IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 3, 2021, accepted August 20, 2021, date of publication August 25, 2021, date of current version September 13, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3107858*

# Speech Emotion Recognition by Late Fusion for Bidirectional Reservoir Computing With Random Projection

**HEMIN IBRAHIM**[1], **CHU KIONG LOO**[1], **(Senior Member, IEEE), AND FADY ALNAJJAR**[2]

[1]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]College of Information Technology, UAE University, Al Ain, United Arab Emirates

Corresponding author: Chu Kiong Loo (ckloo.um@um.edu.my)

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

**ABSTRACT** Many researchers are inspired by studying Speech Emotion Recognition (SER) because it is considered as a key effort in Human-Computer Interaction (HCI). The main focus of this work is to design a model for emotion recognition from speech, which has plenty of challenges within it. Due to the time series and sparse nature of emotion in speech, we have adopted a multivariate time series feature representation of the input data. The work has also adopted the Echo State Network (ESN) which includes reservoir computing as a special case of the Recurrent Neural Network (RNN) to avoid model complexity because of its untrained and sparse nature when mapping the features into a higher dimensional space. Additionally, we applied dimensionality reduction since it offers significant computational advantages by using Sparse Random Projection (SRP). Late fusion of bidirectionality input has been applied to capture additional information independently of the input data. The experiments for speaker-independent and/or speaker-dependent were performed on four common speech emotion datasets which are Emo-DB, SAVEE, RAVDESS, and FAU Aibo Emotion Corpus. The results show that the designed model outperforms the state-of-the-art with a cheaper computation cost.

**INDEX TERMS** Speech emotion recognition, reservoir computing, time series classification, random projection, recurrent neural network.

## I. INTRODUCTION

Emotion can play an important role in many parts of a human's life such as communicating, understanding, helping each other, rational thinking, creativity and sometimes it has a vital part in decision making. However, there has been no general agreement on how to categorize, recognize and analyze it because of the differences among cultures and individuals. Emotion can be detected from various channels such as electroencephalography (EEG) signals, acoustic, visual, text, and gestures. Detecting emotion is a challenging task and it has become a hot field of research topics and covered a wide research area due to the high demand for using it in many practical applications such as healthcare, social robot, and Human-Computer Interaction (HCI) [1], [2]. However, emotions do not have a static categorization, and it is not easy to adapt, which is why some works are done by using unsupervised models for unknown emotions and growing models to deal with the adaptation [3].

Speech is an effective, quick, and important way for individuals to communicate with each other [4] and the speech signal is considered as a fast and useful mechanism for HCI. Emotions have always been a part of normal human conversation which makes the speech more attractive and more effective. Detecting emotions from speech signals is an old yet big challenge in the field of artificial intelligence [5] which makes many researchers inspired to work on it.

For this reason, Speech Emotion Recognition (SER) is playing a significant role in the HCI with great progress in recent years. However, certain aspects of inner feelings remain concealed and are not easily measurable from the speech, particularly when humans want to suppress their emotions. Thus, it cannot be expected for the computer-based

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

system to do beyond what is perceived from the input of the speech sample.

One of the challenges in SER is to determine the most relevant acoustic emotion features which are extracted from the raw speech signal. Researchers endeavor to find more effective features for detecting emotion in speech [6]. Recent studies have shown that emotional information in speech is distributed over multiple types of features [2] and finding the right features which have the most information about human emotion is critical. Two main ways have been used to extract features which are handcrafted features in addition to deep learned emotion features. Therefore, many applications such as speech recognition with time series or sequential data have been shown to achieve state-of-the-art results with some deep learning approaches such as Recurrent Neural Networks (RNNs), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) [7]. However, Zhong *et al.* [8] reviewed data representation research, including traditional feature extraction and deep learning, with the conclusion that the gap between the theory and practical applications of deep learning is still quite big, and deep learning models are not always the best approach, especially in real-world problems.

Multivariate time series emotion feature representation can be able to adapt due to the sparse nature of emotion in speech. To tackle this characteristic, some studies used Echo State Network (ESN) as a special type of RNN, and as a part of the reservoir computing framework. The main reported advantage of ESN is that it has a simple architecture as it contains the input layer, a reservoir layer with sparsely connected neurons that are randomly assigned without training, and the output layer [9]. The temporal dependency of time series data can be handled effectively by ESN since it is successfully applied for chaotic time series prediction models [10], [11]. The simplicity of ESN is represented by assigning a non-trainable randomly weights and avoiding the time complexity of deep recurrent networks [12] which makes ESN an ultimate nominee for tasks involving the real-time processing [13], [14] such as time series forecasting [15].

Some researchers addressed the instability in ESN because of the randomness in weights assigning, which is allocated in the reservoir part and is assigned only once and fixed [9]. However, authors in [16] adopted the use of bidirectional input. Both of the directions of the data feed as an input sequence to the same reservoir in both forward and backward ways to capture different independent versions of information from the input data. Authors in [17] showed that having two different inputs in a straight and reverse order will improve the memorization.

Dimension reduction techniques are transforming the high dimensional data within the feature space into another subspace of lower-dimensional representation to avoid computation and assist in de-correlating the transformed data. Therefore, dimensionality reduction techniques are applied to solve these problems by using a particular transformation map such as Principal Component Analysis (PCA) or Random Projection (RP) [18]. High dimensional sparse output from the reservoir layer makes feature representation intractable and leads to overfitting and high computational resources [17]. In machine learning, dimension reduction is useful to prepare a more informative representation for the classifier. There are studies that used PCA as a powerful tool for dimensional reduction of the output of the reservoir layer [17], [19].

Tuning the hyperparameters in ESN is a common issue since it is significantly affecting the performance of the reservoir. Optimizing these hyperparameters are typically slow and consequently, researchers either assign them manually based on experience [20] or they adopt different optimization approaches such as grid search, random search, and Bayesian optimization [21].

In this work, we proposed a novel reservoir computing approach for SER using bidirectional late fusion, Sparse Random Projection (SRP), and optimizing hyperparameters with a Bayesian optimization method. Additionally, a multivariate time series handcrafted features of which Mel-Frequency Cepstral Coefficients (MFCCs) and Gamma-Tone Cepstral Coefficients (GTCCs) have been used to feed the reservoir layer. The main contributions of the proposed model are: 1) adopting a very sparse random projection [22] approach for dimension reduction which can be more compatible with the sparse data distribution produced by the reservoir; 2) using the bidirectionality approach with the late representation fusion which may improve the memorization capability of ESN.

The rest of this paper is organized as follows: Section II covers literature about the existing methods of SER, while section III presents the proposed model, and section IV shows experiments and results. The discussion work is presented in section V, and finally, the conclusion and future work come in section VI.

## II. LITERATURE REVIEW

Researchers widely use speech signals to detect emotion in the field of HCI to gain a better interaction between them. Therefore, the right design model for classification and relevant emotion features from speech with distinctive information are the two significant aspects in speech emotion recognition models [23].

To extract valuable features, some researchers preferred a handcrafted feature while others used deep learned features. Handcrafted feature representation can be a global feature that represents each sample as one vector or it can be local features extracted from the sequence of the frames. There are a variety of open-source toolkits for extracting features from speech such as openSMILE [24] and COVAREP [25]. A lot of studies [26]–[28], [29] have used openSMILE toolkit as it is one of the most famous tools to extract emotion features from speech. The openSMILE toolkit is extracting non-temporal global features. However, some researchers are using time series features from speech signals to detect the real-time emotion recognition. Scherer *et al.* [14] used spectral features from frames, however, they were not successful

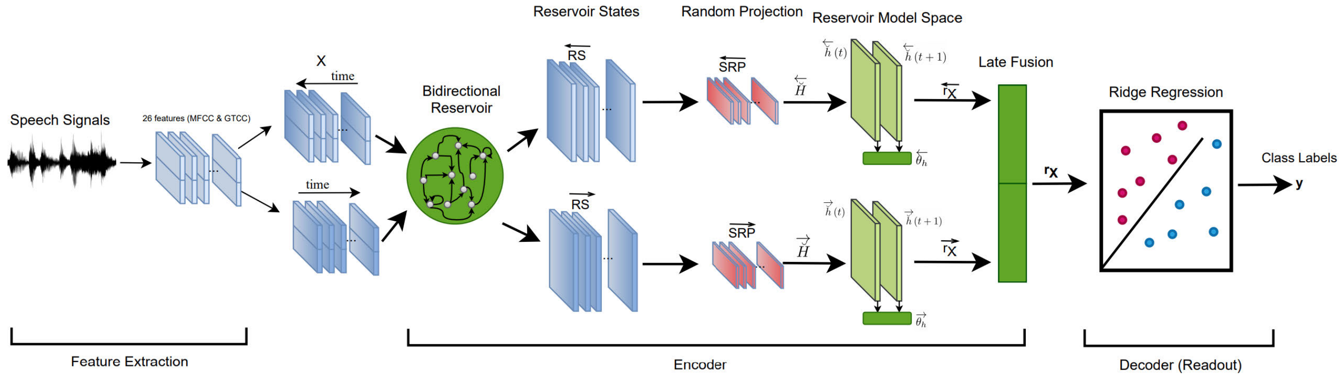


**FIGURE 1.** Reservoir computing with random projection and late bidirectional fusion (The proposed model).

to recognize emotion in real-time. On the other hand, recent works are focusing on learned features directly from the raw speech signal by using deep learning models [8], [30]. Authors in [31], used 1D CNN network for SER systems that can learn features from the speech signal. The time series features representation requires a proper classifier such as RNN which is computationally intensive.

Besides choosing the right features from speech, developing a robust mathematical model is another vital step to the high performance of emotion recognition from speech signals [32]. As mentioned before, frame-based features require a model to support multivariate time series data such as RNN. In [31] and [33] a high-level representation features are used with adopting bidirectional long short-term memory (BLSTM) model. A speech emotion model using both CNN and LSTM proposed in [34] and [35]. The data augmentation techniques are applied in [36] on Acted Emotional Speech Dynamic Database (AESDD) with the use of CNN for continuous speech emotion recognition.

However, few researchers reported the use of ESN for SER, for example, in [14], authors proposed a not fully successful real-time speech emotion recognition model. To participate in Evalita 2014 competition, Gallicchio *et al.* [37] proposed ESN to detect emotion from speech. Additionally, Saleh and Micheli in [38] used ESN for SER, where only neutral and anger emotion classes are used in their model.

The time complexity of RNN-based models (such as LSTM) versus ESN has been investigated and reported frequently. The untrained nature of ESN shows the capability to significantly reduce the time complexity as shown in Table 1. The ESN performance is always comparable to the LSTM. However, we shall see in the discussion section that the proposed ESN in this work can outperform the LSTM for SER. With a competitive performance in time series prediction, ESN with the simplicity of its architecture deterministically raised to propose in many applications [42]. Bidirectionality is applied in ESN by feeding an input sequence into the same reservoir in both forward and backward to capture additional information independently of the input data. For example, authors in [16] and [17] proposed bidirectional reservoir to improve the memorization capability. For the same purpose, Bianchi *et al.* in [43] proposed Bidirectional Deep-readout ESN (BDESN) and multilayer perceptron (MLP) as a classifier. The deep bidirectional LSTM [31], [33] has been used in the SER field to learn the temporal information for detecting the final state of emotion.

**TABLE 1.** The comparison of the training time between LSTM and ESN.

| Method | LSTM (sec.) | ESN (sec.) |
|---|---|---|
| Jirak et al. [39] | 88.9 | 2.6 |
| Gallicchio et al. [40] | 26175 | 677 |
| Variengien & Hinaut. [41] | 410 | 47.1 |
| The proposed model | 1748 (3 epochs) | 50.73 |

High dimensional sparse output from a reservoir layer makes feature representation suffer from the curse of dimensionality which is why dimension reduction step is necessary to prepare a non-sparse representation to feed the classifier. The Principal Component Analysis (PCA) was used with ESN in [17] and [43] to improve the model performance. But [19] used ELM-based Auto-encoder (ELM-AE) beside PCA to reduce dimensionality between reservoirs in their deep ESN approach.

Regarding the hyperparameters in ESN, which have a significant effect on the model performance, some researchers adopted fixing these hyperparameters [17], [20]. However, to improve ESN performance, [44] optimized the hyperparameters by Grasshopper Optimization Algorithm (GOA) approach. ESN is also found to exploit Bayesian optimization [45] approach to tune its hyperparameters [21], [48]. In order to achieve more satisfactory performance for SER, the Bayesian optimization approach has been adopted by [46] to optimize the hyperparameters of k-nearest neighbors, support vector machine and decision tree, and also adopted by [47] to optimize the kernel size for the CNN.

## III. METHODOLOGY

In this section, the model design is presented, and the proposed model is briefly explained. It represents the main components of the solution and explains how the proposed method helps to improve the performance of ESN to recognize emotions from speech. Most of the works on SER have used global features and very few works were working on time series local features. Several works have been found in the literature that used LSTM as a model to feed time series features. In addition, there are few works that used ESN for the SER systems [14], [37], [38], however, none of them reached an outstanding performance. This unconvincing performance may be due to three factors which are: 1) adopting a unidirectional signal processing which results in losing important information between the speech frames in the opposite direction, 2) ESN for temporal data produces a very high dimensional representation that negatively influences the performance of the classifier, and 3) the manual tuning of the ESN hyperparameters instead of optimizing them may not lead to optimum performance of the ESN model. To overcome these drawbacks, we have been inspired by the work of [17], and have used ESN with bidirectional time series features and dimension reduction representation to recognize emotions from speech. Our contribution in this work is to modify the adapted model to improve the performance of SER. The next subsections show the details of the proposed model, which is shown in Figure 1.

### A. FEATURE EXTRACTION

Speech features with discriminative information have a vital role in emotion recognition in speech. Extracting the proper speech emotion features reflects obtainable information about emotion characteristics and the effect of the human's emotional condition on the speech signal.

In this work, frame-based handcrafted features have been adopted to feed the proposed model. The first set of features that have been extracted in this work is 13 MFCC features. MFCC is the most widely used feature for speech emotion recognition because of its simplicity of computation and the good capability of extracting informative features. However, MFCC based models are suffered by decreasing the performance under noisy conditions because MFCCs are biased by noise which triggers mismatched likelihood calculation [49]. Therefore, we extracted 13 GTCC features which have better performance than MFCCs under noisy conditions. Overall, 26 features are used as an input to our model.

The audioFeatureExtractor object method in MATLAB has been used to extract the features with windows of length 30ms overlapped by 20ms. Since the length of the samples vary (See Figure 2), we have equated the length of the samples by padding with zeros or pruning at the start and the end of each row data. Consequently, we have used 500, 600, 400, and 300 frames for Emo-DB, SAVEE, RAVDESS, and FAU Aibo respectively based on the near maximum length for each dataset.

### B. BIDIRECTIONAL RESERVOIR COMPUTING–ESN

Echo State Networks (ESNs) were first proposed by [50] as a special case of RNN for learning nonlinear systems which is also a part of the reservoir computing framework. The Reservoir computing (RC) framework is a kind of RNN model whose recurrent part weights are initiated randomly and then fixed without training, followed by a trainable layer that can be updated with the output [51].

The untrained nature of ESN makes it avoid the complexity available in trained natured networks such as LSTM. It has a sparse nature as it maps the features into a higher dimensional space. In addition, ESN has a simple architecture that contains an input layer, reservoir layer and output layer. Regarding the input layer, a bidirectionality multivariate time series data is applied by feeding an input sequence into the reservoir in both forward and backward. The advantage of the bidirectional approach is to capture additional information independently of the input data and the capability to improve the memorization with straight and reverse inputs. The reservoir layer contains sparsely connected neurons which are randomly assigned and fixed without training.

The temporal dependence of time series data can be handled effectively by ESN which is successfully applied for chaotic time series prediction models. The simplicity of ESN is that most of the weights are randomly assigned and not trainable. The complexity of deep recurrent networks requires an extreme computing time which makes ESN an ultimate nominee for tasks involving real-time processing such as time series forecasting.

The input multivariate time series sample data contains $D$-dimensional feature vector for each time step $t$, where $t = 1, 2, \ldots, T$, and $T$ is the number of time steps. In other words $x(t) \in \mathbb{R}^D$ and $X = [x(1), x(2), \ldots x(T)]^T$. Note that $T$ represents the number of time steps after padding the samples to avoid length differences. As an RNN based model, reservoir model is suitable for the sequential data and for bidirectional approach which has been adopted in this work. The state in the reservoir layer can be updated using the following equations:

$$\begin{aligned}\overrightarrow{h}(t) &= f(\overrightarrow{x}(t), \overrightarrow{h}(t-1); \theta_{enc})\\ \overleftarrow{h}(t) &= f(\overleftarrow{x}(t), \overleftarrow{h}(t-1); \theta_{enc})\end{aligned} \tag{1}$$

where $\overrightarrow{h(.)}$ and $\overleftarrow{h(.)}$ are the RNN states at time $t$ for both bidirectional inputs that can be computed as a function of their previous values ($\overrightarrow{h}(t-1)$, $\overleftarrow{h}(t-1)$) and the current inputs $\overrightarrow{x}(t)$ and $\overleftarrow{x}(t)$. In addition, f is a nonlinear activation hyperbolic tangent function, and $\theta_{enc}$ are the adaptable parameters from the reservoir.

The equation (1) can be presented as the simplest formulation as follows:

$$\begin{aligned}\overrightarrow{h}(t) &= tanh(W_{in}\overrightarrow{x}(t) + W_r\overrightarrow{h}(t-1))\\ \overleftarrow{h}(t) &= tanh(W_{in}\overleftarrow{x}(t) + W_r\overleftarrow{h}(t-1))\end{aligned} \tag{2}$$

where $W_{in}$ is the input weight and $W_r$ is the weight from reservoir connections, and the reservoir states ($\overrightarrow{RS}$ and $\overleftarrow{RS}$) are generated by the reservoir layer over time, where $\overrightarrow{RS} = [\overrightarrow{h}(1), \overrightarrow{h}(2), .., \overrightarrow{h}(T)]^T$ and $\overleftarrow{RS} = [\overleftarrow{h}(1), \overleftarrow{h}(2), .., \overleftarrow{h}(T)]^T$. The $\theta_{enc}$ can be represented as $\theta_{enc} = \{W_{in}, W_r\}$.

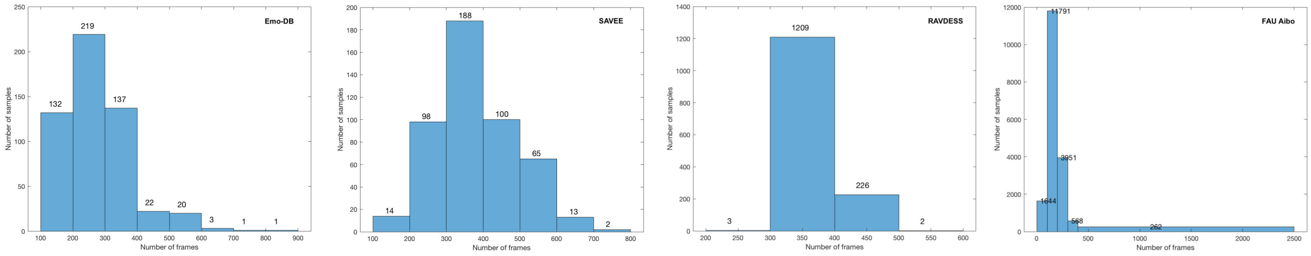


**FIGURE 2.** **The distribution of the audio files based on their lengths (number of frames).**

The reservoir has several hyperparameters that have a significant effect on its performance such as (i) the amount of internal (hidden) units $R$, (ii) spectral radius $\rho$ of reservoir connection weights matrix $W_r$ which helps the system to be stable [52] and normally should be less than 1, (iii) the nonzero connections $\beta$ is used as a percentage of non-zero connection weights, (iv) scaling $\omega$ of the values in $W_{in}$ is another hyperparameter, which controls the total of nonlinearity in handling the hidden units together with $\rho$ and can change the internal dynamics from a chaotic regime to a contractive regime [53], (v) leak as an amount of leakage in the reservoir state update, and (vi) it is also possible to include a dropout regularization and we applied a dropout, particularly for recurrent architectures [17].

### C. RANDOM PROJECTION BASED DIMENSION REDUCTION

The high dimensional sparse output from the reservoir layer makes feature representation intractable and leads to overfitting and high computational cost. Additionally, Sparse Random Projection (SRP) has been used to transform the sparse output into a more compact representation.

Trainable dimension reduction such as PCA is well known, however, because the sparse data distribution produced by the reservoir uses a binomial distribution, adopting a sparse random projection where its values initialized by 1 and $-1$ can be a suitable alternative. In addition, PCA is more time consuming because of the training part inside it. SRP reduces the dimensions and preserves the distances in addition to the fact that random projection has a low complexity since it does not need any training and removes redundancies with minimal loss of information.

In the work, we follow [22] by using a SRP matrix. The SRP matrix R is initialized with 1 and -1 as in the following equation:

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

where $d$ is the dimension of the reservoir output state. This step will reduce the dimension to a specific number that can be fixed or optimized. Reducing the dimensions has a significant impact on implementing the reservoir model space which will be applied nextly. The dimensionality reduction step decreases the number of reservoir output features and produces a new sequences $\overrightarrow{H}$ and $\overleftarrow{H}$ which will be the input to the model space.

### D. RESERVOIR MODEL SPACE AND LATE FUSION

The reservoir model space that has been proposed by [17], distinguishes a generative model of the reservoir sequence and induces a metric relationship between the samples. In this work, we adopted a bidirectional approach with late fusion. Processing of each direction in a separate way can provide richer information about the relation of the time steps in both forward and backward directions. The late fusion will combine more diverse representations of the data and make the characteristics of each individual direction to be more highlighted. Consequently, the formula from a proposed model by [17] has been adapted with two separate outputs from unsupervised dimensionality reduction process from SRP as shown in the following equations:

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

where $\overrightarrow{h}(.)$ and $\overleftarrow{h}(.)$ are the columns of a frontal slice $\overrightarrow{H}$ 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 $D$ is the number of dimension after the reduction process. The late fusion will be applied in this stage by concatenating the generated output from both $\overrightarrow{r_X}$ and $\overleftarrow{r_X}$ where:

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

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

Equation 7 shows that, how the $\overrightarrow{\theta_h}$ and $\overleftarrow{\theta_h}$ can be learned by minimizing a ridge regression loss function:

$$\begin{aligned}\overrightarrow{\theta}_h^* &= \underset{\{\overrightarrow{U_o},\overrightarrow{u_o}\}}{arg\,min}\frac{1}{2}\|\overrightarrow{h}(t)\overrightarrow{U_o} + \overrightarrow{u_o} - \overrightarrow{h}(t+1)\|^2 + \mu\|\overrightarrow{U_o}\|^2\\ \overleftarrow{\theta}_h^* &= \underset{\{\overleftarrow{U_o},\overleftarrow{u_o}\}}{arg\,min}\frac{1}{2}\|\overleftarrow{h}(t)\overleftarrow{U_o} + \overleftarrow{u_o} - \overleftarrow{h}(t+1)\|^2 + \mu\|\overleftarrow{U_o}\|^2\end{aligned} \tag{7}$$

where the $\mu$ is the regularization parameter to adjust the number of the coefficient shrinkage in the reservoir model space. In the classification level ESN adopts a linear model for decoding which is usually formed as in the following

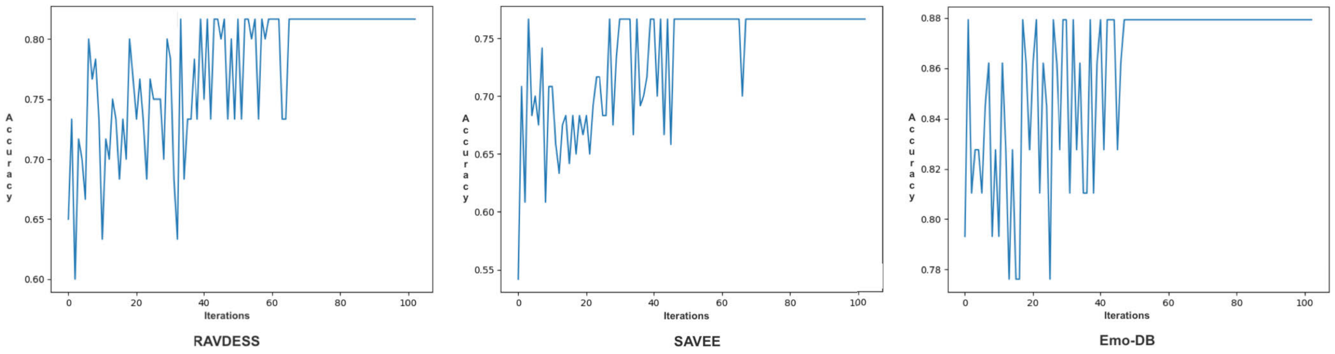


**FIGURE 3.** Samples of optimization process for one of the speakers in RAVDESS, SAVEE, and Emo-DB to find the optimal value for hyperparameters.

equation:

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

This model has a set of parameters $\theta_{dec} = \{V_o, v_o\}$. $\theta_{dec}$ which can be learned by minimizing the loss function in a ridge regression which admits a closed form solution:

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

where $\lambda$ is the regularization parameter for ridge regression and helps to minimize overfitting of the training data. The aim of the linear readout is to perform the final classification that maps the $r_X$ representation into the class labels $y$.

### E. THE BAYESIAN HYPERPARAMETER OPTIMIZATION

Determining the ESN hyperparameters is one of the reported issues due to its effects on the ESN model performance. However, most of the works have assigned ESN parameters manually or based on experiences. In this work, we optimized major ESN hyperparameters such as the size of reservoir state, spectral radius, size of connectivity, input scaling, amount of leakage in the reservoir state update, and the number of dropouts. Furthermore, optimizing the number of resulting dimensions after the dimensionality reduction procedure, and both regularization parameters $\mu$ in modal space and $\lambda$ in ridge regression readout part. Based on the comparison in [48] between Bayesian optimization and grid search, Bayesian optimization shows to be more efficient than a grid search in their experiments. Bayesian optimization is a gradient-free global optimization approach to optimize random functions [54]. It is initiated to minimize loss functions $f(\theta)$ of the models, where their hyperparameters $\theta$ are normally difficult to be tuned [48]. Additionally, the Bayesian optimization method has been used in various applications including SER models [46], [47] to optimize the models' hyperparameters.

**TABLE 2.** The optimized parameters which have been used in the proposed method by Bayesian optimization approach.

| Stages | Parameters |
|---|---|
| Reservoir | Internal units (R), spectral radius ($\rho$), non-zero connections ($\beta$), scaling ($\omega$), leakage and dropout |
| Sparse Random Projection | The size of dimension reduction |
| Model Space | Regularization parameter of the ridge regression ($\mu$) |
| Readout | Regularization parameter of the ridge regression in readout ($\lambda$) |

In this work, Bayesian optimization [54] has been used to tune the parameters of the reservoir layer and the ridge regression, in addition to dimensionality reduction size in SRP as shown in Table 2. The Figure 3 shows a sample of the 100 iterations for the three used datasets. Optimizing these parameters has a significant effect to improve the performance of the model.

### F. SPEAKER NORMALIZATION (SN)

Inspired by the work of Valsenko *et al.* [55] we adopted Speaker Normalization (SN) on each particular speaker sample in speaker-independent experiments. SN is comprehended as subtracting the mean of utterances that belong to one of the speakers in a specific dataset, divided by the standard deviation of those samples. The purpose of using SN is to counteract the samples from specific speaker influences, thus the emotion space is more improved. While SN is a totally unsupervised approach and labels are not necessary. This method has improved the performance of using speaker-independent in our proposed model.

## IV. EXPERIMENTAL SETUP AND RESULTS

In this section, we evaluated and validated the performance of the proposed model to detect emotions from the speech on most public and available SER datasets. So far, there is not much research work that has reported about using ESN for speech emotion recognition. The reason may be the wide use of global features instead of time series features. However, as mentioned in the previous sections, ESN can have a good performance in time series data.

In this work, we have used handcrafted time series features which are 13 MFCCs and 13 GTCCs extracted for each window of length 30ms overlapped by 20ms. The features

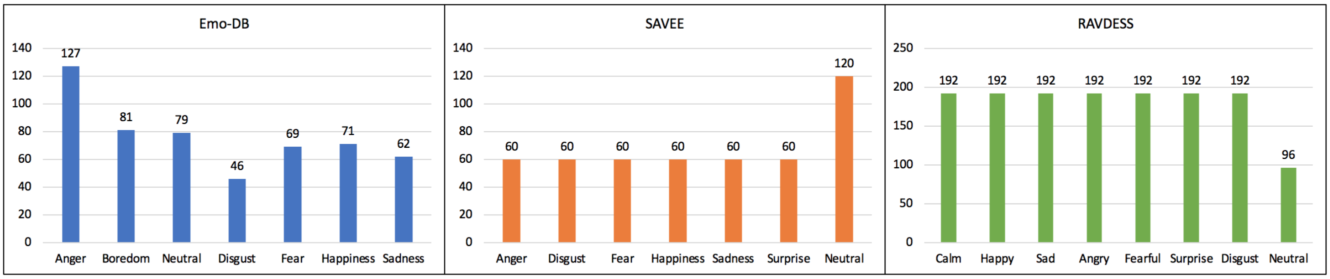


**FIGURE 4.** The total numbers of emotion classes for Emo-DB, SAVEE and RAVDESS datasets.

feed the reservoir layer, where the number of internal units has been optimized using Bayesian optimization. The SRP is applied to transform the high dimensional sparse nature output from the reservoir layer into more compact representation. The reservoir model space distinguishes a generative model of the reservoir sequence and induces a metric relationship between the samples that came from the SRP part. Subsequently, the late fusion of bidirectionality input has been applied with the processing of each direction in a separate way. Bayesian optimization has been used to tune hyperparameters of the reservoir layer, ridge regression, and the size of dimensionality reduction in SRP as shown in Table 2.

The proposed model results are presented in terms of precision, recall, F1 score, unweighted and weighted percentage accuracy. Precision and recall are used to evaluate the performance of classification and F1 score is the weighted average of both precision and recall. The weighted accuracy coincides to the correctly classified emotion divided by the total number of emotion classes, while the unweighted accuracy (UA) means the average of per-class accuracies. The detailed results per each emotion classes of all four datasets are given in Tables 3 - 12.

We applied a speaker-independent approach using Leave One Speaker Out (LOSO), in addition to a speaker-dependent approach using the 5-fold and 10-fold cross-validation techniques on Emo-DB, SAVEE and RAVDESS datasets. In adopting 5-fold and 10-fold cross-validation, the dataset is divided into 5 and 10 folds respectively with mutually exclusive subsets. The model is trained and tested 5 times for 5-fold and 10 times for 10-fold, each time one set is considered as a test set and the remaining sets are considered as a train set. To conduct fair comparison with the state-of-the-art studies of the FAU Aibo dataset, we followed the adopted protocol of the interspeech09 challenge [56].

Since the ESN has no trainable weights in the reservoir layer, but rather it uses fixed weights, it doesn't need any GPU or high resources, therefore, we carried out the experiments using CPU on Google Colab (12 GB RAM) and on PC with 64GB RAM.

Furthermore, our experiments on both speaker-independent on all datasets and speaker-dependent on Emo-DB, SAVEE, and RAVDESS datasets are conducted and have shown better performance as compared to state-of-the-art works. The performance of the proposed model is validated using four well known public speech emotion datasets, which are Berlin Database of Emotional Speech (Emo-DB) [57], Surrey Audio-Visual Expressed Emotion (SAVEE) [58], Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [59] and FAU Aibo Emotion Corpus [60].

## A. EMO-DB

The Emo-DB [57] is a German dataset for emotional speech, produced by the Technical University of Berlin. It covers seven emotion classes: anger, boredom, neutral, disgust, fear, happiness, and sadness. Additionally, 10 actors (5 females and 5 males, between the age of 20 and 35) are involved to take a specific emotion over the memories of their real experience. Emo-DB is the most popular dataset that is used in speech emotion recognition with a total number of 535 utterance files which include anger (127), boredom (81), neutral (79), disgust (46), fear (69), happiness (71) and sadness (62) sentences, see Figure 4. We validated the proposed model based on speaker-independent and speaker-dependent for 5-fold and 10-fold cross-validation.

### 1) SPEAKER-INDEPENDENT

The LOSO method is applied for speaker-independent, as in Emo-DB we set 9 speakers as a train set and one speaker as a test set and this process will be repeated to guarantee the participation of all speakers in the test set.

**TABLE 3.** The proposed model performance (%) for speaker-independent (LOSO approach) SER using Emo-DB dataset.

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Anger | 79.37 | 100 | 88.50 |
| Boredom | 90.48 | 93.83 | 92.12 |
| Disgust | 97.44 | 82.61 | 89.41 |
| Fear | 96.55 | 81.16 | 88.19 |
| Happiness | 93.75 | 63.38 | 75.63 |
| Sadness | 92.31 | 96.77 | 94.49 |
| Neutral | 87.65 | 89.87 | 88.75 |
| **Unweighted** | 91.08 | 86.80 | 88.16 |
| **Weighted** | 89.45 | 88.41 | 88.11 |

Table 3 shows the detailed results of precision, recall, F1 score, unweighted, and weighted percentage accuracy for each emotion class for Emo-DB dataset. The confusion matrix in Figure 5 shows the individual accuracy of each 7 emotion classes of Emo-DB dataset for the speaker-independent approach. The anger class recorded the highest accuracy which is 100% from all speakers while happiness recorded the lowest with only 63%. However, disgust and fear emotions have less accuracy compared with boredom, sadness and neutral.

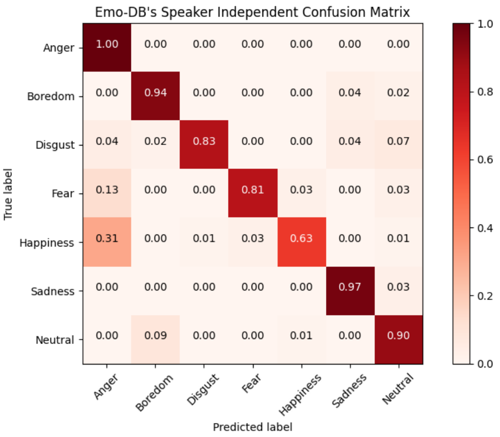


**FIGURE 5.** Speaker-independent confusion matrix of proposed model for Emo-DB dataset.

### 2) SPEAKER-DEPENDENT

For the speaker-dependent approach, we applied the 5-fold cross-validation method on Emo-DB dataset and its results in terms of precision, recall, F1 score, unweighted and weighted accuracy, are shown in Table 4.

**TABLE 4.** The proposed model performance (%) for speaker-dependent (5-fold) SER using Emo-DB dataset.

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Anger | 92.03 | 100 | 95.85 |
| Boredom | 85.71 | 88.89 | 87.27 |
| Disgust | 100 | 91.30 | 95.45 |
| Fear | 92.75 | 92.75 | 92.75 |
| Happiness | 96.61 | 80.28 | 87.69 |
| Sadness | 100 | 98.39 | 99.19 |
| Neutral | 86.59 | 89.87 | 88.20 |
| **Unweighted** | 93.38 | 91.64 | 92.34 |
| **Weighted** | 92.58 | 92.34 | 92.29 |

The confusion matrix for speaker-dependent (5-fold) in Figure 6 shows that the highest accuracy is obtained by the anger emotion, while the lowest accuracy is recorded by the happiness emotion similar to the speaker-independent approach.

The same procedure is applied for the 10-fold cross-validation approach, and the detailed results in terms of precision, recall, F1 score, unweighted and weighted accuracy are shown in Table 5.

The confusion matrix for speaker-dependent (10-fold) in Figure 7 shows that the highest accuracy is achieved by the anger emotion (99%) and similar to the speaker-independent and 5-fold approaches, the lowest accuracy is achieved by happiness emotion.

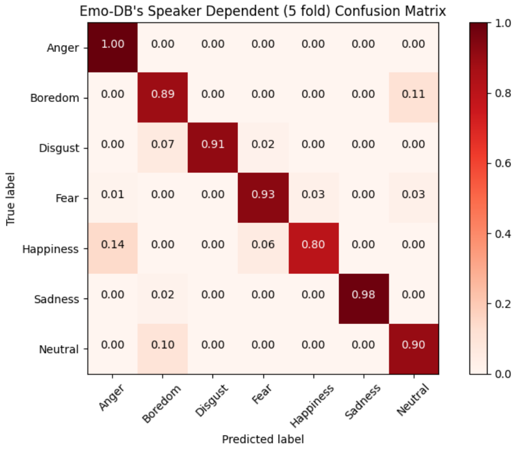


**FIGURE 6.** Speaker-dependent confusion matrix with 5-fold cross-validation of proposed model for Emo-DB dataset.

**TABLE 5.** The proposed model performance (%) for speaker-dependent (10-fold) SER using Emo-DB dataset.

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Anger | 91.97 | 99.21 | 95.45 |
| Boredom | 92.68 | 93.83 | 93.25 |
| Disgust | 97.83 | 97.83 | 97.83 |
| Fear | 94.20 | 94.20 | 94.20 |
| Happiness | 98.31 | 81.69 | 89.23 |
| Sadness | 98.39 | 98.39 | 98.39 |
| Neutral | 93.75 | 94.94 | 94.34 |
| **Unweighted** | 95.30 | 94.30 | 94.67 |
| **Weighted** | 94.72 | 94.58 | 94.51 |

## *B. SAVEE*

SAVEE [58] is a multimodal (Audio and Visual expression) British English voice database that can be used for facial expression and speech emotion recognition. In our study, only the audio speech part has been used. It was recorded from four male native English speakers at the University of Surrey with seven basic emotion categories, which are anger, disgust, fear, happiness, sadness, surprise, and neutral. Each actor recorded 120 utterances which overall speech samples are 480 files, and the total number of utterances of the neutral emotion class is 120 while the other remaining emotion classes comprised of 60 utterances, which is shown in Figure 4. SAVEE dataset also used to validate the proposed model based on speaker-independent and speaker-dependent for 5-fold and 10-fold cross-validation.

### 1) SPEAKER-INDEPENDENT

The LOSO method has been applied for speaker-independent, as in SAVEE we set one speaker as a test set and the remaining speakers as a train set and this process will be repeated to guarantee the participation of all the speakers in the test set. Table 6 shows the detailed results of precision, recall,

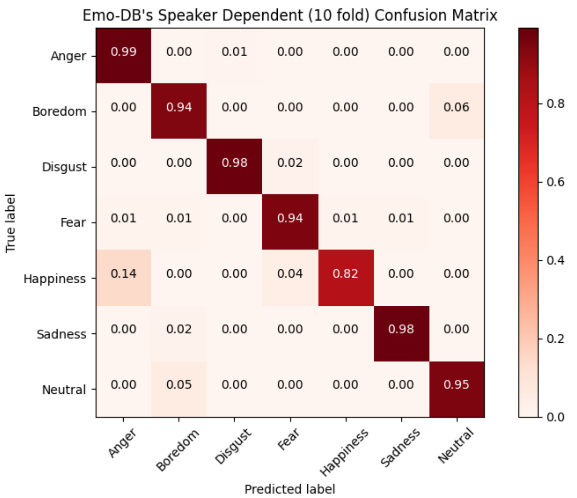


**FIGURE 7.** Speaker-dependent confusion matrix with 10-fold cross-validation of proposed model for Emo-DB dataset.

**TABLE 6.** The proposed model performance (%) for speaker-independent (LOSO approach) SER using SAVEE dataset.

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Anger | 76.81 | 88.33 | 82.17 |
| Disgust | 65.85 | 45.00 | 53.47 |
| Fear | 80.49 | 55.00 | 65.35 |
| Happiness | 73.24 | 86.67 | 79.39 |
| Neutral | 65.68 | 92.50 | 76.82 |
| Sadness | 76.47 | 43.33 | 55.32 |
| Surprise | 74.55 | 68.33 | 71.30 |
| **Unweighted** | 73.30 | 68.45 | 69.12 |
| **Weighted** | 72.35 | 71.46 | 70.08 |

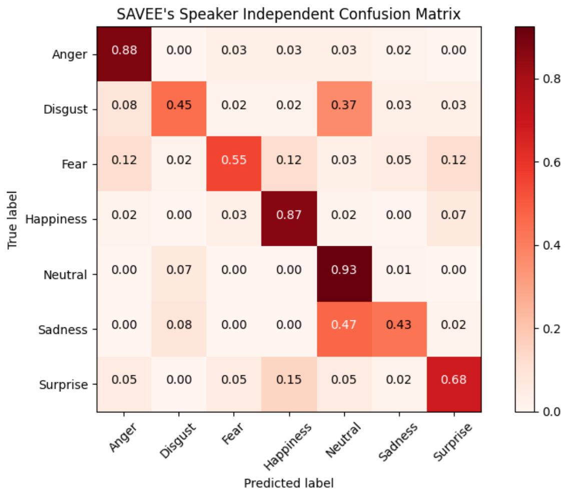


**FIGURE 8.** Speaker-independent confusion matrix of proposed model for SAVEE dataset.

F1 score, unweighted, and weighted percentage accuracy for each emotion class for SAVEE dataset. The confusion matrix in Figure 8 shows the individual accuracy of each 7 emotion classes of SAVEE dataset for the speaker-independent approach. The neutral class recorded the highest accuracy which is 93% from all speakers while sadness recorded the lowest with only 43% while 47% of sadness emotion is considered as a neutral emotion. The same case has happened in disgust emotion which 37% recognized as neutral and only 45% as a current emotion which is disgust.

### 2) SPEAKER-DEPENDENT

For the speaker-dependent method and for getting the most reliable result of our model, same as Emo-DB dataset, we applied 5-fold and 10-fold cross-validation. Table 7 shows the results and statistics in terms of precision, recall, F1 score, weighted, and unweighted percentage accuracy.

**TABLE 7.** The proposed model performance (%) for speaker-dependent (5-fold) SER using SAVEE dataset.

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Anger | 83.02 | 73.33 | 77.88 |
| Disgust | 79.07 | 56.67 | 66.02 |
| Fear | 69.81 | 61.67 | 65.49 |
| Happiness | 69.35 | 71.67 | 70.49 |
| Neutral | 73.12 | 97.50 | 83.57 |
| Sadness | 76.74 | 55.00 | 64.08 |
| Surprise | 69.70 | 76.67 | 73.02 |
| **Unweighted** | 74.40 | 70.36 | 71.51 |
| **Weighted** | 74.24 | 73.75 | 73.01 |

The confusion matrix for speaker-dependent (5-fold) in Figure 9 points out the true emotion label and predicted emotion label. Similar to the speaker-independent approach, the highest accuracy is achieved by the neutral emotion with 97% and while sadness emotion is still the lowest with 55% which 37% are considered as a neutral emotion. However, the disgust emotion obtained 57% which is much higher if we compared it with the speaker-independent accuracy (45%).

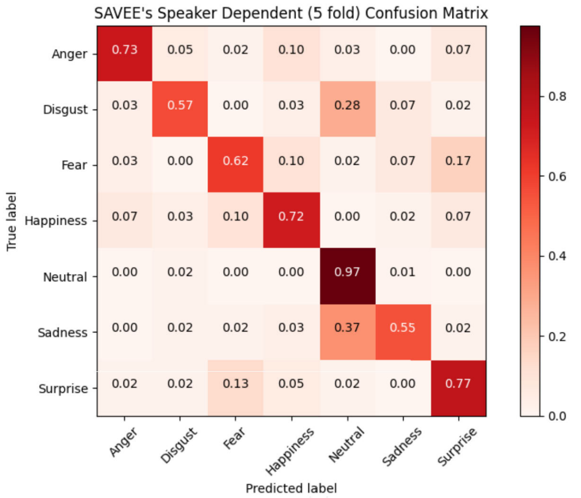


**FIGURE 9.** Speaker-dependent confusion matrix with 5-fold cross-validation of proposed model for SAVEE dataset.

The same procedure is applied for the 10-fold cross-validation approach, and the results of precision, recall, F1 score, unweighted and weighted accuracy are shown in Table 8.

**TABLE 8. The proposed model performance (%) for speaker-dependent (10-fold) SER using SAVEE dataset.**

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Anger | 88.89 | 80.00 | 84.21 |
| Disgust | 81.25 | 65.00 | 72.22 |
| Fear | 86.27 | 73.33 | 79.28 |
| Happiness | 80.33 | 81.67 | 80.99 |
| Neutral | 74.52 | 97.50 | 84.48 |
| Sadness | 79.55 | 58.33 | 67.31 |
| Surprise | 76.92 | 83.33 | 80.00 |
| **Unweighted** | 81.10 | 77.02 | 78.36 |
| **Weighted** | 80.28 | 79.58 | 79.12 |

The proposed model for speaker-dependent (10-fold) evaluation is presented in the given Figure 10. The confusion matrix in Figure 10 shows that the highest accuracy is achieved by neutral emotion (97%) and is the same as speaker-independent and speaker-dependent (5-fold), the sadness emotion has the lowest accuracy (58%) compared with other classes. Additionally, 35% of sadness emotion was recognized as a neutral emotion which is the same situation we have in both previous approaches.

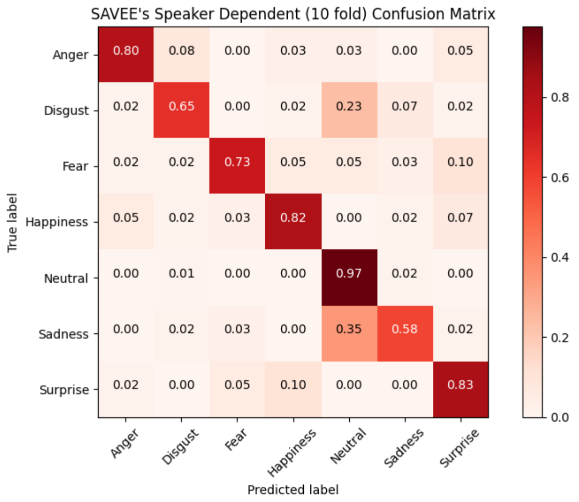


**FIGURE 10. Speaker-dependent confusion matrix with 10-fold cross-validation of proposed model for SAVEE dataset.**

## *C. RAVDESS*

RAVDESS [59] is the third speech emotion dataset that has been used to validate our model. It is a multimodal dataset that contains facial expression and voice data for speech and song. RAVDESS was recorded with a North American accent by 24 professional actors (12 females and 12 males) with eight emotions: calm, happy, sad, angry, fearful, surprise, neutral, and disgust expressions. Additionally, RAVDESS contains overall 7356 files and only 1440 speech files as a voice channel for speech emotion have been used. The total utterances of the neutral emotion class are 96 while the other remaining emotion classes have 192 utterances, which is shown in Figure 4.

The proposed model has been validated on RAVDESS dataset, based on speaker-independent and speaker-dependent for 5-fold and 10-fold cross-validation approaches.

### 1) SPEAKER-INDEPENDENT

LOSO method has been applied for speaker-independent, as in for RAVDESS, we set 23 speakers as a train set and one speaker as a test set and this process will be repeated to guarantee the participation of all speakers in the test set.

There are a few works that applied a speaker-independent approach on RAVDESS dataset, and none of them applied LOSO. However, our work is the same as Emo-DB and SAVEE where we adopted the LOSO approach. Table 9 shows the detailed results of precision, recall, F1 score, unweighted, and weighted percentage accuracy for each emotion class for RAVDESS dataset for speaker-independent.

**TABLE 9. The proposed model performance (%) for speaker-independent (LOSO approach) SER using RAVDESS dataset.**

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Neutral | 69.57 | 50.00 | 58.18 |
| Calm | 74.89 | 88.54 | 81.15 |
| Happy | 78.49 | 70.31 | 74.18 |
| Sad | 61.93 | 63.54 | 62.72 |
| Angry | 80.50 | 83.85 | 82.14 |
| Fearful | 74.73 | 72.40 | 73.54 |
| Disgust | 77.03 | 83.85 | 80.30 |
| Surprised | 76.67 | 71.88 | 74.19 |
| **Unweighted** | 74.23 | 73.05 | 73.30 |
| **Weighted** | 74.54 | 74.58 | 74.31 |

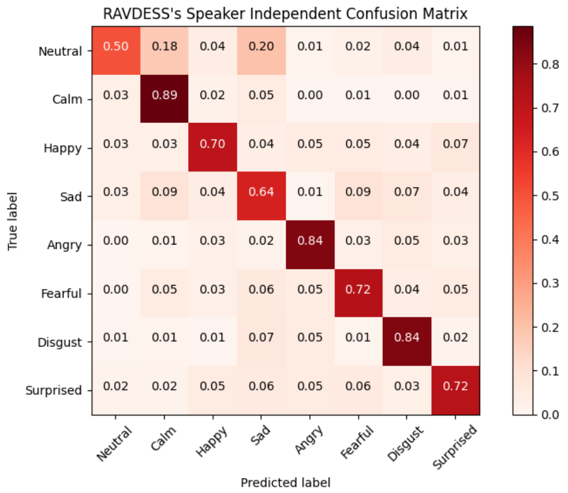


**FIGURE 11. Speaker-independent confusion matrix of proposed model for RAVDESS dataset.**

The confusion matrix in Figure 11 shows the individual accuracy for each of the 8 emotion classes of RAVDESS dataset for the speaker-independent approach. The calm emotion class recorded the highest accuracy which is 89% from all speakers, however, neutral emotion recorded only 50%

accuracy with 18% and 20% recognized as calm and sad emotion classes respectively.

### 2) SPEAKER-DEPENDENT

For the speaker-dependent approach, we applied the 5-fold and 10-fold cross-validation approach. The 5-fold results are shown in Table 10, in terms of precision, recall, F1 score, unweighted and weighted percentage accuracy.

**TABLE 10. The proposed model performance (%) for speaker-dependent (5-fold) SER using RAVDESS dataset.**

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Neutral | 86.52 | 80.21 | 83.24 |
| Calm | 84.11 | 93.75 | 88.67 |
| Happy | 90.00 | 89.06 | 89.53 |
| Sad | 79.10 | 72.92 | 75.88 |
| Angry | 90.58 | 90.10 | 90.34 |
| Fearful | 84.97 | 85.42 | 85.19 |
| Disgust | 85.86 | 88.54 | 87.18 |
| Surprised | 87.23 | 85.42 | 86.32 |
| **Unweighted** | 86.05 | 85.68 | 85.79 |
| **Weighted** | 86.01 | 86.04 | 85.96 |

The confusion matrix for speaker-dependent (5-fold) in Figure 12 shows that the highest accuracy obtained by the calm emotion which is 94% while sad emotion is recorded as the lowest accuracy of 73%. Therefore, the neutral emotion has a significant improvement in the speaker-dependent (5-fold) approach with 80% accuracy, while it was only 50% in speaker-independent.

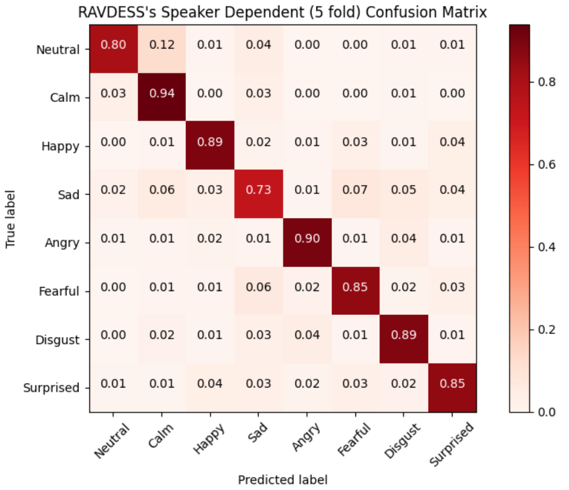


**FIGURE 12. Speaker-dependent confusion matrix with 5-fold cross-validation of proposed model for RAVDESS dataset.**

The same procedure is applied for the 10-fold approach, and the results in terms of precision, recall, F1 score, unweighted, and weighted percentage accuracy are shown in Table 11. The confusion matrix for speaker-dependent (10-fold) in Figure 13 shows that the accuracy for each RAVDESS emotion class, and compared with the 5-fold approach, the accuracy of all emotion classes are higher except happy emotion with 86% accuracy.

**TABLE 11. The proposed model performance (%) for speaker-dependent (10-fold) SER using RAVDESS dataset.**

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Neutral | 85.56 | 80.21 | 82.80 |
| Calm | 87.26 | 96.35 | 91.58 |
| Happy | 89.19 | 85.94 | 87.53 |
| Sad | 86.78 | 78.65 | 82.51 |
| Angry | 94.76 | 94.27 | 94.52 |
| Fearful | 85.64 | 90.10 | 87.82 |
| Disgust | 92.59 | 91.15 | 91.86 |
| Surprised | 89.34 | 91.67 | 90.49 |
| **Unweighted** | 88.89 | 88.54 | 88.64 |
| **Weighted** | 89.11 | 89.10 | 89.03 |

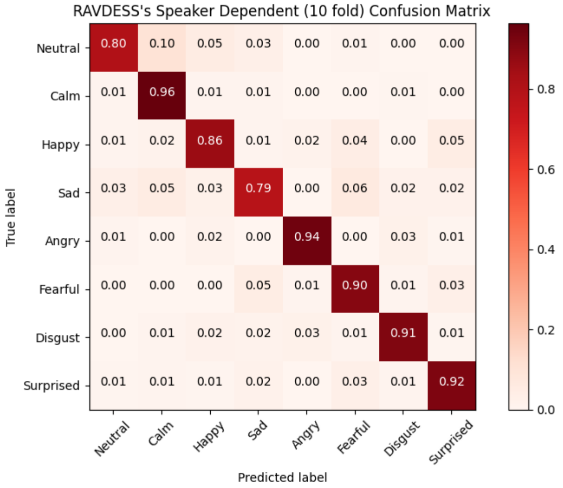


**FIGURE 13. Speaker-dependent confusion matrix with 10-fold cross-validation of proposed model for RAVDESS dataset.**

## D. FAU AIBO EMOTION CORPUS

The fourth dataset that has been used to evaluate the proposed model is the non-acted FAU Aibo Emotion Corpus, which contains 9.2 hours of spontaneous and emotional German speech samples [60]. The dataset was recorded from a total of 51 children (21 male and 30 female) at the age 10-13 years during their interactions with Sony's pet robot Aibo at two different schools, 'Ohm' and 'Mont'. The corpus contains 18216 chunk speech samples where dataset designers labeled each word in the dataset into 10 categories and later, they mapped them into five different emotion classes which are anger, emphatic, neutral, positive, and rest. The final numbers for each emotion class are listed in Figure 15. Following the adopted protocol of the interspeech09 challenge [56], we used 'Ohm' with 9959 chunks from 26 children (13 males, 13 females) as a training set and 'Mont' with 8257 utterances from 25 children (8 males, 17 females) as a testing set.

The number of chunks per class in the FAU Aibo Emotion corpus is extremely unbalanced as shown in Figure 15, where in 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 rest. To overcome the unbalanced issue, we applied random under sampler [61] where under sampling on the

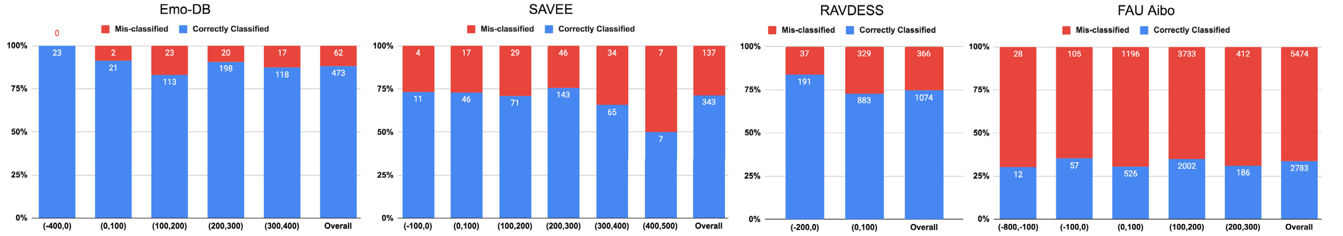


**FIGURE 14.** The impact of equating the length of the samples by padding with zeros or pruning at the start and end of each row data. The x-axis represents the range of zero padding or pruning (positive values refers to padding while negative values refers to length pruning).

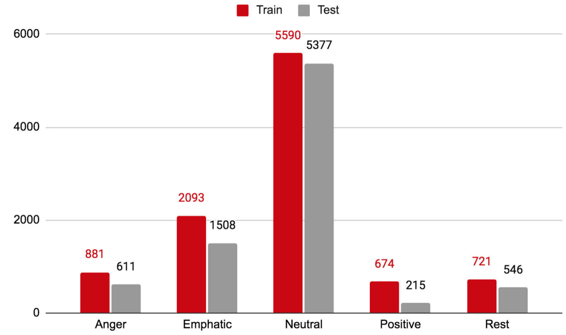


**FIGURE 15.** The total number of emotion classes for FAU Aibo dataset.

**TABLE 12.** The proposed model performance (%) for FAU Aibo dataset.

| Emotion | Precision | Recall | F1 Score |
|---|---|---|---|
| Anger | 19.46 | 64.48 | 29.89 |
| Emphatic | 33.81 | 57.43 | 42.57 |
| Neutral | 83.64 | 23.86 | 37.13 |
| Positive | 9.67 | 65.58 | 16.86 |
| Rest | 14.58 | 18.13 | 16.16 |
| **Unweighted** | 32.23 | 45.90 | 28.52 |
| **Weighted** | 63.30 | 33.70 | 35.67 |

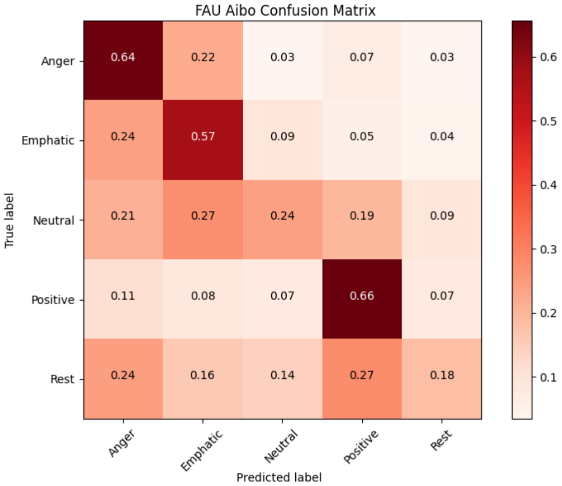


**FIGURE 16.** The confusion matrix of proposed model for FAU Aibo dataset.

majority classes is adopted by randomly picking a fixed number of samples.

Table 12 lists the detailed results of the precision, the 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 Figure 16 shows the accuracy of each 5 involved emotion classes of FAU Aibo. The positive class recorded 66% as the highest accuracy and the rest emotion class with 18% is the lowest accuracy that we have got from the proposed model. The low accuracy of rest may be due to its samples nature where they have different labels but are gathered under the same class.

### E. THE IMPACT OF ZERO PADDINGS

As mentioned in the methodology section, since the length of the samples vary, as shown in Figure 2, we have equated the length of the samples by padding with zeros or pruning at the start and end of each row data. During the experiments as shown in Figure 14, we have found that there is no clear relation between the error and sample length scaling. For example, in Emo-DB dataset, the error of samples when 100-200 zeros are added were 16.91%, however, samples with more added zeros (such as 200-300) recorded lower error (9.17%), while samples with 300-400 added zeros recorded an error of 12.59%. In SAVEE dataset, the low misclassification ratio (50%) where 400-500 zeros are added, may not relate to the zero padding ratio, since 78.5% of the samples in this length range come from one of the speakers of whom its result is 53%. Similar observations can be noticed in RAVDESS and FAU Aibo datasets where noticeable zero padding ratio has been applied. These observations do not highlight any pattern regarding the relation between error increasing and zero paddings.

### F. MODEL EVALUATION

To evaluate the impact of the adopted late fusion instead of early fusion in addition to the use of SRP instead of PCA, we have conducted four speaker independent-based experiments for each dataset, including the use of Early Fusion with PCA (EF-PCA), Early Fusion with RP (EF-RP), Late Fusion with PCA (LF-PCA), and Late Fusion with RP (LF-RP). In the exception of Emo-DB, all of the other

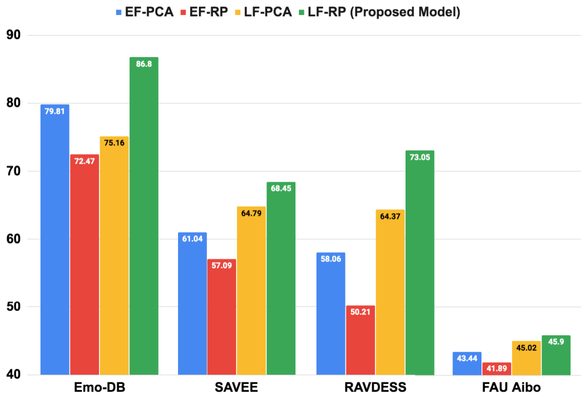


**FIGURE 17.** Comparing the proposed model with speaker-independent method to Early Fusion with PCA (EF-PCA), Early Fusion with RP (EF-RP), and Late Fusion with PCA (LF-PCA).

datasets show that late fusion with both PCA and RP outperforms the early fusion (See Figure 17). However, regarding the Emo-DB dataset, the LF-PCA is not able to outperform EF-PCA, but the RP impact on the late fusion model is significant and records 86.80% of accuracy. Overall, the proposed model (LF-RP) outperformed all other three methods on all four datasets.

**TABLE 13.** The impact of using bidirectional, dimension reduction and optimization method in a speaker-independent approach over basic ESN.

| Dataset | ESN | Proposed model |
|---|---|---|
| Emo-DB | 68.97 | 86.80 |
| SAVEE | 54.29 | 68.45 |
| RAVDESS | 58.33 | 73.05 |
| FAU Aibo | 42.91 | 45.90 |

To show the impact of the adopted bidirectional, dimension reduction and optimization method in the proposed model over a basic ESN (unidirectional, total dimensions, and non-optimized hyperparameters are used). Table 13 shows the outperformance of the proposed model using all the involved datasets in a speaker-independent approach.

## V. DISCUSSION

In this section, we are comparing the proposed model performance with other baseline methods. In order to obtain high classification accuracy, we proposed a novel ESN model which deals with a small size of handcrafted features as an input to the reservoir layer with bidirectional time series representation where its hyperparameters have been optimized. Additionally, we applied sparse random projection to reduce the output feature representation from the reservoir layer, which helped the model to perform better in dealing with sparse representation data. We adopt speaker-independent for all four popular benchmark datasets and speaker-dependent with 5-fold and 10-fold cross-validations on Emo-DB, SAVEE, and RAVDESS datasets to recognize the emotional state from speech signals. The late bidirectional fusion helped to extract more information from the data before feeding it to the ridge regression classifier. This novel proposed approach for SER helped to improve the classification accuracy and because of the simplicity and the trainless nature of ESN, the processing time is reduced as compared to the other deep learning methods such as LSTM and CNN.

In this discussion, we are going to present the overall unweighted accuracy (UA), since the actual performance is more representing, especially when the data is imbalanced in terms of utterance sizes per class, as shown in Figure 4. UA is the sum of all class accuracies divided by the number of classes, without taking into account the number of samples per class. Consequently, UA is a useful evaluation metric for emotion recognition studies due to the imbalanced nature of emotion datasets.

We have compared the performance of our proposed model in speaker-independent and/or speaker-dependent schema with the previously presented methods for Emo-DB, SAVEE, RAVDESS, and FAU Aibo datasets (See Table 14 - 17).

**TABLE 14.** Summary of unweighted accuracies (UA%) achieved by various researchers for Emo-DB, SAVEE and RAVDESS datasets using speaker-independent approach.

| Dataset | Method | UA% |
|---|---|---|
| Emo-DB | [62] (Spectrogram+Deep learning ADRNN) | 84.99 |
| | [30] (Spectrogram+Deep learning PCRN) | 84.53 |
| | [65] (Spectrogram+Deep learning ACRNN) | 82.82 |
| | [63] (SVM-RBF) | 71.02 |
| | [64] (OpenSmile+SVM) | 76.82 |
| | [66] (Handcrafted+GEBF) | 76.81 |
| | [67] (OpenSmile+Deep learning RDBN) | 82.32 |
| | **Proposed model** | **86.80** |
| SAVEE | [66] (Handcrafted+GEBF) | 55.00 |
| | [67] (OpenSmile+Deep learning RDBN) | 53.60 |
| | **Proposed model** | **68.45** |
| RAVDESS | **Proposed model** | **73.05** |

The classification UA of speaker-independent experiments are shown in Table 14. For the Emo-DB dataset, our result achieved 86.80% UA, which performed better compared with various new works that have been conducted recently. Our model when applied on the SAVEE dataset obtained 68.45% which is 13.45% higher than the second-best result. However, our proposed model adopted LOSO, and few works are conducted for RAVDESS speaker-independent method, none of which applied the LOSO approach. For example, [31] used 2 speakers for testing the model while [68] used 19 speakers for training and four other speakers for the testing scenario. However, the proposed model achieved 73.05% by adopting the LOSO approach.

Table 15 shows the summary of unweighted classification accuracies (UA%) achieved by various researchers using 5-fold cross-validation for Emo-DB, SAVEE, and RAVDESS speech datasets.

Among all state-of-the-art methods, our approach performed the best. Considering that ESN is extremely less complicated than LSTM based models, our proposed model when applied to Emo-DB achieved 91.64%, which is slightly

**TABLE 15.** Summary of unweighted accuracies (UA%) achieved by various researchers for Emo-DB, SAVEE and RAVDESS datasets using 5-fold cross-validation approach.

| Dataset | Method | UA% |
|---|---|---|
| Emo-DB | [31] (Spectrogram+Deep learning BiLSTM) | 91.14 |
| | [62] (Spectrogram+Deep learning ADRNN) | 90.37 |
| | [69] (Handcrafted+Deep learning 1D CNN) | 86.10 |
| | [63] (SVM-RBF) | 78.66 |
| | [70] (VGGVox+SVM,kNN) | 80.00 |
| | **Proposed model** | **91.64** |
| SAVEE | [70] (VGGVox+SVM,kNN) | 68.00 |
| | **Proposed model** | **70.36** |
| RAVDESS | [31] (Spectrogram+Deep learning BiLSTM) | 82.01 |
| | [72] (Spectrogram+Deep learning GResNets) | 64.52 |
| | [69] (Handcrafted+Deep learning 1D CNN) | 71.61 |
| | [71] (MFCC+Deep learning BFN, CNA) | 83.00 |
| | [63] (SVM-RBF) | 64.32 |
| | [70] (VGGVox+SVM,kNN) | 71.00 |
| | **Proposed model** | **85.68** |

better than the work in [31] where they used the Deep BiLSTM method and CNN for feature extraction. Additionally, our result in Emo-DB has outperformed the work in [62] by 1.27% and the other works significantly, as shown in Table 15.

The proposed model when validated by SAVEE achieved 70.34% of UA. However, to our best knowledge, there is only one recent work that adopted a 5-fold approach applied to SAVEE dataset [70] and gained 68% of UA with 2% lower than our result.

Regarding the RAVDESS dataset, our proposed model achieves high UA with the 5-fold schema (85.68%) and outperforms the highest achieved results in state-of-the-art studies by 2.68%.

**TABLE 16.** Summary of unweighted accuracies (UA%) achieved by various researchers for Emo-DB, SAVEE and RAVDESS datasets using 10-fold cross-validation approach.

| Dataset | Method | UA% |
|---|---|---|
| Emo-DB | [73] (MFMC,MFCC+SVM) | 81.50 |
| | [29] (OpenSmile+SVM,kNN,MLP) | 84.62 |
| | [64] (OpenSmile+SVM) | 87.66 |
| | **Proposed model** | **94.30** |
| SAVEE | [73] (MFMC,MFCC+SVM) | 75.63 |
| | [29] (OpenSmile+SVM,kNN,MLP) | 72.39 |
| | [74] (MFCC+Deep learning 1D CNN) | 65.83 |
| | **Proposed model** | **77.02** |
| RAVDESS | [73] (MFMC,MFCC+SVM) | 64.31 |
| | [74] (MFCC+Deep learning 1D CNN) | 75.83 |
| | **Proposed model** | **88.54** |

Table 16 shows the performance of the 10-fold cross-validation for the speaker-dependent experiments, and the proposed model achieved 94.3%, 77.02%, and 88.54% of UA for Emo-DB, SAVEE, and RAVDESS datasets respectively.

Based on Table 16 our model has obtained the highest classification UA for all used three datasets. Regarding the Emo-DB dataset, our model has outperformed the closest work in the state-of-the-art [64] by 6.64%.

Regarding the SAVEE dataset, our proposed model achieved 1.39% higher UA than the second highest result. While for RAVDESS dataset, we can clearly see that our novel model has achieved an outstanding UA (88.54%).

To compare the performance of our proposed model with the state-of-the-art studies, a number of recent works that used FAU Aibo dataset and followed the 2009 challenge protocol are presented in Table 17. It is obvious that the proposed model in this work has outperformed these studies by 45.9% of UA accuracy, as an indication of the usefulness of it for non-acted emotional datasets as well besides the acted ones.

**TABLE 17.** Summary of unweighted accuracies (UA%) achieved by various researchers for FAU Aibo dataset followed the 2009 challenge protocol.

| Dataset | Method | UA% |
|---|---|---|
| FAU Aibo | [75] (Spectrogram+Deep learning eResNet) | 41.3 |
| | [78] (Spectrogram+Deep learning BLSTM) | 45.4 |
| | [79] (Spectrogram+Deep learning 2D CNN) | 41.1 |
| | [76] (Handcrafted+SVM,NN,DNN) | 45.3 |
| | [77] (MRA+SVM) | 45.2 |
| | **Proposed model** | **45.9** |

Among the studies mentioned in Table 14, 15 and 17, some of them have adopted spectrogram-based features with deep learning and have achieved distinguished results. In the Emo-DB dataset and speaker-independent approach, authors in [62] and [65] have achieved a classification accuracy of 84.99% and 82.82% respectively using 3-D Log-Mel spectrums from raw speech signals and feed them to 3-D attention-based convolutional recurrent neural networks (ACRNN). Additionally, Jiang *et al.* [30] extracted 3-D log Mel-spectrograms from the speech signal and fed it to a parallelized convolutional recurrent neural network (PCRN) model and recorded 84.53% UA. However, our proposed model is able to outperform these spectrogram-based features with deep learning and achieved UA of 86.80%. On the other hand, when adopting a 5-fold cross-validation method, researches [31], [62] applied their spectrogram-based deep learning model on Emo-DB and have achieved better results than other models. Mustaqeem *et al.* [31] achieved 91.14% UA by using salient features from the speech spectrogram with deep bidirectional LSTM to learn the Spatio-temporal information for detecting the last state of the emotion model. Again, our proposed model achieved 91.64% UA exceeding the model of [31] by 0.5%. Spectrogram-based features with deep learning models are also applied to the RAVDESS dataset [31], [72], however, in spite of its good achievement unlike the EMO-DB, the work of [71] outperformed them.

Regarding the Aibo dataset, one can notice that the highest achieved result in the previous works is using spectrogram-based features with deep learning models [75], [78], [79]. Shih *et al.* [78] achieved 45.4% UA by extracting deep spectrum representations and developing a deep learning model with the attention enhanced FCN and BLSTM networks. Our proposed model is once again able to outperform the mentioned study by achieving UA of 45.9%.

## VI. CONCLUSION AND FUTURE WORK

We proposed a novel recurrent based architecture for time series speech emotion recognition classification by using bidirectional late fusion ESN based on the reservoir model space representation with sparse random projection. Early fusion of the temporal features generated by bidirectional reservoir leads to the loss of independency from both representations. Thus, to avoid the drawback of the linear combination representation of both directional representations produced by dimension reduction, we proposed the late fusion of the representations, which is applied later to the dimension reduction step to overcome this problem.

On the other side, dimensionality reduction of sparse data by using SRP is reported to be useful to prepare a more compact and informative representation for the classifier. SRP reduces the dimensions and preserves the distances in addition to the fact that random projection has a low complexity since it does not need training. Because of the small size of features and a nontrainable ESN method, our model is fast and more robust to achieve better performance. Another factor that has a notable impact on increasing the performance of our model is the use of Bayesian optimization to optimize ESN hyperparameters. The Bayesian optimization in our work has been adopted to fix a large number of parameters in the proposed model and has shown an ability to record a good performance.

This proposed model has come out with the highest classification UA compared to the previous works on SER when using 5-fold and 10-fold speaker-dependent, LOSO speaker-independent on Emo-DB, SAVEE, and RAVDESS datasets, and speaker-independent on FAU Aibo.

A single reservoir suffers from generating a comprehensive representation and from the randomness assigned to it. For this reason, in future work, we intend to use more than one reservoir to create a more typical representation of the input data that captures more information independently of the input data.

## REFERENCES

[1] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, Oct. 2018.

[2] A. Al-Talabani, H. Sellahewa, and S. A. Jassim, "Emotion recognition from speech: Tools and challenges," *Proc. SPIE*, vol. 9497, May 2015, Art. no. 94970N.

[3] P. Barros and S. Wermter, "A self-organizing model for affective memory," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 31–38.

[4] B. Basharirad and M. Moradhaseli, "Speech emotion recognition methods: A literature review," *AIP Conf.*, vol. 1891, no. 1, 2017, Art. no. 020105.

[5] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[6] T. Kathiresan and V. Dellwo, "Cepstral derivatives in MFCCs for emotion recognition," in *Proc. IEEE 4th Int. Conf. Signal Image Process. (ICSIP)*, Jul. 2019, pp. 56–60.

[7] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, Dec. 2018.

[8] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *J. Finance Data Sci.*, vol. 2, no. 4, pp. 265–278, Dec. 2016.

[9] Q. Wu, E. Fokoue, and D. Kudithipudi, "On the statistical challenges of echo state networks and some potential remedies," 2018, *arXiv:1802.07369*. [Online]. Available: http://arxiv.org/abs/1802.07369

[10] Q. Ma, L. Shen, W. Chen, J. Wang, J. Wei, and Z. Yu, "Functional echo state network for time series classification," *Inf. Sci.*, vol. 373, pp. 1–20, Dec. 2016.

[11] Y. Zhang, Y. Yu, and D. Liu, "The application of modified ESN in chaotic time series prediction," in *Proc. 25th Chin. Control Decis. Conf. (CCDC)*, May 2013, pp. 2213–2218.

[12] E. López, C. Valle, H. Allende, E. Gil, and H. Madsen, "Wind power forecasting based on echo state networks and long short-term memory," *Energies*, vol. 11, no. 3, p. 526, Feb. 2018.

[13] S. Ortín, M. C. Soriano, M. Alfaras, and C. R. Mirasso, "Automated real-time method for ventricular heartbeat classification," *Comput. Methods Programs Biomed.*, vol. 169, pp. 1–8, Feb. 2019.

[14] S. Scherer, M. Oubbati, F. Schwenker, and G. Palm, "Real-time emotion recognition from speech using echo state networks," in *Proc. IAPR Workshop Artif. Neural Netw. Pattern Recognit.* Berlin, Germany: Springer, 2008, pp. 205–216.

[15] T. Kim and B. R. King, "Time series prediction using deep echo state networks," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17769–17787, Dec. 2020.

[16] A. Rodan, A. F. Sheta, and H. Faris, "Bidirectional reservoir networks trained using SVM+ privileged information for manufacturing process modeling," *Soft Comput.*, vol. 21, no. 22, pp. 6811–6824, 2017.

[17] F. M. Bianchi, S. Scardapane, S. Løkse, and R. Jenssen, "Reservoir computing approaches for representation and classification of multivariate time series," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2169–2179, May 2021.

[18] A. Al-Talabani, "Automatic speech emotion recognition-feature space dimensionality and classification challenges," Ph.D. dissertation, Univ. Buckingham, Buckingham, U.K., 2015.

[19] Q. Ma, L. Shen, and G. W. Cottrell, "DeePr-ESN: A deep projection-encoding echo-state network," *Inf. Sci.*, vol. 511, pp. 152–171, Feb. 2020.

[20] M. Lukoševičius, *A Practical Guide to Applying Echo State Networks*. Berlin, Germany: Springer, 2012, pp. 659–686.

[21] L. Cerina, M. D. Santambrogio, G. Franco, C. Gallicchio, and A. Micheli, "EchoBay: Design and optimization of echo state networks under memory and time constraints," *ACM Trans. Archit. Code Optim.*, vol. 17, no. 3, pp. 1–24, Aug. 2020.

[22] P. Li, T. J. Hastie, and K. W. Church, "Very sparse random projections," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: Association for Computing Machinery, 2006, p. 287, doi: 10.1145/1150402.1150436.

[23] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on DNN-decision tree SVM model," *Speech Commun.*, vol. 115, pp. 29–37, Dec. 2019.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 1459–1462.

[25] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 960–964.

[26] A. Al-Talabani, H. Sellahewa, and S. Jassim, "Excitation source and low level descriptor features fusion for emotion recognition using SVM and ANN," in *Proc. 5th Comput. Sci. Electron. Eng. Conf. (CEEC)*, Sep. 2013, pp. 156–161.

[27] Z.-T. Liu, B.-H. Wu, D.-Y. Li, P. Xiao, and J.-W. Mao, "Speech emotion recognition based on selective interpolation synthetic minority over-sampling technique in small sample environment," *Sensors*, vol. 20, no. 8, p. 2297, Apr. 2020.

[28] Z. Zhang, "Speech feature selection and emotion recognition based on weighted binary cuckoo search," *Alexandria Eng. J.*, vol. 60, no. 1, pp. 1499–1507, Feb. 2021.

[29] T. Özseven, "A novel feature selection method for speech emotion recognition," *Appl. Acoust.*, vol. 146, pp. 320–326, Mar. 2019.

[30] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019.

[31] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.

[32] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Process.*, vol. 22, pp. 1154–1160, Dec. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200412001133

[33] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. Dresden, Germany: ISCA, 2015, pp. 1537-1540. [Online]. Available: http://dblp.unitrier.de/db/conf/interspeech/interspeech2015.html

[34] Y. Mu, L. A. H. Gómez, A. C. Montes, C. A. Martínez, X. Wang, and H. Gao, "Speech emotion recognition using convolutional-recurrent neural networks with attention model," *DEStech Trans. Comput. Sci. Eng.*, no. cii, pp. 341–350, Nov. 2017.

[35] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+LSTM architecture for speech emotion recognition with data augmentation," 2018, *arXiv:1802.05630*. [Online]. Available: http://arxiv.org/abs/1802.05630

[36] N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas, and G. Kalliris, "Continuous speech emotion recognition with convolutional neural networks," *J. Audio Eng. Soc.*, vol. 68, nos. 1–2, pp. 14–24, Feb. 2020.

[37] C. Gallicchio and A. Micheli, "A preliminary application of echo state networks to emotion recognition," in *Proc. 4th Int. Workshop EVALITA*. Pisa, Italy: Pisa Univ. Press, Dec. 2014, pp. 116–119.

[38] Q. Saleh, C. Merkel, D. Kudithipudi, and B. Wysocki, "Memristive computational architecture of an echo state network for real-time speech-emotion recognition," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl. (CISDA)*, May 2015, pp. 1–5.

[39] D. Jirak, S. Tietz, H. Ali, and S. Wermter, "Echo state networks and long short-term memory for continuous gesture recognition: A comparative study," *Cogn. Comput.*, pp. 1–13, 2020, doi: 10.1007/s12559-020-09754-0.

[40] C. Gallicchio, A. Micheli, and L. Pedrelli, "Comparison between DeepESNs and gated RNNs on multivariate time-series prediction," 2018, *arXiv:1812.11527*. [Online]. Available: http://arxiv.org/abs/1812.11527

[41] A. Variengien and X. Hinaut, "A journey in ESN and LSTM visualisations on a language task," 2020, *arXiv:2012.01748*. [Online]. Available: http://arxiv.org/abs/2012.01748

[42] J. Dan, W. Guo, W. Shi, B. Fang, and T. Zhang, "Deterministic echo state networks based stock price forecasting," *Abstract Appl. Anal.*, vol. 2014, pp. 1–6, Jun. 2014.

[43] F. M. Bianchi, S. Scardapane, S. Løkse, and R. Jenssen, "Bidirectional deep-readout echo state networks," 2017, *arXiv:1711.06509*. [Online]. Available: http://arxiv.org/abs/1711.06509

[44] L. Qin, W. Li, and S. Li, "Effective passenger flow forecasting using STL and ESN based on two improvement strategies," *Neurocomputing*, vol. 356, pp. 244–256, Sep. 2019.

[45] F. Nogueira. (2014). *Bayesian Optimization: Open Source Constrained Global Optimization Tool for Python*. [Online]. Available: https://github.com/fmfn/BayesianOptimization

[46] O. M. Nezami, P. J. Lou, and M. Karami, "ShEMO: A large-scale validated database for Persian speech emotion detection," *Lang. Resour. Eval.*, vol. 53, no. 1, pp. 1–16, Mar. 2019.

[47] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi, and N. Ismail, "Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks," in *Proc. 6th Int. Conf. Wireless Telematics (ICWT)*, Sep. 2020, pp. 1–6.

[48] J. R. Maat, N. Gianniotis, and P. Protopapas, "Efficient optimization of echo state networks for time series datasets," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.

[49] M. Babu, M. A. Kumar, and S. Santhosh, "Extracting MFCC and GTCC features for emotion recognition from audio speech signals," *Int. J. Res. Comput. Appl. Robot*, vol. 2, no. 8, pp. 46–63, 2014.

[50] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science*, vol. 304, no. 5667, pp. 78–80, Apr. 2004. [Online]. Available: https://science.sciencemag.org/content/304/5667/78

[51] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Comput. Sci. Rev.*, vol. 3, no. 3, pp. 127–149, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013709000173

[52] F. M. Bianchi, L. Livi, and C. Alippi, "Investigating echo-state networks dynamics by means of recurrence analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 427–439, Feb. 2018.

[53] L. Livi, F. M. Bianchi, and C. Alippi, "Determination of the edge of criticality in echo state networks through Fisher information maximization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 706–717, Mar. 2018.

[54] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012.

[55] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining frame and turn-level information for robust recognition of emotions within speech," in *Proc. INTERSPEECH*, 2007, pp. 1–4.

[56] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 1–4.

[57] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1–4.

[58] S. Haq and P. Jackson, "Multimodal emotion recognition," in *Machine Audition: Principles, Algorithms and Systems*. Hershey, PA, USA: IGI Global, Aug. 2010, pp. 398–423.

[59] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[60] S. Steidl, *Automatic Classification of Emotion Related User States in Spontaneous Children'S Speech*. Berlin, Germany: Logos-Verlag, 2009.

[61] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-365.html

[62] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[63] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Inf. Sci.*, vol. 563, pp. 309–325, Jul. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025521001584

[64] S. Yildirim, Y. Kaya, and F. Kılıç, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Appl. Acoust.*, vol. 173, Feb. 2021, Art. no. 107721. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X20308252

[65] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

[66] F. Daneshfar, S. J. Kabudian, and A. Neekabadi, "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier," *Appl. Acoust.*, vol. 166, Sep. 2020, Art. no. 107360. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X1931117X

[67] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, "Random deep belief networks for recognizing emotions from speech signals," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–9, Mar. 2017.

[68] M. A. Jalal, R. K. Moore, and T. Hain, "Spatio-temporal context modelling for speech emotion classification," in *Proc. IEEE Automat. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 853–859.

[69] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101894. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809420300501

[70] G. Assunção, P. Menezes, and F. Perdigão, "Speaker awareness for speech emotion recognition," *Int. J. Online Biomed. Eng.*, vol. 16, no. 4, pp. 15–22, 2020.

[71] M. Ezz-Eldin, A. A. M. Khalaf, H. F. A. Hamed, and A. I. Hussein, "Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition," *IEEE Access*, vol. 9, pp. 19999–20011, 2021.

[72] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019, doi: 10.1007/s11042-017-5539-3.

[73] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," *Appl. Acoust.*, vol. 179, Aug. 2021, Art. no. 108046. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X21001390

[74] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative emotion recognition using deep learning for Thai language," in *Proc. Joint Int. Conf. Digit. Arts, Media Technol. With ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng. (ECTI DAMT NCON)*, Mar. 2020, pp. 71–74.

[75] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[76] P.-Y. Shih, C.-P. Chen, and H.-M. Wang, "Speech emotion recognition with skew-robust neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2751–2755.

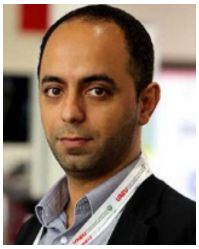
[77] S. Deb and S. Dandapat, "Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 802–815, Mar. 2019.

[78] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019.

[79] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, and B. W. Schuller, "Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition," *Neural Netw.*, vol. 141, pp. 52–60, Sep. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021000939

**HEMIN IBRAHIM** received the bachelor's degree in mathematics from the College of Science, University of Salahaddin, the second bachelor's degree in information technology from the University of Kurdistan–Hewler (UKH), and the master's degree in advanced software engineering from Sheffield University, U.K. He is currently pursuing the Ph.D. degree in artificial intelligence from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. His research interests include speech processing, emotion recognition, and robotics.

**CHU KIONG LOO** (Senior Member, IEEE) received the B.Eng. degree (Hons.) in mechanical engineering from the University of Malaya, Kuala Lumpur, Malaysia, and the Ph.D. degree from Universiti Sains Malaysia, George Town, Malaysia. He was a Design Engineer with various industrial firms and is the Founder of the Advanced Robotics Laboratory, University of Malaya. He has been involved in the application of research into Peruss quantum associative model and Pribram's holonomic brain in humanoid vision projects. He is currently a Professor in computer science and information technology with the University of Malaya. He has led many projects funded by the Ministry of Science in Malaysia and the High Impact Research Grant from the Ministry of Higher Education, Malaysia. His current research interests include brain-inspired quantum neural networks, constructivism-inspired neural networks, synergetic neural networks, and humanoid research.



**FADY ALNAJJAR** received the M.Sc. degree in artificial intelligence systems and the Ph.D. degree in system design engineering from the University of Fukui, Japan, in 2007 and 2010, respectively. From 2010 to 2016, he worked as a Research Scientist with the Brain Science Institute (BSI), RIKEN, Japan, where he has been a Visiting Researcher, since 2016. He had joined the College of Information Technology, UAE University, in 2016, where he has been the Head of the AI and Robotics Laboratory, since 2016. He is working in neuro-robotics study to understand the underlying mechanisms for embodied cognition and mind. He is also exploring the neural mechanisms for motor learning, adaptation, and recovery after brain injury from the sensory- and muscle-synergies perspectives. His research target is to propose an advanced neurorehabilitation methodology for patients with brain impairments, such as children with autism, elderly with cognitive impairments, and post-stroke patients.