Research Article

Journal of Neuro and Spine

Speech Emotion Recognition Using ANFIS and PSO-optimization With Word2Vec

Vahid Rezaie[1*], Amir Parnianifard[2], Demostenes Zegarra Rodriguez[3], Shahid Mumtaz[4], Lunchakorn Wuttisittikulkij[2]

[1]Department of Industrial Engineering, Yazd University, Yazd, Iran.*

[2]Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand.

[3]Department of Computer Sciences, Federal University of Lavras, Lavras, Minas Gerais, Brazil.

[4]Instituto de Telecomunicações Campus Universitario de Santiago P-3810-193 AVEIRO – PORTUGAL.

[*]Corresponding Author

Vahid Rezaei, Department of Industrial Engineering, Yazd University, Yazd, Iran.

Submitted: 10 Feb 2023; **Accepted:** 14 Feb 2023; **Published:** 24 Feb 2023

Citation: Rezaie, V., Parnianifard, A., Rodriguez, D. Z., Mumtaz, S., Wuttisittikulkij, L. (2023). Speech Emotion Recognition Using ANFIS and PSO-optimization With Word2Vec. *J Neuro Spine.* 1(1), 41-56.

Abstract

Speech Emotion Recognition (SER) plays a vital role in human-computer interaction as an important branch of affective computing. Due to inconsistencies in the data and challenging signal extraction, in this paper, we propose a novel emotion recognition method based on the combination of Adaptive Neuro-Fuzzy Inference System (ANFIS) and Particle Swarm Optimization (PSO) with Word to Vector (Word2Vec) models. To begin, the inputs have been pre-processed, which comprise audio and text data. Second, the features were extracted using the Word2vec behind spectral and prosodic approaches. Finally, the features are selected using the Sequential Backward Floating Selection (SBFS) approach. In the end, the ANFIS-PSO model has been used to recognize speech emotion. A performance evaluation of the proposed algorithm is carried out on Sharif Emotional Speech Database (ShEMO). The experimental results show that the proposed algorithm has advantages in accuracy, reaching 0.873 and 0.752 in males and females, respectively, in comparison with the CNNs and SVM, MLP, RF models.

Keywords: Speech Emotion Recognition (SER), Adaptive Neuro-Fuzzy Inference System (ANFIS), Particle Swarm Optimization (PSO), Word2Vec

Introduction

Nowadays, speech signals are effectively and quickly responsible for facilitating communication between humans and machines. Therefore, understanding various characteristics of human speech necessarily requires machines to be intelligent [1]. Many research communities have recently instituted diverse human-computer interaction systems to secure automatic speech recognition (ASR). In essence, enabling humans to communicate with a computer is what drives ASR technology-forward [2]. ASR has many practical applications in speech emotion recognition (SER) in addition to medical diagnosis and call centres, diverse systems such as safe car driving, automatic translation, mobile telecommunication greatly benefit from speech emotion recognition (SER) as a manifestation of ASR [3]. However, these studies significantly suffer from time asymmetry, instability, a low signal-to-noise ratio and uncertain brain areas of specific reactions, resulting in unreliable results [4]. Accordingly, incorporating clustering methods and text mining into ASR has shown to be promising for removing the obstacles mentioned above.

Multi-class classification, which is based on machine learning, has transformed research in this field and has been influential in implementing ASR. This is due to the multi-class classification ability to utilize the relationship among class labels. Besides, in this method, the purpose of providing an automated classification result does not need to be explicitly unveiled [5]. Regardless of its popularity, multi-class classification is harmed by difficulties such as coping with missing data. To demonstrate, missing data affects both the training and classification stages. Also, the practicality of developed algorithms to minimize shortcomings has been hindered by unexpected issues [6].

Integrating the domain adaptation criteria and the Taylor-DBN model proposed by Haidas et al. have effectively addressed these issues [3]. Supported by the Taylor series, the Taylor- Deep Belief Network model updates the weights and the bias for the Deep Belief Network (DBN) training. Since the proposed model used the Multiple Kernel Weighted Mel Frequency Cepstral Coefficient (MKMFCC) scheme with the various kernel functions. It became evident that the model mentioned above achieved 97% accuracy in the Berlin database at 80% training. Another novel emotion recognition method established on the deep learning model has used Electroencephalography (EEG) 's data with Differential Entropy (DE). Each divided segment extracts DE to generate a feature cube, multiple Graph Convolutional Neural Networks (GCNNs) extract graph domain features from each feature cube. Additionally, Long Short-Term Memory (LSTM) cells store the change in the connection between two EEG channels over time and extract temporal data with an accuracy of up to 90.52% for subject-dependent studies.. Transfer learning (TL) and echo state network (ESN) have been applied by a novel classification model, developed by Zhou et al. [7], for addressing the issue of the nonlinearity of EEG samples. Additionally, collecting the high-quality EEG samples is yielded by the Average Frechet Distance (AFD), reaching an accuracy of 68.06% using the proposed model in the EEG signals database. Kwon et al. improved the classification performance of discrete emotions by timbre acoustic features [8]. Also, Sequential forward selection (SFS) finds the most relevant acoustic features among timbre features. Classifying emotions happen through a Support vector machine (SVM) and long short-term memory recurrent neural network (LSTM-RNN). The suggested model's accuracy was estimated in the IEMOCAP dataset at 65.05 percent.. Also, Li et al. Developed the Bi-directional Long-Short Term Memory with Directional Self Attention (BLSTM-DSA) model, which detected emotional statuses such as anger, happiness, sadness, and neutrality[9]. The robustness of the model's structure is thought to be enhanced by the forward and backward LSTM. The model achieved an average accuracy of 71% and 85% using the IEMOCAP and Berlin datasets, respectively. Yildirim et al. focused on feature selection in speech emotion recognition and used a non-dominated sorting genetic algorithm-II (NSGA-II) and Cuckoo Search for redacting the features [10]. Three different classification algorithms, including K-Nearest Neighbours (KNN), Tree-Bagger, and SVM, were used to compare the results. In the EMO-DB and IEMOCAP databases, the suggested model has 87.66 and 69.30 percent accuracy, respectively.

Generating multi-level features and iterative neighbourhood component analysis (INCA) for classifying the model was presented by Tuner et al. as a novel model with Tenable Q wavelet transform (TQWT), resulting in 87.43%, 90.09%, 84.79%, and 79.08% classification accuracies in RAVDESS, Emo-DB (Berlin), SAVEE, and EMOVO databases, respectively [11]. Abdel-Hamid et al. investigated the effects of anger, fear, happiness, and sadness on the linguistic and prosodic features across Arabic and English emotional speeches [12]. The study consists of bilingual linguistic, prosodic features, and non-linear SVM. The new model's is generic 58% and 46.60% accuracy were recorded for females and males, respectively. Irigarayan et al. developed a new model called Recurrent neural network [13]. In their model, combining multimodal SSL features and achieving state-of-the-art results for the task of multimodal emotion recognition were powered by a novel Transformers and Attention-based fusion mechanism. The results showed the proposed model obtained 67.92% and 43.77% of accuracy recorded in IEMOCAP and MELD databases, respectively. studying temporal modulation cues from auditory front-ends and then coming up with a joint deep learning model that combines 3D convolutions and attention-based sliding recurrent neural networks (ASRNNs) for emotion recognition were done by Peng et al. for exploiting the auditory and attention mechanism [14]. The findings reveal that the suggested model has a 62.60 percent accuracy and a 55.70 percent accuracy in the IEMOCAP and MSP-IMPROV databases, respectively. Obviously, there are both advantages and disadvantages associated with the methods, as mentioned earlier. For instance, the necessity for large resources, the time-consuming nature of building an emotion recognition model, cruciality of hyper-parameters such as the number of hidden nodes and layers for them, not responding when the dimensions of emotional features are too many, not being able to model non-linear relationship among emotional features, not working when emotional features are correlated, cruciality of the choice of kernel and its parameters and regularization of parameters for avoiding overfitting for them, and the necessity for a large amount of data are some of the disadvantages of models as mentioned above [15]. So, applying ANFIS methods with Particle Swarm Optimization (PSO) together seems the most reasonable idea to follow. ANFIS-PSO model has overcome those disadvantages and has the advantage of having both numerical and linguistic knowledge, which caused this model to be more transparent to the user and fewer memorization errors. Moreover, the adaptation capability, nonlinear ability, and rapid learning capacity of this model increase the importance of it to solve the above problems[16].

Speech and voice are analysed for detecting emotion; in the same way, text can be used for the same analysis. Detecting emotions from voice/speech and images have been extensively studied, leaving the texts nearly unexplored. This might be since, contrary to multimodal methods, texts might not characterize unusual cues to emotions. In addition, understanding emotions from short texts, emojis, and grammatical errors are time- and energy consuming. Not to mention the rapid and constant evolution of new words, originating from language dynamics. Furthermore, very little is known about detection techniques and the adequacy of emotion dictionaries, which make detecting a challenging task [17]. According to Ahmad et al. [18], the classification performance was effectively improved by cross-lingual word embeddings, and transfer learning which are used between two languages Convolution Neural Network (CNN) and Bi-Directional Long Short-Term Memory (Bi-LSTM) were taken into account for building up a new model, getting an F1-score of 0.53. Six emotions from sentences were extracted by the model suggested by Seal et al. [19]. Furthermore, Phrasal verbs and negative words have been analysed to finetune the results showing that the proposed model has 65% accuracy in the ISEAR database. Singh et al. [20] developed a two-stage text feature selection method. This method, which is based on POS tagger and Chi-square for semantic and statistical text feature extraction and SVM for classification, will select the words by taking semantics and statistics significance into account. As a result, classifying emotions were improved by 34.45% in the ISEAR database. Two models of WordNet were compared by Mozafari et al., suggesting ontological relation of words and vector similarity measure (VSM) [21]. Emotions will be detected from short text using feature vector and cosine similarity by VSM. In comparison to WordNet in the ISEAR database, the VSM approach performs better. Another Logistic Regression (LR)-based model was considered by Alotaibi, according to the proposed model, the pre-processing data has been trained in four classifications [22]. It has been found out that logistic regression outperformed the other methods. For SVM, KNN, and the XG-Boost techniques, accuracy, recall, and F- Score were 86 percent, 84 percent, and 85 percent, respectively. Huang et al. considered the pre-training language model BERT to propose another model that significantly depends on the sentence-level context-aware understanding [23]. The proposed model has been used to predict emotions in words, getting F1-scores of 0.815 and 0.885 for the Friends and Emotion Push datasets, respectively. Poignant et al. illustrated the effectiveness of their classification approach-based model on different datasets on deep neural networks, Bi-LSTM, CNN, and self-attention [24]. According to the proposed model, text-based emotions are detected by a word embedding method that compares the performance of Google Word, Global Vectors for Word Representation (GloVe), and FastText Embedding. The accuracy of the model in the ISEAR dataset is 90.6%. Classifying emotions from textual and emoji utterances into emotions like happy, sad, and angry were conducted by the Bi-LSTM used by Ma et al. [25]. Also, the Bi-LSTM surpassed baseline models with respect to the emotions of happiness and anger, but not sadness. It was concluded that extracting contextual information from texts is possible with Bi-LSTMs. Utilizing deep learning models for understanding both single and multilabel Arabic text categorization was conducted by Elbagir et al. [26]. The study came up with results from comparing different models on different datasets. The model's accuracy was found to be 91.18 percent in the SANAD database and 88.68 percent in the Nadia database, respectively. Thanks to a machine learning-based framework developed by Halim et al., dominant emotions hidden in email text are identifiable [27]. Techniques including Principal Component Analysis (PCA), Mutual Information (MI), and Information Gain (IG) were used for feature selection. Also, ANN, SVM, and RF were applied for classification. In Enron datasets, the model was shown to be 83 percent accurate. There have been shortcomings and constraints, such as ignoring the contextual meaning of words disregarded, high complexity, not impressive context and the semantic information extraction, and disregarding the relation between features [17]. Because of these reasons, we used the Word2vec method was prioritized due to the reasons mentioned above. Furthermore, some properties such as retaining the semantic meaning of different words in a document, the small size of the embedding vector, and transforming the unlabelled raw corpus into labelled data can be increased the advantage of using this model [28].

A combination of the ANFIS-PSO and Word2vec model is proposed and developed in this work, aiming to segment six emotions in the Persian database (ShEMO). Pre-processing the inputs happens prior to the classification of emotion in the SER system. Extracting features will be conducted with 11 methods like Word2vec at the next stage. Thirdly, selecting the features and classifying the emotions were done by the SBFS method and ANFIS-PSO.

Present paper consists of six sections. Second section reviews studies related to the issue of diagnosing speech emotion. Then, the third section discusses the ANFIS and Word2vec methods. The fourth section gives full detail of the proposed method, including its four distinct phases step-by-step. The results from the performance assessment and concluding the paper are discussed in the fifth and sixth sections, respectively.

Background

In this paper, the ANFIS and Word2Vec methods are employed in speech emotion recognition. These two algorithms are introduced in the following subsections to turn this paper into a more self-contained one.

Adaptive Neuro-Fuzzy Inference System (ANFIS)

Jang introduced ANFIS as a systematic hybrid input-output mapping technique in 1993 [29]. Selecting parameters and a membership function is provided by a license from ANFIS, which generates remarkable results concerning fuzzy inference methods [30]. This approach combines the feedforward neural network with the Takagi-Sugano fuzzy inference system. This technique determines the optimal distribution of membership functions in a fuzzy system, regardless of expert knowledge of the system [31]. This model considers two inputs (x, y) and one output, z, to make the determination process more reachable. Two different if-then fuzzy rules configured the matching rule for a first-order Sugano fuzzy model, expressed by (1). In this equation, entries evaluation is by linguistic A1 and B1 variables. A linear combination of the inputs with a constant term entitled r is considered to obtain the results of each rule [32].

Rule 1 = If x is A_1 and y is B_1 Then $Z_1 = p_1x + q_1y + r_1$.
(1) Rule 2 = If x is A_2 and y is B_2 Then $Z_2 = p_2x + q_2y + r_2$ (1)

ANFIS architecture constitutes five layers. An adaptive node is present in the first and fourth layers. In contrast, other layers have a fixed node. Figure 1 depicts the ANFIS structure.

Layer 1: inputs x y, correspond to this layer. The nodes of this layer are adaptive and responsible for calculating the degree of membership of the fuzzy set input $A\ x_i\ (\ )$. Although there are several ways to define membership function, they are all differentiable. For example, bell-shaped represents $A\ x_i\ (\ )$ and B_i (x) . (Eq. (2)) with maximum and minimum equal to 1 and 0, respectively. 1

$$A_i(x) = \frac{1}{1+((\frac{x-c_i}{a_i})^2)^{b_i}}$$

$$B_i(y) = \frac{1}{1+((\frac{y-c_i}{a_i})^2)^{b_i}} \quad (2)$$

Where x is the input and a_i, b_i, c_i is the set of parameters.

Layer 2: This layer is the rule layer. The nodes of this layer are non-adaptive. Also, the output of each node is defined as the product of its inputs.

$$W_i = A_i(x) \times B_i(y) \tag{3}$$

Layer 3: the normalized combination of degrees of membership of all linguistic statements are computed by nodes in this layer. This combination expresses how effectively a rule premise matches a specific input value. That's why it is called degree of rule fulfilment or rule's firing strength. Every node in this layer is labelled as N. the ratio of the I the rule's firing strength to the sum of all rule's firing strengths is calculated by the *I* the node.

$$\overline{W_i} = \frac{W_i}{(W_1 + W_2)} \tag{4}$$

Layer 4: Every node in this layer is a square node with a node function, where Wi is the output of layer 3, and $\{p_i, q_i, r_i\}$ is the parameters set.

$$\overline{W_i} f_i = \overline{w_i}(p_i x + q_i y + r_i) \tag{5}$$

Layer 5: The overall output, composed of a single node. The node of this layer is non-adaptive and its output is defined as the sum of the partial outputs of layer 4 [30], [31], [33], [34]:

$$\sum \overline{W_i} f_i \tag{6}$$

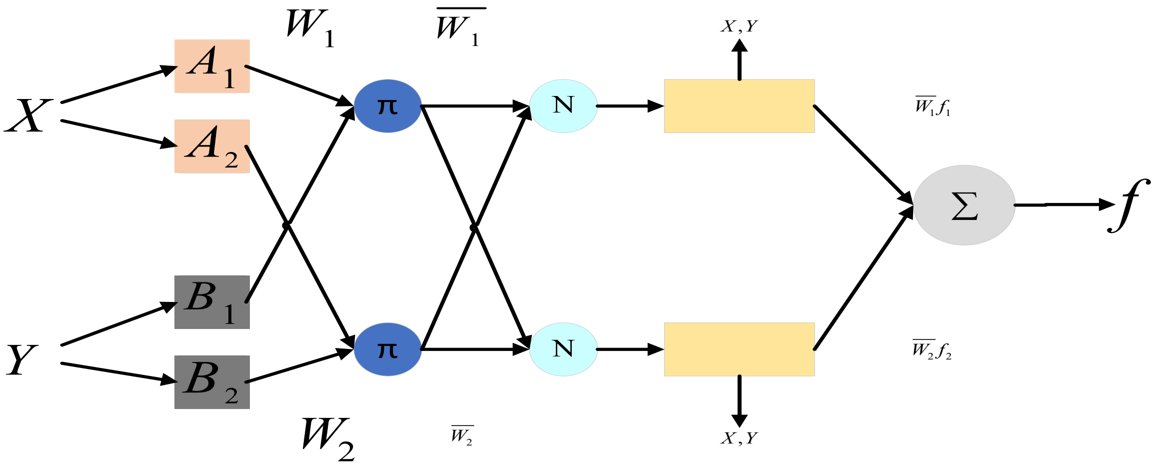

Figure 1: The five layers of ANFIS architecture.

Word to Vector (Word2vec)

Mapping primitive representations of words into high-dimensional numeric vectors in an embedding space with maintaining word distances are conducted by Word embeddings as general approaches. Studies have recently converged on word embeddings. As an example, Word2Vec is considered as one of the most significant text representation models, high correlation of the contexts in the natural language is assumed by Word2Vec. It shows how words can be vectorized according to the contexts. Then, the training corpus obtains word vectors for assessing the semantic similarities between words in natural language, weights of trained language models are generally more responsible for generating word vectors than directing training targets in Word2Vec. Generally, learn distributed representation is learned by Word2Vec, which consists of two types of architectures, including contextual bag-of words (CBOW) and skip-gram (SG) [35]. Algorithmically, those two methods are similar to each other. In Continuous Bag of Words (CBOW) architecture, centre words(target) which are based on the neighbouring words, are predicted by the algorithm. Statistically, CBOW outperforms another distributional corpus, which considers a whole context as one observation. That becomes a positive thing for small corpus. Skip gram is similar to CBOW.

However, it exchanges the output and input. This is the opposite of CBOW which predicts all surrounding words ("context") from one input word. Essentially, discovering word vectors is the training purpose of the skip-gram algorithm. These vectors can effectively find close words in the related contexts. In the skip-gram model, the centre word (opposite of CBOW) predicts nearby context words. The bigger the corpus is, the more influential the skip gram model is because, as a skip-gram model, every context-centre pair is treated as a new observation [36]. The structures of both models are presented in Figure2.

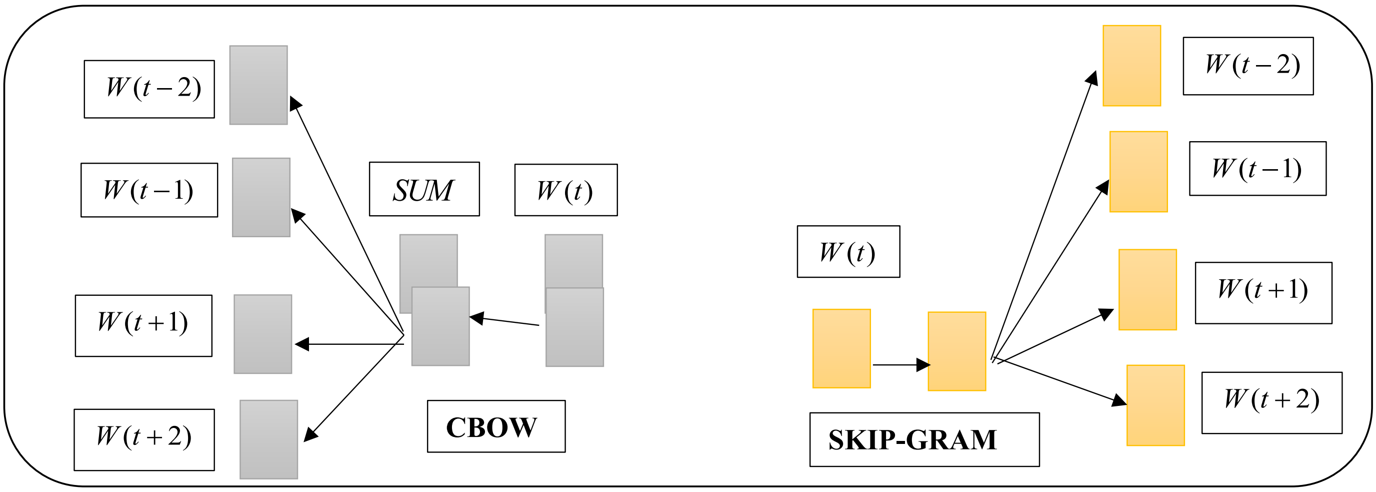

Figure2: The skip-gram and CBOW models

Methodology

Due to high complexity, weak information extraction, and the inability to model non-linear relationships among emotional features, the combination of ANFIS-PSO and Word2vec models can answer these limitations. In this section, a novel model is proposed that is composed by four steps which are shown in figure 3. First, normalizer and noise reduction methods pre-processed both speech and text data from database. Second, audio inputs are used to extract the spectral and prosodic features. Also, the Word2vec model extracted the corresponding text inputs. Third, Sequential Backward Floating Selection (SBFS) selects the most relevant features to solve the problem. Fourth, the ANFIS-PSO model classifies the inputs into six emotions. Finally, the results include the suggested algorithm's steps:

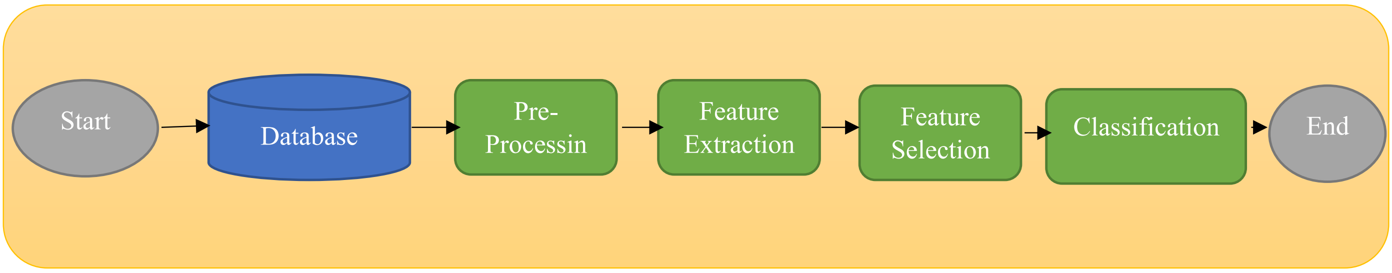

Figure3: The proposed model steps

Pre-processing

Pre-processing is the very first step after collecting data that will be used to train the classifier in an SER system.

The pre-processing models are usually used to normalize speech and text data that have variations in the speakers' voice or have an inferior recording that directly affects the recognition process. Because of using natural speech emotion databases and unifying the voice frequencies spectrum, we used normalization and noise reduction models [37].

Normalization

As an essential step, feature normalization reduces speaker and recording variability while keeping the discriminative strength of the features high. The generalization ability of features is increased by applying feature normalization. The utilized z-normalization by this method is calculated as $Z = \frac{x-\mu}{\sigma}$.

Where μ is the mean and σ is the standard deviation of data [38].

Noise Reduction

Noise has been considered a disruptive factor in SER systems. It alters speech signals and degrades speech quality and intelligibility, leading to considerable damage to human-to-machine communication systems., The noise was reduced using Minimum Mean Square Error (MMSE) approaches.. In MMSE, the clean signal is estimated by comparing it with a noisy signal. Apriorism information of speech and noise spectrum is needed. It is assumed that the additive noise spectrum and estimate of the speech spectrum are available. This method is supposed to minimize the expected distortion measure between clean and estimated speech signals [39].

Feature Extraction

Features are crucial to speech emotion recognition, and each emotion can be successfully characterized by a careful set of features, leading to an increase in the recognition rate. Various features have been used for SER systems; however, an accepted set of features for accurate and distinguished classification is unavailable. According to some studies, prosodic and spectral features have better performance in SER systems, compared with human judges, humans usually perceive prosodic features including intonation and rhythm. We call them para-linguistic features as they cope with the elements of speech [37, 40]. These elements belong to larger units such as syllables, words, phrases, and sentences. Since para-linguistic features are originated from these large units, they are called long-term features. Conveying the most distinguishable properties of emotional content for speech emotion recognition is done by Prosodic features [37, 38].

On the other hand, transforming the time domain signal into the frequency domain signal using the Fourier transform is facilitated by spectral features. Spectral features are extracted from speech segments 20 to 30 milliseconds, partitioned by a windowing method [41]. Affecting the distribution of spectral energy in the speech frequency range by the emotional contents of speech has been mentioned in some studies [42]. So, the existing studies are responsible for utilizing prosodic and spectral features for speech data as well as word2vec for speech transcript data. Table 1 lists 450 features used in this study for emotion recognition. These features include MFCC, F0 hybrid, Energy, Chromocene, Chromic, chromist, Mel spectrogram, Rms, Spectral contrast, spectral_rolloff, Zero_crossing_rate, and Word2vec.

Table 1: List of the candidate features used for emotion recognition

Feature number	Base features	Supplementary features
1-3	Prosodic features	Mean (F0 hybrid), Mean (Energy)
4-150	Spectral features	MFCC, Chroma_cens, Chroma_cqt, chroma_stft, Mel spectrogram, Rms, Spectral_contrast, spectral_rolloff, Zero_crossing_rate
151-450	Word2vec	Word vectors

Feature Selection

Nowadays, dimensional reduction methods are more practical in SER systems, decreasing storage requirements and dimensionality in clustering models. Feature reductions consist of feature selection and feature conversion. Selecting a subset of features with a better result between sets is what feature selection tries to achieve. However, the core subjective in feature conversion is to find a linear or non-linear mapping from the main feature space to a less one. We used the Sequential Backward Floating Selection (SBFS) method to reduce the dimension of features due to many features, enhancing the computational efficiency and minimizing the model's generalization error. SBFS algorithm, a family of greedy search algorithms, reduces an initial d-dimensional feature space to a k-dimensional feature subspace where k < d. Selecting a subset of features automatically relevant to the problem is the main reason for using this algorithm.

The SBFS method is as follows:

- Input: $Y = \{y_1, y_2, ..., y_d\}$
- Output: $X_k = \{x_j \mid j = 1, 2, ..., k; x_j \in Y\}$, where $K = (0, 1, 2, ..., d)$; $K \prec d$
- Initialization: $X_0 = Y, K = d$
- Step 1: In this step, the feature X^- has been removed from the feature subset of X_k

$$X^- = \arg\max J(X_K - x)\text{, where } x \in X_k \tag{7}$$

$$X_{k-1} = X_k - X^- \qquad k = k - 1 \tag{8}$$

Step 2: In Step 2, features that improve the classifier performance have been searched if they are added back to the feature subset. If such features exist, the feature X^+ has been added for which the performance improvement is maximized. Or, If *k*=2 an improvement cannot be made, go back to step 1; else, repeat this step. The features will be added from the feature subset X k until k = p

$$X^{+} = \arg\max J(X_K + x), \text{where } x \in Y - X_k \quad (9)$$

$$ifJ(X_k + X) \succ J(X_k): \quad (10)$$

$$X_{k+1} = X_k + X^{+} \qquad k = k+1$$

Go to step 1 [4]

Classification

Different aspects of neural networks and fuzzy logic are used in characterizing the ANFIS. ANFIS can learn and generalize, making it possible to operate with linguistic variables and incorporate a broader treatment [31]. However, AFIS is disadvantageous because it employs gradient-based techniques to tune the membership functions of the ANFIS model that are more likely to fall in local minima. The error surface is non-convex, highly multi-dimensional, and also contains local minima and flat regions. The gradient-based methods are considered local search approaches, not being able to secure convergence to the global minima. This is mainly because the gradient-based methods apply the chain rule to calculate the error function's gradient. Also, the chain rule cannot discriminate between the local and global minima [43, 44]. It is suggested to deploy non-gradient metaheuristic methods such as Particle Swarm Optimization (PSO). This is because of a reduction in funding the network parameters to an optimization problem to stop local minima. In general, it is suggested to use non-gradient methods of tuning the parameters of the model.

Particle Swarm Optimization (PSO)

As a stochastic population-based optimization algorithm, PSO mimics the grouping behavior. Global communication among members of the algorithm searches for the best solution. The solutions move toward the particle that finds the best position or solution [34]. Assigning an initial random position to each particle is followed by this method. Also, particles move in the multi-dimensional search space with their position and flight speed updated based upon their best-known local position. This position is instructed by other whole particles' best known general position. The process continues until an equilibrium is reached or the computational limitations are exceeded. Imagine a swarm with a population size of N, an initial position x, and movement speed v. The best local position of a particle is denoted as P_{best}, and position of the particle in the swarm, which better minimizes the performance measure, is denoted as G_{best} [31]. The speed and position of the i^{th} particle of the swarm in the next iteration can be formulated as follows:

$$v_i(k) = wv_i(k-1) + \rho_1(x_{P_{best}} - x_i(k)) + \rho_2(x_{G_{bsrt}} - x_i(k)) \quad (11)$$

$$x_i(k) = x_i(k-1) + v_i(k) \quad (12)$$

Where ρ_1 and ρ_2 are random variables defined as $\rho_1 = r_1 c_1$ and $\rho_2 = r_2 c_2$, with r_1 and $r_2 \sim U(0,1)$. The variables c_1 and c_2 are positive acceleration constants that satisfy the condition $c_1 + c_2 \le 4$ and w is the inertial weight that can be calculated using the inertial weight approach (IWA) as follows:

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{Itr_{\max}} Itr \quad (13)$$

Where w_{max} and w_{min} denote the initial and final weight, Itr represents the current iteration number and Itr_{max} is the maximum number of iterations [32].

All steps of the proposed algorithm have been demonstrated in Table 2.

Table2: Summary of the proposed algorithm

Step1: Normalizing the inputs
Step2: Redacting the inputs noise
Step3: Extracting the features X_k
Step4: Removing feature X− from a subset of X_k
Step5: Searching for new features X^{+}
Step6: If the feature X^{+} improves the maximized performance, go to the next step
Else go to step 4
Step7: Initializing the selected features
Step8: Set the number of particles, acceleration coefficient, random vector, and number of fuzzy linguistic set of

ANFIS-PSO

Step9: Generating all initial parameters of each ANFIS-PSO particle
Step10: Apply local search to find fitness function
Step11: Initializing P_{best} and selecting G_{best}
Step12: Evaluating objective functions ():
Step13: If $W_{optimom}$ fine, the process will be ended
Else
Tuning all parameters of each ANFIS-PSO particle
Finding new velocity and position for each particle
Searching new local and fitness function
Updating Pbest for each particle
Obtaining G_{best}
Go to step 11

Results

A combination of ANFIS-PSO and Word2vec algorithms in SER systems was employed in this paper. Therefore, it is necessary to illustrate a better performance of combining these algorithms compared with the others. Also, the combined algorithms could be an appropriate method in SER systems. In this paper, feature selection and classification evaluation assessed the proposed method. At first, the selected features have been specified and evaluated. Secondly, evaluating experimental parameters was conducted by the tested and trained classification model. Moreover, comparing the proposed model with other four well-known models and analyzing the results were conducted. The experiments were implemented with Matlab® software (2016a) and Jupyter Notebook 6.0.3 under a 7th Gen Intel Quadcore 2.53 GHz Processor, 64 GB RAM, NVIDIA Quadro P4000 GPU 4 GB.

Next, the data set and both feature selection evaluation and classification evaluation are explained in detail.

Data set

A standard database called Sharif Emotional Speech Database (ShEMO) was used to evaluate the proposed scheme. ShEMO is a large-scale semi-natural database in Farsi consisting of 3 hours and 25 minutes of speech data from 87 native-Persian speakers (31 females, 56 males). There are 3000 utterances in wave format, 16-bit, 44.1 kHz, and mono, covering five basic emotions: anger, fear, happiness, sadness, neutrality, and surprise. Furthermore, each speech in ShEMO has a transcript attached to it and text mining uses the Ort format.Feature Selection Evaluation.

The feathers selected by the SBFS method were investigated in this part; the linear model called Least absolute shrinkage and selection operator (Lasso) was employed due to existing sparsity between data. Prioritizing the importance of independent variables was specially done by the Lasso too. The objective for finding the minimum by the Lasso differs from the traditional regression approach, which is shown below:

$$Lasso = Minimize_{\beta_{o_i}\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right) \tag{14}$$

Where N is the total number of observations, $\lambda$is a non-negative regularization parameter corresponding to one value of Lambda, y_i is the dependent variable, p is the number of independent variables $x_i = (x_{i1},...,x_{ip})^T$, β_0 is the intercept, and j are the other parameters [45], [46].LASSO method assesses the accuracy of the selected features.

The priority of 50 features in female and male types is shown in Figures 4 and 5. 0.9016 and 0.902 have been calculated as the accuracy values of female and male features, respectively. Additionally, the test and train scores are 0.42 & 0.37 and 0.27 & 0.33 for male and female features, respectively. Both female and male charts apply the MFCC, Word2vec, Chroma_stft, and Mel-spectrogram. Finally, male and female charts selected the Chroma_cens and Spectral_contrast features more, respectively.

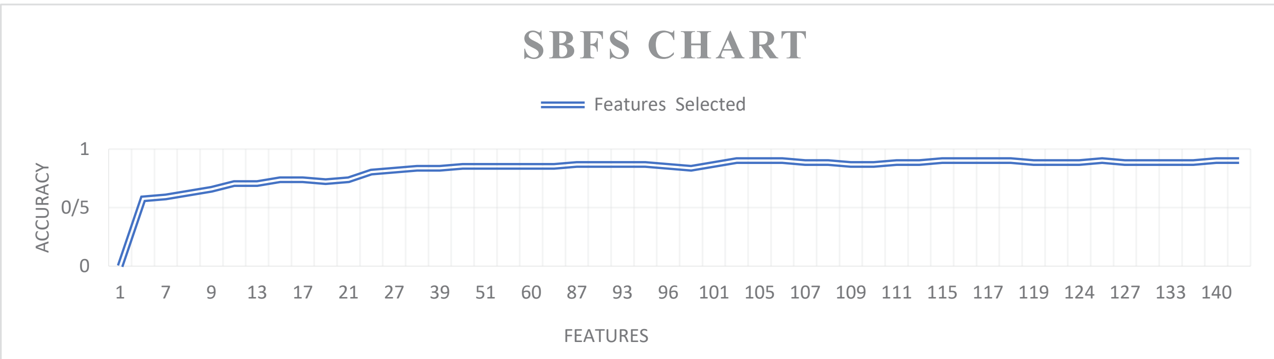

Figure 4: The Accuracy of male SBFS chart.

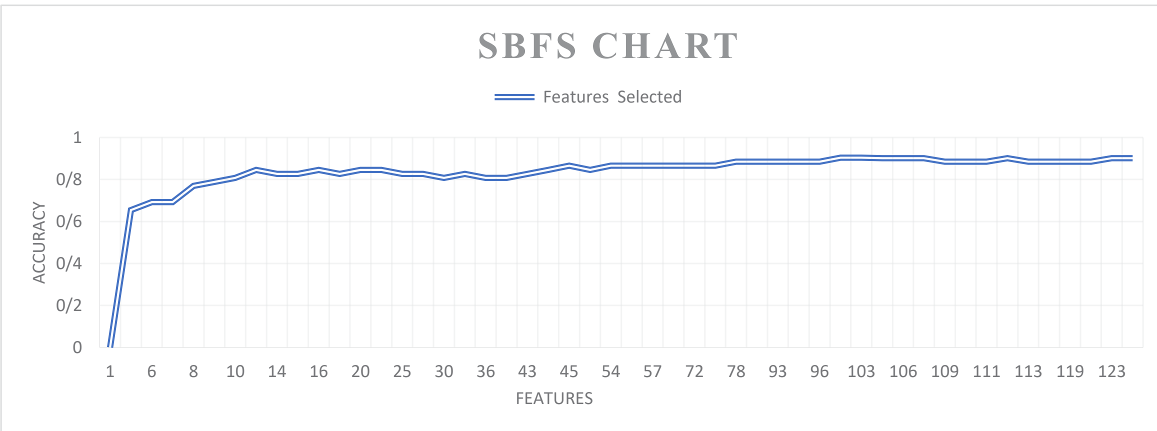

Figure 5: The Accuracy of female SBFS chart.

Classification Evaluation

In this part, we evaluate the classification model in two steps. To begin, we used Mean Squared Error (MSE) and Root-Mean-Square Error (RMSE) in a distinct parameter to test our model. Then, we compare the proposed model with well-known models, such as Random Forest (RF), Convolutional Neural Network (CNN), SVM, and Multilayer. Perceptron (MLP). Moreover, the Precision, Accuracy, Recall, and F1-score of the models are assessed.

Parameter Setting

MSE and RMSE are two statistical indicators that are used to investigate the accuracy of the model. These indicators are introduced as follows [47, 48].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(T\arg et_i - Output_i\right)^2 \quad (15)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(T\arg et_i - Output_i\right)^2} \quad (16)$$

We evaluated the treatment of the proposed model on different types that each part has the control parameters such as Maximum Number of Iterations (MaxIt), Population Size (NPop), Phil1, Phil2, Alpha (Constriction Coefficients), and Inertia Weight Damping Ratio (Wdamp). Table 3 contains four columns as types, parameters, train, and test. The experiment data is divided into 40% test and 60% train, and the process continues until MaxIt . Table 3 shows that with the increase of MaxIt and NPop as 1000, 200 to 2500, 650, For females and men, the MSE and RMSE of the train model drop by 1% and 2%, respectively.. Moreover, the trained model's standard deviation (SD) averagely 0.1% decreased for both the female and the male. On the other hand, the MSE of the test model fluctuated between 2.82 and 5.68 for the female and 2.93 and 5.69 for the male. Also, we have the same treatment in RMSE indicators, the values changed between 1.68 and 2.38 for the female, and 1.71 and 2.38 for the male.

Furthermore, Figure 6 indicates the rate of changing the training and testing process in the male data, and it showed that when the rate of MaxIt and NPop increased, the mean rate changed from 1 to -1, and model type 3 has a lower MSE, RSE, and SD compared to other ones. Likely. A similar treatment can be seen in the female chart, which is shown in Figure 7, but the rate of change is slower than males' rates. Figures 8 and 9 in appendix A are the results of training the female and male data of model type3. It shows that the best performance of the model is between 400 to 600 iterations; also, it has the most errors between 600 to 700 iterations. Also, Figure 9 indicates male model has the most errors between 200 to 600 iterations. The test model of type 3 displays female and male results which have the most errors between 300 to 350 and 200 to 300 iterations, as shown in figures 10 and 11 (appendix A), respectively.

Table 3: parameter setting of ANFIS-PSO model

Type	Parameters	Train					Test			
		Sex	MSE	RMSE	MEAN	SD	MSE	RMSE	MEAN	SD
1	*Alpha=0/000001 MaxIt=1000 NPop=200 Phi1=2/05 Phi2=2/05 Wdamp=1*	Male	1/4073	1/1863	0/005942	1/1867	3/6216	1/903	0/27132	1/8854
		Female	1/3934	1/1804	0/0047675	1/1809	3/54	1/8839	0/32306	1/8578
2	*Alpha=0/000001 MaxIt=2000 NPop=400 Phi1=0/5 Phi2=4 Wdamp=0/1*	Male	1/3381	1/1567	0/009015	1/1521	5/6975	2/3869	0/8021	2/2935
		Female	1/3194	1/1486	0/0088323	1/1491	5/6863	2/3846	0/70342	2/2807
3	*Alpha=0/000001 MaxIt=2500 NPop=400 Phi1=0/5 Phi2=4 Wdamp=0/1*	Male	1/3011	1/140	-0/02125	1/1497	2/9324	1/7123	-0/47823	1/7549
		Female	1/3189	1/1484	-0/020931	1/1489	2/8241	1/6805	-0/30862	1/6541
4	*Alpha=0/000001 MaxIt=2500 NPop=650 Phi1=0/5 Phi2=4 Wdamp=0/1*	Male	1/235	1/111	-0.02254	1/1031	3/3212	1/822	-0/06864	1/7678
		Female	1/243	1/1149	-0/023777	1/1153	3/3489	1/83	-0/073465	1/8309

Clustering Comparison

To analyze the performance of classifiers, we are compared the proposed method with CNNs, SVM, MLP, and RF in the literature. Precision, F-measure, recall, and accuracy are the indexes used to determine the performance of the classifiers, which are defended as follow:

TP (True Positive): Number of data correctly diagnosed under any specific class;

TN (True Negative): Number of data correctly rejected by the classifier;

FP (False Positive): Number of data incorrectly identified by the classifier;

FN (False Negative): Number of data incorrectly discarded by the classifier.

Precision: The probability of the test correctly diagnosed as positive cases given that the number of cases labeled by the system as positive.

$$Precision = \frac{\sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i}}{C} \tag{17}$$

F-Measure (Macro-Averaged F-measure): The weighted combination of recall and precision.

$$F - Measure = \frac{\left(\beta^2 + 1\right) Sensitivity \times precision}{\beta^2 Sensitivity + precision} \tag{18}$$

Average Accuracy: The fraction of test results predicted as correct among all the classes.

$$Accuracy_{Avg} = \frac{\sum_{i=1}^{C} \frac{TP_i + TN_i}{TP_i + FN_i + TN_i + FP_i}}{C} \tag{19}$$

Recall: It is the average probability of complete retrieval averaged over multiple detection queries.[1]

$$\mathrm{Re}call = \frac{TP_i}{TP_i + FN_i} \tag{20}$$

As shown in Table 4 in appendix A, in female data, the RF method with a 0.55 value has the lowest amount, and the proposed model with a 0.72 value has a maximum F-measure. The CNNs, MLP, and SVM methods have 0.63, 0.62, and 0.63 values in the F-measure index, which shows the proposed model has a 5% better performance than others. The suggested model, which has a 0.75 Recall value, outperforms CNNs, SVMs, MLPs, and RF models by 11 percent, 13 percent, and 14 percent, respectively.. In male data, the proposed model with 0.85 value has a better performance than CNNs with 0.74 value, SVM with 0.70 value, MLP with 0.71 value, and RF with 0.65 F-measure value. Moreover, the proposed model with a 0.87 Recall value has 11% and 15% better function than CNNs and SVM, MLP, RF models.

The results of features in six emotions in female and male individuals are shown in Tables 5 and 6 in appendix A, respectively. The results show that each feature's proportion of impact and variation are quite near to each other, indicating the model's effects.

Figure 12 in appendix A demonstrated the proposed model with a 0.873 accuracy in the female and a 0.751 accuracy in the male has a better performance than other models the CNNs, SVM, MLP, and RF models with 11% and 15% in the male and 11%, 13% and 14% have a lees performance than the proposed model. We can claim the proposed model has a better function in male data than the female one. Table 7 present the comparative discussion of the average performance of each classifier in additional research. The result shows that the proposed model in the ShEMO database with 81% accuracy compared to similar research in this database has a better performance. As a result, other data bases are lacking in information and have no free resources. As a result, rather than another study, the suggested model has a sufficient performance.

Conclusion

In this paper, we have proposed and developed a model that can recognize emotion in SER systems. We used hybridmachine learning algorithms to classify the six emotions, anger, fear, happiness, sadness and surprise, and neutral. The proposed method comprises four parts, including Pre-processing, feature extraction, feature selection, and classification. We have developed methods in the pre-processing stage to denoise audio and texts data. Both Znormalization and MMSE methods have been utilized to normalize, and noise reduces the data. Secondly, 450 features have been used to recognize emotions. These features include MFCC, F0 hybrid, Energy, Chroma_cens, Chroma_cqt, chroma_stft, Melspectrogram, Rms, Spectral_contrast, spectral_rolloff, Zero_crossing_rate, and Word2vec. Then we used SBFS techniques to select the features with high accuracy. In the end, ANFIS-PSO models have been utilized to classify emotions. The combination of an ANFIS with Word2vec algorithms in the ShEMO database has allowed us to achieve an average accuracy of 81%, overcoming other methods.

Author Declaration Template

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the

Conflict of Interest

Potential conflict of interest exists:

We wish to draw the attention of the Editor to the following facts, which may be considered as potential conflicts of interest, and to significant financial contributions to this work:

No conflict of interest exists

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Funding

Funding was received for this work.

All of the sources of funding for the work described in this publication are acknowledged below:

No funding was received for this work

References

1. Yin, Y., Zheng, X., Hu, B., Zhang, Y., & Cui, X. (2021). EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. Applied Soft Computing, 100, 106954.
2. J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Fundamentals of speech recognition," in Robust Automatic Speech Recognition, Elsevier, 2016, pp. 9–40.
3. Valiyavalappil Haridas, A., Marimuthu, R., Sivakumar, V. G., & Chakraborty, B. (2020). Emotion recognition of speech signal using Taylor series and deep belief network based classification. Evolutionary Intelligence, 1-14.
4. Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56-76.
5. Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues. J. Basic Appl. Sci, 13, 459-465.
6. Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190.
7. Zhou, J., Chu, S., Li, X., Xiao, F., & Sun, L. (2021). An EEG emotion recognition method based on transfer learning and echo state network for HilCPS. Microprocessors and Microsystems, 87, 103381.
8. Tursunov, A., Kwon, S., & Pang, H. S. (2019). Discriminating emotions in the valence dimension from speech using timbre features. Applied Sciences, 9(12), 2470.
9. Li, D., Liu, J., Yang, Z., Sun, L., & Wang, Z. (2021). Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Systems with Applications, 173, 114683.
10. Yildirim, S., Kaya, Y., & Kılıç, F. (2021). A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. Applied Acoustics, 173, 107721.
11. Tuncer, T., Dogan, S., & Acharya, U. R. (2021). Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. Knowledge-Based Systems, 211, 106547.
12. Abdel-Hamid, L., Shaker, N. H., & Emara, I. (2020). Analysis of linguistic and prosodic features of bilingual Arabic–English speakers for speech emotion recognition. IEEE Access, 8, 72957-72970.
13. Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., & Nanayakkara, S. (2020). Multimodal emotion recognition with transformer-based self supervised feature fusion. IEEE Access, 8, 176274-176285.
14. Peng, Z., Li, X., Zhu, Z., Unoki, M., Dang, J., & Akagi, M. (2020). Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends. IEEE Access, 8, 16560-16572.
15. Fahad, M. S., Ranjan, A., Yadav, J., & Deepak, A. (2021). A survey of speech emotion recognition in natural environment. Digital signal processing, 110, 102951.
16. Şahin, M., & Erol, R. (2017). A comparative study of neural networks and ANFIS for forecasting attendance rate of soccer games. Mathematical and computational applications, 22(4), 43.
17. Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2(7), e12189.
18. Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding," Expert Syst. Appl., vol. 139, p. 112851, Jan. 2020.
19. Seal, D., Roy, U. K., & Basak, R. (2020). Sentence-level emotion detection from text based on semantic rules. In In-

formation and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018 (pp. 423-430). Springer Singapore.
20. Singh, L., Singh, S., & Aggarwal, N. (2018, September). Two-stage text feature selection method for human emotion recognition. In Proceedings of 2nd International Conference on Communication, Computing and Networking: ICCCN 2018, NITTTR Chandigarh, India (pp. 531-538). Singapore: Springer Singapore.
21. Mozafari, F., & Tahayori, H. (2019, January). Emotion detection by using similarity techniques. In 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) (pp. 1-5). IEEE.
22. Alotaibi, F. M. (2019). Classifying text-based emotions using logistic regression.
23. YHuang, Y. H., Lee, S. R., Ma, M. Y., Chen, Y. H., Yu, Y. W., & Chen, Y. S. (2019). EmotionX-IDEA: Emotion BERT--an Affectional Model for Conversation. arXiv preprint arXiv:1908.06264.
24. Polignano, M., Basile, P., de Gemmis, M., & Semeraro, G. (2019, June). A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. In Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (pp. 63-68).
25. Ma, L., Zhang, L., Ye, W., & Hu, W. (2019, June). PKUSE at SemEval-2019 task 3: emotion detection with emotion-oriented neural attention network. In Proceedings of the 13th international workshop on semantic evaluation (pp. 287-291).
26. Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. Information Processing & Management, 57(1), 102121.
27. Halim, Z., Waqar, M., & Tahir, M. (2020). A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. Knowledge-based systems, 208, 106443.
28. J. C. Chen, E. A. Rubin, and G. J. Cornwall, "Natural Language," 2021, pp. 259–281.
29. JJang, J. S. R. (1991, July). Fuzzy modeling using generalized neural networks and kalman filter algorithm. In AAAI (Vol. 91, pp. 762-767).
30. Panahi, M., Gayen, A., Pourghasemi, H. R., Rezaie, F., & Lee, S. (2020). Spatial prediction of landslide susceptibility using hybrid support vector regression (SVR) and the adaptive neuro-fuzzy inference system (ANFIS) with various metaheuristic algorithms. Science of the Total Environment, 741, 139937.
31. Enayatollahi, H., Fussey, P., & Nguyen, B. K. (2020). Modelling evaporator in organic Rankine cycle using hybrid GD-LSE ANFIS and PSO ANFIS techniques. Thermal Science and Engineering Progress, 19, 100570.
32. Noushabadi, A. S., Dashti, A., Raji, M., Zarei, A., & Mohammadi, A. H. (2020). Estimation of cetane numbers of biodiesel and diesel oils using regression and PSO-ANFIS models. Renewable Energy, 158, 465-473.
33. Bensaber, B. A., Diaz, C. G. P., & Lahrouni, Y. (2020). Design and modeling an Adaptive Neuro-Fuzzy Inference System (ANFIS) for the prediction of a security index in VANET. Journal of Computational Science, 47, 101234.
34. Ehteram, M., Yenn, F., & Najah, A. (2020). Performance improvement for infiltration rate prediction using hybridized Adaptive Neuro-Fuzzy Inferences System (ANFIS) with optimization algorithms. Ain Shams Eng J 11: 1665–1676.
35. Sun, J., Luo, X., Gao, H., Wang, W., Gao, Y., & Yang, X. (2020). Categorizing malware via A Word2Vec-based temporal convolutional network scheme. Journal of Cloud Computing, 9(1), 1-14.
36. Yilmaz, S., & Toklu, S. (2020). A deep learning analysis on question classification task using Word2vec representations. Neural Computing and Applications, 32, 2909-2928.
37. Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden Markov models. In Seventh European conference on speech communication and technology.
38. Luengo, I., Navas, E., Hernáez, I., & Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. In Ninth European conference on speech communication and technology.
39. Zhu, Y., Yan, E., & Wang, F. (2017). Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. BMC medical informatics and decision making, 17(1), 1-8.
40. Frick, R. W. (1985). Communicating emotion: The role of prosodic features. Psychological bulletin, 97(3), 412.
41. Wang, X., Zhang, X., Zeng, Z., Wu, Q., & Zhang, J. (2016). Unsupervised spectral feature selection with l1-norm graph. Neurocomputing, 200, 47-54.
42. Jiang, J., Shi, T., Huang, M., & Xiao, Z. (2020). Multi-scale spectral feature extraction for underwater acoustic target recognition. Measurement, 166, 108227.
43. Ghomsheh, V. S., Shoorehdeli, M. A., & Teshnehlab, M. (2007, June). Training ANFIS structure with modified PSO algorithm. In 2007 Mediterranean Conference on Control & Automation (pp. 1-6). IEEE.
44. Shoorehdeli, M. A., Teshnehlab, M., & Sedigh, A. K. (2006, June). A novel training algorithm in ANFIS structure. In 2006 American Control Conference (pp. 6-pp). IEEE.
45. Shi, X., Wang, K., Cheong, T. S., & Zhang, H. (2020). Prioritizing driving factors of household carbon emissions: An application of the LASSO model with survey data. Energy Economics, 92, 104942.
46. Shuku, T., Phoon, K. K., & Yoshida, I. (2020). Trend estimation and layer boundary detection in depth-dependent soil data using sparse Bayesian lasso. Computers and Geotechnics, 128, 103845.
47. Tak, N. (2018). Meta fuzzy functions: Application of recurrent type-1 fuzzy functions. Applied Soft Computing, 73, 1-13.
48. Robati, F. N., & Iranmanesh, S. (2020). Inflation rate modeling: Adaptive neuro-fuzzy inference system approach and particle swarm optimization algorithm (ANFIS-PSO). MethodsX, 7, 101062.
49. van der Burgh, H. K., Schmidt, R., Westeneng, H. J., de Reus, M. A., van den Berg, L. H., & van den Heuvel, M. P. (2017). Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. NeuroImage: Clinical, 13, 361-369.

50. Mohamad Nezami, O., Jamshid Lou, P., & Karami, M. (2019). ShEMO: a large-scale validated database for Persian speech emotion detection. Language Resources and Evaluation, 53, 1-16.

Appendix A

Table 4: Possible outcomes of classifiers in female and male groups.

Models	Female				Male			
	Precision	F-measure	Recall	Accuracy	Precision	F-measure	Recall	Accuracy
CNNs	0.62	0.63	0.64	0.648	0.74	0.74	0.76	0.7601
SVM	0.62	0.63	0.64	0.643	0.70	0.70	0.72	0.7219
MLP	0.61	0.62	0.62	0.620	0.70	0.71	0.72	0.7230
RF	0.58	0.55	0.61	0.610	0.65	0.65	0.72	0.7210
Proposed	0.73	0.72	0.75	0.751	0.84	0.85	0.87	0.8730

Table 5: The results of 6 type of emotions of femail.

Features/Emotions	**Sad**	**Angry**	**Happy**	**Worry**	**Fear**	**Natural**
Chroma_Cens	0/166665082	0/166646479	0/166698279	0/166623126	0/166690825	0/166676209
Chroma_Cqt	0/166653073	0/16661751	0/166715889	0/166637173	0/166724886	0/166651469
Chroma_Stft	0/166378438	0/166373693	0/166836501	0/166569024	0/167042915	0/16679943
Fchroma_Stft	0/036659479	0/192254054	0/192808411	0/192493709	0/193013828	0/192770518
Melspectrogram	0/166972207	0/166959868	0/166429868	0/166950348	0/166312413	0/166375297
Rms	0/166633415	0/166667261	0/166669619	0/166628316	0/166690929	0/16671046
Spectral_Contrast	0/166768166	0/166779656	0/166626449	0/16660113	0/166603238	0/166621362
Spectral_Rolloff	0/192580058	0/192396585	0/040198433	0/192396585	0/19121417	0/19121417
Zero_Crossing_Rate	0/167027628	0/166930558	0/166370419	0/166930558	0/166370419	0/166370419
Text	0/163833875	0/163457426	0/171100868	0/163550025	0/170940579	0/167117226

Table 6: The results of 6 type of emotions of mail.

Features/Emotions	**Sad**	**Angry**	**Happy**	**Worry**	**Fear**	**Natural**
Chroma_Cens	0/16668682	0/166723758	0/166703937	0/16669408	0/166484158	0/166707248
Chroma_Cqt	0/166708653	0/166772186	0/166726917	0/16672266	0/166336149	0/166733435
Chroma_Stft	0/166733154	0/166705969	0/166710691	0/166752517	0/166363482	0/166734186
Fchroma_Stft	0/166563167	0/166529373	0/166736783	0/166673652	0/166795954	0/166701071
Melspectrogram	0/166644955	0/166625027	0/166648272	0/166664202	0/166779005	0/16663854
Rms	0/167465556	0/167436444	0/161773857	0/167430331	0/168467128	0/167426685
Spectral_Contrast	0/166474592	0/166483538	0/166487559	0/16648328	0/167586498	0/166484532
Spectral_Rolloff	0/166536482	0/166405438	0/166486862	0/166342639	0/167879262	0/166349316
Zero_Crossing_Rate	0/166486812	0/16642716	0/166485063	0/166310041	0/167971413	0/166319512
Text	0/166474903	0/165684298	0/165658541	0/166333483	0/169466997	0/166381779

Table 7: Abstract of compered models

Methods	Dataset	Accuracy	Reference
TL-SVM	ESS	68	[7]
SVM-RNN	IEMOCAP	65.05	[8]
LSTM	IEMOCAP	84.4	[49]
NSGA-II	IEMOCAP	69.30	[14]
SVM	ShEMO	58.2	[50]
Proposed	ShEMO	81.2	

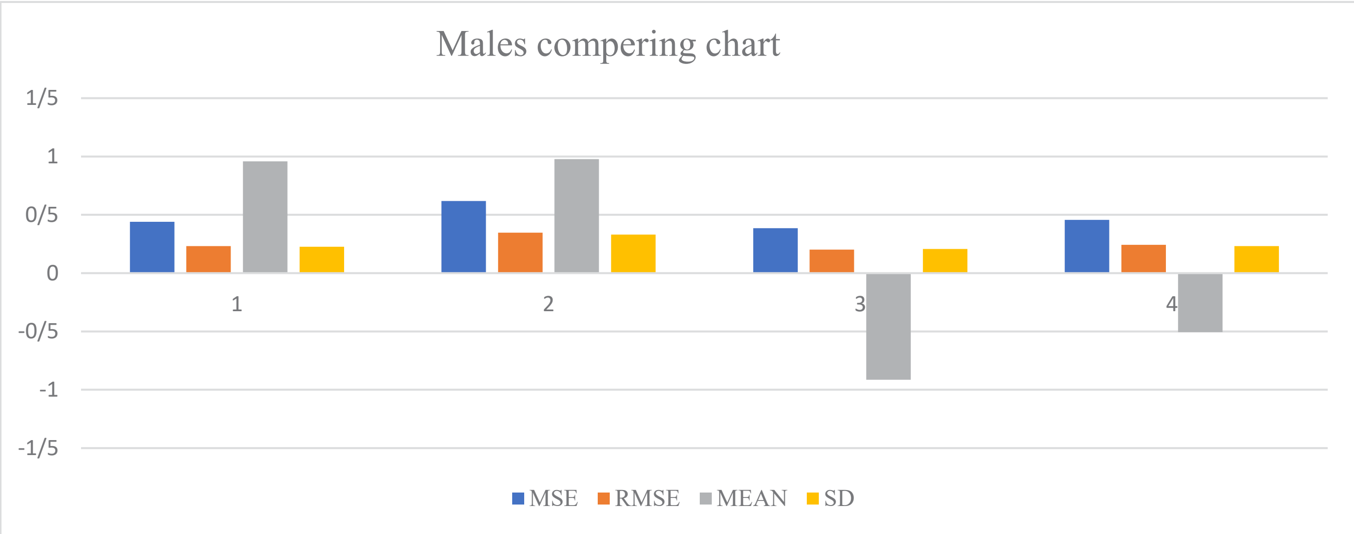

Figure6: The values of males' index in 4 types of models.

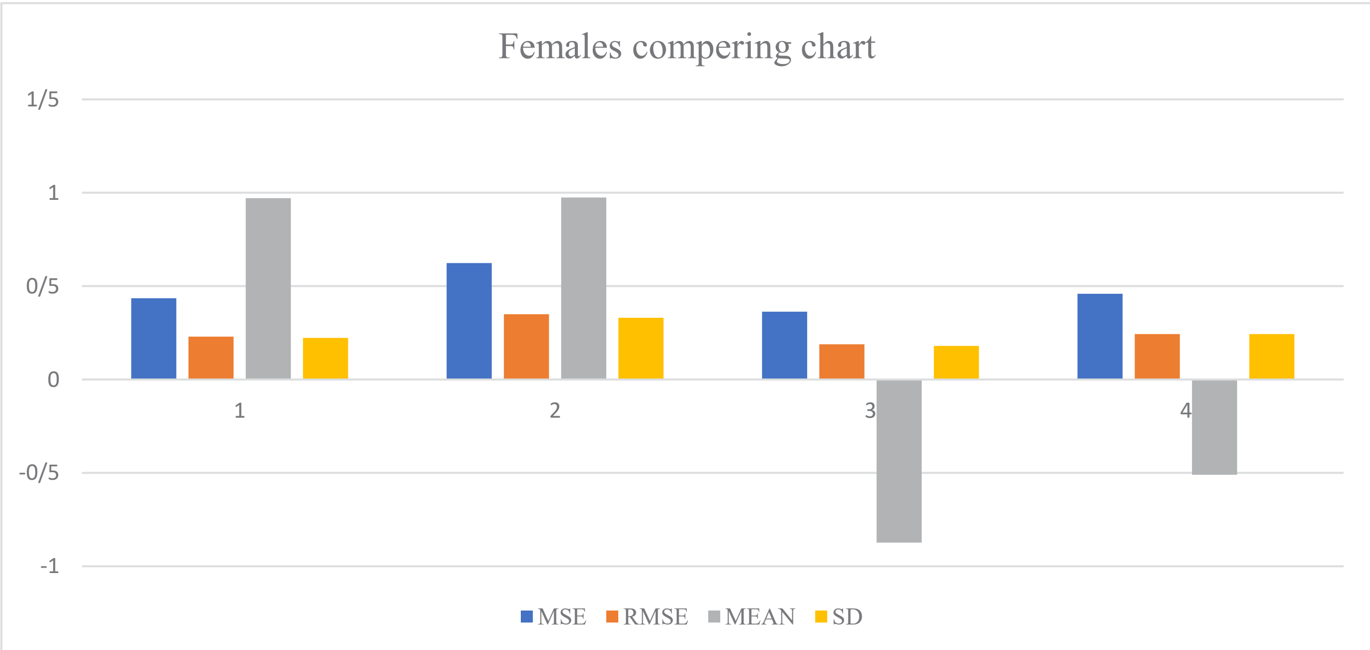

Figure7: Females' 4 types of models indexs' values.

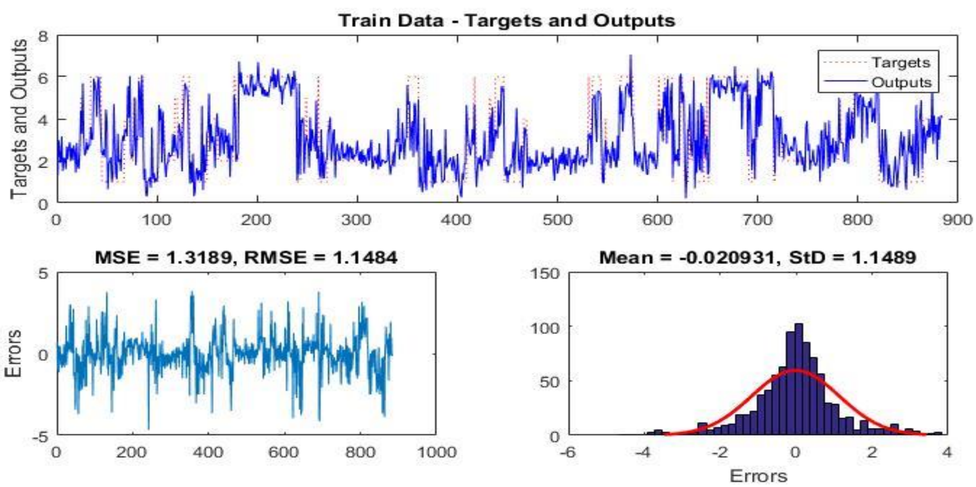

Figure8: The result of female data with 60% training.

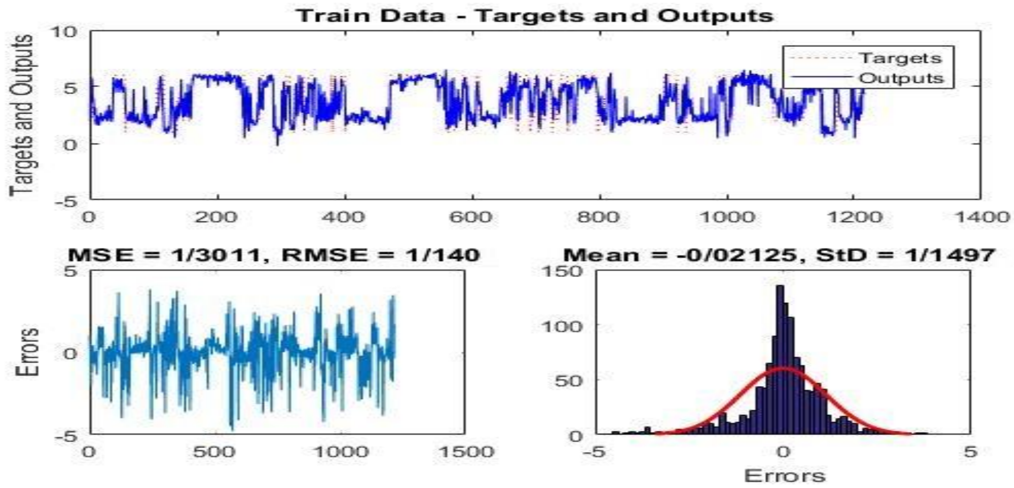

Figure9: The result of male data with 60% training.

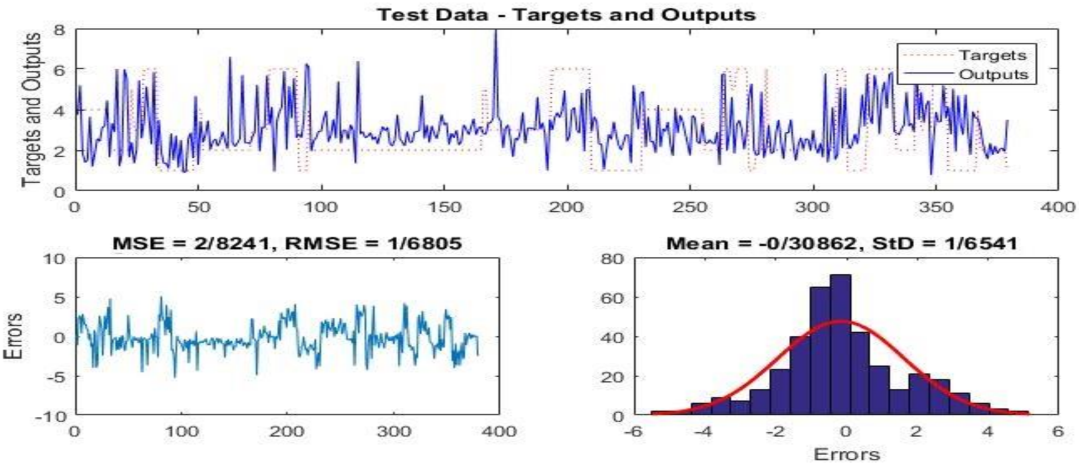

Figure10: The result of female data with 40% testing.

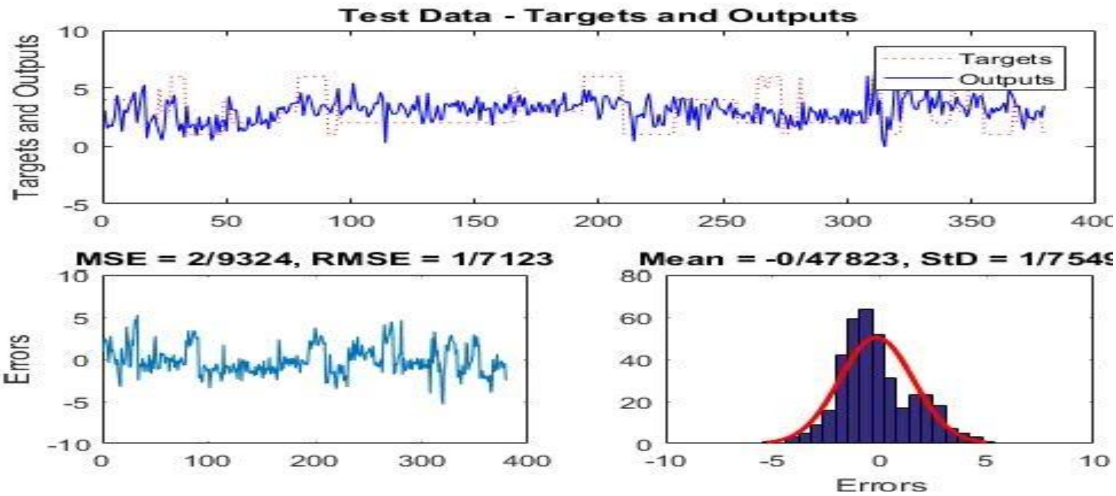

Figure11: The result of male data with 40% testing.

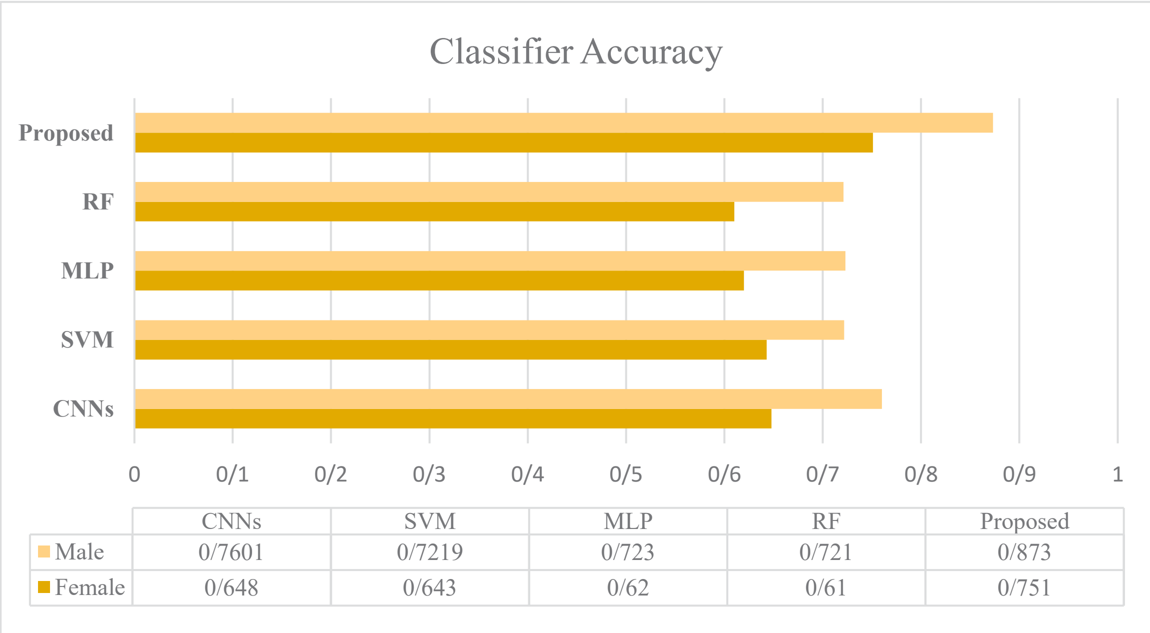

Figure12: Accuracy of classifier in female and male types.