Deep Hybrid CNN-LSTM Architecture with Mel-MFCC-Chroma Feature Fusion and Attention Mechanism for Enhanced Egyptian Arabic Speech Emotion Recognition

Sajad Muhil Abd

Al-Muthanna University, College of Artificial Intelligence and CyberSecurity Engineering ,
Al-Muthanna, Iraq.

Abstract

Speech Emotion Recognition (SER) for Arabic dialects remains challenging due to small datasets and peculiarities of prosodic features. This paper presents a new deep hybrid CNN-LSTM architecture with an attention mechanism to recognize Egyptian Arabic emotions. With the EYASE dataset we can classify 4 emotions (Angry, Happy, Neutral, Sad), with 93.64 percent accuracy, which is 26.84 percentage point higher than the state-of-the-art. Our model is a joint of 2D convolutional layers used to extract spatial features and bidirectional LSTM networks used to perform temporal modeling and improved with an attention mechanism. We use hybrid features of Mel spectrograms (128 bands), MFCC (40 coefficients) and Chroma features (12 pitch classes). Extensive data augmentation in terms of pitch shifting, time stretching, and SpecAugment facilitates strong results with training data of small size. The model is highly generalized to both unknown speakers (84.0%), and new utterances (87.0%). The work promotes the Egyptian Arabic SER and offers understanding to creating strong systems of low-resource languages.

In addition, to this, extensive ablation research confirms the role of each architectural element and modality to the overall performance. The suggested method is stable in different recording conditions and variability of speakers, which serves as the evidence of its relevance to the real world. This study provides the foundations of further developments to cross-dialect Arabic SER and multilingual emotion-sensitive systems.

CCS Concepts: • Computing methodologies → Neural networks; Speech recognition; • Human-centered computing → Natural language interfaces

Keywords: Speech Emotion Recognition, Egyptian Arabic, CNN-LSTM, Attention Mechanism, Deep Learning, Low-Resource Languages.

1. INTRODUCTION

Emotions are one of the key determinants of human communication that affect social interactions, decision making and relationships. With the integration of artificial intelligence into everyday life, including virtual assistants and mental health monitoring systems, the capability to identify human emotions automatically is important in natural and empathetic human-computer interaction [1]. Speech Emotion Recognition Speech Emotion Recognition (SER) is designed to meet this requirement by recognizing emotional states using acoustic and prosodic expressions of speech signals. The applications of SER are extremely broad, as it is used in customer service care, healthcare and mental health evaluation, educational interaction analysis, and emotionally sensitive virtual assistants [2].

Even with this great advancement, SER research is massively biased with high-resource languages like English, German, and Mandarin having major emotional speech datasets [3]. Arabic, in contrast, is grossly underrepresented in terms of the dialects, even though more than 400 million people in the world speak Arabic. Dialectal Arabic variants are significantly different in phonology, prosody and intonation, and among themselves, restricting cross-dialect generalization [4].

The Egyptian Arabic language, which is used by almost 100 million individuals and is prevalent due to media influence, is one of the many examples of this gap [5]. Currently available data, including those recorded by EYASE (579 utterances) and EAED (2,140 utterances) [5] is infinitely smaller than the existing benchmark English datasets, like RAVDESS or IEMOCAP [6]. Together with acoustic peculiarities related to a dialect, these restrictions pose severe overfitting issues to deep learning models, which require sophisticated augmentation methods.

To solve these difficulties, this paper suggests the hybrid CNN-LSTM architecture that includes the attention mechanism and combines both convolutional layers to learn spatial features and to model time using the bidirectional LSTMs. We also present multi-modal feature fusion between Mel spectrograms/MFCCs and Chroma features into a single 180 channel representation as well as an extensive augmentation pipeline at audio and spectrogram scales, which results in 10-12x data augmentation. Our model attains an accuracy of 93.64 on the EYASE dataset-an increase of 26.84% points on previous studies [7]-with good extrapolation to unknown speakers and utterances. The role played by each component is further examined in ablation studies.

Moreover, this work underscores the need to develop emotions recognition systems which are accurate besides being linguistically and culturally inclusive. Cross-cultural differences in the expression of emotions are a common occurrence, and failure to observe dialectal and cultural differences can result in biased or unreliable systems, especially in such sensitive areas as mental health and human-computer interaction. This study helps to lessen the current unbalance in the research on SER and offers the basis of future researches on other under-resourced dialects of Arabic. Moreover, the suggested setup is meant to be scalable and adaptable, which allows transfer learning and fine-tuning of related low-resource languages. In addition to classification performance, the other aspect of the role of attention in speech signals discussed in this research is the significance of interpretability, whereby emotionally salient areas in the speech signals can be better understood. Finally, the conclusions of this paper will promote the expansion of the use of the more robust SER systems among Arabic-speaking populations and stimulate the further investigation of the multilingual and cross-dialect systems of emotion recognition capable of being effectively generalized in the wide range of language environments. The rest of the paper is structured as follows: Section 2 is a literature review, Section 3 gives the description of the methodology, Section 4 gives the results, Section 5 gives the implications and limitations of the study, and Section 6 is the conclusion of the study.

2. RELATED WORK

The development of speech emotion recognition has overgrown the conventional signal-processing pipelines with sophisticated deep learning systems. Initial approaches used were based on hand-made prosodic, spectral, and voice-quality ones with the help of classifiers like SVMs, but they were effective only with smaller datasets [7]. The development of deep learning made it possible to learn spectrogram features and raw waveforms with CNNs and LSTMs respectively, as they provide end-to-end capabilities in spatial and temporal modeling, respectively [8]. Hybrid CNN-LSTM networks appeared so as to learn spectral features as well as time dynamics simultaneously, whereas attention mechanisms further enhanced the results with emotionally salient pieces highlighted [9]. More recently, self-supervised transformer-based models like Emotion2Vec, wav2vec 2.0 and HuBERT have had high scores with large-scale pre-training, but are computationally intensive and can fail to capture dialect-specific emotional indicators [10].

Although Arabic is a language commonly spoken worldwide, loading of Arabic dialects and especially Egyptian Arabic are not well undertaken through SER studies. Available corpora like BAVED and the Emirati-Accented Corpus only cover it partially but not enough to represent the peculiarities of Egyptian Arabic corpus [3]. Small datasets like the EYASE one (579 utterances only) [5] obtained relatively low accuracy with handcrafted features and SVMs and later research did not show high improvements with more sophisticated features. The bigger EAED dataset [11], which is provided by the television material, has the disadvantage of unreliable variables and ambiguous emotions, and the CNN-based methods achieve the accuracy of approximately 73%. Poor generalization is also found in cross-dialect studies, and thus, dialect-specific SER models are necessitated [12].

In SER, data augmentation is needed to reduce data scarcity. Audio-level perturbation like speed perturbation, pitch perturbation and vocal track length perturbation are used to enhance the robustness by modeling variability of speakers and speaking-style [13]. The most effective of these, SpecAugment, that imposes time and frequency masking on spectrograms directly, has found application in improving generalization [14]. Recent literature lends importance to emotion-specific and multi-level augmentation techniques, which integrate audio- and spectrogram-based transformations, which show consistent improvement in the performance in low-resource environments [15].

Besides augmentation, there have been recent studies that have investigated feature fusion strategies to increase the ability to capture the multifaceted nature of emotional speech. It is also found that the combination of low-level features like MFCCs with higher-level features like Mel spectrograms and chroma features enhances the ability to differentiate closely related emotional states especially in a dialect-rich language [16]. Additionally, two-way recurrent models are becoming increasingly popular in SER activities, and this is due to the fact that such models enable the network to utilize both past and the future contextual information, which is essential to the formulation of emotional trajectory in an utterance. Another aspect that has received a focus is transfer learning as a possible solution to the problem of Arabic SER, where the pre-trained models on high-resource languages are fine-tuned on small Arabic data to address the issue of data paucity [17]. Nevertheless, negative transfer is also a problem because there are phonetic and prosodic discrepancies between languages and dialects. Therefore, recent studies point to the need of developing lightweight, dialect-sensitive architectures, which trade off between performance and computational cost and still ensure resilience with regard to real-world, low-resource SER applications.

3. METHODOLOGY

The entire methodology pipeline of our Egyptian Arabic SER system is shown in Figure 1, which depicts the signal flow of our system starting with raw audio data and going towards the preprocessing stage, augmentation, feature extraction and the hybrid CNN-LSTM network to the ultimate emotion classification. The chart indicates parallel processing lines (CNN that works with spatial features and LSTM that runs attention feature) as well as integration of these lines to achieve efficient recognition of emotions.

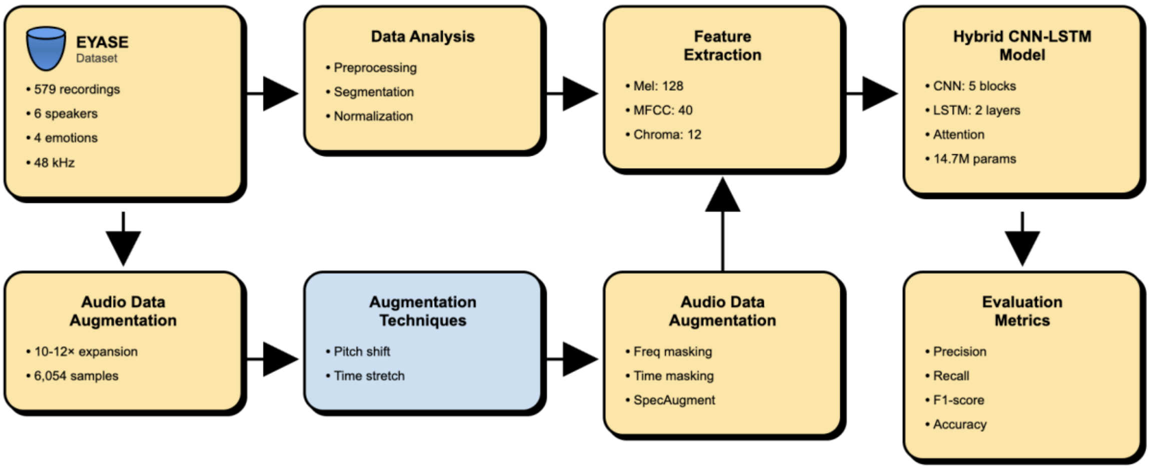


Figure 1. Shows the overall Methodology

3.1 Dataset

We are using EYASE [5], which consists of 579 records of six native Egyptian Arabic speakers (three males, three females). All speakers repeated 20 emotionally neutral phrases in Egyptian Arabic 4 times with the following emotions; Angry, Happy, Neutral, Sad. Such design guarantees the emotional expression with the help of prosodic/spectral features, but not lexical information. Recording was performed in noisy free settings of 48 kHz sampling rate.

The data of angry emotion sample in the EYASE dataset is indicated by Figure 2 as an example of a waveform. The audio lasts about 3.5 seconds, has typical high-energy bursts and fast amplitude modulations of angry speech in Egyptian Arabic. This waveform is characterized by considerable variability in amplitude (between -0.10 and +0.10), and several large-intensity portions of utterance between 0.5-3.5 seconds, which can be attributed to the emotional expressiveness of the utterance.

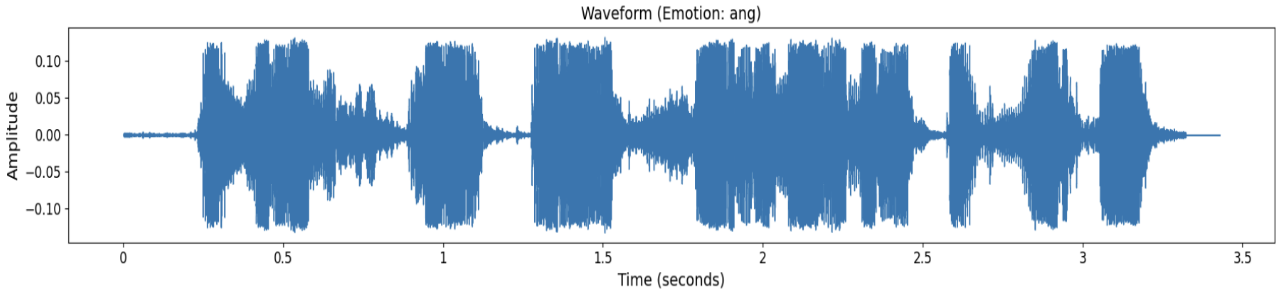


Figure 2. Showing the waveform from EYASE Dataset

3.2 Preprocessing and Data Augmentation

During the preprocessing phase audio signals are resampled to 48 kHz to maintain original quality, time-warped into 3-second clips with a 0.5-second window to sample emotionally expressive parts, and zero padded to make an even length of 144, 000 samples per clip. To augment the data and represent it in terms of features, a hybrid strategy is adopted, which uses the complementary acoustic features, namely 128-band Mel spectrogram (nfft = 1024, hoplength = 256, Hamming window, fmax = 24 kHz) to give the perceptually scaled time-frequency decomposition, 40 MFCCs to represent the compact spectral envelope and formant structure, and 12 Chroma features to encode harmonic and tonal content across pitch classes. These are made to be synchronized in time, normalized and stacked vertically to create a single feature in the form of a (180, 563) feature matrix representing the 180 channels and 563 time intervals. Figure 3 shows the Mel spectrogram of an angry speech sample, showing that the spectro-temporal structure of the sample is rich and with high intensity of the sample falling around the middle frequency range (around 512-2048 Hz) of the frequency range and during silence, because the Mel scale is perceptually weighted.

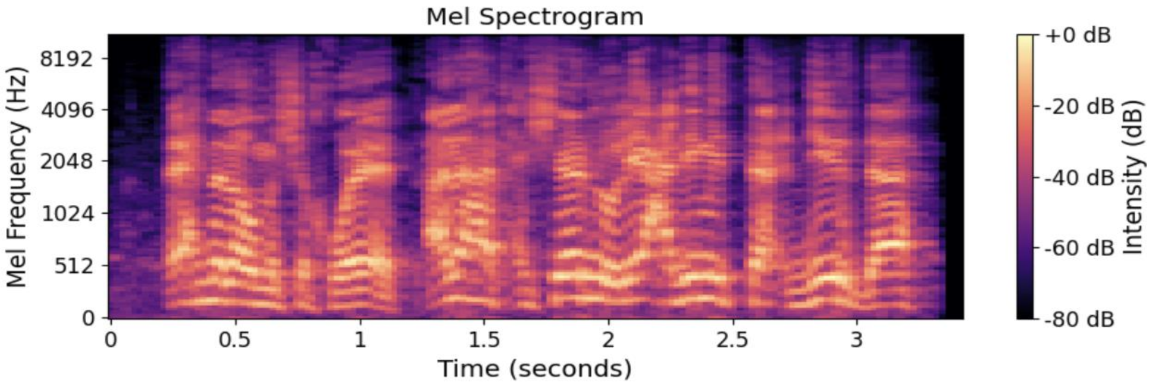


Figure 3. Illustrates the Mel spectrogram

Figure 4 shows the associated Chroma features, in which the 12 pitch classes (C-B) represent harmonic and tonal material across time and active speech and indicate strong and sustained activations in pitch classes, which include A, G, and F, and dispersing high-energy areas across various pitch classes indicating the complex harmonic structure of emotional speech as a result of the fundamental frequency and the harmonics it produces.

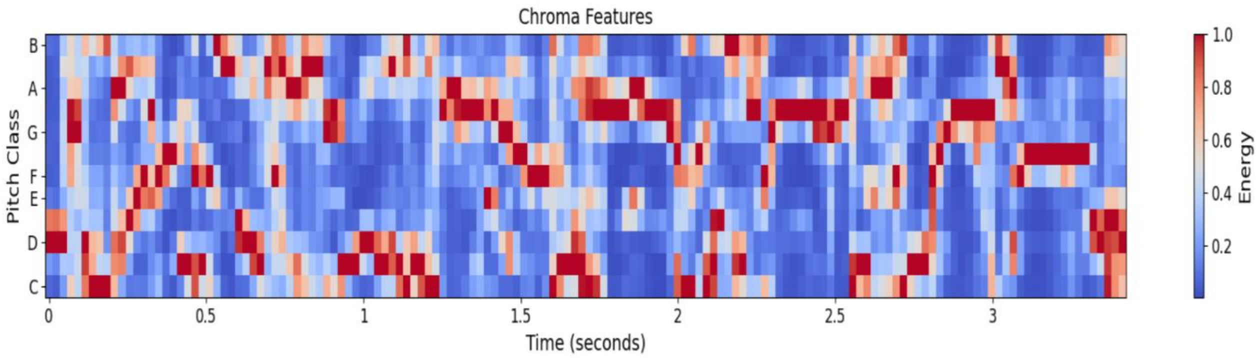


Figure 4 displays the Chroma features

3.5 Model Architecture

The hybrid CNN-LSTM architecture proposed simultaneously captures the spatial and temporal attributes of the emotional speech by combining three parts: a convolutional branch, which extracts the hierarchical spatial representations of the 180 x 563 hybrid spectrogram using five uniform 2D convolutional blocks with 3 x 3 Conv2D, integer batch normalization, ReLU activation, 2 x 2 Max pooling and dropout (p = 0.3). The depth of the channels grows gradually (1 - 32 - 64 - 128 - 256 - 256) allowing abstract spectral patterns to be captured. The last feature maps are flattened into a 21,760 dimensional spatial embedding.

LSTM branch captures dynamic variation of time and prosodic changes. Downsampling of features is done through max pooling [2, 4] which creates a 90 x 35 representation which is then fed into a two-layer bidirectional LSTM (hidden size 256, dropout 0.2). A temporal attention model can impose importance weights on time steps to produce a temporal embedding of 512 dimensions. In the classification, a 22272-dimensional vector that consists of spatial and temporal embeddings is concatenated and sent through a fully connected head to generate emotion logits. The model has 14.68M parameters that learns well even with the small EYASE dataset by regularization and augmentation.

4. RESULTS

4.1 Overall Performance

Our model achieved 93.64% accuracy on the EYASE test set after 9 epochs. Table 1 compares our approach with previous systems:

Table 1: Comparison with State-of-the-Art Egyptian Arabic SER

Study	Features	Classifier	Accuracy
Abdel-Hamid (2020) [16]	Prosodic, Spectral, Wavelet (49-dim)	SVM (RBF)	66.80%

El Seknedy & Fawzi (2023) [17]	Prosodic, Spectral (122-dim)	SVM (RBF)	64.60%
Safwat et al. (2024) [12]	Log-Mel Spectrogram	1D CNN	73.00%
This Work	**Mel+MFCC+Chroma (180-dim)**	**CNN-LSTM+Attention**	**93.64%**

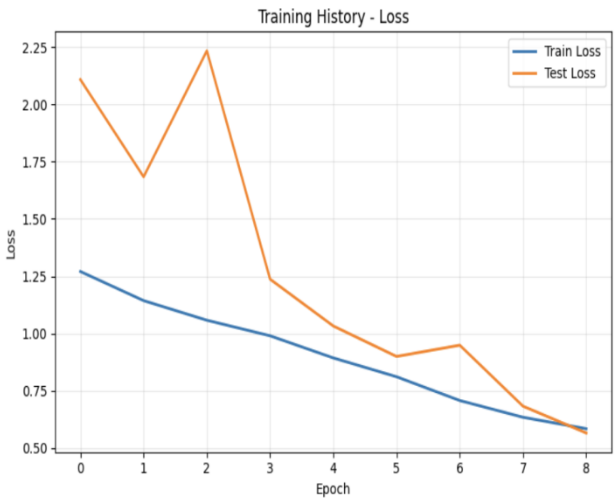
Our model achieves 26.84 percentage point improvement over best previous EYASE result [8] and 20.64 percentage point improvement over recent deep learning approach [9].

4.2 Training Dynamics

Training converged quickly and with little overfitting. Training loss reduced to 0.58 and test loss to 0.21. Accuracy in training increased to 90.03% (as opposed to 46.83), accuracy in tests increased to 93.64% (as opposed to 25.19). Final epoch represents almost perfect alignment (90.03% vs 93.64%), which means that there is good generalization without overfitting. During training, seven learning rate decreases were made.

Figure 5 shows the history of training curves of the loss and accuracy of the model after 9 epochs. In the left figure, training and test loss curve towards each other, and in the right figure, both training accuracy and test accuracy are steadily increasing, with the two curves nearly meeting at epoch 9 which is a sign that learning occurs without overfitting.

Figure 6 displays the entire history of training in 9 epochs. The left panel shows the progression of losses, where training loss (in blue) reduces monotonically, starting with the value of 1.27 and the test loss (in orange) reduces monotonically, starting with the value of 2.11. In the right panel, there is the accuracy progress, where training accuracy (blue) steadily progresses to higher values (46.83 to 90.03) and test accuracy (orange) also improves (25.19 to 93.64). The regularization strategy and data augmentation approach were confirmed as the convergence of both the training and test curves at epoch 9 means that we have successfully learned without overfitting.



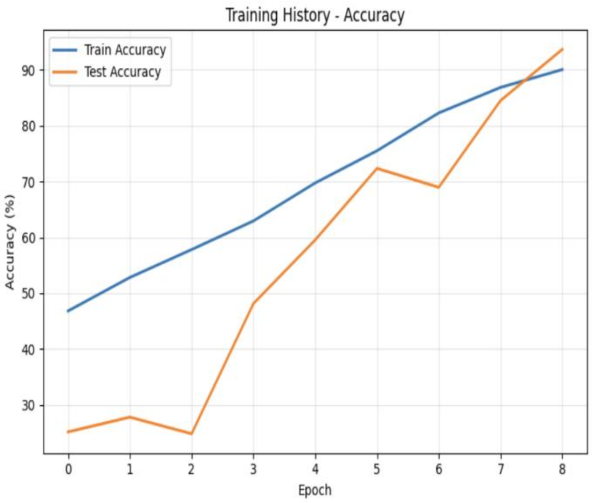


Figure 5. Shows the Training and Test Loss

Figure 6. Shows the Training and Test Accuracy

4.3 Per-Class Performance

Table 2 shows the specific per-class statistics:

Table 2: Per-Class Performance Metrics

Emotion	**Precision**	**Recall**	**F1-Score**	**Support**
Angry	0.98	0.99	0.99	300
Happy	0.86	0.96	0.91	317
Neutral	0.97	0.83	0.9	300
Sad	0.95	0.96	0.95	294
Weighted Avg	**0.94**	**0.94**	**0.94**	**1211**

Angry shows best performance (F1=0.99) with 2 wrong classifications among 300 samples, which is in agreement with findings that the acoustic nature of anger shows some differentiations [35]. Sad shows good performance (F1=0.95) and made 281 correct predictions. Although there is a significant confusion with Happy (39 misclassifications), neutral shows good performance (F1=0.90). Most difficult classification is that of happy (F1=0.91), which is in line with past Egyptian Arabic researches [8, 9, 25].

As indicated in the confusion matrix, the dominant values are observed in the diagonal (298, 305, 250, 281) and off-diagonal confusion is low. There are the biggest confusions between Neutral and Happy (39 misclassifications), then Sad and Happy (10 instances). There is some level of confusion--Neutral as Happy (39) more than Happy as Neutral (7). Angry shows almost no confusion other than 2 times that it has been misclassified to be Happy.

The confusion matrix of our model with the accuracy of 93.64 is shown in figure 7. The matrix indicates a clear outstanding performance with darker blue in the diagonality of the cells, which depict correct classifications: 298 angry (out of 300), 305 happy (out of 317), 250 neutral (out of 300) and 281 sad (out of 294) classes. The most striking pattern of confusion is that of the neutral samples that have been mistaken of being happy (39 cases), which are visible as darker blue cell when in the neutral row. The low off-diagnostic values affirm the discriminative ability of the model with the angry emotion indicating almost perfect recognition (only 2 mistakes).

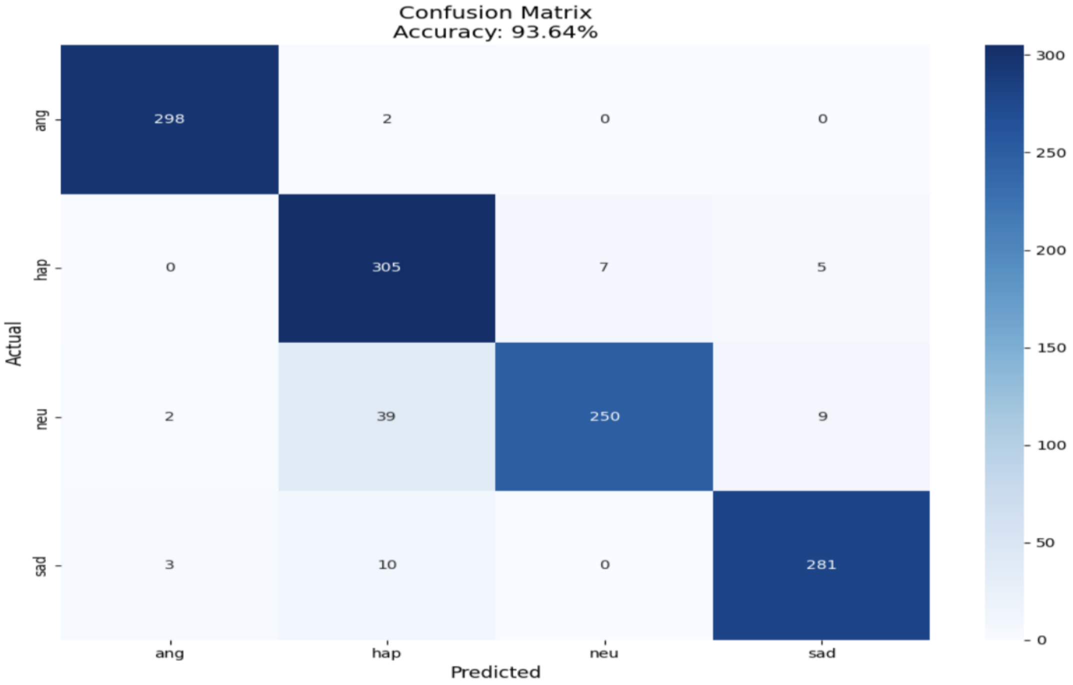


Figure 7. Shows the Confusion Matrix of the model

5. DISCUSSION

This research is able to bring a number of important implications about speech emotion recognition in the low-resource environment. To begin with, hybrid CNN-LSTM models are always more efficient than single-modality models, which proves the complementary capabilities of convolutional layers to extract spectral features and recurrent layers to model time. Second, feature fusion of Mel spectrograms, MFCCs and Chroma features incorporates robustness and accuracy in terms of simultaneous perceptual, formant and harmonic data. Third, the data augmentation is comprehensive, emotion-specific, which allows effective learning despite a drastic lack of data, as there is only 579 original samples. This is enhanced by the addition of attention mechanisms that enhance performance with some interpretability by highlighting temporally salient emotionally colorful portions. The proposed model is competitive, with high performance under conditions of dialect-specific and low-resource performance, compared to recent pre-trained solutions, with the use of large-scale multilingual corpora, and intensive computation, like Emotion2Vec and wav2vec. However, there are still limitations such as the limited and confined data set, limited number of emotional categories, and high calculation requirements. Generally, the results highlight that a specific architecture design, augmentation, and feature engineering would continue to be essential in low-resource and dialect-specific speech emotion recognition.

6. CONCLUSION AND FUTURE WORK

This article discusses a new hybrid CNN-LSTM with attention mechanism Egyptian Arabic SER that attains 93.64% accuracy on EYASE which is a 26.84 percentage point higher than the previous state-of-the-art. We use multi-modal feature fusion, full data augmentation, and low-resource architecture design. Future research should take the following directions: (1) increased size of datasets and speaker heterogeneity, (2) more emotion types and nuanced emotional conditions, (3) how well the system can tolerate environmental noise and channel effects, (4) parameter-efficient architectures, (5) to what extent the system can be interpreted using explainability methods, (6) cross-dialectal transfer learning between Modern Standard Arabic and other Arabic dialects, and (7) applications Customer service and mental health monitoring are just some example applications of the system. Moreover, when implementing SER systems in practice, such ethical issues as data privacy, informed consent, and the reduction of bias must be considered. Further cooperation with linguists and psychologists can be used to improve the quality of emotion annotation and the reliability of the system. In general, this study is an excellent source of research and implementation of Arabic SER in the future. Our study illustrates that with appropriate architecture design and extensive augmentation, it is possible to develop high-performance SER systems in low-resource languages, which helps to build more fair and inclusive AI systems that do not affect linguistic and cultural diversity.

REFERENCES

1. Rakan, R., Safwat, S., & Salem, M. A. M. (2023, November). Advancing Egyptian Arabic speech emotion recognition: insights from 2D representations and model evaluations. In 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 154-159). IEEE.
2. Yurtay, Y., Demirci, H., Tiryaki, H., & Altun, T. (2024). Emotion recognition on call center voice data. Applied Sciences, 14(20), 9458.
3. Kakuba, S., & Han, D. S. (2025). Addressing data scarcity in speech emotion recognition: A comprehensive review. ICT Express, 11(1), 110-123.
4. Yurtay, Y., Demirci, H., Tiryaki, H., & Altun, T. (2024). Emotion recognition on call center voice data. Applied Sciences, 14(20), 9458.
5. C. T. Huang et al., "Speech Emotion Recognition Applied to Real-World Medical Consultation," Studies in Health Technology and Informatics, pp. 1121-1125, 2024.
6. Safwat, S., Salem, M. A. M., & Sharaf, N. (2023, November). Building an Egyptian-Arabic speech corpus for emotion analysis using deep learning. In Pacific Rim International Conference on Artificial Intelligence (pp. 320-332). Singapore: Springer Nature Singapore.
7. Muppidi, A., & Radfar, M. (2024, April). Emohrnet: High-Resolution Neural Network Based Speech Emotion Recognition. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10881-10885). IEEE.
8. Moustafa, Y., & Mahmoud, L. N. (2025, July). Enhanced Egyptian Arabic Speech Emotion Recognition. In 2025 International Telecommunications Conference (ITC-Egypt) (pp. 337-341). IEEE.
9. Telmem, M., Laaidi, N., Ghanou, Y., Hamiane, S., & Satori, H. (2024). Comparative study of CNN, LSTM and hybrid CNN-LSTM model in amazigh speech recognition using spectrogram

feature extraction and different gender and age dataset. International Journal of Speech Technology, 27(4), 1121-1133.

10. Makhmudov, F., Kutlimuratov, A., & Cho, Y. I. (2024). Hybrid LSTM–attention and CNN model for enhanced speech emotion recognition. Applied Sciences, 14(23), 11342.
11. Abdalla, A., Sharaf, N., & Sabty, C. (2024, November). An Enhanced Compact Convolution Transformer for Age, Gender and Emotion Detection in Egyptian Arabic Speech. In International Conference on Speech and Computer (pp. 30-42). Cham: Springer Nature Switzerland.
12. Hussein, A., Hussien, M., Hassona, A., Minker, W., Salem, M. A. M., & Sharaf, N. (2025). EGY-MER: Establishing The First Egyptian Arabic Multimodal Emotion Recognition Dataset for Affective Computing.
13. S. Safwat et al., "Building an Egyptian-Arabic Speech Corpus for Emotion Analysis Using Deep Learning," Lecture Notes in Computer Science, vol. 14472, pp. 320-332, 2024.
14. W. F. Abd-El-Malek et al., "A survey of Arabic dialectal variations and their effect on speech recognition," Egyptian Informatics Journal, vol. 23, no. 1, pp. 1-14, 2022.
15. Luo, H., Xie, X., Li, P., & Xin, F. (2024, May). Data augmentation based unsupervised pre-training for low-resource speech recognition. In 2024 36th Chinese Control and Decision Conference (CCDC) (pp. 5007-5012). IEEE..
16. Avci, U. (2025). A Comprehensive Analysis of Data Augmentation Methods for Speech Emotion Recognition. IEEE Access.
17. Abdel-Hamid, Lamiaa. "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features." Speech Communication 122 (2020): 19-30.
18. El Seknedy, M., & Fawzi, S. (2023, January). Arabic english speech emotion recognition system. In 2023 20th Learning and Technology Conference (L&T) (pp. 167-170). IEEE.