# Deep Learning, Ensemble and Supervised Machine Learning for Arabic Speech Emotion Recognition

**Wahiba Ismaiel**
Department of Science and Technology, University College of Ranyah, Taif University, Saudi Arabia
w.wahiba@tu.edu.sa (corresponding author)

**Abdalilah Alhalangy**
Department of Computer Engineering, College of Computer, Qassim University, Saudi Arabia
a.alhalangy@qu.edu.sa

**Adil. O. Y. Mohamed**
Department of Computer Science, College of Computer, Qassim University, Saudi Arabia
adi.mohamed@qu.edu.sa

**Abdalla Ibrahim Abdalla Musa**
Department of Computer Science, College of Computer, Qassim University, Saudi Arabia
ab.musa@qu.edu.sa

*Received: 24 February 2024 | Revised: 10 March 2024 | Accepted: 12 March 2024*

 *DOI: https://doi.org/10.48084/etasr.7134*

## ABSTRACT

**Today, automatic emotion recognition in speech is one of the most important areas of research in signal processing. Identifying emotional content in Arabic speech is regarded as a very challenging and intricate task due to several obstacles, such as the wide range of cultures and dialects, the influence of cultural factors on emotional expression, and the scarcity of available datasets. This study used a variety of artificial intelligence models, including Xgboost, Adaboost, KNN, DT, and SOM, and a deep-learning model named SERDNN. ANAD was employed as a training dataset, which contains three emotions, "angry", "happy", and "surprised", with 844 features. This study aimed to present a more efficient and accurate technique for recognizing emotions in Arabic speech. Precision, accuracy, recall, and F1-score metrics were utilized to evaluate the effectiveness of the proposed techniques. The results showed that the Xgboost, SOM, and KNN classifiers achieved superior performance in recognizing emotions in Arabic speech. The SERDNN deep learning model outperformed the other techniques, achieving the highest accuracy of 97.40% with a loss rate of 0.1457. Therefore, it can be relied upon and deployed to recognize emotions in Arabic speech.**

***Keywords-Arabic speech emotion recognition; ANAD; SERDNN; SOM; Xgboost; Adaboost; DT; KNN***

## I. INTRODUCTION

Speech is a common and natural mode of human communication and is considered the simplest, quickest, and most natural type of human connection but does not apply to machines [1-3]. Although human-machine-speech connections are developing, machines cannot grasp human emotions and are incapable of engaging in a genuine and natural debate [4]. Emotion recognition is the first step in improving human-computer interaction with speech-based computing systems. Recognizing emotions in Arabic speech is considered challenging due to several issues, such as cultural disparities, the wide range of Arabic dialects, and how they are circulated among Arab people. Accurate identification of emotions in speech remains challenging due to the absence of a specific definition of emotion and the extensive and complicated influence of creating and expressing emotions. Machine Learning (ML) is a breakthrough technology in which significant amounts of data are provided to computer systems [5]. This involves making computers identify human speech and capable of detecting emotions. Computers use data to enhance their algorithm and better process other data in the future [2]. Researchers are looking for techniques to extract the emotional state of a speaker from their words and to successfully understand human emotions [6]. Emotions may be

represented in speech and then exploited to extract useful connotations from spoken words, therefore improving speech recognition implementations [4, 7].

Today, artificial emotional intelligence, often known as emotional AI, is a popular topic of study. People communicate their emotions through nonverbal indications, such as external facial expressions, gestures, and nonverbal communication, using body language and tone of voice. Emotional AI seeks to identify emotions in the same way that humans do, through many channels. Emotional AI is frequently used in intelligent security, intelligent contact centers, customer assistance, medical applications, forensics, banking, and stress and anxiety treatment [8]. In [9], facial expression-based surveillance was employed to detect neurodegenerative diseases. Utilizing the LSTM model, emotions are identified from both speech and text [10]. In [11], a smart scarf was presented to help people who struggle to communicate their feelings. In [12], Speech Emotion Recognition (SER) was compared in Arabic and English, suggesting that the introduction of particular acoustic elements ameliorates emotion recognition for Arabic words.

Automatic emotion identification from voice data without language clues is an important and developing study field. In [13-14], emotion identification in spoken Arabic data was investigated using five ensemble models on a voice emotion identification system. The SMO achieved the highest accuracy (95.52%) among single classifiers in recognizing the "happy", "angry", and "surprised" emotions in Arabic real speech. Due to the lack of Arabic speech emotion datasets, few studies have built Arabic SER systems. In [15], emotion identification in Egyptian Arabic speech was investigated deploying prosodic, spectral, and wavelet data, implementing an Egyptian television series to generate a semi-natural Egyptian Arabic speech emotion dataset. Identifying and classifying the emotional content of a speech signal is a key problem for researchers. In [3], the KS Arabic spoken emotion corpus was engaged along with classification approaches, such as K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). Convolutional Neural Networks (CNNs) can emerge as the prevailing method within this domain. In [16], two neural architectures were applied to identify emotions in Arabic voice, and the results showed that the proposed strategy may lead to considerable improvement (2.2%) over a powerful deep CNN baseline system, due to the use of sophisticated layers of deep learning paradigms, namely Hubert, wav2vec2.0, wav2vecU, and WavBERT, which allow better representation learning and greater information capture. In [17], a deep learning model of emotional recognition was introduced for spoken Arabic dialogue.

In the field of emotion detection, when a person becomes emotional, their voice alters to reflect their emotional state. In [18], a deep learning-based emotional recognition model was presented for spoken Arabic dialogues. In [19], a multi-head attention multi-learning model (ABMD) was investigated, using Dilated Convolution (DC) and Residual Dilated Causal Convolution (RDCC) blocks for emotion recognition from speech patterns. Based on three datasets, EMODB, RAVDESS, and SAVEE, the ABMD model achieved accuracy rates of 95.93, 85.89, and 93.75%, respectively. In [20], a feature vector, which consisted of 28 measures, was submitted for emotion recognition. This study adopted a Moroccan Arabic Dialect Emotional Database (MADED) with a KNN classifier to detect four emotions: "joy", "anger", "neutral", and "sadness". This KNN obtained an accuracy rate of 64.65% for all emotions. In [21], a recognition and classification model, named TLBOML-ERC, was suggested to detect emotional expressions and sentiments expressed in Arabic tweets. This study employed a denoising autoencoder to classify emotional expressions in an Arabic tweet dataset into four emotion classes, "anger", "sadness", "joy", and "fear". Three different ratios were put into service for the training and testing sets, 90% train and 10% test, 80% train and 20% test, and 70% train and 30% test set, and TLBOML-ERC achieved 97.86, 98.84, and 98.24% accuracy, respectively. In [22], a CNN model was proposed for Arabic speech emotion recognition based on the BASE-DB training dataset with four emotions (anger, neutral, sadness, and joy) and five features (Mel-scaled spectrogram, MFCC, tonal centroid, chromatogram, and spectral contrast), achieving an accuracy rate of 98.18%.

In [23], a 1DCNN was integrated with LSTM and an attention model to recognize emotions in speech. This model utilized RMSE, MFC, and ZCR features and was evaluated on four datasets: BAVED, ANAD, SAVEE, and EMO-DB. The accuracy of the proposed model varied across datasets, with results of 88.39, 96.72, 97.13, and 96.72%, accordingly. In comparison, the accuracy rates obtained by the 2DCNN model were 83.39, 90.16% 51.30, and 50%, correspondingly. In [24], GWO-KNN, which employed the Grey Wolf Optimizer (GWO) to extract SER features in both Arabic and English languages, was presented. This study adopted the Emirati dialect Arabic speech database, the RAVDESS, and the SAVEE datasets. In [25], two sets of features were generated by combining spectral and prosodic characteristics from the Egyptian Arabic semi-natural emotion speech dataset, which consists of 579 utterances. Five machine learning classifiers, namely Random Forest (RF), Multilayer Perceptron (MLP), Ensemble learning, Logistic Regression (LR), and Support Vector Machine (SVM), were implemented to predict four emotions: "happy", "sad", "angry", and "neutral". The results showed that ensemble learning obtained the highest rate of 87.59%, whereas it achieved 64% for predicting all multi-emotions. In [26], SRE was deployed for the Algerian dialect, based on an Algerian speech emotion dataset with four emotions: "happy", "sad", "neutral", and "angry". This study applied various machine learning methods, and LSTM-CNN had the highest accuracy rate of 93.34%. In [27], an Arabic SER system was built based on supervised machine learning, using ANAD as a training set with three emotions: "angry", "happy", and "surprised". The proposed model was named C-SVM and its highest score was achieved for "angry" with 98.1, 98.7, and 98.3% precision, recall, and F1-score, respectively. In [28], SER was performed on Arabic speech utilizing the ANAD dataset with three emotions and implementing 35 classifiers.

Many problems remain unsolved for effective AI systems, including poor accuracy rates, high computing complexity of hybrid classifier models, and a paucity of natural datasets. Multiple studies have suggested different ways to recognize

vocal emotions. This study aims to fill the existing gap employing the Arabic Natural Audio Dataset (ANAD) for SER, relying on four supervised machine learning classifiers: extreme boosting (Xgboost), Adaptive boosting (Adaboost), KNN, and DT. The purpose of this research is to assist apps and institutions that rely on speech to make correct and ideal judgments, which are free from prejudice. Table I depicts the contributions of related works and their suggestions for identifying emotions in Arabic speech. This study also presents and evaluates a simple deep-learning model to precisely recognize and classify emotions from Arabic speech. The following is a list of the contributions of this study:

- Present the most recent advances in Arabic SER.
- Propose a simple and effective deep learning model.
- Implement five machine learning techniques.
- Examine the accuracy, recall, and F1-score of the models.
- Evaluate the accuracy of the proposed models against the results of previous studies.

TABLE I. CONTRIBUTION OF RELATED WORKS

| Reference | Year | Models | Datasets | Language | Emotion types |
|---|---|---|---|---|---|
| [20] | 2021 | KNN | MADED | Arabic/Moroccan | KNN |
| [21] | 2023 | TLBOML-ERC | Arabic tweets | Arabic | Joy, angry, fear, sadness |
| [22] | 2022 | CNN | BASE-DB | Arabic | Joy, angry, neutral, sadness |
| [23] | 2023 | 1DCNN integrated with LSTM and attention, 2DCNN | BAVED, ANAD, SAVEE, EMO-DB | Arabic | high, low, neutral, sad, fear, bored, disgust, happy, angry, surprised |
| [24] | 2023 | GWO-KNN | SAVEE, RAVDESS, Emirati Dialect Arabic Speech Database | Arabic/English | Surprised, neutral, angry, happy, fear, disgust, sad, calm |
| [25] | 2023 | MLP, Ensemble learning, RF, MLP&SVM | Egyptian Arabic Semi-Natural Emotion speech | Arabic/Egyptian | happy, sad, angry, neutral |
| [26] | 2021 | LSTM, BLSTM, CNN | Algerian speech emotion | Arabic/Algerian | happy, sad, neutral, angry |
| [27] | 2022 | C-SVM | ANAD | Arabic | angry, happy, surprised |
| [28] | 2018 | SOM | ANAD | Arabic | angry, happy, surprised |

## II. METHODOLOGY

Four supervised machine learning methods (Xgboost, Adaboost, KNN, and DT) and one unsupervised model (SOM) were used to detect and recognize emotions in Arabic speech. Figure 1 portrays the methodology of this study.

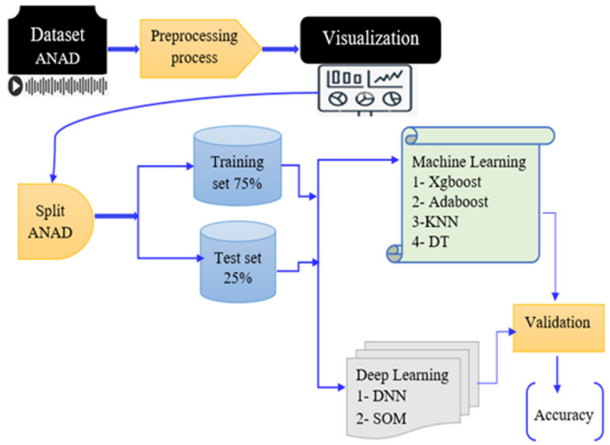


Fig. 1. Methodology.

### A. Data Collection and Description

The ANAD and Arabic Speech Emotion Recognition (ASER) datasets were used [29]. The ANAD dataset consists of 846 attributes. The first attribute contains 1383 speech files, the second one includesw the target field of emotion types, and the rest 844 attributes are speech features. This dataset provides 25 ANAD features that are listed in Table II. Eighteen (18) listeners participated in reporting their feelings, listening to eight different videos and choosing from surprised, happy, and angry. The dataset contained 137 surprised, 505 happy, and 741 angry rows, and Figure 2 displays the emotion categories.

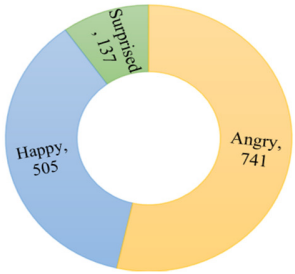


Fig. 2. Emotion categories.

### B. Data Preprocessing and Visualization

This process is considered one of the most basic steps in dealing with machine and deep learning algorithms to achieve prediction accuracy based on data free from bias and noisy values. After examination, it was determined that ANAD has no null values but its data consist of different formats: float, int, and categorical (object type). The categorical format refers to two attributes: speech name and label. The label attribute represents three emotions: "angry", "happy", and "surprised". This format is not compatible with the model, so it was converted into numerical values of 0 or 1, split into three columns named 0, 1, and 2, corresponding to the emotions "angry", "happy", and "surprised", respectively. The audio files were processed to exclude any noise, quiet, and laughter. All audio files had a length of 1 second and a signal rate of 22050. Figures 3-5 exhibit the waveform visualization of each emotion's sampling, highlighting the distinction between the three emotions. All 844 characteristics were extracted from the

dataset and divided into a training set, including 75% of the data, and a test set, entailing the remaining portion. The training set consisted of 1037 records, while the test set consisted of 346 records. Following the process of data splitting, the subsequent step involved adjusting the feature values to fall within a specific range using the normalization method provided by the scikit-learn module in Python. The normalization process depends on (1), which calculates the new value ($V$) based on the minimum feature value ($f_{min}$), the maximum feature value ($f_{max}$), and the current value of the feature ($v$):

$$V = \frac{v - f_{min}}{f_{max} - f_{min}} \quad (1)$$

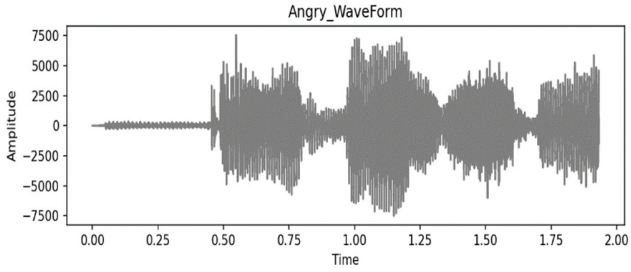


Fig. 3. Angry sampling waveform.

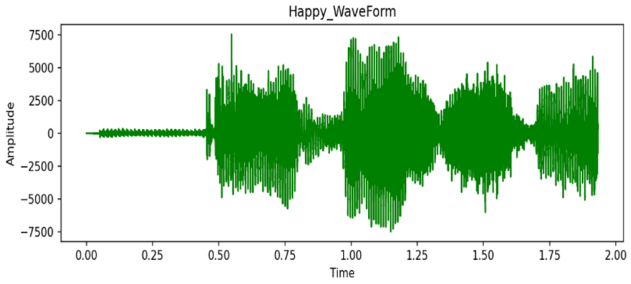


Fig. 4. Happy sampling waveform.

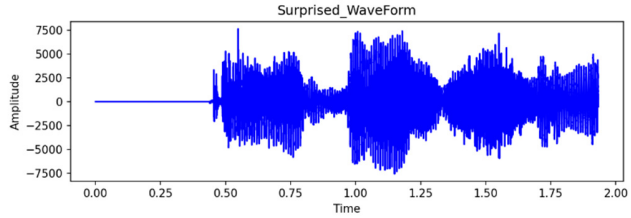


Fig. 5. Surprised sampling waveform.

### C. Proposed Speech Emotion Recognition Deep Neural Network (SERDNN)

SERDNN was built predicated on the Keras module in Python language and dense class. SERDNN had 844 features as the input stage and three hidden layers that implemented a ReLU activation function. The last layer had three neurons to output the emotion with a sigmoid function. The model was compiled using the Adam optimizer function and categorical cross entropy for the loss function. Figure 6 illustrates the configuration of the SERDNN model. Table III demonstrates the configuration details of SERDNN and the activation size calculation. The train parameters for each layer were estimated according to the previous layer neuron *Pn* and the current layer neuron *Cn* (2). The SERDNN model was trained for 200 epochs with a fixed batch size of 64, to obtain the prediction of speech emotion results.

$$No\ of\ Parameters = Pn * Cn + 1 * Cn \quad (2)$$

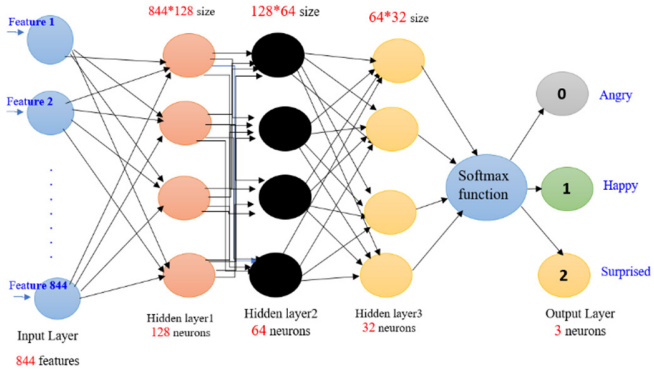


Fig. 6. SERDNN model configuration.

TABLE II. CONFIGURATION DETAILS OF SERDNN

| Layer | output shape | Activation size | #Parameters |
|---|---|---|---|
| dense | (none, 128) | 844×128=108032 | 844×128+128=108160 |
| dense _1 | (none, 128) | 128×128=16384 | 128×128+128=16512 |
| dense _2 | (none, 64) | 128×64=8192 | 128×64+64=8256 |
| dense _3 | (none, 32) | 64×32=2048 | 64×32+32=2080 |
| dense_4 | (none,3) | 32×3=96 | 32×3+3=99 |

### D. Proposed Artificial Neural Network (SOM)

Self-Organizing Map (SOM) is an unsupervised learning technique that accepts higher dimensional training data grouped according to their similarity to decrease their dimensions. It contains two layers, one for the input and a second for the output, which entail 2D neurons [30]. SOM is trained on the basis of the following steps:

- Initialize neurons by assigning a small random value to weight 00 and initialize the learning rate $a$
- Find the Euclidean distance value using:

$$distance(j) = \Sigma\,(weight_{ij} - x_i)^2 \quad (3)$$

where $j = 1 \ldots m$ and $i = 1 \ldots n$

- Assign the winning index with the smallest value of distance weight ($D(W_j)$)
- Determine the neighborhood of $j$ for all $i$ utilizing:

$$weight_{ij}(current\ value) = weight_{ij}(old - value) + a[x_i - weight_{ij}(old - value)] \quad (4)$$

- Update the learning rate at time $t$ employing:

$$a(t+1) = 0.5 \times t \quad (5)$$

- Stop training at the iteration end.

This study deployed the Minisom library in Python to build an SOM of 12×12 neurons, with 844 features as input values.

The SOM model implemented the Gaussian function to calculate the distance between the neighborhood, the learning rate was 0.5 and the sigma value was assigned to 1.25. The sigma is the radius of a different neighborhood in the neural network. This model was trained in 3000 iterations. Figure 7 presents the proposed SOM architecture.

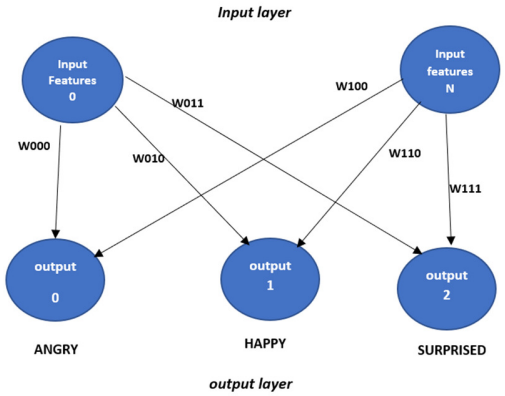


Fig. 7. SOM architecture.

### *E. Proposed Supervised Machine Learning Classifiers*

Four supervised classifiers Xgboost, Adaboost, KNN, and DT were also utilized to classify and detect speech emotions. Xgboost and Adaboost are ensemble classifiers.

#### *1) K-Nearest Neighbor (KNN)*

KNN can perform regression and classification processes and K represents the number of nearest neighbors. KNN focuses on finding the smallest distance between neighbors, so it predicts the target based on the Euclidean distance (6). Figure 8 depicts the KNN classifier. In this study, K was assigned to 3.

$$Distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (6)$$

where $(x_1, y_1)$ and $(x_2, y_2)$ are the coordinates of the two points. The goal is to find the smaller distance between the two points, supposing that these points are two features.

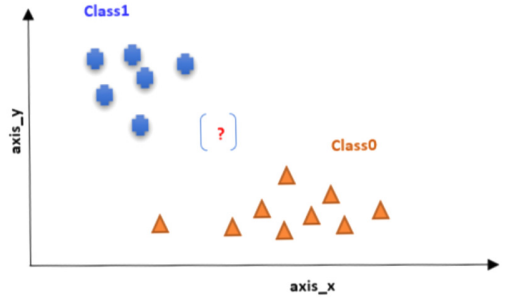


Fig. 8. KNN classifier.

#### *2) Decision Tree Classifier (DT)*

DT is commonly employed to solve classification and regression problems. The root of the tree is a point to the particular dataset where the DT classifier starts the prediction process based on the compared results of the root value with the rows of a dataset, taking the branch direction for the next node.

#### *3) Extreme Gradient Boosting (Xgboost)*

Xgboost is an ensemble classifier that can train a large dataset of different domains [31].

#### *4) Adaptive Boosting (Adaboost)*

Adaboost can also be considered an ensemble learning technique, such as Xgboost, for classification and regression problems. This technique merges multiple weak classifiers to provide a strong one and depends on the weight used to solve the problem of incorrect prediction [32]. Equation (7) illustrates the Adaboost method, where $N$ corresponds to the weight, and $f_N$ explains the $N$ weak classifiers.

$$F = sin\left(\sum_{N=1}^{N} \theta_N \ f_N(X)\right) \quad (7)$$

## III. EXPERIMENTAL RESULTS

All the proposed models were run on an MSI laptop with a Core™ i7-11th Intel® @ 3.00GHz CPU, and 32GB of RAM, using Python and Jupyter Notebook on Windows 10 Pro 64-bit. Three experiments were carried out to obtain the results of the proposed models on the ANAD dataset. The first experiment involved building the SERDNN deep learning model, the second was utilized for the supervised learning algorithms (Xgboost, Adaboost, KNN, and DT), and the final experiment employed the SOM model. Four metrics were applied: accuracy, precision, recall, and F1-score. Furthermore, a confusion matrix was also put into service to measure the performance of the proposed models in Arabic SER. The confusion matrix displays the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) results. TP represents the cases where the model predicts a positive result and an actual positive result. TN signifies cases where the model predicts a negative result for an actual negative result. FP denotes cases where the model forecasts a positive result but the actual result is negative. FN represents cases where the model predicts a negative result but the actual result is positive. The following equations demonstrate how these metrics are calculated.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (9)$$

$$\text{Recall}= \frac{TP}{TP+FN} \quad (10)$$

$$F1-score = \frac{2*(Precision)*Recall}{Precision+Recall} \quad (11)$$

Figure 9 exhibits the train and test accuracy results of the SERDNN model. Figure 10 demonstrates the train and test loss of the SERDNN model.

Figure 11 illustrates the train and test accuracies of Xgboost, Adaboost, KNN, DT, SOM, and SERDNN. Figure 12-14 provide the precision, F1-score, and recall scores of all models tested to detect the "angry", "happy", and "surprised" emotions. Figure 15 presents the confusion matrices.

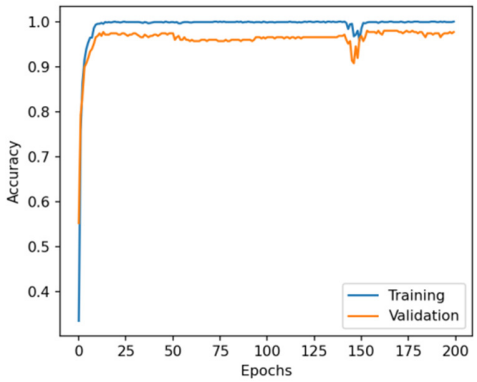


Fig. 9. Accuracy of the SERDNN model.

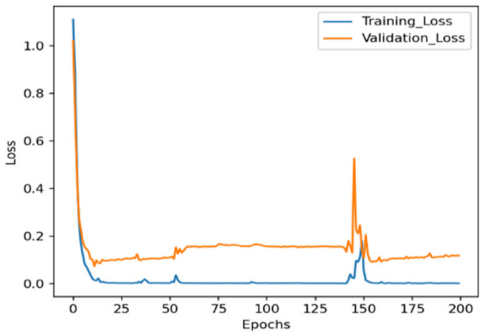


Fig. 10. Loss of the SERDNN model.

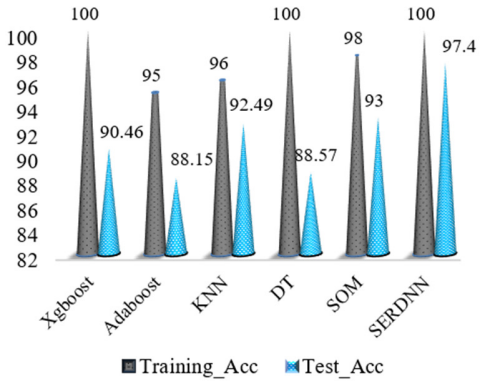


Fig. 11. Train and test accuracies of proposed models.

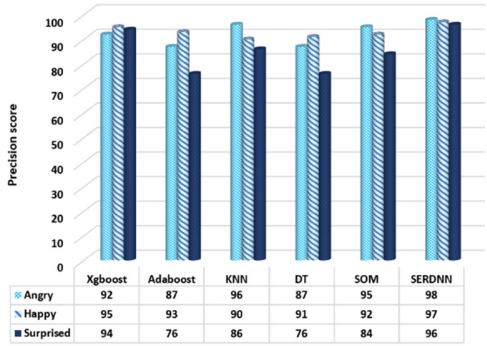


Fig. 12. Precision on surprised, happy, and angry emotions.

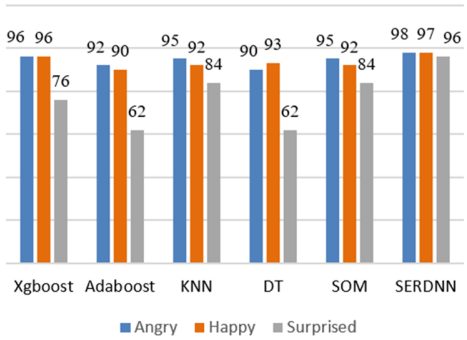


Fig. 13. F1-score of surprised, happy, and angry emotions.

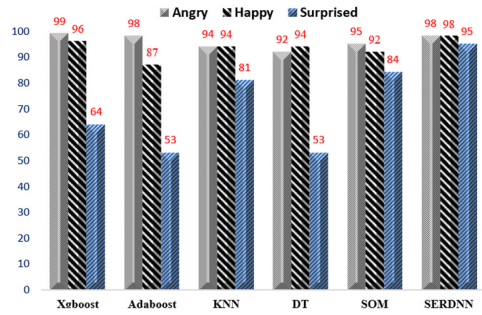


Fig. 14. Recall on surprised, happy, and angry emotions.

## IV. DISCUSSION

The highest training accuracy result was obtained by SERDNN, DT, and Xgboost with a 100% rate, while the highest prediction accuracy was acquired by SERDNN with 97.40%, followed by SOM with 93%, and KNN with 92.49%, while Adaboost had the lowest accuracy at 88.15%. SERDNN achieved the highest precision for all emotions (98, 97, and 96% for "angry", "happy", and "surprised", respectively). The lowest precision score for "angry" was obtained by Adaboost and DT (87%), the lowest precision for "happy" was achieved by KNN (90%), and DT and Adaboost had the same precision (76%) for the "surprised" emotion. SERDNN achieved the highest F1-scores (98, 97, and 95% for "angry", "happy", and "surprised", respectively). The lowest F1-score for "angry" was accomplished by DT (90%), the lowest F1-score for "happy" was acquired by Adaboost (90%), and DT and Adaboost both rated 62% for the "surprised" emotion. The highest recall score was obtained by Xgboost, reaching 99% for "angry", followed by Adaboost and SERDNN (both at 98%). SERDNN had the highest recall rates for "happy" and "surprised" (98 and 95%, respectively). DT had the lowest recall score for "angry" (92%), Adaboost had 87% recall for "happy", and Adaboost and DT obtained the lowest recall for "surprised" (53%). The results disclose that all models have a satisfactory ability to recognize these three emotions. SERDNN, Xgboost, and KNN were the best at detecting the "angry" emotion. For the "happy" emotion, SERDNN and Xgboost were the best models. The emotion prediction rates of "surprised" were low compared to other emotions. Table III displays a comparison of the results between the models in this study and previous studies that used the same dataset.

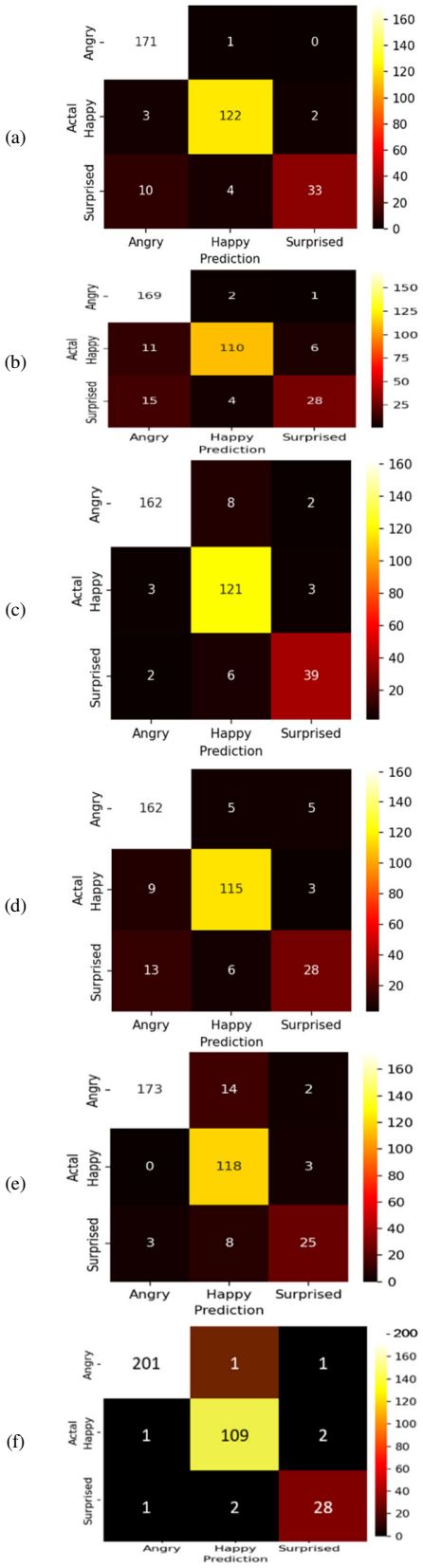


Fig. 15. Confusion matrices of (a) Xgboost, (b) Adaboost, (c) KNN, (d) DT, (e) SOM, (f) SERDNN.

TABLE III. PREVIOUS STUDIES WITH PROPOSED MODELS BASED ON ANAD DATASET

| Reference | Models | Highest accuracy |
|---|---|---|
| [23] | 1DCNN with LSTM & attention | 96.72%, |
| | 2DCNN | 90.16% |
| [27] | C-SVM | Precision 98.1% |
| | | Recall 98.7% |
| | | F1-score 98.3% |
| [28] | SOM | 95.52% |
| **This study** | SERDNN | 97.40% |
| | SOM | 95% |
| | DT | 88.57% |
| | KNN | 92.49% |
| | Xdgboost | 90.46% |
| | Adaboost | 88.15% |

## V. CONCLUSION

This study developed and evaluated models that can recognize emotions in Arabic speech. A variety of artificial intelligence techniques, including Xgboost, Adaboost, KNN, DT, and SOM, were adopted. In addition, this study developed a simple deep-learning model, called SERDNN, to fill the gap in this domain. These models were trained and tested on the ANAD dataset to assess their ability to recognize three different emotions: "angry", "happy", and "surprised". The results manifesterd that all models were able to recognize the "angry" and "happy" emotions with high accuracy, while the precision, F1-score, and recall were high for the "surprised" emotion by SERDNN, SOM, Xgboost, and KNN compared to the DT and Adaboost models. The highest accuracy for all emotions was achieved by the proposed deep learning model SERDNN, at 97.40%. In the future, the SERDNN model will be improved and evaluated on other datasets, while other machine learning models will also be investigated to extract emotional features.

## REFERENCES

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

[2] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," *Applied Sciences*, vol. 13, no. 8, Jan. 2023, Art. no. 4750, https://doi.org/10.3390/app13084750.

[3] A. H. Meftah, M. Qamhan, Uks.-A. 22nd I. C. on C. M. and S. Alotaibi, and Y. A. Zakariah, "Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus," presented at the UKSim-AMSS 22nd International Conference on Computer Modelling and Simulation, Mar. 2020, https://doi.org/10.5013/IJSSST.a.21.02.21.

[4] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic Speech Emotion Recognition From Saudi Dialect Corpus," *IEEE Access*, vol. 9, pp. 127081–127085, 2021, https://doi.org/10.1109/ACCESS.2021.3110992.

[5] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism," *Electronics*, vol. 12, no. 20, Jan. 2023, Art. no. 4376, https://doi.org/10.3390/electronics12204376.

[6] S. Akinpelu and S. Viriri, "Speech emotion classification using attention based network and regularized feature selection," *Scientific Reports*, vol. 13, no. 1, Jul. 2023, Art. no. 11990, https://doi.org/10.1038/s41598-023-38868-2.

[7] Muljono, M. R. Prasetya, A. Harjoko, and C. Supriyanto, "Speech Emotion Recognition of Indonesian Movie Audio Tracks based on MFCC and SVM," in *2019 International Conference on contemporary*

*Computing and Informatics (IC3I)*, Singapore, Dec. 2019, pp. 22–25, https://doi.org/10.1109/IC3I46837.2019.9055509.

[8] S. Hamsa, I. Shahin, Y. Iraqi, and N. Werghi, "Emotion Recognition From Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier," *IEEE Access*, vol. 8, pp. 96994–97006, 2020, https://doi.org/10.1109/ACCESS.2020.2991811.

[9] S. Xefteris, N. Doulamis, V. Andronikou, T. Varvarigou, and G. Cambourakis, "Behavioral Biometrics in Assisted Living: A Methodology for Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 6, no. 4, pp. 1035–1044, Aug. 2016, https://doi.org/10.48084/etasr.634.

[10] S. C. Venkateswarlu, S. R. Jeevakala, N. U. Kumar, P. Munaswamy, and D. Pendyala, "Emotion Recognition From Speech and Text using Long Short-Term Memory," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11166–11169, Aug. 2023, https://doi.org/10.48084/etasr.6004.

[11] W. Almukadi, "Smart Scarf: An IOT-based Solution for Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10870–10874, Jun. 2023, https://doi.org/10.48084/etasr.5952.

[12] A. Meftah, Y. Alotaibi, and S.-A. Selouani, "Emotional speech recognition: A multilingual perspective," in *2016 International Conference on Bio-engineering for Smart Technologies (BioSMART)*, Dubai, United Arab Emirates, Sep. 2016, https://doi.org/10.1109/BIOSMART.2016.7835600.

[13] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Emotion recognition in Arabic speech," *Analog Integrated Circuits and Signal Processing*, vol. 96, no. 2, pp. 337–351, Aug. 2018, https://doi.org/10.1007/s10470-018-1142-4.

[14] R. Zantout, S. Klaylat, L. Hamandi, and Z. Osman, "Ensemble Models for Enhancement of an Arabic Speech Emotion Recognition System," in *Advances in Information and Communication*, 2020, pp. 174–187, https://doi.org/10.1007/978-3-030-12385-7_15.

[15] L. Abdel-Hamid, "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Communication*, vol. 122, pp. 19–30, Sep. 2020, https://doi.org/10.1016/j.specom.2020.04.005.

[16] A. Ali and Y. Hifny, "Efficient Arabic emotion recognition using deep neural networks." arXiv, Oct. 31, 2020, https://doi.org/10.48550/arXiv.2011.00346.

[17] O. Mohamed and S. A. Aly, "Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset." arXiv, Oct. 08, 2021, https://doi.org/10.48550/arXiv.2110.04425.

[18] O. A. Mohammad and M. Elhadef, "Arabic Speech Emotion Recognition Method Based On LPC And PPSD," in *2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, Jan. 2021, pp. 31–36, https://doi.org/10.1109/ICCAKM50778.2021.9357769.

[19] S. Kakuba, A. Poulose, and D. S. Han, "Attention-Based Multi-Learning Approach for Speech Emotion Recognition With Dilated Convolution," *IEEE Access*, vol. 10, pp. 122302–122313, 2022, https://doi.org/10.1109/ACCESS.2022.3223705.

[20] A. Agrima, I. Mounir, A. Farchi, L. Elmaazouzi, and B. Mounir, "Emotion recognition from syllabic units using k-nearest-neighbor classification and energy distribution," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5438–5449, Dec. 2021, https://doi.org/10.11591/ijece.v11i6.pp5438-5449.

[21] I. Alwayle *et al.*, "Parameter Tuned Machine Learning Based Emotion Recognition on Arabic Twitter Data," *Computer Systems Science and Engineering*, vol. 46, no. 3, pp. 3423–3438, 2023, https://doi.org/10.32604/csse.2023.033834.

[22] M. Tajalsir, S. M. Hernandez, and F. A. Mohammed, "ASERS-CNN: Arabic Speech Emotion Recognition System based on CNN Model," *Signal & Image Processing : An International Journal*, vol. 13, no. 1, pp. 45–53, Feb. 2022, https://doi.org/10.5121/sipij.2022.13104.

[23] W. Alsabhan, "Human–Computer Interaction with a Real-Time Speech Emotion Recognition with Ensembling Techniques 1D Convolution Neural Network and Attention," *Sensors*, vol. 23, no. 3, Jan. 2023, Art. no. 1386, https://doi.org/10.3390/s23031386.

[24] I. Shahin, O. A. Alomari, A. B. Nassif, I. Afyouni, I. A. Hashem, and A. Elnagar, "An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer," *Applied Acoustics*, vol. 205, Mar. 2023, Art. no. 109279, https://doi.org/10.1016/j.apacoust.2023.109279.

[25] M. El Seknedy and S. A. Fawzi, "Emotion Recognition System for Arabic Speech: Case Study Egyptian Accent," in *Model and Data Engineering*, Cairo, Egypt, 2023, pp. 102–115, https://doi.org/10.1007/978-3-031-21595-7_8.

[26] R. Y. Cherif, A. Moussaoui, N. Frahta, and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," in *2021 International Conference of Women in Data Science at Taif University (WiDSTaif )*, Taif, Saudi Arabia, Mar. 2021, https://doi.org/10.1109/WiDSTaif52235.2021.9430224.

[27] W. G. S. Al Fadahli, R. K. S. Al Hinai, P. C. Sherimon, V. Sherimon, and R. K. Remya, "An Automated Emotion Recognition from Arabic Speech Using Machine Learning Technique," *International Journal of Creative Research Thoughts*, vol. 10, no. 10, pp. a435–a438, Oct. 2022.

[28] S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, "Enhancement of an Arabic Speech Emotion Recognition System," *International Journal of Applied Engineering Research*, vol. 13, no. 5, pp. 2380–2389, 2018.

[29] "Arabic Natural Audio Dataset." [Online]. Available: https://www.kaggle.com/datasets/suso172/arabic-natural-audio-dataset.

[30] E. de Bodt, M. Cottrell, P. Letremy, and M. Verleysen, "On the use of self-organizing maps to accelerate vector quantization," *Neurocomputing*, vol. 56, pp. 187–203, Jan. 2004, https://doi.org/10.1016/j.neucom.2003.09.009.

[31] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, May 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[32] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997, https://doi.org/10.1006/jcss.1997.1504.