RESERVOIR COMPUTING WITH TRUNCATED NORMAL DISTRIBUTION FOR SPEECH EMOTION RECOGNITION

Hemin Ibrahim[1], Chu Kiong Loo[2*]

[1,2]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

Email: hemin.ibrahim@siswa.um.edu.my[1], ckloo.um@um.edu.my[2*](corresponding author)

DOI: https://doi.org/10.22452/mjcs.vol35no2.3

ABSTRACT

Speech is an effective, quick, and important way for communicating and exchanging complex information between humans. Emotions have always been a part of normal human conversation which makes the speech more attractive. Because of this major role of both speech and emotion, many researchers are inspired by studying Speech Emotion Recognition (SER) which still has plenty of challenges. In this study, we proposed a novel reservoir computing approach with the initialization of random connection weights for the input weight by the truncated normal distribution. Furthermore, Population-Based Training (PBT) is adopted to optimize the hyperparameters of the whole Echo State Network (ESN) model which have a significant impact on the model performance. The proposed model has adopted bidirectional reservoir input to increase the memorization capability, and Sparse Random Projection (SRP) was applied for dimensional reduction as a simple, unsupervised, and low complexity approach. The speaker-independent strategy was employed on EMODB and SAVEE datasets as an acted speech emotion dataset and Aibo as a non-acted dataset. The model achieved 84.8%, 65.95%, and 45.99% unweighted average recalls on the EMODB, SAVEE, and Aibo datasets respectively. The results show that the proposed model outperforms the recent state-of-the-art studies with a cheaper computational cost.

Keywords: Reservoir Computing, Truncated Normal Distribution, Population-Based Training, Speech Emotion Recognition, Recurrent Neural Network

1.0 INTRODUCTION

There are many ways of communication among humans, but speech is one of the fastest, natural, and effective ways to communication. Speech with emotion makes communication more natural between humans and helps them understand each other. Recognizing emotions from speech is a great demand to improve the user experience between humans and machines. However, emotion recognition from the speech is a big challenging task in machine learning and still a challenging research topic to many vital applications. Speech Emotion Recognition (SER) system can be used in a variety of applications, such as in-car board for driver safety, call center services for customer satisfaction, and children in care to detect their emotional status.

For any SER system, selecting the optimal emotion features from speech signals, a robust and cheap computational model is required to detect emotions in real-time. In the past decade, some common classification models have been adopted in SER systems, such as Support Vector Machine (SVM) [1], k-Nearest Neighbor (k-NN) [2], Gaussian Mixture Model (GMM) [2], and Hidden Markov Model (HMM) [3]. Additionally, the deep learning approach is also being paid extensive attention in the SER area to achieve better results compared to the traditional models, despite the complexity and expensive computations. Therefore, most of the traditional models for detecting emotions from the speech were using handcrafted global features that present each sample as one vector [4]. Some recent deep learning models such as Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) are used for feeding temporal frame-based features from speech data. However, researchers in [5] concluded that deep learning models are not always the best option especially in real-world applications due to the noticed big gap between the theory and practical applications.

Some studies prefer to use the Echo State Network (ESN) approach as a compatible model with multivariate time series features and overcome the computation complexity. The simplicity, the untrained nature, and the sparsely connected neurons in the reservoir layer make the ESN a good choice to deal with the sparse nature of emotion in speech [6]. Although the ESN has advantages, some problems still need to be solved, such as the instability with initializing fixed weights randomly and some hyperparameters which have a big impact on the ESN performance. The input weight and the reservoir weight that are generated randomly and fixed, are initialized in the reservoir layer

which has a vital impact on the ESN [7]. Additionally, tuning the ESN hyperparameters is a challenge that has a significant impact on the model performance. Some researchers choose the fixed optimal values based on experience [8] and others adopt various optimization methods [9].

In this study, we propose a novel reservoir computing approach with truncated normal distribution [10] for initializing random connection weights for the input weight, in addition to optimizing the hyperparameters by Population-Based Training (PBT). Moreover, the handcrafted multivariate time series features are adopted as an input to the model which contains Mel-Frequency Cepstral Coefficients (MFCCs) and Gamma-Tone Cepstral Coefficients (GTCCs) features.

The main contributions of this work are 1) the use of input weight initializing method in the reservoir layer that is able to minimizes the chance of having non-trainable and repeated weights due to the use of tanh activation function. 2) adopt an optimization method to train and optimize a sequence of networks parallelly with less computational cost.

The rest of this work is sorted into the following sections: literature about the existing methods of SER and the different weight initialization are presented in Section 2. The proposed reservoir computing model is presented and explained in Section 3. In Section 4 the datasets and experimental setup are given, and in Section 5 the experiment results of the proposed method are shown and the results with the state of the arts are discussed. Lastly, conclusion and future work is presented in Section 6.

2.0 LITERATURE REVIEW

Emotion recognition from speech is a difficult task in the speech research area due to the complexity of emotion, consequently, classification method and discriminative emotion features selection are a key role to build a robust SER system. In recent years, many studies focus on finding a proper way to select the optimal emotion feature from speech and some researchers propose a different model design to take advantage to get the state-of-the-art recognition accuracy.

Deep learning approaches are used in recent works for SER systems and robust features learning. Researchers in [11] used Convolutional Neural Network (CNN) to learn high-level distinguishing features from 3-D log Mel-spectrogram. However, the classical way to extract the handcrafted features is still used by many researchers, Patni et al. [12] used energy, pitch, chromagram, MFCC, and Gammatone frequency spectrum coefficients (GFCC) handcrafted features and 2D CNN as a classifier.

Typical approaches for supporting temporal data are LSTM and ESN. Authors in [13] used Bi-LSTM deep learning with two heterogeneous branches where the left side has two dense layers and the right side has a convolution layer. Additionally, the handcrafted time-series features with 512 frames are used in [14] for feeding CNN and Bi-LSTM model. However, the time complexity of LSTM has been reported to to be more time consuming compared to the ESN (see Table 1). Limited studies have adopted ESN for detecting emotion from speech, Scherer et al. [15] explores the use of ESN for real-time emotion recognition from speech signals. The direct use of time series features from speech signals and avoiding a need for features extraction with the ESN model are proposed by [16].

Table 1: The comparison of the training time between LSTM and ESN

Method	**LSTM (sec.)**	**ESN (sec.)**
Gallicchio et al. [17]	26175	677
Jirak et al. [18]	88.9	2.6
Variengien & Hinaut. [19]	410	47.1

Unlike the LSTM and RNN models, the weights in ESN are fixed and remain unchanged once initialized randomly. For this reason, both input weight and reservoir weight are unlikely to be ideal when they are generated inside the reservoir layer and later trained in the readout part [20]. The uniform distribution in the interval [-1,1] was adopted to initialize the input and the reservoir weights in the traditional ESN [21]. To evaluate the ESN performance with a different method, the singular value decomposition (SVD) is used in [22] to generate weight matrix, and the ESN-DE proposed in [23], where differential evolution algorithm is used to improve the randomly initialized weights

inside reservoir layer for the possible improvement. Additionally, authors in [24] adopted Cauchy inequality to conduct an optimal state of initial weights.

Regarding the selection of an optimal value for ESN hyperparameters, which significantly have a vital role in the model performance, some works were using optimization methods [9] and others were fixing the values based on experience [8]. Tuning hyperparameters by Grasshopper Optimization Algorithm (GOA) approach has improved the ESN performance in [25], and [26] adopted the Bayesian optimization approach to optimize and select the right value for ESN parameters. However, the Population-Based Training method for tuning ESN hyperparameters can be another option to select the optimal value quicker by training a sequence of networks at the same time with no computational overhead.

3.0 PROPOSED SPEECH EMOTION RECOGNITION MODEL

In this section, the ESN model is presented and explained in detail including the extraction of the handcrafted frame-based features, reservoir computing weight initialization, dimension reduction, and the classifier used for performance evaluation. Fig. 1 shows general architecture of the SER model design which contains the input layer for extracting features, reservoir layer for general reservoir state, and readout layer for applying dimension reduction, multivariate time series representations, and model classifier.

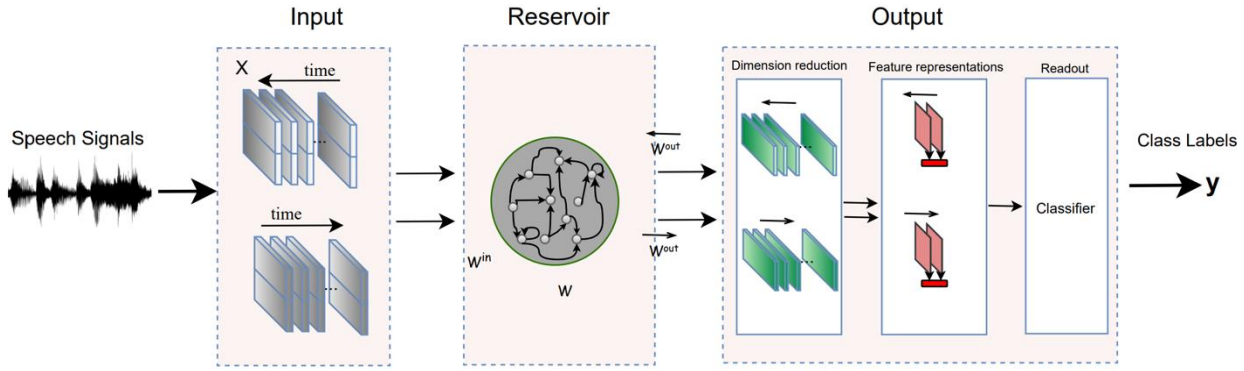

Fig. 1: The general design of the proposed model

3.1 Input Layer

The input layer contains bidirectional form of the extracted features with both backward and forward directions to feed the reservoir layer later.

The most relevant emotion features and the way for extracting them are the important stage that impacts the complete model performance and is still a challenge in SER systems. In this study, two sets of handcrafted frame-based features are extracted, which include 13 MFCCs features and 13 GTCC features. MFCC features are the commonly used features in many SER applications because of the capability of having significant emotional information and exhibiting remarkable results [27]. On the other hand, the GTCC features are better at handling noisy conditions than MFCC. Consequently, an overall of 26 features (13 GTCC and 13 MFCC) are used to feed the reservoir layer.

The ESN model suffers from the randomness and the instability since the weights are randomly initialized only once and fixed in the reservoir layer [6]. The bidirectionality approach helps in overcoming this problem which feeds the data to the reservoir layer in both backward and forward directions to capture additional information independently of the input data and increase the memorization ability [28].

3.2 Echo State Network

Echo State Networks (ESNs) as a simple type and powerful network structure of RNNs was first proposed by [21] to learn nonlinear systems. As a special type of RNN family, the modeling and learning procedures are different in ESN compared to the traditional RNN models. The ESN model contains three main layers, input layer, reservoir layer, and output layer as shown in Fig. 2. The main concept of ESN is that input time series data are fed into a fixed

nonlinear system called reservoir which randomly assigns weights without training. Only the output weights, the readout, are then trained by a ridge regression classifier.

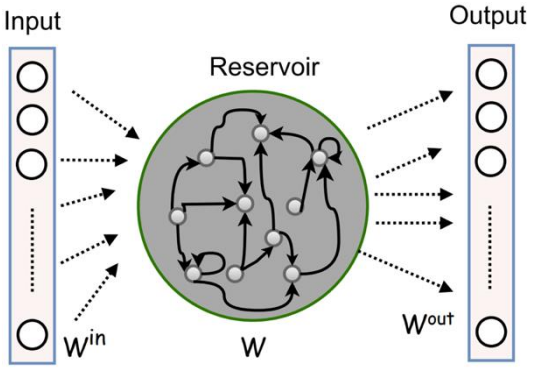

Fig. 2: Basic structure of the ESN model

As shown by Fig. 2 the input layer contains multivariate time series data which will be multiplied by the input weight (W^{in}) the output will be processed inside the reservoir layer based on the nodes and their consequent sparse weights W. The frame-based data from the input layer has L-dimensional size for each time step t where $t = 1,2,3,\dots,T$ and T is the total number of frames (time steps) for each sample. Therefore, the $x(t) \in R^L$ where $x(t)$ is a feature vector row from the t time step and $X = [x(1), x(2), \dots, x(T)]^T$. The input weight $W^{in} \in R^{N \times L}$ where N is an internal unit size, and the reservoir weight $W \in R^{N \times N}$ are randomly initialized once and remained fixed. The reservoir state output $W^{out} \in R^N$ generated from the reservoir layer over time. In our case, bidirectional data inputs to the reservoir, the two $\overrightarrow{W}^{out}$ and $\overleftarrow{W}^{out}$ are generated based on the following equations:

$$\overrightarrow{rs}(t) = f(W^{in} * \vec{x}(t) + W * \overrightarrow{rs}(t-1))$$
$$\overleftarrow{rs}(t) = f\left(W^{in} * \overleftarrow{x}(t) + W * \overleftarrow{rs}(t-1)\right) \quad (1)$$

where the $\overrightarrow{rs}(.)$ and $\overleftarrow{rs}(.)$ are the output reservoir state in time t for forward and backward input data. To comput the $\overrightarrow{rs}(.)$ and $\overleftarrow{rs}(.)$, the generated input weight is multiplied by the current input $\vec{x}(t)$ and $\overleftarrow{x}(t)$, respectively, and added to the result of the reservoir weight multiplied to the their previous values $\overrightarrow{rs}(t-1)$ and $\overleftarrow{rs}(t-1)$, respectively. The $f(.)$ is a non-linear activation hyperbolic tangent function. The $tanh(.)$ is a common activation function that is used on many ESN models, and it has been adopted in this study too. The overall output $\overrightarrow{W}^{out}$ and $\overleftarrow{W}^{out}$ are the output states where $\overrightarrow{W}^{out} = [\overrightarrow{rs}(1), \overrightarrow{rs}(2), \dots, \overrightarrow{rs}(T)]^T$ and $\overleftarrow{W}^{out} = [\overleftarrow{rs}(1), \overleftarrow{rs}(2), \dots, \overleftarrow{rs}(T)]^T$.

3.2.1 Weight Initialization

Nontrainable weights in the reservoir layer have significant advantages because it makes it computationally cheap, which nominates ESN as an ideal suggestion for many real-time applications. Additionally, the applied approach for initializing weights is one of the important aspects to increase the learning speed and the performance of any neural network model. The traditional ESN models are mostly adopted uniform distribution in the interval [-1,1] to initialize the input weight and the reservoir weight.

In this work, to initialize the input weight we proposed the use of truncated normal distribution as one of the most important distribution [29] to initialize the input weight. Truncated normal distribution was presented more than a century ago [30], however, it has not been used widely in academia until recent years because of the complexity of truncated normal distribution numeric characteristics [31]. A truncated normal distribution was proposed by [10] to examine trotting horse speeds in order to exclude records that were less than a definite known time. The truncated normal distribution helps for adjusting parameters to fit the data by selecting a standard probability distribution.

The truncated normal distribution function is presented in the equation below:

$$f(x;\mu,\sigma,a,b) = \frac{\frac{1}{\sigma}\,\phi\left(\frac{x-\mu}{\sigma}\right)}{\varphi\left(\frac{b-\mu}{\sigma}\right) - \varphi\left(\frac{a-\mu}{\sigma}\right)} \quad (2)$$

where μ is the mean value, σ is the standard deviation, and (a, b) are upper and lower limits of the truncated normal distribution. The $\phi(.)$ is a probability density function and

$$\phi(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

and $\varphi(.)$ is a cumulative distribution function and

$$\varphi(x) = \frac{1}{2}\left(1 + ERF\left(\frac{x}{\sqrt{2}}\right)\right)$$

In our study, the value of the mean is adopted to be zero while the standard deviation is fixed as one.

One of the reasons for using the truncated normal distribution for initializing the fixed weights is that the use of the tanh activation function will flatten the outputs of values far from the mean which makes it not useful to be trained in the next readout. Truncated normal distribution minimizes the chance to have non-trainable weights.

3.2.2 Reservoir Hyperparameters

The reservoir layer has several parameters that have an important role in the model performance such as (1) the size of internal unit size N where the output weight is based on the size of this parameter, (2) the spectral radius of the reservoir weight W which helps the stability of the model [32] and affects the short-term memory as values close to one, provide a longer memory and the value should be between 0 to 1, (3) the input scaling of the input weight where the small value makes the reservoir act around the zero point of their hyperbolic tangent function, (4) the connectivity percentage of non-zero connection of the reservoir weight W which helps remove the connection between neurons in the reservoir, (5) the leaking rate of the reservoir neurons which control the time-scale variance between the input weight and reservoir weight [33], (6) the number of drops remove a number of frames at the beginning which has little information, and (7) the level of noise as a Gaussian noise added in the state update function in Eq. (1) for regularization reasons [34]. To find the optimal values for all the reservoir parameters, a Population-based training method has been adopted.

3.3 Output Layer

Three stages are applied in this layer, firstly, the dimensionality reduction method is applied to reduce the high-dimensional from output weight, secondly, the reservoir model space is adopted for feature representation and classification step to map the reservoir model space output into the emotion class labels y.

3.3.1 Dimensionality Reduction

The high-dimensional sparse reservoir output weights $\overrightarrow{W}^{out}$ and $\overleftarrow{W}^{out}$ are projected to adopt Sparse Random Project (SRP) as an unsupervised dimensional reduction approach. The sparse output weights from the reservoir layer lead to overfitting and high computational cost, therefore, the SRP method helps to transform the sparse data into a more compact representation and reduce the dimensionality of the reservoir output weights. The SPR is an unsupervised approach and has a low complexity while it deletes redundancies with minimal loss of information and it has nontrainable nature. Additionally, SPR is considered as a powerful approach for dimension reduction in machine learning areas [35]. With a focus on developing the efficiency of the projection phase, authors in [36] proposed SRP matrix which is defined by:

$$r_{n.m} = \sqrt{s} \begin{cases} +1 \ with\ probability & \frac{1}{2s} \\ 0 \ with\ probability & 1 - \frac{1}{s} \\ -1 \ with\ probability & \frac{1}{2s} \end{cases} \quad (3)$$

where s can be 1 or 3, but in our work $s = \frac{1}{\sqrt{D}}$ and D is the size of the original feature from the reservoir output weight. For instance, if $s = 3$ means one-third of the data are sampled at random and two-thirds of the whole data are removed. Therefore, the low-dimensional multivariate time series output from SRP has a significant impact on the next stage of the proposed model which is the feature representations from the reservoir model space.

3.3.2 Feature Representations and Readout

The multivariate time series classification aims to assign a label to sequence data. Some studies tried to find a suitable representation for time series data and fed it to classification models. Other studies have adopted a traditional distance metric for sequences, Dynamic Time Warping (DTW), from the raw time series data and used a suitable learning model. To deal with the multivariate time series output weights from the dimension reduction stage, we adopted the reservoir model space as proposed by [8]. Reservoir model space is trained for each time series as one-step-ahead prediction before readout to produce a timeless representation of the time series data by accounting for all the weight output from SRP. The backward and forward data from SRP are fed to the reservoir model space separately. For each time series, $h(t)$ is trained to predict the next data input $h(t+1)$ from the $u(t)$ as a current data from SRP output as shown in the equation below:

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

and $R_h = [\vec{R}_h; \overleftarrow{R}_h]$ where $\vec{R}_h = \left[vec\left(\vec{U}_h\right); \vec{u}_h\right] \in \mathbb{R}^{M(M+1)}$, and $\overleftarrow{R}_h = \left[vec\left(\overleftarrow{U}_h\right); \overleftarrow{u}_h\right] \in \mathbb{R}^{M(M+1)}$, M is the size of the SRP output feature size, and $\vec{R}_h$ and $\overleftarrow{R}_h$ can be learned by minimizing a ridge regression loss function. The final feature representations from both directions are concatenated as R_h and fed to the classifier which shows in the below equation:

$$y = g(R_h) \quad (5)$$

The linear readout is the last step to perform the classification based on R_h which maps into the emotion class labels y.

3.4 Hyperparameter optimization

Selecting a significant value for ESN hyperparameters is still a challenge that has a huge impact on the model performance. In this study, we adopted Population Based Training (PBT) [37] algorithm in the Sherpa library [38] for optimizing hyperparameters in the proposed model. The PBT method is a fusion of the two widely used algorithms for hyperparameter optimization which are random search and hand-tuning. It trains and optimizes a sequence of networks parallelly and with no computational cost as quickly as traditional methods. The PBT has the capability to an automatic finding of hyperparameter schedules, which leads to better performance and stable training. In this work, the PBT has been used for optimizing ten hyperparameters, seven from the reservoir layer, the dimension size of SRP, regularization parameter of the ridge regression in both reservoir model space and ridge regression classifier as shown in Table 2.

Table 2: The list of the parameters that have been optimized in the ESN model

Part	**Parameters**
Reservoir	Internal unit size, spectral radius, connectivity, input scaling, leak, dropout, and noise level.
Sparse Random Projection	The dimensionality of the target projection space.
Reservoir model space	Regularization parameter of the ridge regression in feature representation.
Readout	Regularization parameter of the ridge regression classifier (readout).

4.0 DATASETS AND EXPERIMENTAL SETUP

In this section, the three most common speech emotion datasets are presented which they used to evaluate the performance of our proposed model and the experimental setup of implementing the proposed ESN model.

4.1 Datasets

The proposed ESN model is evaluated on two acted datasets EMODB [39] and SAVEE [40] and one non-acted Fau Aibo Emotion Corpus [41].

Berlin Database of Emotional Speech (EMODB) is the most widely used speech emotion dataset for validating SER models. It is an acted German speech emotion dataset that involves 7 emotion states 1) anger; 2) boredom; 3) anxiety; 4) happiness; 5) sadness; 6) disgust; and 7) neutral. Five professional speakers (5 females and 5 males) from the Institute of Communication Science, at Technical University are recorded with a total of 535 utterances. All 7 emotions and 10 speaker samples are involved to evaluate our proposed model.

Surrey Audio-Visual Expressed Emotion (SAVEE) is a multimodal emotion dataset that was recorded by four English native postgraduate students and researchers from the University of Surrey. SAVEE dataset involves 7 emotion states such as anger, disgust, fear, happiness, sadness, surprise, and they added a neutral class to provide 7 emotion categories, and each speaker recorded 120 utterances.

Fau Aibo Emotion Corpus (Aibo) is a non-acted speech emotion dataset that contains 18216 spontaneous and emotional German speech samples. A total of 51 samples are recorded from children at ‘Ohm’ and ‘Mont’ schools where they interacted with Sony’s pet robot Aibo. First, the dataset had 10 labels and later they mapped it to five emotion classes such as anger, emphatic, neutral, positive, and rest. Our work, followed by the protocol of the interspeech09 challenge [42], we adopted the ‘Ohm’ samples (9959) as a training set and ‘Mont’ samples (8257) as a testing set.

4.2 Experimental setup

The speaker-independent Leave-One-Speaker-Out (LOSO) has been adopted for EMODB and SAVEE datasets and speaker-independent based on the protocol of the interspeech09 for the Aibo dataset. The number of samples per emotion class in the Aibo dataset is extremely unbalanced. To overcome this problem, the random under-sampling [43] approach is applied on the majority classes by randomly picking the fixed number of samples. Meanwhile, the test set classes are imbalanced, the performance of all adopted datasets are reported as unweighted average recall (UAR).

The Aibo dataset is trained on a PC with 64GB RAM and the other two datasets are trained on 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, and all experiments are carried out on the CPU.

5.0 RESULTS AND DISCUSSION

In this section, we present the results of this study in terms of precision, recall, and F1 score, in addition to the model weighted and unweighted accuracy for each emotion class. All results in this work are shown as unweighted accuracy which is more realistic for accurate measurement especially when the test set of datasets are imbalanced.

Additionally, the match and mismatch between predicted and truth labels are presented as a confusion matrix for all experiments. Finally, the results are compared with the recent state of the arts.

5.1 Proposed model results

For the EMODB dataset, the LOSO approach is applied by setting 9 speakers as a train set and one speaker out as a test set and repeating the same procedure to assure the contribution of all speakers in the test set. Table 3 shows the percentage performance for the EMODB dataset for each emotion class.

Table 3: The performance of the proposed model using the EMODB dataset.

Emotion	Precision	Recall	F1 Score
Anger	77.44	100	87.29
Boredom	92.50	91.36	91.93
Disgust	100	84.78	91.76
Fear	89.29	72.46	80.00
Happiness	87.23	57.75	69.49
Sadness	90.62	93.55	92.06
Neutral	87.06	93.67	90.24
Unweighted	89.16	84.80	86.11
Weighted	87.44	86.54	86.06

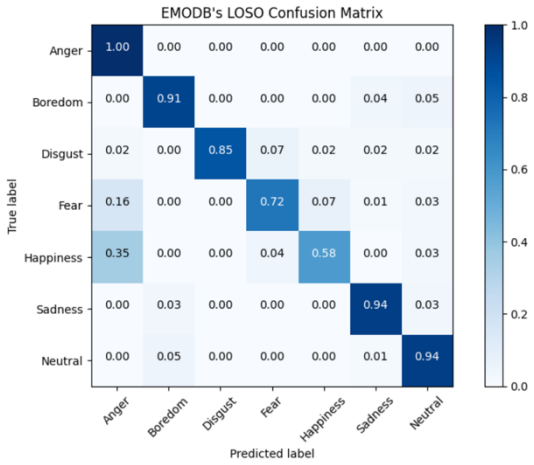

Fig. 3: The confusion matrix of proposed model for EMODB dataset.

The anger emotion class has the highest accuracy where all anger samples were recognized correctly, however, in the happiness class almost half the samples were recognized correctly. Fig. 3 presents the EMODB dataset confusion matrix between true labels and predicted labels. The 35% of the happiness class were recognized as anger which may be a sign that the happiness expressed in high arousal shares similarities with the anger class.

The classification result for each emotion class for the SAVEE dataset is shown in detail in Table 4, where the result of each class are shown in terms of F1 Score, recall, and precision, in addition to weighted and unweighted classification accuracy. The gap between weighted and unweighted accuracy shows that the test set from the SAVEE dataset is unbalanced where unweighted accuracy is 4.05% less than the weighted accuracy. Additionally, Fig. 4 presents the confusion matrix for the SAVEE dataset, which clearly shows that the disgust emotion samples are mostly recognized as neutral and 40% of the neutral emotion class is recognized as sadness. The neutral emotion recorded the highest accuracy, which may contain the doubled samples compared to other classes.

Table 4: The performance of the proposed model using the SAVEE dataset.

Emotion	Precision	Recall	F1 Score
Anger	76.81	88.33	82.17
Disgust	71.88	38.33	50.00
Fear	78.05	53.33	63.37
Happiness	72.41	70.00	71.19
Neutral	66.29	98.33	79.19
Sadness	63.16	40.00	48.98
Surprise	68.75	73.33	70.97
Unweighted	71.05	65.95	66.55
Weighted	70.46	70.00	68.13

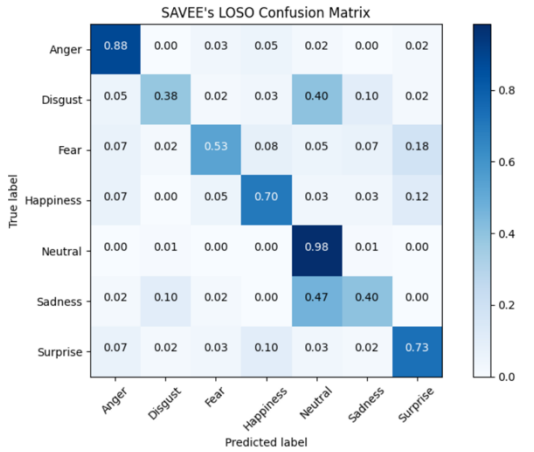

Fig. 4: The confusion matrix of proposed model for SAVEE dataset.

Table 5 lists the detailed classification results based on precision, recall, and F1 score for the Aibo dataset. We notice that the gap between the weighted and unweighted in Aibo is almost doubled compared to the SAVEE dataset due to the high imbalance of data in the Aibo dataset. The low accuracy in Aibo dataset compared to EMODB and SAVEE reflects that recognizing emotion from speech in a spontaneous dataset is still a big challenge. The confusion matrix of the Aibo dataset shows that 33% of the rest class are recognized as positive and 24% as anger, where only 17% of samples were recognized correctly, which may be due to having different labels which are grouped under the same class (see Fig. 5).

Table 5: The performance of the proposed model using the Aibo dataset.

Emotion	Precision	Recall	F1 Score
Anger	19.28	67.59	30.00
Emphatic	38.08	48.41	42.63
Neutral	82.05	28.57	42.38
Positive	08.99	68.37	15.88
Rest	13.48	17.03	15.05
Unweighted	32.38	45.99	29.19
Weighted	62.94	35.35	39.01

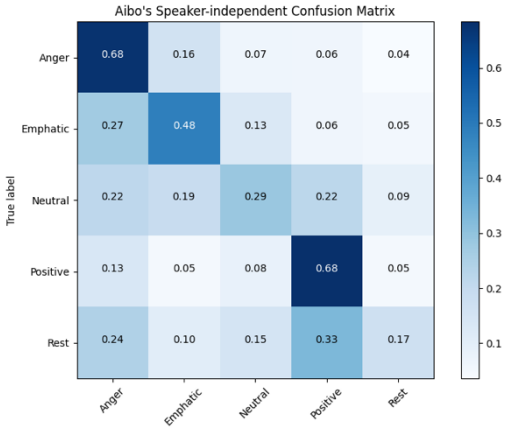

Fig. 5: The confusion matrix of proposed model for Aibo dataset.

5.2 Comparison with the State-of-the-Art

The use of truncated normal distribution to initialize the input weight, population-based training for optimizing the whole 10 hyperparameters, adopting SRP for reducing the high dimensional output weight, and a bidirectional approach for capturing more information in our proposed ESN model has helped improve the classification accuracy.

The overall results are shown as an unweighted accuracy (UA) which is more realistic than weighted accuracy when the datasets are imbalanced. Our proposed model for the EMODB dataset achieved 84.8%, SAVEE dataset 65.95%, and Aibo 45.99% UA which outperforms the recent state-of-the-art models that are applying the LOSO approach and unweighted accuracy.

Table 6: The Comparison of unweighted accuracies (UA%) of the proposed model for EMODB, SAVEE, and FAU Aibo datasets with recent works

Dataset	Method	UA%
EMODB	[13] Heterogeneous Parallel Convolution Bi-LSTM	84.65
	[11] Parallelized convolutional recurrent neural network	84.53
	[44] Random Deep Belief Networks	82.32
	[45] 3-dimensional attention-based convolutional recurrent neural networks	82.82
	[46] SVM-RBF	71.02
	[47] GEBF	76.81
	Proposed Model	**84.80**
SAVEE	[13] Heterogeneous Parallel Convolution Bi-LSTM	56.50
	[47] GEBF	55.00
	[44] Random Deep Belief Networks	53.60
	Proposed Model	**65.95**
Aibo	[48] Deep learning eResNe	41.3
	[49] Bi-LSTM	45.4
	[50] SVM,NN,DNN	45.3

	[51] MRA+SVM	45.2
	Proposed Model	**45.99**

The comparison between our work with various new studies that have been conducted recently for classification UA speaker-independent experiments are shown in Table 6. Some studies that used deep learning obtained remarkable results, for example researchers in [11] [13] [44] [45] have used the EMODB dataset to evaluate the deep learning approach and they achieved distinguished results. However, the proposed ESN model is able to outperform these deep learning models by achieving 84.80% UA. Authors in [13] adopted the Heterogeneous Parallel Convolution Bi-LSTM model and applied speaker-independent for SAVEE dataset and they achieved 56.5% of UA, and Random Deep Belief Networks model [44] performed 53.60% UA for SAVEE, however, our method obtained 65.95% UA. Unlike EMODB and SAVEE, one can notice the big challenge to gain higher accuracy for the Aibo dataset. The highest UA (45.4%) was achieved by using Bidirectional LSTM with attention enhanced FCN [49], however, once again our proposed model outperforms the mentioned work by obtaining UA of 45.99%.

6.0 CONCLUSION AND FUTURE WORK

In this study, we proposed a novel ESN architecture for multivariate time series classification by adopting truncated normal distribution for generating random connection weights for the input weight and a population-based training approach for optimizing the model hyperparameters. Furthermore, the handcrafted frame-based features are adopted as input to the model which contains 13 MFCC features and 13 GTCC features. The proposed model used the bidirectional reservoir input to increase the memorization capability, and SRP was applied for dimensional reduction as a simple, unsupervised, and low complexity approach. Because of the nontrainable nature of ESN, adopting a small size of features, applying a fast optimizer, and using the truncated normal distribution to minimize the chance of having non-trainable weights, our model is fast and more robust to achieve better performance. The proposed model is validated by adopting a speaker-independent approach and the most widely used speech emotion datasets, such as EMODB and SAVEE as acted datasets and Aibo as a non-acted dataset. For future work, the parameters of a truncated normal distribution which may affect the model performance can be optimized using the same adopted optimizer. Additionally, the reservoir model space for feature representations which suffers from producing high dimension representation, need to be more investigated to be replaced with model that can produce more convenient representation for the classification models.

7.0 ACKNOWLEDGMENT

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

REFERENCES

[1] Y. Pan, P. Shen and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home,* vol. 6, no. 2, pp. 101-108, 2012.

[2] R. B. Lanjewar, S. Mathurkar and N. Patel, "Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques," *Procedia computer science*, pp. 50-57, 2015.

[3] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[4] A. Al-Talabani, "Automatic Speech Emotion Recognition- Feature space Dimensionality and Classification Challenges," University of Buckingham, Buckingham, 2015.

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

[6] Q. Wu, E. Fokoué and D. Kudithipudi, "On the Statistical Challenges of Echo State Networks and Some Potential Remedies.," *CoRR*, 2018.

[7] H. Wang, C. Ni and X.-f. Yan, "Optimizing the echo state network based on mutual information for modeling fed-batch bioprocesses," *Neurocomputing*, vol. 225, pp. 111-118, 2017.

[8] F. M. Bianchi, S. Scardapane, S. Løkse and R. Jenssen, "Reservoir computing approaches for representation and classification of multivariate time series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2169-2179, 2020.

[9] H. Hu, L. Wang and R. Tao, "Wind speed forecasting based on variational mode decomposition and improved echo state network," *Renewable Energy*, vol. 164, pp. 729-751, 2021.

[10] F. Galton, "An examination into the registered speeds of American trotting horses, with remarks on their value as hereditary data," *Proceedings of the Royal Society of London*, vol. 62, no. 1, p. 310–315, 1898.

[11] J. Pengxu, 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.

[12] H. Patni, A. Jagtap, V. Bhoyar and A. Gupta, "Speech Emotion Recognition using MFCC, GFCC, Chromagram and RMSE features," in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2021.

[13] H. Zhang, H. Huang and H. Han, "A Novel Heterogeneous Parallel Convolution Bi-LSTM for Speech Emotion Recognition," *Applied Sciences*, vol. 11, no. 21, 2021.

[14] C. Fu, T. Dissanayake, K. Hosoda, T. Maekawa and H. Ishiguro, "Similarity of Speech Emotion in Different Languages Revealed by a Neural Network with Attention," in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, San Diego, CA, USA, 2020.

[15] S. Scherer, M. Oubbati, F. Schwenker and G. Palm, "Real-Time Emotion Recognition Using Echo State Networks," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, Berlin, Heidelberg, 2008.

[16] C. Gallicchio and A. Micheli, "A Preliminary Application of Echo State Networks to Emotion Recognition," in *Fourth International Workshop EVALITA 2014*, Pisa, Italy, 2014.

[17] C. Gallicchio, A. Micheli and L. Pedrelli, "Comparison between DeepESNs and gated RNNs on multivariate time-series prediction," in *ESANN 2019 - Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges (Belgium), 2019.

[18] D. Jirak, S. Tietz, H. Ali and S. Wermter, "Echo State Networks and Long Short-Term Memory for Continuous Gesture Recognition: a Comparative Study," *Cognitive Computation*, 2020.

[19] A. Variengien and X. Hinaut, "A journey in ESN and LSTM visualisations on a language task," *CoRR*, 2020.

[20] J. Liu, T. Sun, Y. Luo, S. Yang, Y. Cao and J. Zhai, "An echo state network architecture based on quantum logic gate and its optimization," *Neurocomputing*, vol. 371, pp. 100-107, 2020.

[21] H. JAEGER and H. HAAS, "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication," *Science*, vol. 304, no. 5667, pp. 78-80, 2004.

[22] J. Qiao, F. Li, H. Han and W. Li, "Growing Echo-State Network With Multiple Subreservoirs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 2, pp. 391 - 404, 2017.

[23] L. Wang, H. Hu, X.-Y. Ai and H. Liu, "Effective electricity energy consumption forecasting using echo state network improved by differential evolution algorithm," *Energy*, vol. 153, pp. 801-815, 2018.

[24] J. QIAO, L. WANG, C. YANG and K. GU, "Adaptive Levenberg-Marquardt Algorithm Based Echo State Network for Chaotic Time Series Prediction," *IEEE Access*, vol. 6, pp. 10720 - 10732, 2018.

[25] 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, 2019.

[26] J. R. Maat, N. Gianniotis and P. Proto, "Efficient Optimization of Echo State Networks for Time Series Datasets," in *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, 2018.

[27] N. Sato and Y. Obuchi, "Emotion Recognition using Mel-Frequency Cepstral Coefficients," *Journal of Natural Language Processing*, vol. 14, no. 4, pp. 83-96, 2007.

[28] A. Rodan, A. F. Sheta and H. Faris, "Bidirectional reservoir networks trained using SVM+ privileged information for manufacturing process modeling," *Soft Computing*, vol. 21, p. 6811–6824, 2017.

[29] J. Pender, "The truncated normal distribution: Applications to queues with impatient customers," *Operations Research Letters*, vol. 43, pp. 40-45, 2015.

[30] J. Cha and B. R. Cho, "Rethinking the truncated normal distribution," *International Journal of Experimental Design and Process Optimisation*, vol. 3, no. 4, pp. 327-363, 2013.

[31] S. Chen and W. Gui, "Estimation of Unknown Parameters of Truncated Normal Distribution under Adaptive Progressive Type II Censoring Scheme," *Mathematics*, vol. 9, no. 49, 2021.

[32] F. M. Bianchi, . L. Livi and C. Alippi, "Investigating echo state networks dynamics by means of recurrence analysis," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 2, p. 427–439, 2018.

[33] M. Dale, S. O'Keefe, A. Sebald, S. Stepney and M. A. Trefzer, "Reservoir computing quality: connectivity and topology," *Natural Computing*, vol. 20, p. 205–216, 2021.

[34] H. Jaeger, "The “echo state” approach to analysing and training recurrent neural networks-with an erratum note," GMD - German National Research Institute for Computer Science, 2001.

[35] C. Shi and W. Lu, "A Sparse Random Projection-based Test for Overall Qualitative Treatment Effects," *Journal of the American Statistical Association,* vol. 115, no. 531, pp. 1201-1213, 2020.

[36] P. Li, T. J. Hastie and K. W. Church, "Very sparse random projections," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[37] M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi and O. Vinyals, "Population based training of neural networks," CoRR, 2017.

[38] L. Hertel, J. Collado, P. Sadowski, J. Ott and P. Baldi, "Sherpa: Robust hyperparameter optimization for machine learning," *SoftwareX,* vol. 12, 2020.

[39] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, "A Database of German Emotional Speech," in *INTERSPEECH*, Lisbon, Portugal, 2005.

[40] S. Haq and P. J. Jackson, "Multimodal emotion recognition," in *Machine Audition: Principles, Algorithms and Systems*, IGI Global, 2010, pp. 398 - 423.

[41] S. Steidl, A. Batliner, B. Schuller and D. Seppi, "The hinterland of emotions: facing the open-microphone challenge," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009.

[42] B. Schuller, S. Steidl and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Com- munication Association*, 2009.

[43] G. Lemaˆıtre, F. Nogueira and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research,* vol. 8, no. 17, p. 1–5, 2017.

[44] G. Wen, H. Li, J. Huang, D. Li and E. Xun, "Random deep belief networks for recognizing emotions from speech signals," *Computational Intelligence and Neuroscience,* vol. 2017, 2017.

[45] M. Chen, X. He, J. Yang and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters,* vol. 25, no. 10, p. 1440–1444, 2018.

[46] 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," *Information Sciences,* vol. 563, pp. 309-325, 2021.

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

[48] A. Triantafyllopoulos, S. Liu and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China, 2021.

[49] 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.

[50] P.-Y. Shih, C.-P. Chen and H.-M. Wang, "Speech emotion recognition with skew-robust neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017.

[51] "Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, p. 802–815, 2019.