Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/190107

Exploring Deep Learning Methods for Audio Speech Emotion Detection: An Ensemble MFCCs, CNNs and LSTM

Shaik Abdul Khalandar Basha[1,*], *P. M. Durai Raj Vincent*[1], *Suleiman Ibrahim Mohammad*[2,3], *Asokan Vasudevan*[4,5], *Eddie Eu Hui Soon*[4], *Qusai Shambour*[6], *and Muhammad Turki Alshurideh*[7]

[1]School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, 632014 Vellore, India
[2]Electronic Marketing and Social Media, Economic and Administrative Sciences, Zarqa University, 13110 Zarqa, Jordan
[3]Research follower, INTI International University, 71800 Negeri Sembilan, Malaysia
[4]Faculty of Business and Communications, INTI International University, Persiaran Perdana BBN Putra Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
[5]Wekerle Business School, Budapest, Jázmin u. 10, 1083 Hungary
[6]Software Engineering Department, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, 19111 Amman, Jordan
[7]Department of Marketing, School of Business, The University of Jordan, 11942 Amman, Jordan

Received: 23 Jul. 2024, Revised: 12 Oct. 2024, Accepted: 18 Oct. 2024
Published online: 1 Jan. 2025

Abstract: Our world relies entirely on the gadgets we use every day, making the world heavily materialized.The human-machine interactions that are currently available are not supported under line-of-sight (LOS). The proposed emotional communication is based on non-line-of-sight (NLOS) to break away from Conventional human-machine interactions. This emotional communication is defined as interactive, similar to the usual video and voice media we use daily; similarly, the information is transmitted over long distances. We proposed the EAS framework, another ensemble technique for an emotional communication protocol for real-time communication requirements. This framework supports the communication of emotional realization. They also designed. Finally, which are developing CNN-LSTM architectures for feature extraction, implementing an attention mechanism for selecting relevant features, creating for selecting relevant features, and creating for real-time scenarios, performance-evaluated matrices are applied CNN-LSTM networks with and without attention mechanisms. DCCA feature extraction is used to extract attributes and find correlations among different labels in the dataset. Toanalyze the real-time performance of the process in emotional communication with long-distance communications between others. The proposed CNN-LSTM model achieves the highest accuracy with 87.08% accuracy, while existing models, such as CNN baseline and LSTM models, showed 81.11% and 84.01%, respectively. Our approach shows improved Accuracy compared to existing works, especially for real-time applications.

Keywords: MFCC, CNN-LSTM, Communication protocol, RMSE. Emotion Detection, Markov state transfer, RNN, Gender distribution

1 Introduction

The internet plays a massive role in our world. We use the internet everywhere—on our phones and laptops. Many results from human-machine interactions (HMI) are available [1,2,3]. These interactions are divided into four different categories. They are analyzed using the audio-visual information given to the machine, which is low because of the lack of Accuracy. For example, emotion recognition would be interfered with when a user carries such equipment on stairs [4,5,6]. We live in a society where we must communicate with those around us; speech is essential for that cause. Man has continually evolved for the greater good; humans want to be given an easy way out, leading to machines' invention [7,8,9]. For humans to communicate with the machines and let them know what the end product or output needs to be produced, the machine needs to understand and recognize

* Corresponding author e-mail: pmvincent@vit.ac.in

human voices, but that would require sufficient intelligence [10,11,12,13]. Tremendous research over the years, since the fifties, has been made to convert human speech to a sequence of words understandable to the machine; despite having made significant progress, we still need to have a natural conversation with a machine [14,15,16]. The main reason for the above hurdle is that the machine does not recognize the human's emotional state; it may be able to understand the meaning of the words spoken, but it might miss the undertone of the speech [17,18].

It can be used for domestic use by therapists' It needs to be clarified to the system which emotion is of greater importance/priority when distinguishing speech features. Different speakers have tone, acoustic variability, styles, and speaking rates [19,20,21]. The majority of the effects mentioned above There might be more than one emotion being perceived in the same phase, so it becomes difficult to put boundaries between these portions [22,23,24,25]. The cultural and environmental aspect also plays an important role depending on their vicinity. Most of the work done on the application has been done considering that most speakers have no cultural differences, which is invalid. However, the case of multilingual classification is being investigated. The paper gives us a complete run-through of the REVDESS dataset.

As we all know, the number of older adults living alone has rapidly increased. As their age grows, they need complete monitoring to recognize their unusual activities, which means they might need emergency help or care quickly. Hospitalization and care would only be possible at some times due to high costs and limited resources; as technology in today's world is rapidly increasing, humans have created many ways to overcome the following problem. One of the solutions through the research Lately, multiple cameras have been utilized to get a 3D and a 2D view and shapes, but these cameras increase the cost and complexity.

2 Related Work

This section discusses recent studies in speech emotion recognition, including all the feature extraction and classification methods used previously. The earliest research in the area of SER was performed by Batziou et al. [26] in their survey paper, which included all the feature classification schemes and available datasets. The emotions contained in a speech signal can be described into two categories, i.e., definite [27,28,29], where emotions are discrete classes of disgust, happiness, sadness, fear, anger, and surprise. The other is dimensional approaches as in [30] where emotions are defined as numerical values over distinct emotions. This approach can classify complex emotions easily. The speech emotion recognition task consists of 2 kinds of research work. The first one is choosing the best method for feature extraction, i.e., choosing among MFCC, RMSE, Timber, Zero Frequency crossing rate, etc. The second one is choosing the best feature classification method. In literature, we have multiple algorithms like support vector machine (SVM) as in [31], CNN algorithm as in [32], LSTMs algorithms in [33], and Hidden Markov Models (HMMs) as in [33,34].

Since 2012, various research has been conducted to explore different techniques and features of speech signals by [33], in which the attributes are examined, and hidden patterns are retrieved in the audio dataset. The characteristics of the source will increase the recognition rate by 80%. Another research performed by [35] with the Beihang University Database of Emotional Speech (BHUDES) in Mandarin extracts features like MFCC, pitch, and correlation for each frame. It performs feature reduction using Fisher's Linear Discriminant Analysis (LDA) and PCA methods and gives a 3-stage Emotion recognition model with an average rate of 87%, 69%, and 50% at each stage.

Recent research in SER has been performed using deep learning algorithms. The research work proposed by [36,37] used a Convolutional Neural Network with two two-stage SER models. The first stage extracts local features using an auto-encoder for a spectrogram of a speech signal, and the second stage extracts features using PCA for recognition. The SER model is tested on four datasets with an average recognition rate of 79%. Due to the lack of datasets for the SER task [36], TESS [37], SAVEE [38], and CREMA-D [39] since they contain all collections of emotions that allow direct comparison of classification results [40,41,42].

3 Materials and Methods

This section gives a detailed analysis of the datasets used in our experiments. Speech audio datasets are created in multiple systems. These audio frequency datasets are categorized into three types: acted (or represented data), invoked (or appealed), and spontaneous (or impulsive). Acted datasets provide easier modeling of data and detecting emotions since each audio file represents a different kind of emotion, and these files are not changed by environmental noise. Hence, it does not require a pre-processing phase, which causes loss of information. Since the acted datasets have combined emotions, the models built will tend to overfit the different emotions in real-world audio voices.

Gender distribution is the primary objective implemented in our classification model so that the model is not weighted more towards one particular gender. We found that the RAVDESS (12 male and 12 female actors) and CREMA-D (48 male and 43 female actors) datasets have the exact composition of male and female actors. Furthermore, the TESS dataset only contains audio files of male actors, and the SAVEE dataset contains audio files of female actors.

Table 1: There are eight labels with intensity on the dataset

Class	Emotion	Intensity Level
1	neutral	normal
2	calm	normal, strong
3	happy	normal, strong
4	sand	normal, strong
5	angry	normal, strong
6	fearful	normal, strong
7	disgust	normal, strong
8	surprise	normal, strong

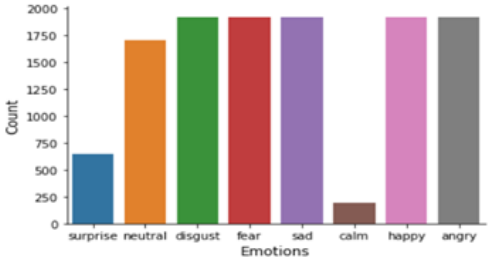


Fig. 1: Count of Emotions

From Table 1 and Figure 1, it can be seen that all datasets do not contain the same categories of emotions. The sampling rate, amplitude(dB), and RMSE are specific to each dataset. The recording duration of each audio file ranges from 2 to 5 seconds. The CREMA-D dataset has more variation in volume with more echoes. Unlike RAVDESS, which contains enough actors to provide gender independence, the RAVDESS dataset does not subject the audio files to environmental noise, which degrades the audio quality. Hence, the RAVDESS dataset is preferred for our experimental analysis.

4 Methodology

In this study, we are proposing a framework for speech Emotion Recognition.

1.Pre-processing
2.Classification
3.Experimentation

Input speech signal to be processed is given to the pre-processing unit, and features are collected from it. Then, using any classification method, the speech signal is classified according to the emotion [28]. The classifier is trained using the features. Once the Training is completed, the system will become familiar with the feelings and classify the new speech signal.

The Framework

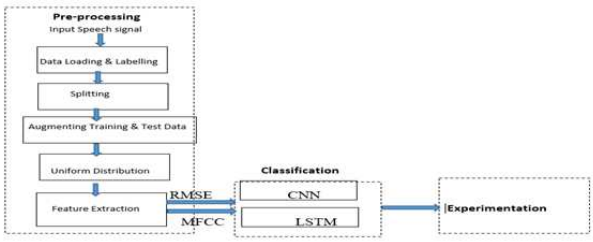


Fig. 2: Multi-Level Proposed Framework to Extract Audio Recognition Feature

4.1 Pre-Processing

Pre-processing speech signals is an essential step in the efficient development of SER tasks. It will improve the quality of the input signal. Pre-processing transforms the speech signal so that it can be processed well to extract its features. It also separates the voiced region from the silent portions of the input signal from the database. The steps involved in the pre-processing are explained below

4.1.1 Data Loading and Labelling

We performed loading and sampling of the audio files through the librosa python package for analyzing audio files. It is a fundamental building block for extracting features like spectrogram, spectral contrast, centroid, etc., from an audio signal. The sampling rate is fixed at 16 KHz, which forms the exact Nyquist sampling rate for human-generated speech signals. Then, the samples are binary and classified into two categories, i.e., disruptive (anger, grief, disgust) and non-disruptive (calm, neutral, happy) emotions. After aggregating these labels, the datasets will have balanced emotions.

4.1.2 splitting

To perform the SER task, the entire dataset must be divided into three categories: Training, validation, and Test data to build an ML model. Instead of splitting randomly, which leads to the partial evaluation of the model, the split can be made such that equal numbers of male and female speakers come in a split, and each speaker will go in only one split. The process of labeling and splitting is given in Figure 2.

4.1.3 Augmenting Training & Test Data

The typical datasets used in the SER task are defined in a noiseless environment, but the practical use of Speech Emotion Recognition Noise plays a significant role in detecting emotions. Thus, we are creating new synthetic training data by adding AWGN noise to the audio

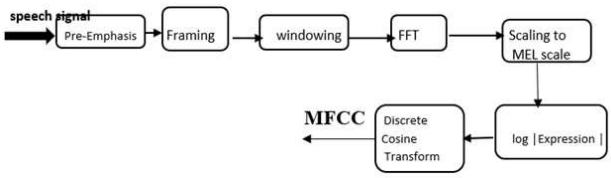


Fig. 3: MFCC Extraction Process

samples. Then, the corrupted samples are merged with the original samples, doubling the Training data amount. The Test data is augmented by combining the audio samples of multiple speakers, which enables the simulation of the SER task in a multi-user environment. The final distribution of test data and Training data after this overlapping is shown in Figure 2.

4.1.4 Uniform Distribution

Since most recordings will have unequal timings and accompany the silence periods, the tapes must be cut and trimmed to have a uniform distribution. This is done by truncating the longer ones and padding with the median value of the audio file for the shorter ones. The average duration of each recording is fixed at 5 seconds, consisting of 80000 samples with a sampling frequency of 16 KHz.

4.1.5 Feature Extraction

Attribute Extraction is an essential step in the SER task process. Speech signals consist of spectral features and prosodic features. Prosodic features are enough to distinguish the emotions as they represent stress, rhythm, intonation, and pauses. However, few feelings are too close and cannot be distinguished by prosodic features. Feature extraction is done by converting the audio signal to a form with a lesser data rate for further processing and analysis of the speech content. Feature extraction approaches usually represent a speech signal into a multi-dimensional feature vector. In this paper, we have used MFCC in Figure 2 to extract spectral properties in a speech signal.

4.2 Classification

Emotion Detection in a speech signal will be performed in 2 stages. The first is the processing unit for extracting all the relevant attributes from audio features, and the next is the classification phase, which decides the different emotion attributes in the uttered Audio feature. Selecting the classifier for the emotion recognition task depends on the geometry of the feature vector. Various classifiers, such as the ANN, RNN, LSTM, and CNN, are used for SER tasks.

Convolutional Neural Network (CNN): A mimic of the human brain is the basic structure of the neural network, in which data is induced as input, processed with hidden layers, and predicted as output. The neural network is part of deep learning, a subset of machine learning. Many neutrons, i.e., nods, are well connected from the input to the output layer. Where These neurons represent hidden layers, which use weight. When errors shappen automatically, weights are adjusted and improved continuously. Neural networks are more suitable for solving complex problems like recognizing audio, summaries of documents, face recognition, and object detection. Neural networks are more appropriate for making decisions and helping humans. Any deep learning models learn relations among non-linear and linear complex problems.

LSTM Networks: It's also part of deep learning models. Which helps to predict a sequential manner. The basic structure of LSTM is the RNN model. LSTM has to address the drawbacks of RNN. The phenomena RNN is used to predict the next group of words for the long term, which is not possible in RNN. However, LSTM is developed to address long-term memorization and avoid long-term dependency. LSTM processes three gates called units. Input data is passed through the forget gate, a second unit is passed through the input gate, and output is predicted through the output gate. The basic structure of LSTM follows a forward network.

4.3 The Proposed Algorithm

In this paper, algorithm 1 (see Figure 4), we present two different algorithms to execute the proposed framework. EAS is applied to extract attributes from the Revdess dataset to predict eight classes mentioned in Figure. Another algorithm (LSTM-SER) is proposed to compute datasets to achieve emotion recognition. In EAS, the proposed algorithm reviews data provided as input R and attribute N as output.

In algorithm 2 (see Figure 5), then applying MFCC as D1, STFT as D2, and Mel as D3, attribute D is formed. C is mapped to attribute and repeat F1, F2, and F3 to retrieve from MFCC, Mel, and STFT features mentioned in Figure 2. Extracted feature D collects the result added to C and repeats up to N. C is considered an overall attribute in the second proposed algorithm, A1 is trained, and A2 tests data. B is the output function, and C is the excused data from EAS, which is applied to the MLP method through C to obtain values. Save the overall model in C, then test A2 with EAS, which is applied and combined to C. Predict the class label in c and update until B. Performance metrics are used to validate, and the final results are displayed. Building model:

– Defining a neural network as a series of sequential layers.

Input: REVDESS dataset R
Output: Attribute N
1. start
2. Apply MFCC attribute value D_1
3. Apply STFT attribute D_2
4. Apply Mel attribute value D_3
5. Apply Attribute D
6. Apply attribute C
7. For each d in R
8. F_1←retrieves the MFCC attribute
9. F_2 ←retrieves the STFT attribute
10. F_3← retrieves Mel attribute
11. **D←D_1+D_2+D_3**
12. ADD d and D to C
13. termination
14. remit N
15. close iterations

Fig. 4: Ensemble Attribute Selection Algorithm (EAS)

Input: Raydees train data A_1
Raydess test data A_2
Output: Speech Recognition output B
1. Start
2. Consider attribute map C
3. Consider speech recognition output B
4. C←execute[EAS] A_1
5. c← Apply the MLP method through C and obtain values
6. Save and conclude model c
7. C←Execute [EAS] A_2
8. Label←predict speech emotion(c)
9. Update B
10. End for
11. Result B
12. Apply matric evaluation
13. Display the final result
14. End

Fig. 5: LSTM-Based Speech Emotion Recognition (LSTM-SER) Algorithm

–Compiling the network, which converts the layers into an optimized matrix.
–After building a model, it needs to be tuned by adjusting the weights of trained data.
–After Training, the model must be evaluated using a separate data set and predicted performance metrics.
–After assessing the performance metrics, we can use them to predict new data.

4.4 Evaluation Methodology

The proposed framework is determined for its interpretation by applying a confusion matrix. Confusion

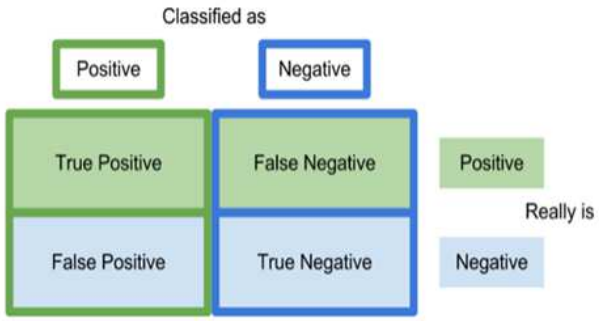


Fig. 6: Confusion Matrix

Metrics are formulated as false positive, false negative, true negative, and true positive. These help to determine the interpretation of the given algorithm in terms of recall, accuracy, precision, and F1 score.

The performance metrics are computed based on an ML model's correct and wrong predictions. Precision and recall are calculated as in Equations 1, 2, 3, and 4.

$$F1\text{–}Score = 2\frac{q*t}{q+t} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

5 Results and Discussion

In this result module, the authors tried to explore the hidden patterns of the SER dataset in detail. Our empirical study is made with an implementation of the framework and the underlying algorithm. The results are provided in terms of exploratory data insights, detection of human emotions, and performance evaluation. Splitting of train and test data is considered to be a logical process because it identifies independent and dependent features. Our dataset consists of 1440 audio files with eight classes. So, Splitting will be a challenging task. To overcome this process, we have applied the random_split method, which is used to divide the train and test set to achieve stratified split [Train 0.7, Test 0.2, and cross-validation 0.1]

5.1 Exploratory Data Analysis

In this sub-section, Figure 7 plots the total number of male vs. female counts for various attributes in the dataset. Female participants are higher than male

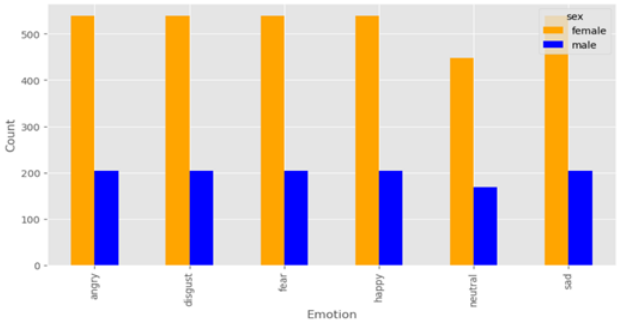


Fig. 7: Distribution of Male and Female Gender in Emotion

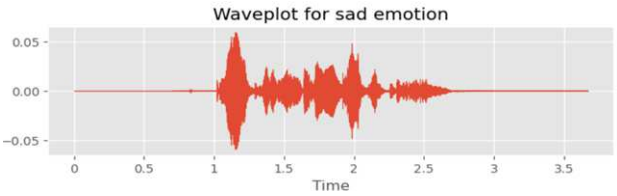


Fig. 8: Example of the Audio Frequency with Sad Signal

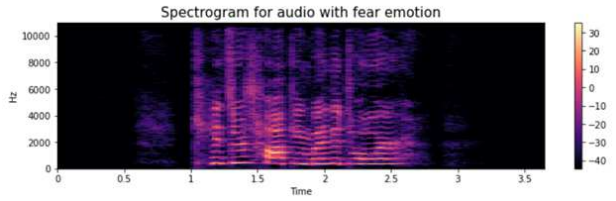


Fig. 9: Example of the Audio Spectrogram with Fear Signal

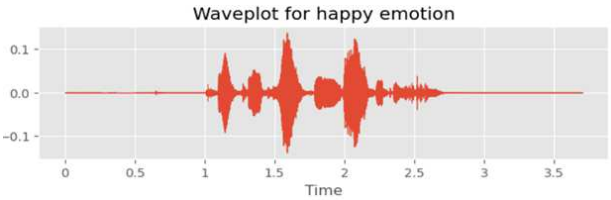


Fig. 10: Example of the Audio Frequency with Happy Signal

participants. In Figure 8, the wave plot for the sad emotion sample exhibits time as the X-axis and frequency as the Y-axis.

In Figure 9, A sample of fear emotion is plotted using a spectrogram using a Fourier transformer and wavelength, time in the X-axis, and Hz in the Y-axis, respectively. Figure 9 plots a sample of wave plots for happy emotion. Spectrograms for audio with angry emotion are plotted in Figure 11.

Figure 14 applies the CNN- LSTM algorithm through 50 epochs. A variation of train and test accuracy in red

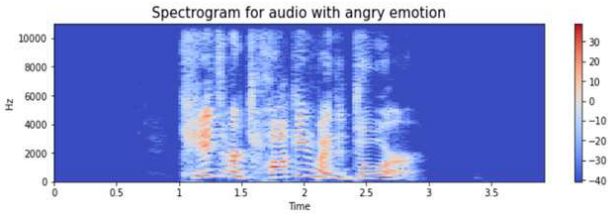


Fig. 11: Example of the Audio Spectrogram with Angry Signal

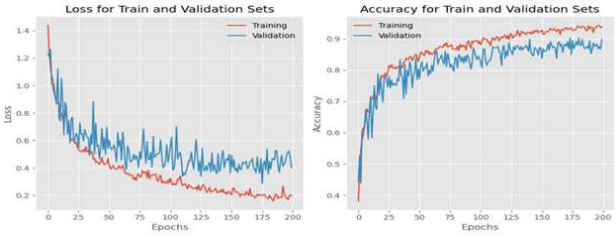


Fig. 12: Difference Between Train and Loss Validation

Layer (type)	Output Shape	Param #
lstm_4 (LSTM)	(None, 352, 64)	20480
lstm_5 (LSTM)	(None, 128)	98816
dense_3 (Dense)	(None, 128)	16512
dense_4 (Dense)	(None, 64)	8256
dense_5 (Dense)	(None, 6)	390

Total params: 144,454
Trainable params: 144,454
Non-trainable params: 0

Fig. 13: LSTM Difference Dense Layer with Params

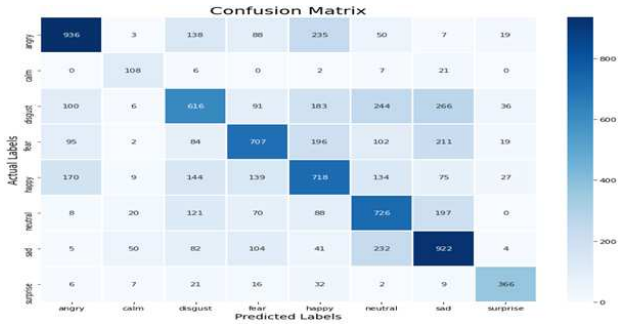


Fig. 14: The Confusion Matrix Reflects Eight Classes of Correlation

and blue graph epochs is plotted on the X-axis, and LOSS accuracy is on the Y-axis. The proposed algorithm performed well in both graphs in the training and testing phase and was validated using Accuracy on both sides.

Table 2: Comparative Table of Proposed Research and the Existing Work

Authors	Feature section techniques (fst)	Audio features	Emotion Lebals classes	ML and DL Algorithms	Evaluation Metrics	Results
Liu et al.	fst by Fisher Criterion and correlation analysis	spectrum features, prosodic and quality shimmer, pitch,	Sad, neutral, Happy, fear, surprise and angry	NN-BP, SVM, KNN	Average Recognition Rate	SVM:79% NN BP:80.1%
Koolagudi et al.	Extraction of Framing with 20 Ms	Spectral, intensity, and jitter	sadness, anger, neutrality, fear, and happiness	VQ, K, ANN, and GMM	Accuracy	ANN: 72%, GMM: 79% VQ: K 57%
Huang et al.	Fourier-transform-based filter bank	LLDs	Valence (Positive and negative)	PCASS, MT, KMM	Emo-DBy, Accuracy	62.52%-MT,63.36 KMM, 63.75%SHLA
Lim et al.	Segmenting signals	MFCCs and log-energy	neutral, state, anger, sadness, boredom, and fear	Baseline: LSTM, RNN, CNN	Accuracy, F1-score and precision	LSTM:81.11%, RNN:78.83%, CNN:78.31%
Jalal et al.	Not mentioned	energy augmented and MFCCs	Disgust -neutral, fearful, calm, angry, happy, and sad	Baselines-CNN	Accuracy	CNN: 71.64%
Proposed CNN-LSTM	**MFCC, Chroma, Mel-spectrogram**	**windowing and framing**	**Neutral state, anger, sadness, boredom, and fear; Neutral, Disgust, Surprised, Fearful, happy, calm, and sad**	**Enhanced CNN**	**Accuracy**	**CNN-LSTM:87.8%**

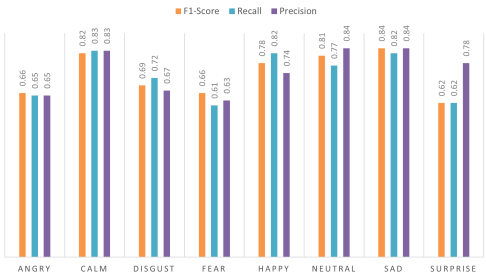


Fig. 15: Emotion Recognition Performance Comparison

5.2 Result and Evaluation

In Table 2, A proposed algorithm has a total number of input layers, hidden layer, and output layer configured. The total output of 144454 params is included as input, as 144454 are trainable params in the CNN-LSTM proposed algorithm. Where non-trainable params are zero, in Figure 10, the confusion matrix is plotted with the correlation between 8 attributes. The SER dataset is higher, and the correlation-densecolor will be a thinker. Table 5 The proposed algorithm has performed precision, F1-score, and Recall, respectively, through the different classes of SER datasets. In Figure 14, a comparison of other performances and SER recognition is plotted.

Table 3: Presenting Evaluation Metrics of Different Classes for Speech Emotion Identification

	F1-Score	Recall	Precision
Sad	0.66	0.65	0.65
Calm	0.82	0.83	0.83
Fear	0.69	0.72	0.67
Neutral	0.66	0.61	0.63
Disgust	0.78	0.82	0.74
Surprise	0.81	0.77	0.84
Angry	0.84	0.82	0.84
Happy	0.62	0.62	0.78

As presented in Figure **??**, the accuracy of the proposed enhanced CNN-LSTM model is compared against existing models such as the CNN baseline and LSTM. The proposed model outperforms the existing ones. The accuracy of the CNN baseline is 81.11%, while LSTM showed 84.01% accuracy. The proposed enhanced CNN-LSTM model exhibits the highest accuracy, with 87.08

Table 4: Comparison of Different Speech Recognition Models

Models	Accuracy (%)
Bhavan et al. [2] CNN baseline	81.11
Kattel, et al. [6] LSTM	84.01
Proposed	87.08

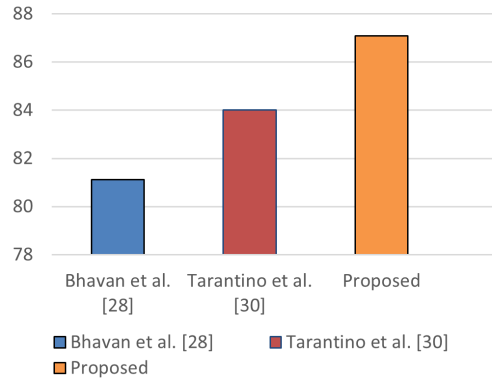


Fig. 16: Different Accuracy of Models

6 Conclusion and Future works

The first investigation uses prosodic features such as pitch, and a neural network is used as a classifier. This approach achieves higher Accuracy than existing algorithms despite having a smaller feature vector dimension. The study explores the use of magnitude and phase spectral features, demonstrating that phase information influences SER performance. Bottleneck features are obtained using a UNET-based autoencoder for information and magnitude, and the final feature vector is constructed by fusing these features using Deep

Canonical Correlation Analysis (DCCA). Future scope for research includes exploring more advanced multi-modal approaches combining audio and visual features and exploring transfer learning techniques for improving performance across different datasets or domains. The final contribution of this paperwork consists of three parts: developing CNN-LSTM architectures for feature extraction, implementing an attention mechanism for selecting relevant features, creating an application for selecting relevant features, and creating an application for real-time scenarios. We also evaluated the performance of your system using multichannel CNN-LSTM networks with and without attention mechanisms: fused magnitude and phase spectral features using DCCA.The proposed CNLSTM model achieves the highest accuracy with 87.08% accuracy, while existing models, such as CNN baseline and LSTM models, showed 81.11% and 84.01%, respectively.Our approach shows improved Accuracy compared to existing works, especially for real-time applications.

Acknowledgement

The authors thank all the respondents who provided valuable responses and support for the survey. They offer special gratitude to INTI International for publishing the research work, particularly to INTI International University for funding its publication, and acknowledge the partial funding support provided by the Electronic Marketing and Social Media Department, Economic and Administrative Sciences, Zarqa University.

Funding

The authors offer special gratitude to INTI International University for the opportunity to conduct research and publish the research work. In particular, the authors would like to thank INTI International University for funding the publication of this research work. Also, we extend our heartfelt gratitude to all research participants for their valuable contributions, which have been integral to the success of this study.

Conflict of Interest

The authors have no conflict of interest to declare.

References

[1] A.M. Al-Adamat, M.K. Alserhan, L.S. Mohammad, D. Singh, S.I.S. Al-Hawary, A.A. Mohammad, M.F. Hunitie, The Impact of Digital Marketing Tools on Customer Loyalty of Jordanian Islamic Banks. In *Emerging Trends and Innovation in Business and Finance* (pp. 105-118). Singapore: Springer Nature Singapore (2023).

[2] M.S. Al-Batah, E.R. Al-Kwaldeh, M. Abdel Wahed, M. Alzyoud, N. Al-Shanableh, Enhancement over DBSCAN Satellite Spatial Data Clustering. *Journal of Electrical and Computer Engineering*,**2024**, 2330624 (2023).

[3] M.S. Al-Batah, M.S. Alzboon, M. Alzyoud, N. Al-Shanableh, Enhancing Image Cryptography Performance with Block Left Rotation Operations. *Applied Computational Intelligence and Soft Computing*, **2024**, 3641927 (2024).

[4] M. Bouazza, A. AlSsaideh, The impact of the digital economy on enhancing the quality of banking services an application study on Islamic banks operating in Jordan. *Al-Balqa Journal for Research and Studies*,**26**, 89-107 (2024).

[5] F.M. Aldaihani, A.A. Mohammad, H. AlChahadat, S.I.S. Al-Hawary, M.F. Almaaitah, N.A. Al-Husban, A. Mohammad, Customers' perception of the social responsibility in the private hospitals in Greater Amman. In *The effect of information technology on business and marketing intelligence systems* (pp. 2177-2191). Cham: Springer International Publishing (2023).

[6] F.A. Al-Fakeh, M.S. Al-Shaikh, S.I.S. Al-Hawary, L.S. Mohammad, D. Singh, A.A. Mohammad, M.H. Al-Safadi, The Impact of Integrated Marketing Communications Tools on Achieving Competitive Advantage in Jordanian Universities. In *Emerging Trends and Innovation in Business and Finance* (pp. 149-165). Singapore: Springer Nature Singapore (2023).

[7] D.A. Al-Husban, S.I.S. Al-Hawary, I.R. AlTaweel, N.A. Al-Husban, M.F. Almaaitah, F.M. Aldaihani, D.I. Mohammad, The impact of intellectual capital on competitive capabilities: evidence from firms listed in ASE. In *The effect of information technology on business and marketing intelligence systems* (pp. 1707-1723). Cham: Springer International Publishing (2023).

[8] M.M. Abualhaj, Q.Y. Shambour, A. Alsaaidah, A. Abu-Shareha, S. Al-Khatib, M.O. Hiari, Enhancing Spam Detection Using Hybrid of Harris Hawks and Firefly Optimization Algorithms. *Journal of Applied Data Sciences*,**5**, 901-911 (2024).

[9] M.I. Alkhawaldeh, F.M. Aldaihani, B.A. Al-Zyoud, S.I.S. Al-Hawary, N.A. Shamaileh, A.A. Mohammad, O.A. Al-Adamat, Impact of internal marketing practices on intention to stay in commercial banks in Jordan. In *The effect of information technology on business and marketing intelligence systems* (pp. 2231-2247). Cham: Springer International Publishing (2023).

[10] N. Al-shanableh, M. Alzyoud, R.Y. Al-husban, N.M. Alshanableh, A. Al-Oun, M.S. Al-Batah, S. Alzboon, Advanced Ensemble Machine Learning Techniques for Optimizing Diabetes Mellitus Prognostication: A Detailed Examination of Hospital Data. *Data and Metadata*,**3**, 363-363 (2024).

[11] N. Al-shanableh, M. Alzyoud, E. Nashnush, Enhancing Email Spam Detection Through Ensemble Machine Learning: A Comprehensive Evaluation Of Model Integration And Performance. *Communications of the IIMA*,*22*, 2 (2024).

[12] A.A. Mohammad, I.A. Khanfar, B. Al Oraini, A. Vasudevan, I.M. Suleiman, Z. Fei, Predictive analytics on artificial intelligence in supply chain optimization. *Data and Metadata*,**3**, 395-395 (2024).

[13] A.A. Mohammad, F.L. Aityassine, Z.N. Al-fugaha, M. Alshurideh, N.S. Alajarmeh, A.A. Al-Momani, A.M. Al-

Adamat, The Impact of Influencer Marketing on Brand Perception: A Study of Jordanian Customers Influenced on Social Media Platforms. In *Business Analytical Capabilities and Artificial Intelligence-Enabled Analytics: Applications and Challenges in the Digital Era* (pp. 363-376). Cham: Springer Nature Switzerland (2024).

[14] M.S. Alshura, S.S. Tayeh, Y.S. Melhem, F.N. Al-Shaikh, H.M. Almomani, F.L. Aityassine, A.A. Mohammad, Authentic leadership and its impact on sustainable performance: the mediating role of knowledge ability in Jordan customs department. In *The effect of information technology on business and marketing intelligence systems* (pp. 1437-1454). Cham: Springer International Publishing (2023).

[15] A.M. Alsaaidah, Q.Y. Shambour, M.M. Abualhaj, A.A. Abu-Shareha, A novel approach for e-health recommender systems. *Bulletin of Electrical Engineering and Informatics*,**13**, 2902-2912 (2024).

[16] A.A. Mohammad, I.A. Khanfar, B. Al Oraini, A. Vasudevan, I.M. Suleiman, M. Ala'a, User acceptance of health information technologies (HIT): an application of the theory of planned behavior. *Data and Metadata*,**3**, 394-394 (2024).

[17] L. Shokr, D. AlAgry, S. Al-Sagga, Critical Assessment of Core Self-Evaluations Theory. *Al-Balqa Journal for Research and Studies*,**25**, 162-184 (2022).

[18] A.A. Mohammad, M.Y. Barghouth, N.A. Al-Husban, F.M. Aldaihani, D.A. Al-Husban, A.A. Lemoun, S.I.S. Al-Hawary, Does Social Media Marketing Affect Marketing Performance. In *Emerging Trends and Innovation in Business and Finance* (pp. 21-34). Singapore: Springer Nature Singapore (2023).

[19] A.A. Mohammad, M.M. Al-Qasem, S.M. Khodeer, F.M. Aldaihani, A.F. Alserhan, A.A. Haija, S.I.S. Al-Hawary, Effect of Green Branding on Customers Green Consciousness Toward Green Technology. In *Emerging Trends and Innovation in Business and Finance* (pp. 35-48). Singapore: Springer Nature Singapore (2023).

[20] B. Ewaida, The Status of Arabic Language in the Social media (Challenges and Proposals): Facebook as a case study. *Al-Balqa Journal for Research and Studies*,**25**, 93-112 (2022).

[21] L.H. Baniata, S. Kang, M.A. Alsharaiah, M.H. Baniata, Advanced Deep Learning Model for Predicting the Academic Performances of Students in Educational Institutions. *Applied Sciences*,**14**, 1963 (2024).

[22] N. Al-Shanableh, M. Al-Zyoud, R.Y. Al-Husban, N. Al-Shdayfat, J.F. Alkhawaldeh, N.S. Alajarmeh, S.I.S. Al-Hawary, Data Mining to Reveal Factors Associated with Quality of life among Jordanian Women with Breast Cancer. *Appl. Math.*,**18**, 403-408 (2024).

[23] N. Al-shanableh, S. Anagreh, A.A. Haija, M. Alzyoud, M. Azzam, H.M. Maabreh, S.I.S. Al-Hawary, The Adoption of RegTech in Enhancing Tax Compliance: Evidence from Telecommunication Companies in Jordan. In *Business Analytical Capabilities and Artificial Intelligence-enabled Analytics: Applications and Challenges in the Digital Era* (pp. 181-195). Cham: Springer Nature Switzerland (2024).

[24] M.A. Aljubouri, M.Z. Iskandarani, Comparative analysis of coding schemes for effective wireless communication. *Indonesian Journal of Electrical Engineering and Computer Science*,**34**, 926-940 (2024).

[25] F. AlFaouri, M. Afif, Punishment and precautionary measures. *Al-Balqa Journal for Research and Studies*,**24**, 158-173 (2021).

[26] E. Batziou, E. Michail, K. Avgerinakis, S. Vrochidis, I. Patras, I. Kompatsiaris, *Visual and audio analysis of movies video for emotion detection*. In Emotional Impact of Movies task MediaEval (2018).

[27] M. Chen, P. Zhou, G. Fortino, Emotion communication system. *IEEE access*,**5**, 326-337 (2016).

[28] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*,**59**, 103-126 (2020).

[29] U.R. Vamsi, B.Y. Chowdhary, M. Harshitha, S.R. Theja, J.D. Udayan, Speech emotion recognition (ser) using multilayer perceptron and deep learning techniques. *IEEE Access*,**27**, 386-394 (2021).

[30] A. Shah, M. Kattel, A. Nepal, D. Shrestha, Chroma feature extraction. Chroma Feature Extraction using Fourier Transform, *Encyclopedia of GIS*, no. January, pp. 1–9 (2019).

[31] C.A. Frantzidis, C. Bratsas, M.A. Klados, E. Konstantinidis, C.D. Lithari, A.B. Vivas, P.D. Bamidis, On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*,**14**, 309-318 (2010).

[32] L. Tarantino, P.N. Garner, A. Lazaridis, Self-Attention for Speech Emotion Recognition. In *Interspeech* (pp. 2578-2582) (2019).

[33] D. Jiang, K. Wu, D. Chen, G. Tu, T. Zhou, A. Garg, L. Gao, A probability and integrated learning based classification algorithm for high-level human emotion recognition problems. *Measurement*,**150**, 107049 (2020).

[34] M. Xu, F. Zhang, W. Zhang, Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. *IEEE Access*,**9**, 74539-74549 (2021).

[35] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, G. Dai, Pre-trained deep convolution neural network model with attention for speech emotion recognition. *Frontiers in Physiology*,**12**, 643202 (2021).

[36] S.R. Livingstone, F.A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*,**13**, e0196391 (2018).

[37] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, A. Košir, Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*,**46**, 184-192 (2019).

[38] J.Z. Lim, J. Mountstephens, J. Teo, Emotion recognition using eye-tracking: taxonomy, review and current challenges. *Sensors*,**20**, 2384 (2020).

[39] Y. Li, T. Zhao, T. Kawahara, Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In *Interspeech* (pp. 2803-2807) (2019).

[40] A. Koduru, H.B. Valiveti, A.K. Budati, Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*,**23**, 45-55 (2020).

[41] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, X. Li, Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv*,1909.05645 (2019).

[42] A. Bhavan, P. Chauhan, R.R. Shah, Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*,**184**, 104886 (2019).

Shaik A. K. Basha is a faculty member at the School of Computer Science Engineering and Information Systems at Vellore Institute of Technology, Vellore, India. His expertise encompasses Information Extraction, Data Mining, Knowledge Discovery, Social Network Analysis, Clustering, and Large-Scale Data Analysis. With a focus on advanced data mining techniques and distributed data analytics, Dr. Basha has contributed to the field through six publications, demonstrating his commitment to research and innovation in data-driven methodologies. His Orcid ID is orcid.org/0000-0003-4742-8699.

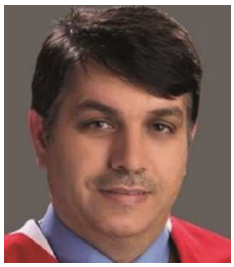
P. M. D. R. Vincent is an Associate Professor in the School of Computer Science Engineering and Information Systems at Vellore Institute of Technology (VIT), Vellore, India. With over 13 years of teaching and research experience, he holds a B.E. in Electronics, an M.E. in Computer Science from Anna University, and a Ph.D. from VIT (2015). He specializes in machine learning, IoT, data analytics, and security, and has published more than 50 papers indexed in Scopus. Known for his strong industry connections, Dr. Vincent has led training sessions for Wipro employees, served on doctoral committees, and given keynote addresses at various conferences. He is also a reviewer for esteemed journals, including IEEE Access. His Orcid ID is orcid.org/0000-0002-7598-1363.

Suleiman Ibrahim Mohammad is a Professor of Business Management at Al al-Bayt University, Jordan (currently at Zarqa University, Jordan), with more than 17 years of teaching experience. He has published over 100 research papers in prestigious journals. He holds a PhD in Financial Management and an MCom from Rajasthan University, India, and a Bachelor's in Commerce from Yarmouk University, Jordan. His research interests focus on supply chain management, Marketing, and total quality (TQ). His ORCID ID is orcid.org/0000-0001-6156-9063.

Asokan Vasudevan is a distinguished academic at INTI International University, Malaysia. He holds multiple degrees, including a PhD in Management from UNITEN, Malaysia, and has held key roles such as Lecturer, Department Chair, and Program Director. His research, published in esteemed journals, focuses on business management, ethics, and leadership. Dr. Vasudevan has received several awards, including the Best Lecturer Award from Infrastructure University Kuala Lumpur and the Teaching Excellence Award from INTI International University. His ORCID ID is orcid.org/0000-0002-9866-4045.

Eddie Eu Hui Soon is a Senior Lecturer at INTI International University with over 20 years of experience in academia and the animation industry. Before academia, he worked as a Technical Director in Malaysian production houses, contributing to TV commercials, series, feature films, and corporate videos. He continues to consult in the animation and gaming industry, specializing in 3D cinematic design. His research spans transdisciplinary topics, including Graph Theory, Systems Design, and digital frameworks. Dr. Soon is also involved in prototyping and visualization at the university's fabrication lab and supports research initiatives through journal and website management. His ORCID ID is orcid.org/0000-0002-3154-3943.

Qusai Shambour is affiliated with the Laboratory of Decision Systems and e-Service Intelligence, within the Centre for Quantum Computation and Intelligent Systems at the University of Technology Sydney. He is part of the School of Software in the Faculty of Engineering and Information Technology. His research primarily focuses on recommender systems, collaborative filtering, multi-criteria decision-making, and fuzzy logic. Dr. Shambour explores topics such as recommendation accuracy, semantic similarity, user preferences, and the cold-start problem in recommendation approaches. His work is pivotal in addressing issues like information overload and enhancing the quality of personalized recommendations in online services and social networks. His Orcid ID is orcid.org/0000-0002-3026-845X.



Muhammad Turki Alshurideh is a faculty member at the School of Business at the University of Jordan and the College of Business Administration, at the University of Sharjah, UAE. He teaches a variety of Marketing and Business courses to both undergraduate and postgraduate students. With over 170 published papers, his research focuses primarily on Customer Relationship Management (CRM) and customer retention. His ORCID ID is orcid.org/0000-0002-7336-381X.