
Neutral
A2
EMO-DB with random oversampling
Anger
Boredom
Disgust
Fear
Happiness
Sadness
Neutral
B1
SAVEE without sampling
Anger
Disgust
Fear
Happiness
Neutral
Sadness
Surprise
B2
SAVEE with random undersampling
Anger
Disgust
Fear
Happiness
Neutral
Sadness
Surprise
C1
RAVDESS without sampling
Neutral
Calm
Happy
Sad
Angry
Fearful
Disgust
Surprised
C2
RAVDESS with random oversampling
Neutral
Calm
Happy
Sad
Angry
Fearful
Disgust
Surprised
D1
FAU Aibo without sampling
Anger
Emphatic
Neutral
Positive
Rest
D2
FAU Aibo with random undersampling
Anger
Emphatic
Neutral
Positive
Rest

◀**Fig. 4** Figures A1, B1, C1, and D1 show the four involved datasets before resampling. Figures A2 and C2 present the over-sampling applied to Emo-DB and RAVDESS datasets, respectively. The points that have the dot inside refer to the duplicated samples. Figure B2 and D2 show the effect of random under-sampling applied to the SAVEE and FAU Aibo datasets. We may observe that the number of neutral samples has been reduced by half in SAVEE dataset

issue. ESN is a promising approach for multivariate time series classification.

In our work, the multivariate time series data are applied by feeding an input sequence into two reservoirs parallelly in each forward and backward direction to catch further information separately of the input data with different random weights from both reservoir layers. These two parallel reservoirs are initializing sparse high-dimensional connected neurons without being trained. A fixed random weights in each reservoir guarantee transforming the data into two independent sub-spaces, since each set of random weights represent an independent bases of an new sub-space. We shall see that two reservoirs improve the SER performance over a single reservoir as an indication of having a complementary information in one of the reservoir to the other one.

Choosing the number of reservoirs is an open problem and depends on the application task [53]. However, researchers reported that adding more than two reservoirs is mostly not improving the performance of the model [54, 55]. Moreover, authors of [56] reported that adding layers to a deep reservoir architecture, resulting in driving toward (equally or) less stable behaviors. Consequently, in this work, a twin of reservoirs is adopted, which also helps in avoiding the time complexity resulting in adopting more than two reservoirs.

The ESN model can effectively handle time series data, since the temporal dependence data can be handled and successfully adopted by chaotic time series data prediction models. Our input data are a multivariate time series data which has D-dimensional size feature for each time step t, where $t = 1, 2, \ldots, T$, and T is considered as a number of time steps, where time t is given as $\mathrm{x}(t) \in R^D$ and the X can defined as $X = [\mathrm{x}(1), \mathrm{x}(2), \ldots, \mathrm{x}(T)]^T$. By pruning and padding samples, to prevent length differences, the number of time steps is unified to T. The states in the reservoirs are updated based on the following equations:

$$\begin{aligned} \overrightarrow{\mathrm{h_i}}(t) &= f(\overrightarrow{\mathrm{x}}(t), \overrightarrow{\mathrm{h_i}}(t-1); \theta_i^{enc}) \\ \overleftarrow{\mathrm{h_i}}(t) &= f(\overleftarrow{\mathrm{x}}(t), \overleftarrow{\mathrm{h_i}}(t-1); \theta_i^{enc}) \end{aligned} \tag{1}$$

where $i = 1, \ldots, N$, N is a number of reservoirs (in our study the $N = 2$ where we have only two reservoirs) and $h_i(.)$ is the RNN states for reservoirs at time t which compute as a function of their preceding values $(\overrightarrow{\mathrm{h_i}}(t-1), \overleftarrow{\mathrm{h_i}}(t-1))$ for reservoirs and the present input $x(t)$. Furthermore, the function $f(.)$ is a nonlinear activation function, in addition to θ_i^{enc} is the adjustable parameters from the reservoirs.

Equation (1) can be shown as the following:

$$\begin{aligned} \overrightarrow{\mathrm{h_i}}(t) &= \tanh(W_i^{in} * \overrightarrow{\mathrm{x}}(t) + W_i^{\mathrm{res}} * \overrightarrow{\mathrm{h_i}}(t-1)) \\ \overleftarrow{\mathrm{h_i}}(t) &= \tanh(W_i^{in} * \overleftarrow{\mathrm{x}}(t) + W_i^{res} * \overleftarrow{\mathrm{h_i}}(t-1)) \end{aligned} \tag{2}$$

where $i = 1, \ldots, N$, W_i^{in} is the input weights, W_i^{res} is the weights from reservoirs connections, and the reservoir states ($\overrightarrow{\mathrm{RS_i}}$ and $\overleftarrow{\mathrm{RS_i}}$) are produced by the *ith* reservoir over time, where $\overrightarrow{\mathrm{RS_i}} = [\overrightarrow{\mathrm{h_i}}(1), \overrightarrow{\mathrm{h_i}}(2), \ldots, \overrightarrow{\mathrm{h_i}}(T)]^T$ and $\overleftarrow{\mathrm{RS_i}} = [\overleftarrow{\mathrm{h_i}}(1), \overleftarrow{\mathrm{h_i}}(2), \ldots, \overleftarrow{\mathrm{h_i}}(T)]^T$. The adjustable parameters from the reservoirs can be represented as $\theta_i^{enc} = \{W_i^{in}, W_i^{\mathrm{res}}\}$.

There are some hyperparameters that are affecting the performance of the ESN model significantly, such as the internal units size, the spectral radius, the nonzero connections, scaling of the values in the input weights W_i^{in}, the leak as an amount of leakage to control the timescale in the reservoir states update and apply dropout regularization, particularly for recurrent architectures.

3.4 Dimension reduction and direction concatenation

Each reservoir has its own two-directional outputs as a result of the bidirectional input data. In this work, four generated states which are two output weights from the first reservoir in both forward and backward ($\overrightarrow{RS1}$ and $\overleftarrow{RS1}$) and two output weights from the second reservoir in both forward and backward ($\overrightarrow{RS2}$ and $\overleftarrow{RS2}$) are fed to the dimension reduction stage. The output weights have high-dimensional sparse feature representation which leads to high computational cost, over-fitting, and redundancy. The very sparse random projections technique has been applied to transform the high-dimensional sparse data into a more convenient representation and reduce the dimensions. To reduce the complexity, the SRP method minimizes the dimensions of the reservoir output without training, which is able to remove redundant data with minimal loss of information. This dimension size of reduced data can be optimized or set based on experience. This step has an important effect on applying the reservoir model space in the next stage.

The forward dimension reduction outputs ($\overrightarrow{SRP_1}$ and $\overrightarrow{SRP_2}$) from the first reservoir and second reservoir are concatenated separately from the concatenation of the backward direction produced by both reservoirs in the

dimension reduction stage. We concatenate the forward direction $\overrightarrow{SRP_1}$ with $\overrightarrow{SRP_2}$ and the backward direction $\overleftarrow{SRP_1}$ with $\overleftarrow{SRP_2}$ from both reservoirs to generate the $\overrightarrow{SRP}_{(1,2)}$ and $\overleftarrow{SRP}_{(1,2)}$. This sort of concatenation aims to process the information regarding each direction using multiple reservoirs independently from the other direction. This approach will expectedly ease modeling each direction for any time series classification application. The output from the dimensionality reduction concatenation in both forward and backward directions produce new sequences $\overrightarrow{H}$ and $\overleftarrow{H}$ which can be an input for the reservoir model space.

The sparse projection matrix R is set with 1 and -1 which are equiprobable as shown in the following equation:

$$P_r(R_{i,j} = 1) = P_r(R_{i,j} = -1) = \frac{1}{2\sqrt{d}} \tag{3}$$

where P_r refers to the probability, and d is the original feature dimensionality from the reservoir output state. This dimension size of reduced data can be optimized or set based on experience. This step has an important effect on applying the reservoir model space in the next stage. The output from the dimensionality reduction method produces new sequences $\overrightarrow{H}$ and $\overleftarrow{H}$ which can be an input for the reservoir model space.

3.5 Reservoir model space and the bidirectional fusion

Researchers in [21] proposed a reservoir model space as a self-supervised method based on reservoir computing. In this stage, for each time series date from reservoirs, ESN is trained to predict the next step in the time series. It helps to characterize a generative model of the reservoir output data and makes a metric relationship between the samples. Reservoir model space is able to predict the next reservoir states and provides the most definitive characterization of the time series feature representations. Hence, in this work, we used the reservoir model space to generate each representation of SRP output sequence data into one-dimensional feature vector, which then is processed by the readout stage.

Finally, the output of both reservoir model spaces will be fused to produce a characterization that highlights both directions separately. Therefore, the formula of the reservoir model spaces that are applied on the outputs of the unsupervised SRP dimensionality reduction method can be presented as the equations below:

$$\begin{aligned} \overrightarrow{h}(t+1) &= \overrightarrow{U_h}\,\overrightarrow{h}(t) + \overrightarrow{u_h} \\ \overleftarrow{h}(t+1) &= \overleftarrow{U_h}\,\overleftarrow{h}(t) + \overleftarrow{u_h} \end{aligned} \tag{4}$$

where the columns of a frontal slice $\overrightarrow{H}$ and $\overleftarrow{H}$ are represented by $\overrightarrow{h}(.)$ and $\overleftarrow{h}(.)$, respectively, $\overrightarrow{U_h}, \overleftarrow{U_h} \in \mathbb{R}^{D\times D}$ and $\overrightarrow{u_h}, \overleftarrow{u_h} \in \mathbb{R}^{D}$, where the size of dimension after the reduction process is represented as D. The data representation is synchronizing with the parameters as follows:

$$\begin{aligned} \overrightarrow{r_X} &= \overrightarrow{\theta_h} = [vec(\overrightarrow{U_h}); \overrightarrow{u_h}] \\ \overleftarrow{r_X} &= \overleftarrow{\theta_h} = [vec(\overleftarrow{U_h}); \overleftarrow{u_h}] \end{aligned} \tag{5}$$

The produced data output from $\overrightarrow{r_X}$ and $\overleftarrow{r_X}$ is concatenated as shown in the following equation:

$$r_X = [\overrightarrow{r_X}; \overleftarrow{r_X}] \tag{6}$$

where the r_X feeds to the ridge regression and Eq. 7 shows that how the $\overrightarrow{\theta_h}$ and $\overleftarrow{\theta_h}$ can be learned by minimizing a ridge regression loss function:

$$\begin{aligned} \overrightarrow{\theta_h^*} &= \underset{\{\overrightarrow{U_h}, \overrightarrow{u_h}\}}{\arg\,min} \frac{1}{2}\| \overrightarrow{h}(t)\,\overrightarrow{U_h} + \overrightarrow{u_h} - \overrightarrow{h}(t+1)\|^2 + \alpha\| \overrightarrow{U_h}\|^2 \\ \overleftarrow{\theta_h^*} &= \underset{\{\overleftarrow{U_h}, \overleftarrow{u_h}\}}{\arg\,min} \frac{1}{2}\| \overleftarrow{h}(t)\,\overleftarrow{U_h} + \overleftarrow{u_h} - \overleftarrow{h}(t+1)\|^2 + \alpha\| \overleftarrow{U_h}\|^2 \end{aligned} \tag{7}$$

where α is the reservoir model space regularization strength parameter to set the number of the coefficient shrinkage. The readout stage (which is a classification level in ESN) is a linear model for decoding that can be formed by the following equation:

$$y = g(r_X) = V_o r_X + v_o \tag{8}$$

The $\theta_{\text{dec}} = \{V_o, v_o\}$. θ_{dec} is a set of parameters in this model, and it can be admitted a closed form solution when the ridge regression can be learned by minimizing the loss function:

$$\theta_{\text{dec}}^* = \underset{\{V_o, v_o\}}{\arg\,min} \frac{1}{2}\|r_X V_o + v_o - y\|^2 + \mu\|V_o\|^2 \tag{9}$$

where μ is the regularization penalty parameter in ridge regression to minimize overfitting of the training set. The linear readout step is to perform the classification based on the reservoir model space output features which maps the r_X data representation into the emotion class labels y.

In order to use the ridge regression to classify the final r_X representations, the training process sets the dimension of y to be equal to the number of emotions k. The target of y has the same dimension as y, where the value of its components is all equal to zero except the dimension that

refers to the correct target of the sample, which is set to be one. In other words, when the problem is a multi-class classification, the model will have multi-output regression, and the output with the maximum value will be considered as the predicted class.

3.6 Hyperparameter optimization

Identifying hyperparameters in ESN has been reported as an issue and its effects on the performance of the ESN model. Additionally, some researchers prefer to assign parameters manually or based on experience, however, authors of [15] optimized the whole ESN parameters. In this study, we have used Bayesian optimization [57] to optimize the size of internal units R and dropouts from ESN, while fixing the value of spectral radius (ρ), size of connectivity (β), input scaling (ω), and the amount of leakage to control the timescale of its neurons inside reservoirs. The reason for fixing some of the ESN parameters is to reduce the complexity since multiple reservoirs are adopted in this work which doubles the number of parameters that need optimization. Moreover, we optimized the number of dimensions resulted from the SRP, in addition to the regularization strength parameters α in the reservoir model space and μ in the ridge regression readout stage. Optimized parameters have an important effect to improve the model performance.

3.7 Normalization

Speaker Normalization (SN) is a widely-known unsupervised approach to improve speech recognition performance tasks [58]. As performed in [59], we applied SN on each specific speaker sample in our experiments for all the four datasets which analyze emotion of the speaker independently. The SN is comprehended by a normalization of utterances by its mean and standard deviation which belong to one of the speakers. The aim behind adopting SN in our experiments is to compensate for speaker variations and prevent all samples from specific speaker influences, which help to improve emotion recognition performance.

4 Datasets

The performance of our proposed model has been evaluated and validated for detecting emotions from speech signals using three acted-speech emotion datasets, EMO-DB [60], SAVEE [61], and RAVDESS [62]. These three datasets are widely used to evaluate SER systems and are recorded by professionals where they acted to express different emotions in addition to FAU Aibo [63] as a non-acted dataset.

Berlin Database of Emotional Speech (EMO-DB) [60] is a German speech emotion dataset, which contains seven emotions: anger, boredom, neutral, disgust, fear, happiness, and sadness. EMO-DB is a widely used dataset for SER models with a total number of 535 utterances. Additionally, five males and five females participated to record their emotional states over the memories of their real-life experiences.

The second to validate our model is Surrey Audio-Visual Expressed Emotion (SAVEE) [61]. It is an English audio and visual expression acted dataset used for SER and facial expression systems, although in this work, only the audio channel has been used. The utterances are recorded by four native English speakers with seven emotional states, which are anger, disgust, fear, happiness, sadness, surprise, and neutral. There are 480 video files where each male participated by recording 120 videos.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [62] dataset is used to validate the proposed model. RAVDESS is a multimodal emotion dataset that includes voice files for song and speech with facial expressions. It was recorded by 24 professional actors of equal genders. It has eight emotional states: calm, happy, sad, angry, fearful, surprise, neutral, and disgust expressions. In this work, 1440 speech files have been used from the sum of 7356 files.

The FAU Aibo Emotion Corpus as a non-acted dataset has been used to validate the proposed model, which contains spontaneous and emotional German speech samples [63]. Through their interactions with Sony's pet robot Aibo, the 51 children between the ages of 10–13 years in 'Ohm' and 'Mont' schools were participated to record 18216 emotional speech samples. In the beginning, the dataset had 10 emotion labels, and later, they mapped into five emotion classes such as anger, emphatic, neutral, positive, and rest. In this work, we followed the adopted protocol of the interspeech09 challenge [64], which the training set contains 9959 samples from Ohm and the testing set contains 8257 samples from Mont school.

5 Experimental setup

The proposed model has adopted Leave One Speaker Out (LOSO) as a speaker-independent approach for EMO-DB, SAVEE, and RAVDESS datasets. To conduct a fair comparison with the state-of-the-art studies of the FAU Aibo dataset, this study followed the adopted protocol of the interspeech09 challenge [64]. In this work, we have adopted the use of handcrafted multivariate time series data which contains 26 features from MFCCs and GTCCs for each window of length 30 milliseconds overlapped by 20 milliseconds. The random over-sampling method was

applied on EMO-DB and RAVDESS, while random under-sampling was applied on SAVEE and FAU Aibo. Regarding the parameters that have not been optimized in the proposed model, it has been fixed as follows: spectral radius (ρ) = 0.6, non-zero connections (β) = 0.25, input scaling (ω) = 0.1, the leakage percentage = 0.6, and the level of noise = 0.01.

The proposed model used CPU to carry out all experiments on a PC with 64GB RAM and Google Colab with 12GB RAM. Since ESN has a simple architecture without any training in the reservoir layer, it does not need GPU or high PC resources.

6 Results and discussion

This section showcases the results of the evaluation proposed model for EMO-DB, SAVEE, RAVDESS, and FAU Aibo speech emotion datasets. The speaker-independent cross-validation technique is adopted as a more applicable method in emotion recognition from speech which is more challenging than the speaker-dependent approach. The results of this study have been shown in terms of precision, recall, and F1 score, in addition to the model weighted and unweighted accuracy.

The weighted accuracy computes the correctly classified samples in the test set for all the emotion classes divided by the whole number of classes in the testing set, and it can be used properly for the balanced dataset. The weighted accuracy is presented by the following mathematical form [65]:

$$\text{Weighted Average} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} TP_i + FN_i} \tag{10}$$

where $i = 1, 2, \ldots, C$ introduces the number of emotion classes used, TP (True Positive) refers to the number of positive samples that were recognized correctly as positive samples from the classification model.

However, the unweighted accuracy (UA) refers to the average of per-class accuracies. In this work, all results have been shown as unweighted accuracy which is more realistic for accuracy measurement especially when the test set of datasets is imbalanced. The unweighted accuracy is presented by the following mathematical form:

$$\text{Unweighted Accuracy} = \frac{1}{|C|} \sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FN_i} \tag{11}$$

Additionally, we presented the confusion matrix for all experiments to show the match and mismatch between predicted and actual labels.

6.1 Experimental evaluations

In this study, we have evaluated the impact of the two adopted reservoirs instead of the single reservoir in addition to the impact of using and without using over-sampling and under-sampling. We have handled different experiments to evaluate our proposed model as shown in Table 1. Following are the models that have been investigated using all the involved datasets:

- **Model-1, single ESN with sampling:** The model has been evaluated based on having one reservoir with a sampling technique to overcome the imbalanced issue. The conducted experiments show that sampling methods (random over-sampling and random under-sampling) are able to increase the performance of a single reservoir compared to the use of a single reservoir without sampling [15] by 1.35% UA for EMO-DB, 0.36% UA for SAVEE, and 1.56% UA for RAVDESS (see Table 1). However, our proposed model, where parallel reservoirs are adopted, outperformed this model (Model-1) in all datasets except FAU Aibo, which is an indication that both reservoirs in the proposed ESN model can have a better SER performance than the single reservoir.
- **Model-2, Parallel ESN without sampling:** To show the impact of the use of sampling data to resolve the imbalanced issue using parallel ESN, we conducted an experiment applied to all the datasets as shown in Table 1. Feeding two reservoirs with the bidirectional data without solving the unbalancing issue leads to a decrease in the performance using all of the involved datasets. We notice that in both EMO-DB and SAVEE datasets, the proposed model is able to improve the performance significantly over the current Model-2. However, the proposed model improvement when applied to RAVDESS dataset is not more than 1.11% of accuracy.
- **Model-3, Parallel ESN with SMOTE:** In order to investigate the impact of the random over-sampling and random under-sampling approach over the well-known SMOTE method, we have adopted the parallel ESN with SMOTE. In the SMOTE method, new samples are created between two randomly chosen samples with a random distance. On the other side, random over-sampling duplicates samples in the minority classes randomly. The results showed that adopting random over-sampling and random under-sampling outperformed the same model when using SMOTE. As shown in Table 1, the result of the use of SMOTE for RAVDESS is very close to the random over-sampling model where the difference is only 0.13%. However, the proposed model outperformed by 3.56%, 1.9%, and

Table 1 Evaluating and comparing the proposed model to other methods with different techniques

Method	EMO-DB	SAVEE	RAVDESS	FAU Aibo
Model-1 (SingleESN+Sampling)	88.15	68.81	74.61	**45.90**
Model-2 (Parallel ESN+Without sampling)	82.19	63.45	74.28	34.11
Model-3 (Parallel ESN+SMOTE)	85.34	67.62	75.26	38.65
Model-4 (Parallel ESN+Reservoir Fusion)	82.45	65.83	69.53	45.03
Proposed model	**88.9**	**69.52**	**75.39**	45.51

The results are shown as unweighted percentage accuracy for Emo-DB, SAVEE, RAVDESS, and FAU Aibo datasets

The highest accuracies are marked in bold

6.86% for Emo-DB, SAVEE, and FAU Aibo, respectively.

- **Model-4, Parallel ESN with reservoir fusion:** To conduct how the fusion based on directions in both reservoirs has an impact on the model performance, we adopted the reservoir fusion method. Here, we deal with the output of each reservoir independently from the other one. Both directions that output from a single reservoir fused together to produce a representation to feed to the classifier as shown in Fig. 5. The result shows that mixing both directions and feeding them to the reservoir model space will affect the performance negatively as shown in Table 1, while fusion of forward and backward directions from both reservoirs separately will contribute to enhancing the SER accuracy. The fusion based on directions achieved considerably higher performance in Emo-DB, SAVEE, and RAVDESS datasets compared to the fusion based on reservoirs. However, the differences between this model and the proposed model in FAU Aibo accuracy are only 0.48%.

The overall results show that the parallel ESN with random over-sampling for EMO-DB and RAVDESS and random under-sampling for SAVEE has a significant improvement on the proposed model.

Although the improvement of the proposed model over the Model-1 is not as much as the improvement over the other models, however, this improvement is statistically significant. Additionally, the time consuming cost of the proposed model is linearly increased, i.e., in this case, it is doubled.

6.2 Proposed model results

Table 2 shows the detailed speaker-independent results measured by precision, recall, and F1 score for each emotion, in addition to the model weighted and unweighted percentage accuracy in the EMO-DB dataset. We pursued the cross-validation method for EMO-DB using LOSO, where 9 speakers are chosen as a training set and one speaker as a testing set, then the procedure is repeated 10 times to give the chance for all the other 9 speakers to be chosen for testing the model.

Table 2 presents the performance per class of the proposed model for the EMO-DB dataset. The sadness class has the best accuracy where all samples were recognized correctly, and the anger class has 99.21%. However, happiness has a very weak performance (64.79% accuracy).

Moreover, Fig. 6 illustrates the confusion matrix for EMO-DB dataset results between the true label and the predicted label for all of the seven emotion classes. Since the happiness emotion has recorded the lowest accuracy, one can observe from the confusion matrix that 28% of happiness emotion samples are recognized as anger class.

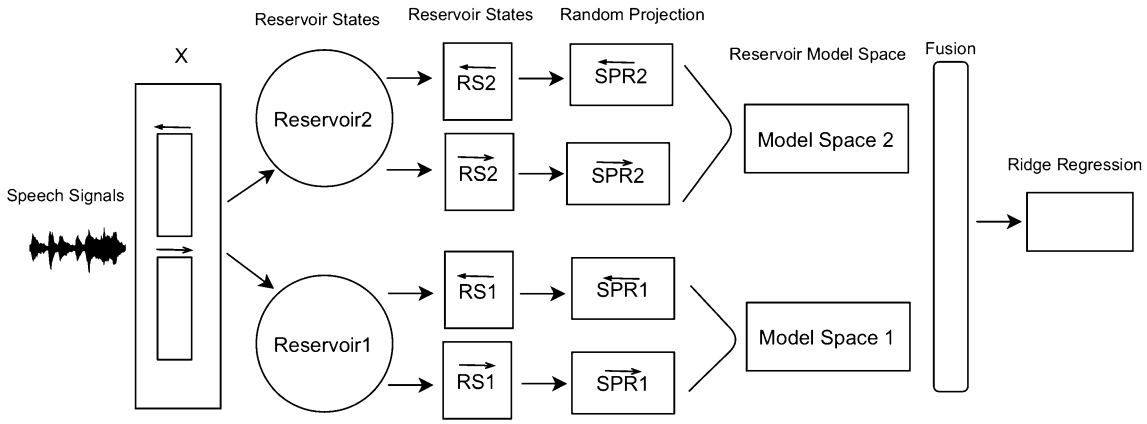


Fig. 5 The parallel ESN with reservoir fusion (Model-4)

Table 2 The proposed model performance for EMO-DB dataset in terms of the model weighted and unweighted accuracy, in addition to precision, recall and F1 score for each emotion class

Emotion	Precision	Recall	F1 score
Anger	82.89	99.21	90.32
Boredom	92.86	96.30	94.55
Disgust	95.24	86.96	90.91
Fear	93.44	82.61	87.69
Happiness	90.20	64.79	75.41
Sadness	95.38	100	97.64
Neutral	91.25	92.41	91.82
Weighted	90.47	90.09	89.76
Unweighted	91.61	88.90	89.76

Table 3 The proposed model performance for SAVEE dataset in terms of the model weighted and unweighted accuracy, in addition to precision, recall and F1 score for each emotion class

Emotion	Precision	Recall	F1 score
Anger	72.60	88.33	79.70
Disgust	58.62	56.67	57.63
Fear	78.95	50.00	61.22
Happiness	82.14	76.67	79.31
Neutral	71.90	91.67	80.59
Sadness	74.29	43.33	54.74
Surprise	71.64	80.00	75.59
Weighted	72.75	72.29	71.17
Unweighted	72.88	69.52	69.82

This may be an indication that the happiness emotion expressed in a high arousal which is shared with the anger emotion.

The classification result per each emotion class of the SAVEE dataset is shown in Table 3. Table 3 presents the details of the speaker-independent approach performance measured by precision, recall, and F1 score for each class, in addition to the model weighted and unweighted percentage accuracy in the SAVEE dataset. The neutral class recorded the highest performance by recognizing 91.67% of its samples, while sadness, fear, and disgust emotions have a lower performance by 43.33%, 50%, and 56.67%, respectively. The rest of the emotion classes which include anger, surprise, and happiness showed 88.33%, 80%, and 76.67%, respectively.

From Fig. 7, the performance of the proposed model for the SAVEE dataset has been shown by the confusion matrix. The sadness class was mostly recognized as neutral and disgust by 37% and 15%, respectively. In addition, 20% of disgust emotion was recognized as neutral, and

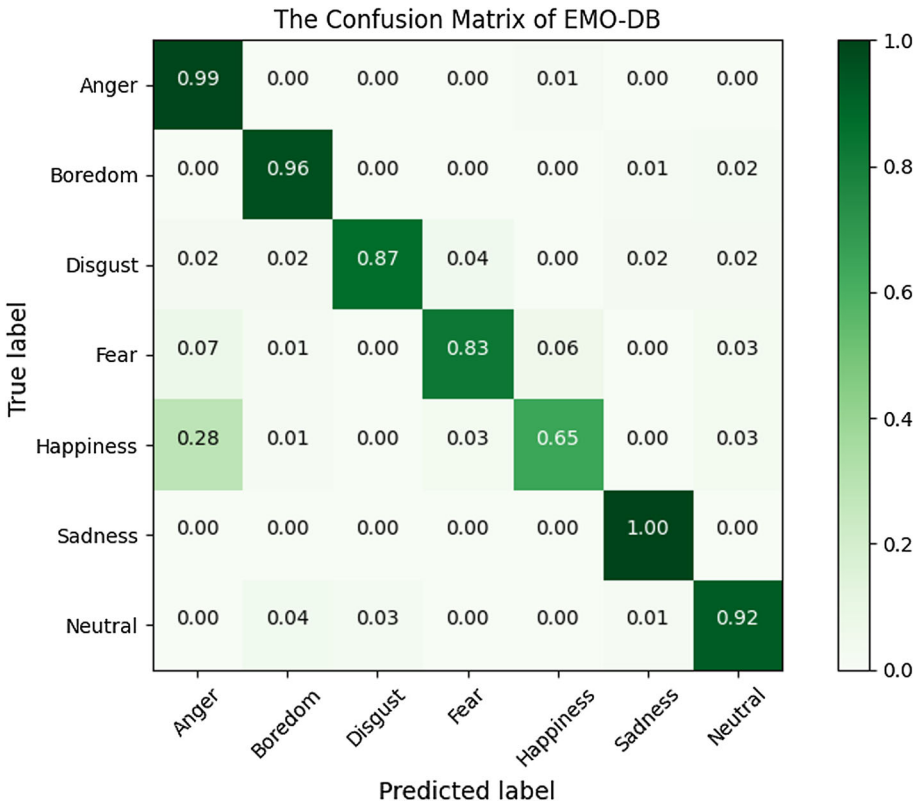


Fig. 6 The confusion matrix of the proposed model for the EMO-DB dataset

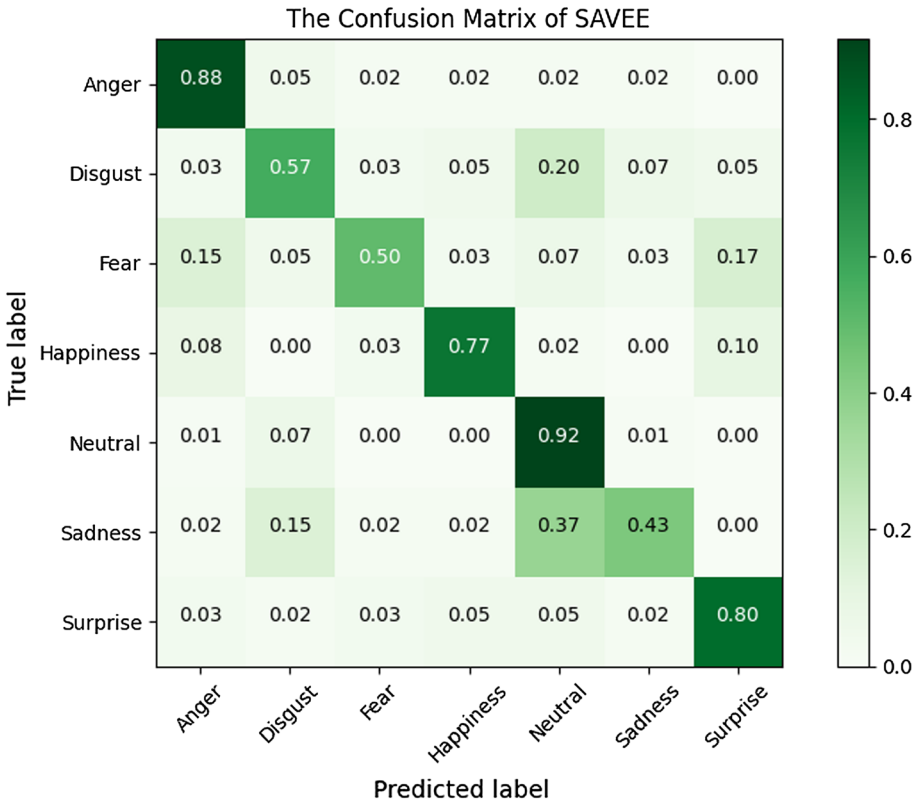


Fig. 7 The confusion matrix of the proposed model for SAVEE dataset

50% of fear emotion was mainly recognized as surprise and anger classes.

Table 4 shows the detailed results in terms of precision, recall, F1 score, weighted, and unweighted accuracy for each individual emotion class for the RAVDESS dataset. We followed the LOSO approach where 23 speakers are chosen for the training set, while a single speaker is used as a testing set. This process is repeated 24 times such that each speaker can represent the testing set separately from the remaining speakers. The neutral and sad emotions recorded the lowest performance (57.29% and 58.33%, respectively). However, calm and disgust have the highest performance by gaining 88.54% of accuracy for each of them.

Table 4 The proposed model performance for RAVDESS dataset in terms of the model weighted and unweighted accuracy, in addition to precision, recall and F1 score for each emotion class

Emotion	Precision	Recall	F1 score
Neutral	70.51	57.29	63.22
Calm	74.89	88.54	81.15
Happy	80.92	72.92	76.71
Sad	65.50	58.33	61.71
Angry	86.56	83.85	85.19
Fearful	74.41	81.77	77.92
Disgust	77.98	88.54	82.93
Surprised	78.41	71.88	75.00
Weighted	76.52	76.60	76.29
Unweighted	76.15	75.39	75.48

The confusion matrix in Fig. 8 shows the accuracy of the eight emotion classes individually of the RAVDESS dataset. The neutral emotion recorded the lowest performance, where 17% and 16% are recognized as calm and sad, respectively. Additionally, 12% of the sad emotion samples in the test set are recognized as calm emotion, where both emotions have low arousal characteristics.

To conduct a fair comparison with the state-of-the-art studies, we have followed the same interspeech09 challenge [64] protocol adopted in the previous chapter. Table 5 lists the detailed classification results of the precision, recall, F1 score, unweighted, and weighted percentage accuracy for each emotion class for FAU Aibo dataset. It can be observed that there is a big gap between the weighted and unweighted accuracy due to the high imbalance of data. The low accuracy of this dataset compared to the others reflects the challenge of emotion recognition in a spontaneous dataset.

The confusion matrix in Fig. 9 shows the accuracy of each five involved emotion classes of FAU Aibo dataset. The anger class recorded 66% as the highest accuracy and the rest emotion class with 17% is the lowest accuracy that

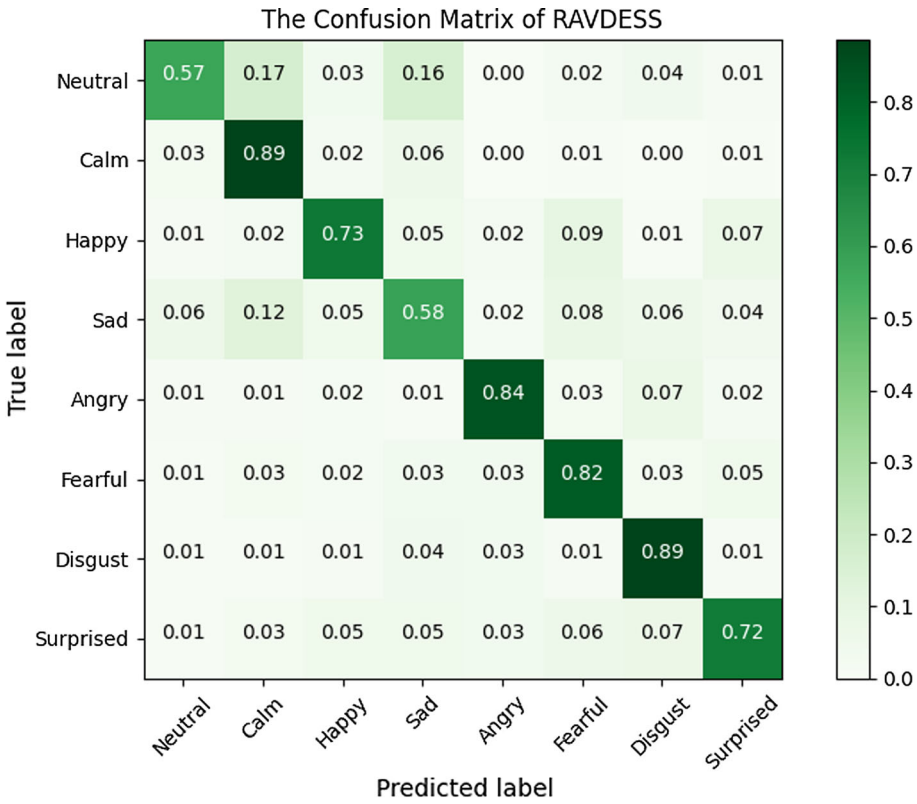


Fig. 8 The confusion matrix of the proposed model for the RAVDESS dataset

Table 5 The proposed model performance (%) for FAU Aibo dataset in terms of the model weighted and unweighted accuracy, in addition to precision, recall and F1 score for each emotion class

Emotion	Precision	Recall	F1 score
Anger	21.31	65.96	32.21
Emphatic	35.29	57.69	43.80
Neutral	83.12	30.31	44.43
Positive	09.56	56.74	16.36
Rest	13.86	16.85	15.21
Unweighted	32.63	45.51	30.40
Weighted	63.32	37.75	40.74

we have got from the proposed model. The low accuracy of rest class may be due to its samples nature where they have different labels but are gathered under the same class.

6.3 Comparison with the state-of-the-art

The proposed model used two reservoirs to create a more typical representation of the input data and capture independent information of the input data. Due to the unbalancing of the used speech emotion datasets, we applied over-sampling and under-sampling techniques to reduce the negative effect of data unbalancing at the sample level by removing or duplicating the majority and minority samples, respectively. Moreover, SRP has been adopted to decrease the dimensionality of the output feature representation from the reservoir layer. This novel proposed model to recognize emotion from speech with its simple structure and the trainless nature assists to improve the classification accuracy.

Due to the imbalanced nature of the SER datasets, we present the overall unweighted accuracy (UA) for our experiments, since its more realistic than weighted accuracy. Our proposed model for EMO-DB achieved 88.9% UA, for SAVEE dataset obtained 69.52% UA, and for RAVDESS, the achieved performance is 75.39%. Among all the state-of-the-art methods that are adopting the LOSO approach, our proposed model was able to outperform them, in the exception of the result for the proposed single reservoir for FAU Aibo, which outperform the proposed model. The comparison between our work with various new studies that have been conducted lately for classification UA LOSO experiments is shown in Table 6.

For the EMO-DB dataset, we can observe in Table 6 that the proposed model outperformed the remaining studies significantly including the LSTM models [25, 67, 68]. We can notice that the performance of the model with two reservoirs and resampling data for EMO-DB has improved the ESN model by 2.1% compared to the ESN model with a single reservoir. Our model for the SAVEE dataset

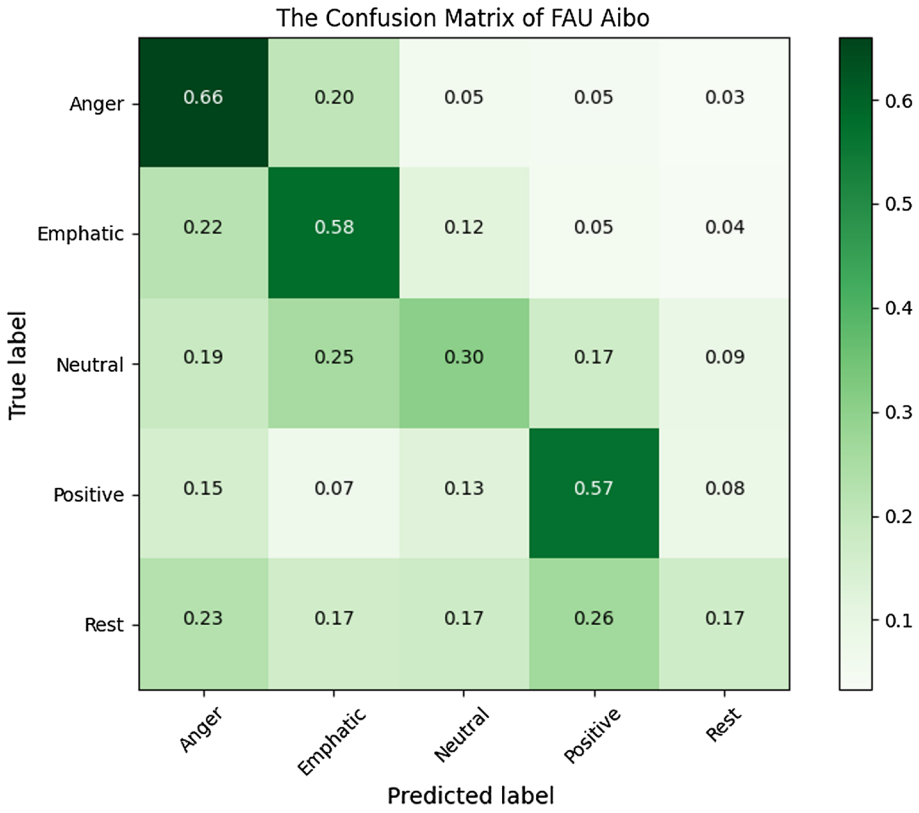


Fig. 9 The confusion matrix of the proposed model for the FAU Aibo dataset

achieved 69.52% which is 14.52% and 15.92% higher than the GEBF model [12] and RDBN deep learning model [66], respectively. Additionally, the proposed model outperforms the use of a single reservoir by 1.07%. Regarding the RAVDESS dataset, there are few works conducted using the LOSO approach [15]. Our model recorded a new high accuracy of the LOSO method in RAVDESS by obtaining 75.39%. The deep learning and ESN models in Table 6 have achieved distinguished results. Besides the work of [15] who are using a single reservoir in bidirectional late fusion ESN model, researchers in [68] and [67] have used 3-D Log-Mel spectrums from speech signals and fed them to the deep learning model with a classification UA of 84.99% and 82.82%, respectively, for EMO-DB dataset. Moreover, authors in [25] adopted deep learning parallelized convolutional recurrent neural network (PCRN) model with 3-D log Mel-spectrograms features from EMO-DB and they achieved 84.53% UA. However, our model outperformed the single reservoir ESN model and the deep learning model by obtaining 88.90% UA.

Regarding the FAU Aibo dataset, one can notice that the highest achieved result in the our previous works is the single reservoir model proposed in [15]. The proposed model is able to outperform the previous works by the use of spectrogram-based features with deep learning models [71–73] by achieving UA of 45.51%, however, this achieved result is slightly lower than the proposed single reservoir model in [15].

7 Conclusion and future work

The novel recurrent-based architecture for multivariate time series SER classification by having two reservoirs and resampling data to overcome the imbalanced datasets has been proposed. The proposed model adopted bidirectional data with two different stages of fusion and dimension reduction by using SRP. We can conclude from the obtained experimental results that the fusion of the same direction produced from multi-reservoirs to be fused lately with the other direction can have a more informative representation for SER application. Additionally, the random over-sampling and random under-sampling that adopt duplicating the samples or removing them randomly show distinguished ability over SMOTE method. The proposed model achieved the highest classification unweighted accuracy compared to the recent studies on speech emotion recognition using speaker-independent by applying LOSO on EMO-DB, SAVEE, and RAVDESS datasets and speaker-independent on FAU Aibo dataset.

For future work, we can improve the current model by using different approaches of sampling data and optimizing hyperparameters that may have a significant influence for

Table 6 The comparison summary of unweighted accuracies (UA%) obtained by various studies for EMO-DB, SAVEE and RAVDESS datasets using LOSO approach

Dataset	Method	UA%
EMO-DB	[66] (Deep learning RDBN)	82.32
	[67] (Deep learning ACRNN)	82.82
	[68] (Deep learning ADRNN)	84.99
	[25] (Deep learning PCRN)	84.53
	[12] (GEBF)	76.81
	[69] (SVM-RBF)	71.02
	[70] (OpenSmile+SVM)	76.82
	[15] (ESN-Single reservoir)	86.80
	Proposed model	**88.90**
SAVEE	[66] (Deep learning RDBN)	53.60
	[12] (GEBF)	55.00
	[15] (ESN-Single reservoir)	68.45
	Proposed model	**69.52**
RAVDESS	[15] (ESN-Single reservoir)	73.05
	Proposed model	**75.39**
FAU Aibo	[71] (Deep learning eResNet)	41.3
	[72] (Deep learning BLSTM)	45.4
	[73] (Deep learning 2D CNN)	41.1
	[74] (Handcrafted+SVM,NN,DNN)	45.3
	[75] (MRA+SVM)	45.2
	[15] (ESN-Single reservoir)	**45.9**
	Proposed model	45.51

The FAU Aibo dataset results followed the 2009 challenge protocol
The highest accuracies are marked in bold

improving the SER system. Additionally, different architectures of having parallel and sequential multi-reservoirs can be put under investigation, since the result in this work indicates a promising performance for such models.

Acknowledgements This work was supported in part by the COVID-19 Special Research Grant under Project CSRG008-2020ST, Impact Oriented Interdisciplinary Research Grant Programme (IIRG), IIRG002C-19HWB from University of Malaya, and the AUA-UAEU Joint Research Grant 31R188.

Declarations

Conflict of interest The authors declare there is no conflict of interest.

References

1. Bojanić M, Delić V, Karpov A (2020) Call redistribution for a call center based on speech emotion recognition. Appl Sci 10(13):4653
2. Katsis CD, Rigas G, Goletsis Y, Fotiadis DI (2015) Emotion recognition in car industry. Emot Recognit A Pattern Anal Approach 515–544
3. Al-Talabani A (2015) Automatic speech emotion recognition-feature space dimensionality and classification challenges. PhD thesis, University of Buckingham
4. Pérez-Espinosa H, Gutiérrez-Serafín B, Martínez-Miranda J, Espinosa-Curiel IE (2022) Automatic children's personality assessment from emotional speech. Expert Syst Appl 187:115885. https://doi.org/10.1016/j.eswa.2021.115885
5. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans Multimed 16(8):2203–2213
6. Kathiresan T, Dellwo V (2019) Cepstral derivatives in mfccs for emotion recognition. In: 2019 IEEE 4th international conference on signal and image processing (ICSIP), pp 56–60. IEEE
7. Abbaschian BJ, Sierra-Sosa D, Elmaghraby A (2021) Deep learning techniques for speech emotion recognition, from databases to models. Sensors. https://doi.org/10.3390/s21041249
8. Mustaqeem Kwon S (2021) Mlt-dnet: speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. Expert Syst Appl 167:114177. https://doi.org/10.1016/j.eswa.2020.114177
9. Li D, Liu J, Yang Z, Sun L, Wang Z (2021) Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Syst Appl 173:114683. https://doi.org/10.1016/j.eswa.2021.114683
10. Ma Z, Yu H, Chen W, Guo J (2019) Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features. IEEE Trans Veh Technol 68(1):121–128. https://doi.org/10.1109/TVT.2018.2879361
11. Fayek HM, Lech M, Cavedon L (2017) Evaluating deep learning architectures for speech emotion recognition. Neural Netw 92:60–68
12. Daneshfar F, Kabudian SJ, Neekabadi A (2020) Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and gaussian elliptical basis function network classifier. Appl Acoust 166:107360. https://doi.org/10.1016/j.apacoust.2020.107360
13. Ma Q, Shen L, Chen W, Wang J, Wei J, Yu Z (2016) Functional echo state network for time series classification. Inf Sci 373:1–20. https://doi.org/10.1016/j.ins.2016.08.081
14. Chen Y, Keogh E, Hu B, Begum N, Bagnall A, Mueen A, Batista G (2015) The ucr time series classification archive
15. Ibrahim H, Loo CK, Alnajjar F (2021) Speech emotion recognition by late fusion for bidirectional reservoir computing with random projection. IEEE Access 1:18. https://doi.org/10.1109/ACCESS.2021.3107858
16. Wu Q, Fokoue E, Kudithipudi D (2018) On the statistical challenges of echo state networks and some potential remedies. arXiv:1802.07369
17. Shoumy NJ, Ang L-M, Rahaman DM, Zia T, Seng KP, Khatun S (2021) Augmented audio data in improving speech emotion classification tasks. In: International conference on industrial, engineering and other applications of applied intelligent systems, pp 360–365. Springer
18. López E, Valle C, Allende H, Gil E, Madsen H (2018) Wind power forecasting based on echo state networks and long short-term memory. Energies 11(3):526
19. Scherer S, Oubbati M, Schwenker F, Palm G (2008) Real-time emotion recognition from speech using echo state networks. In: IAPR workshop on artificial neural networks in pattern recognition, pp 205–216. Springer
20. Rodan A, Sheta AF, Faris H (2017) Bidirectional reservoir networks trained using svm + privileged information for manufacturing process modeling. Soft Comput 21(22):6811–6824

21. Bianchi FM, Scardapane S, Løkse S, Jenssen R (2020) Reservoir computing approaches for representation and classification of multivariate time series. IEEE Trans Neural Netw Learn Syst
22. Gallicchio C, Micheli A (2019) Reservoir topology in deep echo state networks. In: International conference on artificial neural networks, pp 62–75. Springer
23. Sun L, Zou B, Fu S, Chen J, Wang F (2019) Speech emotion recognition based on dnn-decision tree svm model. Speech Commun 115:29–37
24. Zhong G, Wang L-N, Ling X, Dong J (2016) An overview on data representation learning: from traditional feature learning to recent deep learning. J Financ Data Sci 2(4):265–278
25. Jiang P, Fu H, Tao H, Lei P, Zhao L (2019) Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. IEEE Access 7:90368–90377. https://doi.org/10.1109/ACCESS.2019.2927384
26. Dai D, Wu Z, Li R, Wu X, Jia J, Meng H (2019) Learning discriminative features from spectrograms using center loss for speech emotion recognition. In: ICASSP 2019—2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 7405–7409. https://doi.org/10.1109/ICASSP.2019.8683765
27. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: The munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on multimedia. MM '10, pp 1459–1462. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1873951.1874246
28. Amiriparian S, Gerczuk M, Ottl S, Cummins N, Freitag M, Pugachevskiy S, Baird A, Schuller B (2017) Snore sound classification using image-based deep spectrum features. In: Interspeech 2017, pp 3512–3516. ISCA
29. Al-Talabani A, Sellahewa H, Jassim S (2013) Excitation source and low level descriptor features fusion for emotion recognition using svm and ann. In: 2013 5th computer science and electronic engineering conference (CEEC), pp 156–161. https://doi.org/10.1109/CEEC.2013.6659464
30. Liu Z-T, Wu B-H, Li D-Y, Xiao P, Mao J-W (2020) Speech emotion recognition based on selective interpolation synthetic minority over-sampling technique in small sample environment. Sensors 20(8):2297
31. Ooi CS, Seng KP, Ang L-M, Chew LW (2014) A new approach of audio emotion recognition. Expert Syst Appl 41(13):5858–5869. https://doi.org/10.1016/j.eswa.2014.03.026
32. Zhou S, Jia J, Wang Y, Chen W, Meng F, Li Y, Tao J (2018) Emotion inferring from large-scale internet voice data: A multimodal deep learning approach. In: 2018 first Asian conference on affective computing and intelligent interaction (ACII Asia), pp 1–6. https://doi.org/10.1109/ACIIAsia.2018.8470311
33. Fu C, Dissanayake T, Hosoda K, Maekawa T, Ishiguro H (2020) Similarity of speech emotion in different languages revealed by a neural network with attention. In: 2020 IEEE 14th international conference on semantic computing (ICSC), pp 381–386. https://doi.org/10.1109/ICSC.2020.00076
34. Chen L, Mao X, Xue Y, Cheng LL (2012) Speech emotion recognition: features and classification models. Digit. Signal Process. 22(6):1154–1160. https://doi.org/10.1016/j.dsp.2012.05.007
35. Lee J, Tashev I (2015) High-level feature representation using recurrent neural network for speech emotion recognition. In: INTERSPEECH, pp 1537–1540. ISCA, Dresden, Germany. http://dblp.uni-trier.de/db/conf/interspeech/interspeech2015.html
36. Vryzas N, Vrysis L, Matsiola M, Kotsakis R, Dimoulas C, Kalliris G (2020) Continuous speech emotion recognition with convolutional neural networks. J Audio Eng Soc 68(1/2):14–24
37. Gallicchio C, Micheli A (2014) A preliminary application of echo state networks to emotion recognition. In: Fourth international workshop EVALITA 2014, pp 116–119. Pisa University Press, Pisa, Italy
38. Saleh Q, Merkel C, Kudithipudi D, Wysocki B (2015) Memristive computational architecture of an echo state network for real-time speech-emotion recognition. In: 2015 IEEE symposium on computational intelligence for security and defense applications (CISDA), pp 1–5. https://doi.org/10.1109/CISDA.2015.7208624
39. Wang Z, Yao X, Huang Z, Liu L (2021) Deep echo state network with multiple adaptive reservoirs for time series prediction. IEEE Trans Cognit Dev Syst. https://doi.org/10.1109/TCDS.2021.3062177
40. Gallicchio C, Micheli A, Pedrelli L (2017) Deep reservoir computing: A critical experimental analysis. Neurocomputing **268**, 87–99. https://doi.org/10.1016/j.neucom.2016.12.089. Advances in artificial neural networks, machine learning and computational intelligence
41. Huang Z, Yang C, Chen X, Zhou X, Chen G, Huang T, Gui W (2021) Functional deep echo state network improved by a bi-level optimization approach for multivariate time series classification. Appl Soft Comput 106:107314. https://doi.org/10.1016/j.asoc.2021.107314
42. Wcisło R, Czech W (2021) Grouped multi-layer echo state networks with self-normalizing activations. In: International conference on computational science, pp 90–97. Springer
43. Attabi Y, Dumouchel P (2013) Anchor models for emotion recognition from speech. IEEE Trans Affect Comput 4(3):280–290
44. Bianchi FM, Scardapane S, Løkse S, Jenssen R (2017) Bidirectional deep-readout echo state networks. arXiv:1711.06509
45. Li P, Hastie TJ, Church KW (2006) Very sparse random projections. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. KDD 06, pp 287–296. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1150402.1150436
46. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28(4):357–366. https://doi.org/10.1109/TASSP.1980.1163420
47. Babu M, Kumar MA, Santhosh S (2014) Extracting mfcc and gtcc features for emotion recognition from audio speech signals. Int J Res Comput Appl Robot 2(8):46–63
48. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from imbalanced data sets vol. 10. Springer
49. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239
50. Menardi G, Torelli N (2012) Training and assessing classification rules with imbalanced data. Data Min Knowl Discov 28:92–122
51. Jaeger H, Haas H (2004) Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. Science 304(5667):78–80. https://doi.org/10.1126/science.1091277
52. Lukoševičius M, Jaeger H (2009) Reservoir computing approaches to recurrent neural network training. Comput Sci Rev 3(3):127–149. https://doi.org/10.1016/j.cosrev.2009.03.005
53. Xue Y, Yang L, Haykin S (2007) Decoupled echo state networks with lateral inhibition. Neural Netw. **20**(3), 365–376. https://doi.org/10.1016/j.neunet.2007.04.014. Echo State Networks and Liquid State Machines
54. Malik ZK, Hussain A, Wu QJ (2017) Multilayered echo state machine: a novel architecture and algorithm. IEEE Trans Cybern 47(4):946–959. https://doi.org/10.1109/TCYB.2016.2533545

55. Chouikhi N, Ammar B, Alimi AM (2018) Genesis of basic and multi-layer echo state network recurrent autoencoders for efficient data representations. arXiv:1804.08996
56. Gallicchio C, Micheli A (2017) Echo state property of deep reservoir computing networks. Cognit Comput 9(3):337–350
57. Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems, vol. 25. Curran Associates, Inc
58. Wu S, Falk TH, Chan W-Y (2011) Automatic speech emotion recognition using modulation spectral features. Speech Commun **53**(5), 768–785. https://doi.org/10.1016/j.specom.2010.08.013. Perceptual and Statistical Audition
59. Vlasenko B, Schuller B, Wendemuth A, Rigoll G (2007) Combining frame and turn-level information for robust recognition of emotions within speech. In: INTERSPEECH
60. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: INTERSPEECH
61. Haq S, Jackson PJB (2010) Multimodal emotion recognition. In: Wang W (ed) Machine audition: principles, algorithms and systems. IGI Global, Hershey PA, pp 398–423
62. Livingstone S, Russo F (2018) The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north American english. PLoS ONE **13**
63. Steidl S (2009) Automatic classification of emotion related user states in spontaneous children's speech. Logos-Verlag
64. Schuller B, Steidl S, Batliner A (2009) The interspeech 2009 emotion challenge. In: Tenth annual conference of the international speech communication association
65. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. **45**(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
66. Wen G, Li H, Huang J, Li D, Xun E (2017) Random deep belief networks for recognizing emotions from speech signals. Comput Intell Neurosci **2017**
67. Chen M, He X, Yang J, Zhang H (2018) 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process Lett 25(10):1440–1444. https://doi.org/10.1109/LSP.2018.2860246
68. 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.2019.2938007
69. Liu Z-T, Rehman A, Wu M, Cao W-H, Hao M (2021) Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. Inf Sci 563:309–325. https://doi.org/10.1016/j.ins.2021.02.016
70. Yildirim S, Kaya Y, Kılıç F (2021) A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. Appl Acoust 173:107721. https://doi.org/10.1016/j.apacoust.2020.107721
71. Triantafyllopoulos A, Liu S, Schuller BW (2021) Deep speaker conditioning for speech emotion recognition. In: 2021 IEEE international conference on multimedia and expo (ICME), pp 1–6. https://doi.org/10.1109/ICME51207.2021.9428217
72. Zhao Z, Bao Z, Zhao Y, Zhang Z, Cummins N, Ren Z, Schuller B (2019) Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. IEEE Access 7:97515–97525. https://doi.org/10.1109/ACCESS.2019.2928625
73. Zhao Z, Li Q, Zhang Z, Cummins N, Wang H, Tao J, W. Schuller B, (2021) Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition. Neural Netw 141:52–60. https://doi.org/10.1016/j.neunet.2021.03.013
74. Shih P-Y, Chen C-P, Wang H-M (2017) Speech emotion recognition with skew-robust neural networks. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2751–2755. https://doi.org/10.1109/ICASSP.2017.7952657
75. Deb S, Dandapat S (2019) Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification. IEEE Trans Cybern 49(3):802–815. https://doi.org/10.1109/TCYB.2017.2787717