Hybrid CNN-LSTM Architecture with MFCC-Based Class-Balanced Augmentation for Arabic Speech Emotion Recognition

Sarmad H. Alfarag
Electrical Engineering Department, Wasit University, Wasit, Iraq
Email: sarmad.hamad@uowasit.edu.iq (S.H.A.)

Manuscript received January 1, 2026; revised April 1, 2026; accepted April 21, 2026

***Abstract*—Arabic Speech Emotion Recognition (SER) is also associated with serious challenges, including the variety of dialects, excessive unbalanced classes, and the lack of sufficient data. In order to resolve these concerns, the paper provides a comprehensive methodology, which integrates hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architecture with strategic data augmentation. Our method is evaluated with the help of two Arabic emotional speech data sets: Eastern Youngsters Arabic Speech Emotions (EYASE) which is the independent validation set and Berlin Arabic Vocal Emotions Dataset (BAVED) which is the actual training data. Class-balanced augmentation techniques such as pitch shifting, time stretching, and noise injection are used to enhance the grotesquely unbalanced Basic Arabic Vocal Emotions Dataset (BAVED) dataset with approximately 1,600 samples of each emotion class. Our hybrid architecture integrates both the bidirectional Long Short-Term Memory (LSTM) networks and attention mechanism of modelling a temporal sequence, and the convolutional neural network of extracting spatial features of Mel-Frequency Cepstral Coefficient (MFCC) representations. The experimental findings demonstrate that the performance of the POL2 is increased substantially, as the general accuracy on BAVED increases to 97.23% as compared to baseline (89%). Cross-dataset testing of EYASE indicates that the generalization of EYASE is high across different speakers and recording conditions and that this accuracy is 87%. The findings demonstrate that balanced augmentation of data quality has a greater impact on performance than architectural complexity alone, providing useful guidance for the development of Arabic emotion recognition systems in environments with limited resources.**

***Index Terms*—Arabic language, attention mechanism, class imbalance, Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM), data augmentation, low-resource languages, Mel-Frequency Cepstral Coefficient (MFCC) features, speech emotion recognition**

I. Introduction

Speech recognition of human emotion has become a highly significant area of research both at the boundary of affective computing, signal processing, and artificial intelligence due to its potentially transformative interest in a wide variety of diverse applications, such as human-computer interaction, mental health diagnosis, customer service analytics, educational technology, and intelligent virtual assistants. An automatic process of detecting and classifying emotional states by relying on vocal cues allows machines to respond in a more natural and correct manner to human users and build more empathetic and situation-aware systems capable of modifying their own behavior in response to the emotional condition of the user. Speech Emotion Recognition (SER) systems are systems that attempt to determine the underlying emotional state of the speaker using acoustic and prosodic speech signal characteristics taking into account the fact that, each emotion creates an imprint on the basic frequency, energy distribution, speaking rate, voice quality, and spectral properties [1]. Emotion recognition is valuable in various areas of practical activities that make a difference in the society. In medical use, specifically mental health surveillance, the capacity to identify depression, anxiety, or emotional distress based on speech patterns is a non-invasive and scalable screening and surveillance method to supplement the usual clinical evaluation protocols [2].

In the case of depression, an example of acoustic patterns is the lack of difference in pitches, a slower code of speaking, and lower energy levels, and speech analysis is one potential area of research to detect the disease early and provide continuous monitoring. Emotion recognition is one of the aspects in human-robot interaction that allows robots to be aware of human emotive states and respond appropriately, which makes human-machine cooperation more natural and effective [3]. Emotion recognition feature in robots can modify their behavior, communication approach, and strategies of executing certain tasks depending on the emotional state of their human companions, which results in increased user satisfaction, less frustration, and better completion of tasks. Another potential area of application of emotion recognition is educational technology, which is another promising area of application where adaptive systems have the potential to be improved by detecting the engagement, confusion, or frustration of the student and changing the instructional content or pace of learning.

In much the same way, in entertainment uses like video games, emotion recognition can be used to adaptively control the difficulty of the game, plot, or audiovisual display depending on the emotional reaction to the player, producing more entertaining and customized experiences [4]. Emotion recognition is useful in customer service and call center applications to monitor quality in real-time, offer electronic help to agents, and predict customer

satisfaction, which allows organizations to understand what problems they have in service, how they can improve the quality of interaction, and how they can improve customer experiences. Although there has been a massive advancement in the study of speech emotion recognition in the last twenty years, the discipline still experiences major challenges that restrict the effectiveness and application of the existing systems.

The most basic problem is the subjective and complex nature of human emotions that cannot be categorized into discrete and well-defined categories instead existing in continuous dimensions with blurred boundaries and often overlapping and co-existing multiple emotional states [5]. Issues of culture and language make the matter even more complex, since the expression of emotions is significantly different among languages, dialects, and cultures, and various societies also use various prosodic patterns, norms of intensity, and rules of display in the expression of emotions. The variability of speakers is another long-standing problem: different people have different acoustic fundamentals and also different ways of expressing specific emotions, and there has to be systems to differentiate the acoustic difference related to emotions, and the acoustic difference that are related to the speaker.

Accessibility and quality of data is arguably the most important bottleneck of emotion recognition studies, especially concerning low-resource languages and demographic groups. The majority of available emotion recognition studies have been more inclined to study English and other major languages with little attachment given to the Arabic language although it is the fifth most spoken language in the world with more than 420 million native speakers in 22 countries. The Arabic language poses certain challenges and prospects to emotion recognition research because the language has a rich morphological structure, different dialectal variants, different prosodic peculiarities, and cultural factors that condition the occurrence of particular pattern of emotional expression [6]. The lack of high-quality and large-scale Arabic emotional speech corpora has limited the creation of strong Arabic emotion recognition systems, and the available ones are either small, biased in classes, limited in their emotional range, or not diversified in dialects [7, 8] class imbalance.

The problem of class imbalance is especially critical in existing emotion recognition datasets in which some of the emotional categories are over-represented and others are severely under-represented. This imbalance has a direct effect on model training where systems become biased to the majority classes and not capable of learning sufficient representations to be successful on the minority classes leading to systems that seem statistically successful, due to aggregate accuracy measures but disastrous on underrepresented emotions. Of particular concern are especially infrequent emotions that may be of especial importance in specific uses, e.g., high intensity emotions in crisis detection systems, or fine emotional gradients in mental health monitoring. Data augmentation has also become an active area of research to overcome the limitation of the size of datasets as well as the imbalance between the classes and methods like pitch shifting, time stretching and noise injection have proven effective in increasing the size of training corpora whilst maintaining the emotional content [9−11].

Recent developments in deep learning have transformed speech emotion recognition with Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid architecture having been shown to perform better than traditional machine learning methods which used hand-crafted feature engineering and shallow classifiers [12, 13]. CNNs are good at acquiring hierarchical spectrogram-like representations of space, and automatically identifying meaningful acoustic patterns without and without explicitly designing features. Sequential networks such as Long Short-Term Memory (LSTM) networks and other recurrent networks represent the temporal variations of the prosodic change and temporal movement that define emotional expression [14]. Attention mechanisms allow models to pay attention to emotionally salient parts of utterances enhancing recognition accuracy because of the most informative temporal parts of utterances [15].

Hybrid networks that use CNNs to extract spatial features and LSTMs to model time have also been of particular interest, with each method using the advantages of the other to complement each other [16]. The problematic issues discussed in this research are the recognition of speech emotion in Arabic language based on a thorough approach that combines the strategic data augmentation technique with the advanced hybrid neural network design.

The main goals are to: 1) show effective methods for tackling severe class imbalance in Arabic emotional speech datasets using balanced augmentation techniques like pitch shifting, time stretching, and noise injection; 2) create and test a hybrid CNN-LSTM architecture that combines spatial feature learning from convolutional processing with temporal sequence modeling from recurrent processing and attention mechanisms; 3) evaluate how data augmentation compares to architectural design in affecting overall system performance; and 4) assess how well the system generalizes across different datasets to ensure reliability with various speakers, recording conditions, and dialects.

The methodology is tested on two complementary Arabic emotional speech datasets: Berlin Arabic Vocal Emotions Dataset (BAVED), which is the main training corpus, and Eastern Youngsters Arabic Speech Emotions (EYASE), which is used as an independent validation set for assessing cross-dataset generalization.

The contributions of this work are as follows.

1) A clear demonstration of how data augmentation significantly improves Arabic emotion recognition, especially for underrepresented emotion classes, with high arousal recognition rising from nearly random performance (50% F1−Score) to perfect classification (100% F1−Score);
2) Validation that the hybrid CNN-LSTM architecture outperforms CNN-only approaches, showing that temporal modeling provides meaningful performance

improvements beyond spatial feature learning;
3) A breakdown of performance gains showing that data quality improvements from augmentation account for about 74% of total gains while architectural enhancements contribute roughly 26%;
4) Establishment of a strong baseline performance on BAVED (97.23% accuracy) with solid generalization to EYASE (87% accuracy);
5) Practical insights into the importance of data quality compared to model complexity in developing Arabic emotion recognition systems.

II. Related Work

The field of the speech emotion recognition has made a substantial developmental progress in the last 20 years, shifting the earlier concept of the traditional feature engineering to the advanced deep learning approaches. The initial studies have determined that emotions are represented by acoustic features, which are measurable and are expressed in terms of pitch, energy, speaking rate, voice quality, and spectral distribution [1]. Conventional methods were based on hand-engineered attributes including Mel-Frequency Cepstral Coefficients (MFCCs), prosodic features, and voice quality and were used with shallow classifiers like Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and Gaussian mixture models (GMMs) [1]. Although these were moderately useful, they demanded a lot of domain experience, and had difficulty with complicated hierarchical patterns that define emotion. Deep learning has radically changed the way emotion recognition is performed since it allows automatic feature learning using raw or low-level processed speech signals. CNNs perform well because they take spectrograms in the form of image-like features, learn feature representations in a hierarchy, which represent increasingly abstract acoustic patterns [14, 17].

Pan and Wu suggested a deep 1D and 2D CNN-LSTM network structure and was able to show that parallel processing of raw audio and spectrogram representations by independent CNN streams and subsequent LSTM-based temporal modeling and late fusion were more effective than one-pathway network structures [14]. Their study found that many complementary speech signal views obtained in parallel and combined together were able to achieve strong emotion predictions. The end-to-end deep convolutional recurrent networks, proposed by Khan *et al*. [17] are a major advancement in doing away with hand-crafted feature engineering. Their work established that the deep learning architectures acquired task specific representations directly using spectrograms and performed better than systems using pre-defined acoustic features. Zhang *et al*. [18] explored spontaneous speech emotion recognition based on multiscale deep convolutional LSTM networks, which the investigation of subtle emotion manifestations in naturalistic, unscripted speech. Their multiscale methodology took into consideration emotional cues that occur at varying time resolutions.

Deep learning methods have always shown great superiority over the classic machine learning in several benchmark tasks and show that hierarchical learning of features makes sense in general to learn more intricate acoustic patterns [18]. Recurrent architectures, especially the LSTM networks are good at capturing time-dependent and prosodic change on time scales. The seminal paper by Ghojogh and Ghodsi on the topic of Long Short-Term Memory networks gave the bases of modeling long-range relationships in a series of data, overcoming the vanishing gradient problem that afflicted the previous recurrent networks [19]. Gating mechanisms of selective retention and forgetting information conducted by the LSTM architecture would make it ideal to use in emotion recognition, where emotional content can be carried out over long periods based on the time-dependent development of prosodic characteristics [19].

Late research has examined transfer learning in which models already trained on large scale speech recognitions are refined to emotion recognition by using acquired acoustic representations to enhance performance in cases where there is limited emotion speech data that is labeled [20, 21]. Payandeh *et al*. have given a deep representation learning in speech processing a thorough analysis of how pre-trained models obtain generalizable acoustic features that are transferable across tasks [20]. Jain *et al*. [21] created Wav2Vec, a self-supervised learning model which is trained on unlabeled audio by contrastive predictive coding. Self-supervised approaches such as Wav2Vec and HuBERT have been found to be especially promising in under-resourced languages, where no large-scale labeled emotion datasets are available, but unlabeled speech data is more readily available [22].

Shi *et al*. [22] suggested HuBERT (hidden-unit BERT), viewing speech representations as masked prediction, and attaining good results on downstream emotion recognition downstream downstream emotion recognition downstream emotion recognition Shi *et al*. [22] proposed HuBER, which learns speech representations by performing masked prediction, and achieves good performance on downstream emotion recognition tasks. Nevertheless, it still has problems in terms of the lack of data, dialectal heterogeneity, and the lack of cross-dataset testing. The King Saud University emotions (KSUEmotions) corpus is an Arabic emotion recognition in Saudi dialect created by Meftah *et al*. [23] reached baseline performance of about 70% to 75% with the use of K-nearest neighbors classifier, Nasr *et al*. [24] overall construction and analysis showed that there was a high level of acoustic differences between emotions: anger was characterized by high pitch and high energy, sadness was characterized by low pitch range and low energy, and happy by high pitch change and higher rate of speech.

The corpus comprises of 12 speakers (6 men, 6 women) who utter emotionally in Saudi Arabic. Aljuhani *et al*. paid particular attention to the emotion recognition in Saudi dialect, indicating that emotion recognition systems trained on one Arabic dialect might not be very generalizable to the other Arabic Arabic dialects because of the phonetic and prosodic variations [7]. On Saudi dialect corpus, they reached a 76 percent accuracy with deep learning, which is to the effect that dialect-specific training yielded many better results than Modern Standard Arabic or mixed dialectal corpus. This shows the need to

have a variety of training information that covers a wide range of Arabic dialects in broad applicable systems. The article by Mahmoudi and Bouami investigated effective Arabic emotion recognition with the help of the deep neural network and reached the accuracy of 85 percent without compromising the computational efficiency that is, in turn, efficient enough to use in real-time applications [25].

Their efforts emphasized the need to tackle Arabic challenges that are unique such as emphatic consonants, pharyngeal sounds, and unique prosodic patterns. Shahin *et al*. suggested hybrid methods of using Gaussian mixture models and deep neural networks, and with the assistance of traditional probabilistic modeling, as well as modern deep learning, the accuracy reached 82% on Arabic databases of emotional speech [26]. Hamid [27] examined Egyptian Arabic emotion recognition in terms of prosodic, spectral and wavelet features, with 85.3 per cent accuracy of the recognition with wavelet-based features, which can be used to complement prosodic features and MFCCs. Prlinčević *et al*. examined cross-lingual recognition of emotions and discovered that although certain acoustic cues of emotion (elevated energy as an expression of anger) can be generalized across languages, that there are cross-lingual limitations to cross-lingual transfer, with recognition decreasing to 52 percent compared to 78 percent within the same language [28].

Paul *et al*. [29] created Tunisian Arabic emotion recognition system that showed dialect-specific models to be most accurate with 81 percent compared to Modern Standard Arabic with 68 or other dialectal varieties with 72 percent. A major contribution is the BAVED which has the recording of the different arousal levels of various speakers [6, 30]. Aouf designed BAVED to deal with the lack of publicly available corpora of Arabic emotional speech, and organized the recordings according to arousal levels (low, neutral, high) as opposed to discrete categories of emotion, as dimensional models of emotion [6]. Mohamed and Aly tested emotion recognition on BAVED with pre-trained models such as Wav2Vec2.0 and HuBERT at 89% and 91% respectively, showing that transfer learning could partially offset a lack of labeled Arabic emotional speech examples as well as that class imbalance and domain adaptation were a problem.

Boukherouba [8] studied natural Arabic resources for recognizing emotions in the Algerian dialect. They pointed out the diversity in how emotions are expressed in different regional varieties of Arabic. Their research stated that "Arabic emotion recognition" covers a wide range of dialects, each needing specific attention. Recent studies have looked into combining different modes, such as speech with facial expressions, text, or physiological signals. The article by Okoye has explored the utility of fusing both the acoustic and linguistic characteristics to enhance emotion recognition during educational conversations [31]. Their results indicated that the accuracy rate increased with this multimodal strategy, as compared to the performance when acoustic information was used (76% to 84% with acoustic and linguistic information respectively). Lin *et al*. [32] presented a Interpretable autoencoder-based approach to domain adaptation to recognize emotional speech, demonstrating how it is possible to enhance cross-corpus generalization.

Augmenting the amount of data has emerged as a technique of significant importance to address the problem of a small supply of labeled emotional speech data. Several augmentation methods are useful to increase training data without reducing or deteriorating model generalization. One of the first researchers to use audio augmentation to recognize speech was Sun *et al*. [10]. They added speed and volume alterations, and reverberation. The method expanded the variety of training data without additional recording. Such techniques have since been utilized in the successful adaptation to emotional recognition tasks. The models help them to be less sensitive to small acoustical variations and prioritize patterns of emotion relevancy. The speed perturbation that modifies the playback speed to 0.9, 1.0, and 1.1 times was essentially a tripling of the training data but with realistic variation in the speaking rate.

Park *et al*. [9] created the SpecAugment which is a successful technique that operates on spectrograms directly by using time masking, frequency masking and time warping. SpecAugment has been a popular speech processing technique because of its effectiveness and the constant performance enhancement. The approach makes random masks to time and frequency components of spectrograms, and it stimulates models to create strong representations that do not depend on only one particular region. SpecAugment is carefully used in emotion recognition to make the learning of representations robust without relying on individual regions very much. This increases the generalization to other speakers and conditions. Issa *et al*. [33] utilized SpecAugment to perform speak emotional recognition with deep CNNs with 7−12 percent improvement in the accuracy of various benchmark datasets. Nevertheless, excessive augmentation will conceal important acoustic elements of emotion. This needs to be fine-tuned to take into consideration the positive and negative aspects of regularization versus the danger of an individual losing valuable emotive data.

The study of Jakubec *et al*. [11] devoted to transfer learning and spectrogram augmentation to achieve enhanced emotion recognition. Their findings revealed that it was better to use a combination of pre-trained representations and smart augmentation when compared to either of the two methods. Using transfer learning alone they were able to achieve 85% accuracy, augmentation alone 82% and both augmentation and transfer learning together 91%. The aspect that was investigated by Paraskevopoulou *et al*. [34] emotion recognition through specific augmentation techniques. They investigated the performance of models under different strategies. Their critical analysis also indicated what transformations retain emotion content and introduce constructive variation, and what can diminish traits that are relevant to emotions. They discovered that pitch shifting and time stretching were the most useful to increase F1−Scores by 8−15%. Too much noise, however, had a negative impact by depriving emotional information.

This highlighted why augmentation strategies in

emotion recognition needed to be designed with serious consideration, rather than simply applying methods that have been applied in other speech tasks, as the features of emotion acoustics are likely to be more responsive to some changes. Atmaja and Sasou [35] performed extensive experiments in order to examine the influence of different augmentation methods on emotion recognition performance. Their complete comparison of techniques such as pitch shifting, time stretching, speed perturbation, noise addition, and vocal tract length perturbation was very valuable in determining which strategies to be used depending on the attributes of the dataset and the emotion to be elicited. They observed that augmentation performance was dependent on particular emotions and the original training data properties and emphasized that it is necessary to design task-specific augmentations.

Pitch shifting was the most effective in emotions with distinct pitch patterns, like anger and happiness whereas time stretching was useful in emotions with distinct timing patterns, like sadness and boredom. They found that a combination of different complementary augmentation methods gave the highest performance and increased their accuracy to 83. Various studies have offered evidence that data augmentation, when carefully designed and applied, can be of tremendous help in speech emotion recognition. It increases the effective training set size, increases model robustness and assists in the class imbalance.

The selection and configuration of augmentation methods must be dependent on the specifics of the target data set and acoustic properties of identified emotions, and the validation of the augmented samples should be done cautiously so that they still have their true emotional attributes and also to add beneficial variation. A clever advance achieved through class-balanced augmentation, which augments minority emotion classes selectively, rather than augmenting all of the classes uniformly, is a response to the large class imbalance in many emotion datasets, such as BAVED. This guarantees equal representation of all categories of emotions in the training and maintenance of real emotional properties.

III. Methodology

In this section, we shall present our integrated perspective of Arabic speech emotion identification, such as the aspects of the dataset, preprocessing pipelines, augmenting mechanisms, feature extraction, and recommended neural architecture. It is suggested that the systematic pipeline (Fig. 1) is the part of the given methodology, the analysis of the BAVED and EYASE data is conducted systematically, and MFCC features are extracted. The techniques of augmentation such as pitch shifting, time stretching, and the addition of colored noise are used to increase the dataset. The results of augmented features are used as inputs into a hybrid CNN-LSTM to classify emotions. The precision, recall, F1−Score, and accuracy metrics are used in performance evaluation to evaluate the model efficacy in all categories of emotions.

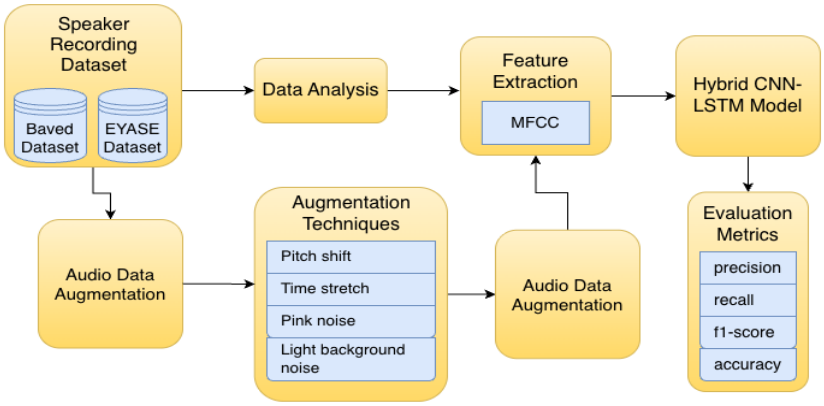

Fig. 1. Shows the methodology used in the study.

A. Dataset Selection and Preparation

Two Arabic emotional speech datasets are used to complement each other and create and test the emotion recognition system. The main training corpus is the BAVED (Berlin Arabic vocal emotion database), whereas the EYASE (eastern youngsters Arabic speech emotions) training material is used as an evaluation set to determine the ability to cross-test data sets. The two-dataset model guarantees the acquisition of strong emotion-relevant features which can be successfully transferred to different speakers, recording conditions and dialectal changes.

The first step is related to the thorough analysis of both datasets in order to study their peculiarities, learn about possible difficulties, and make the right preprocessing plans. The original BAVED data has a severe imbalance of classes in general, specifically, the three arousal levels low, neutral, and high. As can be seen, initially, there are about 1,600 samples of low arousal emotions, about 300 samples of neutral emotions, and less than 10 samples of high arousal emotions as represented in Fig. 2. Such a large imbalance presents a major threat to model training, where it is possible to be biased to the majority group and the performance of recognition to the underrepresented emotional states is worse. Therefore, it was augmented in order to make it balanced as depicted in Fig. 3.

The EYASE data in Fig. 4 is more evenly distributed in four categorical emotions of anger, happiness, neutrality, and sadness, each group consisting of 130 to 150 samples. The relatively small size of the overall samples of around 477−600, however, causes the need to apply augmentation techniques to make the model training strong and avoid

overfitting that is depicted in Fig. 5.

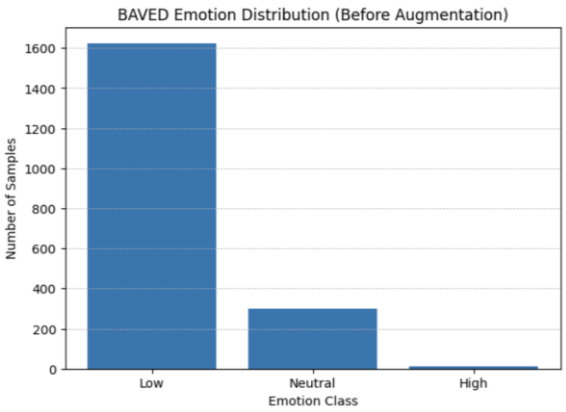

Fig. 2. BAVED dataset emotion distribution.

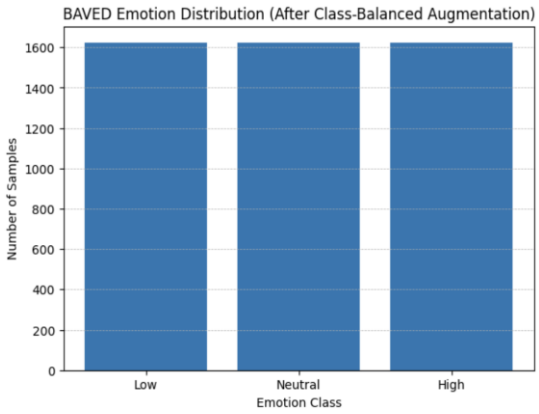

Fig. 3. Augmented BAVED dataset emotion distribution.

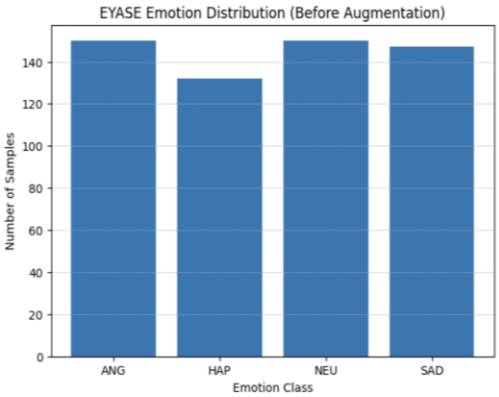

Fig. 4. EYASE dataset emotion distribution.

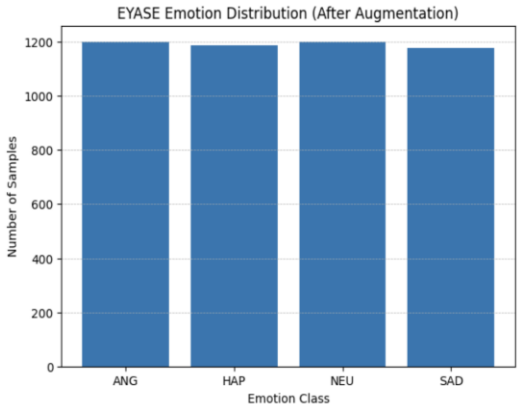

Fig. 5. EYASE dataset emotion distribution after augmentation.

B. Data Augmentation Strategy

To deal with the imbalance of classes and increase the size of the datasets, a methodical data augmentation pipeline is applied in which several acoustic transformation methods are used to retain emotional content and add a controlled variation. The augmentation plan is aimed at producing the realistic variations of the original records which preserve the basic affective properties but diversify the acoustic patterns which the model experiences in the course of training. Fig. 6 represents the original waveform where augmentation is to be accomplished. The initial augmentation method is pitch shifting, whereby the basic frequency of the speech signal is adjusted but there is no change in the time aspects. Fig. 7 indicates that positive pitch shifts of +2 semitones pitch up the utterance, which mimics changes in vocal range of the speaker or intensity of emotions, whereas negative pitch shifts of –2 semitones pitch down as in Fig. 8, producing variants of deeper voice.

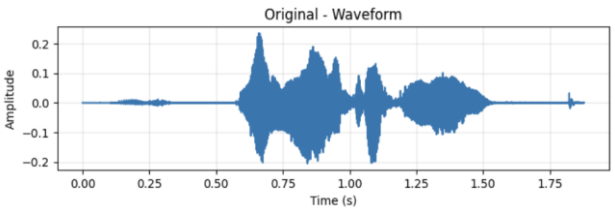

Fig. 6. Original waveform.

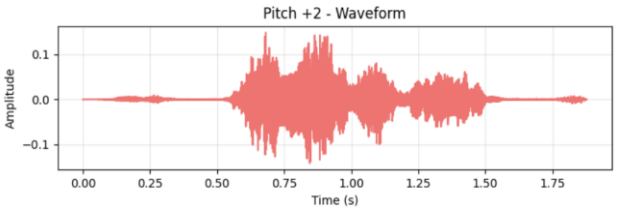

Fig. 7. Positive pitch shifts of +2 semitones.

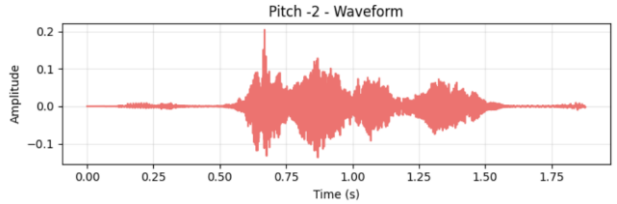

Fig. 8. Negative pitch shifts of -2 semitones.

These changes of pitch do not alter the speech level of intelligibility or emotional expression but add acoustic diversity that enables the model to generalize to diverse speaker features. Time stretching is the second augmentation strategy which varies the time length of utterance without variability in pitch feature. Fig. 9 displays the fact that a time stretch factor of 1.05 makes versions that are 5% slower than the original, reproducing the effect of deliberate or emphatic speech patterns, whereas a factor of 0.95 generates versions that are 5% faster as seen in Fig. 10, a more representative of quicker and more excited speech delivery. These timing changes maintain the spectral property and pitch contours which transmit emotional content but bring up variation in speech rate and rhythm.

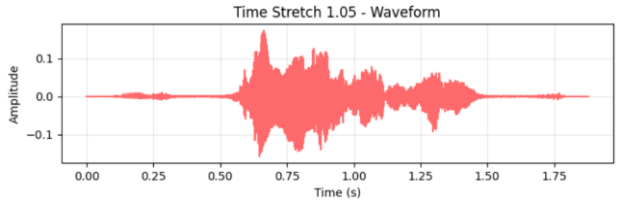

Fig. 9. Time stretch factor of 1.0.

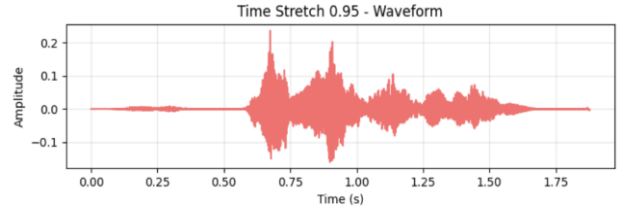

Fig. 10. Time stretch factor of 0.95.

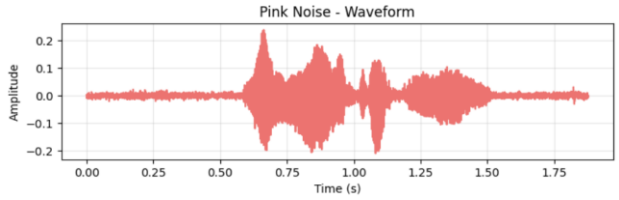

Fig. 11. Pink noise injection.

The third form of augmentation method is noise injection as illustrated in Fig. 11 wherein noise is injected through pink noise (1/*f* noise) and light background noise such that the additions to the original records are taken into account. Pink noise has the properties of spectral representations that resemble the natural acoustical settings, and light background noise can be used to replicate the realistic recording environment by ambient noise. Noise levels are also put under great control and do not harm model robustness at the expense of the speech quality or the words conveying emotions.

Application of the augmentation process is class-balanced in nature, whereby after augmentation, an equal number of samples in each emotion classes exist. In the case of BAVED, the effect of this balancing is that the three arousal levels (low, neutral, and high) have an equal number of samples per class, i.e., 1,600 samples per class, and thus the effect of balancing converts the unbalanced dataset to a perfectly balanced training corpus. In EYASE, every category of emotion is augmented to around 1,200 samples, which significantly escalates the size of the dataset and evenly balances the number of samples of the classes of anger, happiness, neutrality, and sadness.

C. Feature Extraction

After augmenting the data, Mel-Frequency Cepstral Coefficients (MFCCs) are obtained on all the audio samples to generate powerful acoustic feature representations that will be used as deep learning model inputs as indicated in Fig. 12.

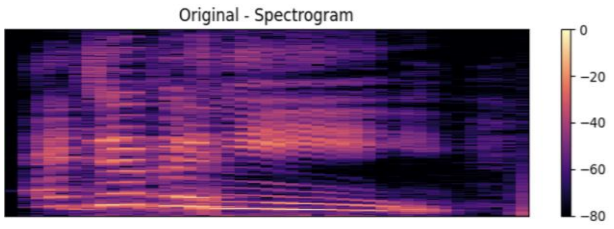

Fig. 12. MFCC spectrogram.

MFCC extraction algorithm converts raw audio waveforms to small representations that encode the spectral envelope features which are especially important when it comes to speech and emotion recognition tasks. The pipeline of feature extraction starts with the pre-emphasis filtering of the audio signal to highlight the high-frequency parts of the audio signal, and the frame-based windowing of the continuous audio signal into small blocks of the analysis. The frames are processed by fast Fourier transform to transform the time domain signal into the frequency domain signal and the spectrum is plotted onto the mel scale which is a linear scale below 1000 Hz and logarithmic scale above. The energies of melt-scale filterbank are calculated and logarithmically compressed then transformed using discrete cosine transform into the final MFCC coefficients. The MFCC extraction is set in a way that it produces 13 fixed coefficients at each frame and these coefficients represent the general spectral shape that defines the linguistic content as well as the emotional expression.

These constant coefficients are extended with first order derivatives (delta features) and second order derivatives (delta-delta features) in order to add the information on the time dynamics of the spectral characteristics. The rate of change in the spectral characteristics with time is captured in delta features, whereas the acceleration of the change is captured in delta-delta features, which are essential in resolving the difference between distinct emotional states that are similar in their static spectral properties, but differ in their time dynamics. The resulting feature matrices have dimensions of batch size × 1 channel × 13 coefficients × 161-time frames. The arrangement has a fine-grained representation which is computationally accurate and information rich. The dimensionality captures enough temporal context to model prosodic patterns linked to emotional expression. At the same time, it stays manageable for deep neural network processing. The 161-time frames match utterances of typical duration found in the datasets. We apply padding or truncation as needed to maintain consistent input dimensions across all samples.

D. Hybrid CNN-LSTM Architecture

The strength of such a methodology is the use of a hybrid neural network architecture, which integrates Convolutional Neural Networks (CNNs) to learn spatial-based features and long Short-Term Memory Networks (LSTMs) to model a temporal sequence, which are then processed in parallel and then finally fused to make final classification. The resulting parallel architecture enables this model to extract complementary kinds of information using the input features and provides a spatial pattern in the time-frequency representation via the CNN pathway and temporal dependencies in the sequential evolution of acoustic properties via the LSTM pathway.

In Fig. 13, we present our architecture that consists of three complementary parts to extract the most information out of limited data, amounting 569,028 parameters, which is intentionally moderate and allows avoiding overfitting but at the same time is large enough. The suggested architecture uses a hybrid CNN-BiLSTM-attention system that is meant to generate compatible spectral-temporal features of speech signals. The model input is in the form of log-Mel spectrogram or MFCC

feature maps, $1 \times M \times T$, where M is the number of feature coefficients and T is the time axis. The general system comprises a convolutional pathway, which is a spatial feature extraction, a recurrent pathway, which is a temporal modeling, an attention mechanism, which is a dynamic weighting, and a dense classifier which is an emotion recognition. The convolutional branch is used to extract local spectral variations and higher-level features of the input features.

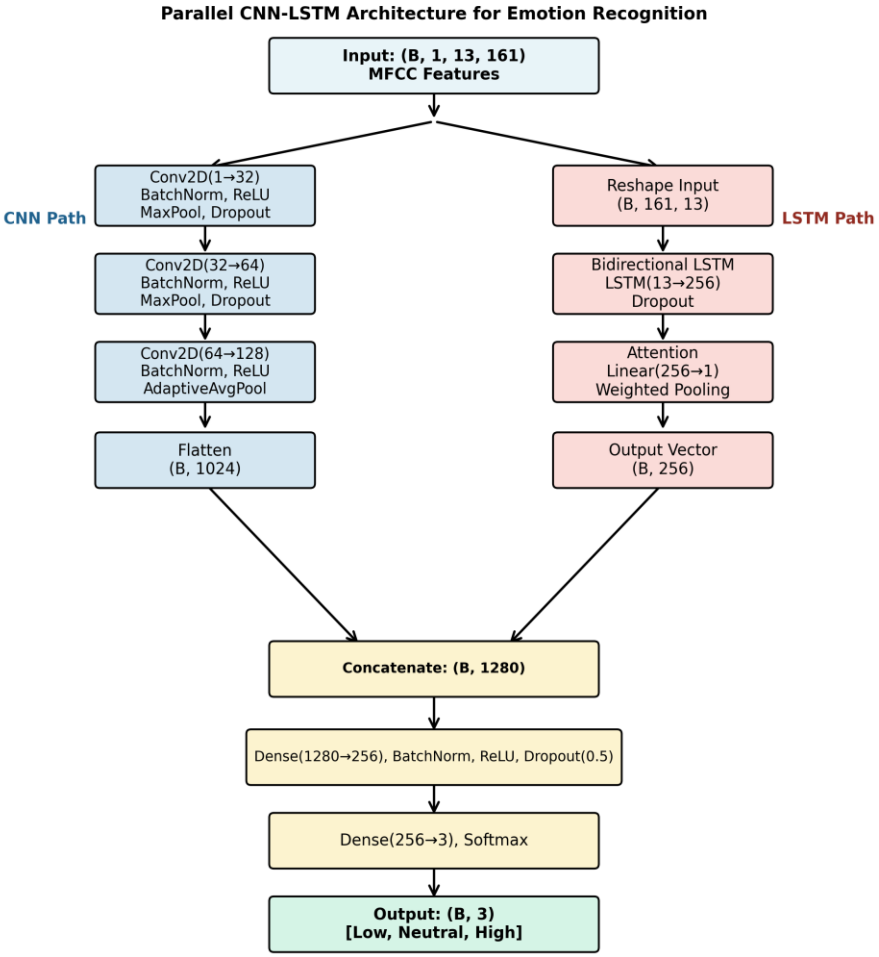

Fig. 13. Hybrid CNN+LSTM architecture.

The network starts with a Conv2D (1-32) layer (with 3 −3 kernel) then batch normalization followed by a ReLU 2×2 MaxPooling layer then decreases the spatial resolution but at the same time retains the main spectral textures. An underfitting problem is avoided by applying a dropout rate of 0.3. This is further repeated by a more convolutional block, a Conv2D layer (32−64) with nearly the same 3×3 filters, BatchNorm, and ReLU. Once more, maxPooling and dropout are utilized. The last convolution block makes use of Conv2D layer (64−128) to acquire high-level discriminative features. A pooling layer is adaptive and compresses the output to a fixed spatial size (1×8) to allow input lengths of different lengths to be compatible.

The last dropout (0.3) is the one that is applied and then the feature map is flattened into the CNN embedding vector. The temporal dependencies are modeled by using a parallel recurrent pathway to complement the spectral features that were obtained by the CNN. The input array is rearranged and restructured into the shape $T \times M$, allowing a serial application. This channel starts with a bidirectional LSTM (hidden size = 128) periodicity that acquires forward and backward time trends to obtain a context both past and future frames. The LSTM outputs are scaled by dropout rate of 0.3 in order to increase generalization. The sequence is repeated through an attention mechanism which computes the learned attention weights and produces a weighted sum over the temporal features in order to highlight the most informative time steps.

This will give a 256−dimensional attention LSTM embedding. To take advantage of the complementary information on the two pathways, the flattened CNN embedding and LSTM-attention embedding are fused and the resulting single feature (1280) is obtained. This combination allows the combination of both localized spectral stimuli and global temporal dynamics, which make it possible to represent emotional properties of speech more comprehensively the classification network. The fused representation is run through a dense transformation block which includes a fully onnected layer (1280−256), batch normalization, ReLU activation, and Dropout (0.5). It is a block that decreases the dimensionality but increases the feature robustness. Lastly, a dense output layer (256-C) is used, with C representing the classes of emotions.

E. Training Configuration

The model training uses the categorical cross-entropy loss which is operated using AdaGrad and RMSprop optimizers in order to overcome high-dimensional learning. Various dataset split ratios (75:25, 80:20, 90:10) are used to test the performance of various training ratios. BAVED is the training corpus, whereas EYASE is used as independent validation in comparing cross-generalization. The sample sizes usually are in the range of 32−64 samples, which is a compromise between memory and gradient stability. Overfitting between network layers is prevented by regularization by dropout (rates 0.3−0.5). The dropout is still active in the training of neurons that are randomly deactivated. Early stopping is a process that validates loss of monitors, stopping training once the performance has reached a specified limit in terms of epochs, it optimizes the selection of a model and it avoids overfitting to training data without losing generalization abilities.

F. Evaluation Metrics

Model performance is fully assessed on the basis of several metrics that give various points of view on classification quality. Precision is used to determine the accuracy of positive prediction in each of the classes which is computed as a ratio of true positives to the total of the true positives and the false positives, meaning how many of the utterances that are classified as a certain emotion are actually that emotion, as given

$$Precision = \frac{TP}{(TP+FP)} \quad (1)$$

Recall measures the completeness of positive identification and it can be expressed as the ratio of the true positives to the total of true positives and false negatives:

$$Recall = \frac{TP}{(TP+FN)} \quad (2)$$

Eq. (2) marks the percentage of utterances which actually belonged to a group that were indeed recognized.

The F1−Score offers a more balanced measure that both evaluates the precision and the recall by the harmonic mean of the two and it equally relies on both features of classification performance and offers a single score that can be used as a measure of both the accuracy and completeness:

$$F1 = \frac{2\times(Precision \times Recall)}{(Precision + Recall)} \quad (3)$$

Overall accuracy is a ratio of all correct predictions made of all the classes and gives a simple aggregate measure of all classification performance:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

True *Positives* (*TP*) in this case are those samples in which the model makes accurate prediction of the actual emotion label. *False Positives* (*FP*) occur when the model forecasts an emotion which the sample is not a part of. *True Negatives* (*TN*) on the other hand are those in which the model is right about the absence of a certain emotion. *False Negatives* (*FN*) refer to instances when the model does not recognize the right emotion and gives some other label in its place. The performance is computed per-class (specifically, evidence-based metrics of each of the emotion categories) to determine which classes the model performs poorly and at the aggregate level (macro-averaged metrics using unweighted average across classes) and weighted-average metrics (weighted by class support). Confusion matrices help visualize the full pattern of predictions. They display correct classifications along the diagonal. They also show specific misclassification patterns in the off-diagonal elements. This shows the most commonly confused pairs of emotions and provides information about the acoustic similarity of various expressions of emotions.

IV. Results and Discusssion

The proposed hybrid CNN-LSTM model of Arabic speech emotion recognition was experimentally evaluated and demonstrated strong and repeated enhancement of its performance. These advances were realized in many areas, which affirmed both data augmentation plan and architectural decisions. The findings are clear that there is an improvement in the baseline models to the final optimized system with the most significant improvement observed in cases where the issue of class imbalance is addressed by focusing on augmentation.

The subsequent subsections are the detailed performance analysis in case of various experimental conditions, dataset settings and model architecture. This is a combined quantitative and qualitative analysis of the model behavior.

Fig. 14 provides a comparison between the performance of four experimental structures on the BAVED dataset: CNN, CNN + augmentation, Hybrid CNN-LSTM, and CNN-LSTM + augmentation. The metrics of evaluation to be considered are the precision, recall, F1−Score, and accuracy.

The findings from Fig. 14 show quite clearly that data augmentation is needed to enhance the performance of the classification, especially in such an imbalanced dataset as BAVED. Both CNN and Hybrid models cannot effectively experience a minority classification, a problem not resolved by augmentation, which leads to worse macro-level results. Once the suggested class-balanced augmentation strategy has been employed, the performance of both of the models improves significantly. The CNN accuracy increases to 0.94 as compared to 0.84 whereas the Hybrid model accuracy increases to 0.97 as compared to 0.89. It is noteworthy that the Hybrid CNN-LSTM architecture achieves higher performance than the standalone CNN, and the benefits of the convolutional feature extraction approach in combination with the time modeling can be observed. On the whole, this figure underlines that balanced augmentation and the hybrid architecture integration led to the most highly performing and reliable emotion recognition.

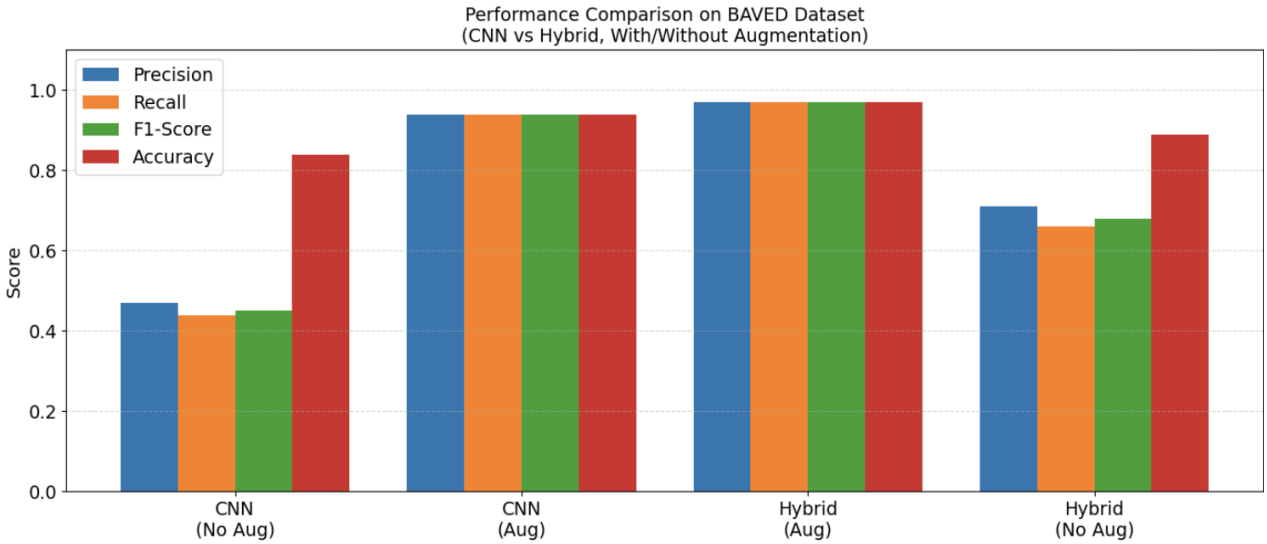

Fig. 14. Comparative performance of BAVED dataset with and without augmentation.

Fig. 15 offers a comparative analysis of the Hybrid CNN-LSTM model when using three traintest split ratios namely 75:25, 80:20, and 90:10. This evaluation helps evaluate the impact of the size of training data on the performance of models and confirm the strength of the architecture offered. The model has a very high performance in all formats, and the precision, the recall, the F1−Score, and the accuracy are tightly concentrated in the range of 0.96 to 0.97. This minimal variation is an indication that the model is generalizable and its behavior is not a pronounced variant of training set size. The 80:20 split is the most suitable setting among the tested ones, with the highest measures (0.97 per measure), with a sufficient amount of training data and sufficient test data to have a robust test set. The 75:25 split that contains the largest test set has a more statistically rigorous evaluation of generalization, whereas the 90:10 split provides the greatest exposure of training data that assists the model to acquire the finer details of the patterns even when the test set is smaller. The same performance of these splits proves that these gains are real and not just the data arrangement artifacts.

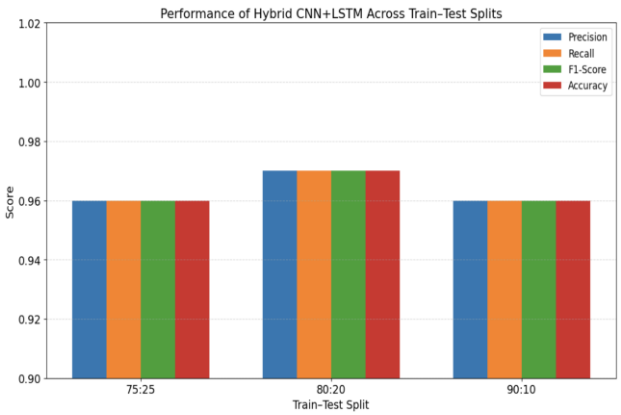

Fig. 15. Comparative performance of BAVED dataset with different splits.

The confusion matrix in Fig. 16 of the 80:20 assessment of the hybrid CNN-LSTM model on the BAVED dataset demonstrates high and well-balanced classification of all 3 classes; low, neutral, and high. The model was able to classify 303, 319 and 325 cases of low, neutral and high intensity respectively, which show a high level of reliability. There were very minimal misclassifications made with only 21 low cases wrongly classified as neutral and 6 neutral cases wrongfully classified as low which indicates some overlap between these two closely related emotional extremes. Notably, there were no misclassifications between low and high, and there were no instances of neutral which were not correctly predicted as high, which could indicate that the model can be used successfully to draw the demarcation between classes that are far apart semantically. In general, the confusion matrix demonstrates that the hybrid architecture is able to focus on local and sequential trends in the text, and achieve a steady performance and good discrimination between the categories of the BAVED datasets.

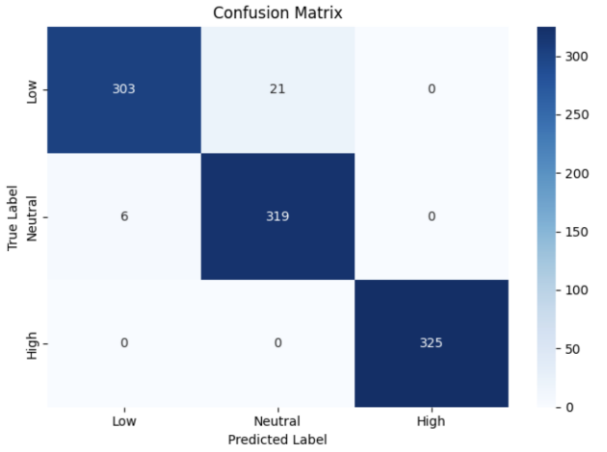

Fig. 16. Confusion matrix from BAVED dataset for hybrid CNN-LSTM model.

A middle experimental stage used to test the data augmentation role alone of the hybrid architecture and tested a CNN-only model on the augmented BAVED

dataset. An architecture using CNN alone reached an accuracy of 94.35% which is an 8.05 percentage point better than a baseline (probably 86.30%). This finding gives valuable information about the role of various methodological elements in overall performance. The 94.35% accuracy of CNN-only model proved that even simple architectures initiated with the data augmentation strategy helped to reach the desired results as it is clear that the balanced training data played a central role in the model to discover distinguishing features. The 8.05% enhancement over baseline indicated that augmentation would be sufficient to restore much of the lost performance on a system due to class imbalance, turning a moderately successful system into a highly successful one. Nonetheless, the transition between the CNN-only structure (94.35% accuracy) and the hybrid CNN-LSTM structure (97.23% accuracy) offered the added performance of 2.88 percentage points. This significant progression showed that the time modeling function of the LSTM pathway and attention mechanism were value added to the spatial feature learning of CNN pathway only. The capability of the LSTM to record the evolution of prosodic and temporal changes and sequence in the utterances of emotion also provided further hints to enhance the classification performance particularly when it comes to an emotion with a distinctive temporal characteristics such as continuous arousal or slowing emotional change in natural speech. These findings imply a hierarchical contribution model: data augmentation provided the background by making sure that there were sufficient training instances in all classes, and the hybrid CNN-LSTM architecture enhanced it by learning complementary details of emotional expression. The practical value of this finding on designing a system is that, in situations where resources are scarce, data quality and balance can be more beneficial than architectural complexity. But, to perform optimally, balanced data are required as well as advanced architectures.

The results of the Hybrid CNN-LSTM model presented in Fig. 17 on the EYASE dataset indicate the difficulties of the cross-dataset generalization and the strengths of data augmentation to promote the robustness. The model performed a general accuracy of 83 without augmentation and the four-emotional groups have equal and weighted averages of 0.83. Anger was the most identified emotion (F1 = 0.90), and happiness had the worst result (F1 = 0.77), which means a clear confusion with acoustically close classes. Neutral and sad emotions had more moderate recognition quality (F1 = 0.82 and 0.84, respectively) which demonstrates the adequate yet not perfect generalizability of the model to speech of other speakers and recording settings. The performance of the model also improved steadily following augmentation which raised the number of samples to approximately 4,800 samples, raising the overall accuracy to 87 percent and the macro average to 0.87.

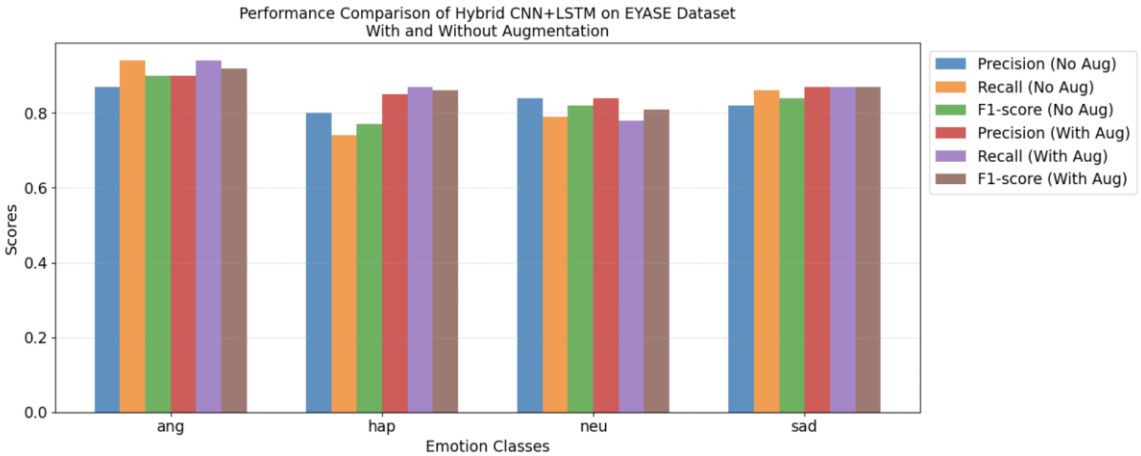

Fig. 17. Shows performance of Hybrid CNN-LSTM model on EYASE dataset.

All the categories of emotions were augmented, with an increase in happiness being the most (F1 rising by 0.77 to 0.86) and this could indicate that the increased diverse examples would be able to cover the high acoustic variability of happiness. Anger and sadness also improved (F1 increased to 0.92 and 0.87), whereas neutrality did not change, since there were only slight differences which could fall under the boundaries of statistical fluctuations. These findings demonstrate that augmentation significantly enhances the generalizability of the model to emotional characteristics in diverse datasets especially in high expressive variability emotions which supports the overall stability of the hybrid architecture in cross-corpus scenarios. The results have a number of significant implications to Arabic Speech Emotion Recognition. Above all, consideration of the considerable enhancement brought by augmentation demonstrates that numerous current systems have failed due to the inadequate supply of limited or unbalanced information, as opposed to the inefficiency of models.

The augmentation methods, like pitch shifting, time stretching, and noise injection, which are categorized as simple methods, were enough to significantly enhance the performance that made them not necessary to gather data on large scales, which cost more money. The findings also attest to the usefulness of temporal modelling, the hybrid CNN-LSTM architecture was better than CNN-only models, and it is necessary to consider the emotional dynamics during time. Besides, good cross-dataset generalization of BAVED to EYASE indicates that emotion acoustic cues are relatively constant across the Arabic speakers and across the Arabic dialects.

Nonetheless, the other gap in generalization shows that multi-dataset training or domain adaptation might benefit.

Although these are promising outcomes, a number of limitations suggest opportunities of future research. The datasets on which they train are small according to the standards of deep learning and the augmentation techniques that are implemented are relatively simple as compared to the current generative approaches. Future studies ought to seek more sophisticated augmentation methods, larger and more multifarious datasets, and wider performance measures such as real-time performance and noise or spontaneous speech resistance. Furthermore, the discretized emotion naming can be replaced by continuous or hierarchical emotion descriptions, which can give more accurate indicators of human emotions. Assessment of models on spontaneous, naturalistic emotional speech is another important step to be taken towards the real world.

A. Comparative Analysis

According to the comparative results, our model without augmentation can also perform the same at 89% whereas the previous BAVED models produced accuracy ranging between 89% to 92.63%. CNN-only architecture as well as Hybrid CNN-LSTM model increases the accuracy by 94.35, and 97.23 percent respectively, demonstrating the importance of temporal modelling as shown in Table I. With cross-dataset validation, even though the data and speaker are different, 87% is achieved with EYASE, which is a very high level of generalisation.

TABLE I: COMPARATIVE STUDY

Model	Accuracy (%)
BAVED SER [30]	89
BAVED SER [36]	92.63
BAVED SER [37]	73.13
Our Model SER without augmentation (BAVED)	89
Our Model SER with CNN-only + augmentation (BAVED)	94.35
Our Model SER with Hybrid CNN-LSTM + augmentation (BAVED)	97.23
Our Model SER cross-dataset validation (EYASE)	87

V. CONCLUSION

The experimental findings fully confirm the use of a proposed methodology to recognize the emotions of Arabic speech because the class-balanced data augmentation and hybrid CNN-LSTM architecture are shown to perform excellently on the BAVED dataset (97.23% accuracy) and exhibit good cross-dataset generalizations to EYASE (87% accuracy). Contribution breakdown shows that most performance improvement (73.7% of total gains) is to contribute by data quality enhancement by augmentation, and further performance contribution (26.3% of total gains) is to be made by architectural sophistication by hybrid design. The radical change in high arousal recognition of random performance to perfect recognition with augmentation demonstrates the importance of balanced training data and the steady improvement by hybrid architecture confirms the significance of temporal modeling in emotion recognition. These results provide a solid foundation of the Arabic speech emotion recognition and indicate the definite paths of future improvement by the means of larger data sets, more advanced augmentation, and longer testing on spontaneous emotional speech under real-life circumstances.

CONFLICT OF INTEREST

The author declares no conflict of interest

REFERENCES

[1] G. Alhussein, I. Ziogas, S. Saleem, and L. J. Hadjileontiadis, "Speech emotion recognition in conversations using artificial intelligence: A systematic review and meta-analysis," *Artificial Intelligence Review*, vol. 58, no. 7, 198, 2025.

[2] Y. Wei, S. Qin, F. Liu, R. Liu, Y. Zhou, Y. Chen *et al*., "Acoustic-based machine learning approaches for depression detection in Chinese university students," *Frontiers in Public Health*, vol. 13, 1561332, 2025.

[3] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers in Robotics and AI*, vol. 7, 532279, 2020.

[4] P. Kozlov, A. Akram, and P. Shamoi, "Fuzzy approach for audio-video emotion recognition in computer games for children," *Procedia Computer Science*, vol. 231, pp. 771–778, 2024.

[5] L. Yang and S. Zhao, "AI-induced emotions in L2 education: exploring EFL students' perceived emotions and regulation strategies," *Computers in Human Behavior*, vol. 159, 108337, 2024.

[6] Arabic-Vocal-Emotions-Dataset. GitHub repository. (2024). [Online]. Available: https://github.com/40uf411/Basic-Arabic-Vocal-Emotions-Dataset

[7] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic speech emotion recognition from Saudi dialect corpus," *IEEE Access*, vol. 9, pp. 127081–127085, 2021.

[8] M. R. Boukherouba, "Sentiment analysis of Algerian dialect using machine learning techniques," 2024. https://dspace.univ-guelma.dz/jspui/handle/123456789/16457

[9] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.

[10] Y. Sun, K. Xu, C. Liu, Y. Dou, H. Wang, B. Ding, and Q. Pan, "Automated data augmentation for audio classification," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 2716−2728, 2024, 10.1109/TASLP.2024.3402049

[11] M. Jakubec, E. Lieskovska, R. Jarina, M. Spisiak, and P. Kasak, "Speech emotion recognition using transfer learning: Integration of advanced speaker embeddings and image recognition models," *Applied Sciences*, vol. 14, no. 21, 9981, 2024.

[12] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech emotion recognition using convolutional neural networks with attention mechanism," *Electronics*, vol. 12, no. 20, 4376, 2023.

[13] N. Banskota, A. Alsadoon, P. W. C. Prasad, A. Dawoud, T. A. Rashid, and O. H. Alsadoon, "A novel enhanced convolution neural network with extreme learning machine: Facial emotional recognition in psychology practices," *Multimedia Tools and Applications*, vol. 82, no. 5, pp. 6479−6503, 2023.

[14] S. T. Pan and H. J. Wu, "Performance improvement of speech emotion recognition systems by combining 1D CNN and LSTM with data augmentation," *Electronics*, vol. 12, no. 11, 2436, 2023.

[15] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, 7530, 2021.

[16] M. R. Ahmed, S. Islam, A. M. Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," *Expert Systems with Applications*, vol. 218, 119633, 2023.

[17] M. Khan, A. El Saddik, F. S. Alotaibi, and N. T. Pham, "AAD-Net: Advanced end-to-end signal processing system for human emotion detection and recognition using attention-based deep echo state network," *Knowledge-Based Systems*, vol. 270, 110525, 2023.

[18] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion

recognition using multiscale deep convolutional LSTM," *IEEE Trans. on Affective Computing*, vol. 13, no. 2, pp. 680−688, 2022.
[19] B. Ghojogh and A. Ghodsi, "Recurrent neural networks and long short-term memory networks: Tutorial and survey," arXiv preprint arXiv:2304.11461, 2023.
[20] A. Payandeh, K. T. Baghaei, P. Fayyazsanavi, S. B. Ramezani, Z. Chen, and S. Rahimi, "Deep representation learning: Fundamentals, technologies, applications, and open challenges," *IEEE Access*, vol. 11, pp. 137621−137659, 2023.
[21] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46938−46948, 2023.
[22] J. Shi, H. Inaguma, X. Ma, I. Kulikov, and A. Sun, "Multi-resolution HuBERT: Multi-resolution speech self-supervised learning with masked unit prediction," arXiv preprint arXiv:2310.02720, 2023.
[23] A. H. Meftah, M. A. Qamhan, Y. Seddiq, Y. A. Alotaibi, and S. A. Selouani, "King Saud University emotions corpus: Construction, analysis, evaluation, and comparison," *IEEE Access*, vol. 9, pp. 54201–54219, 2021.
[24] L. I. Nasr, A. Masmoudi, and L. H. Belguith, "Survey on Arabic speech emotion recognition," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 53-68, 2024.
[25] O. Mahmoudi and M. F. Bouami, "Arabic speech emotion recognition using deep neural network," in *Proc. Int. Conf. on Digital Technologies and Applications*, Cham: Springer Nature Switzerland, Jan. 2023, pp. 124−133.
[26] C. Hema and F. P. G. Marquez, "Emotional speech recognition using CNN and deep learning techniques," *Applied Acoustics*, vol. 211, 109492, 2023.
[27] L. Abdel-Hamid, "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features," *Speech Communication*, vol. 122, pp. 19−30, Sept. 2020.
[28] B. Prlinčević, Z. Milivojević, V. Stojanović, D. Kostić, and Z. Veličković, "Estimation of emotion from speech through analysis of fundamental frequency derivative," in *Proc. 2025 24th Int. Symposium Infoteh-Jahorina (INFOTEH)*, Mar. 2025. doi: 10.1109/INFOTEH64129.2025.10959287
[29] B. Paul, S. Bera, T. Dey, and S. Phadikar, "Machine learning approach of speech emotions recognition using feature fusion technique," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 8663–8688, 2024.
[30] P. Jafarzadeh, A. M. Rostami, and P. Choobdar, "Speaker emotion recognition: Leveraging self-supervised models for feature extraction using Wav2Vec2 and HuBERT," arXiv preprint arXiv:2411.02964, 2024.
[31] K. Okoye, "The impact of emotional valence on students learning performance and evaluation: A text mining of students' opinion data," in *Proc.* 2024 4th *Int. Conf. on Electrical, Computer, Communications and Mechatronics Engineering*, Nov. 2024. doi: 10.1109/ICECCME62383.2024.10796989
[32] W. C. Lin, K. Sridhar, and C. Busso, "An interpretable deep mutual information curriculum metric for a robust and generalized speech emotion recognition system," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, pp. 5117−5130, Nov. 2024.
[33] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, 101894, 2020.
[34] G. Paraskevopoulou, E. Spyrou, and S. Perantonis, "A data augmentation approach for improving the performance of speech emotion recognition," in *Proc. 19th Int. Conf. Signal Processing and Multimedia Applications (SIGMAP 2022)*, Lisbon, Portugal, Jul. 2022, pp. 61−69.
[35] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, 5941, 2022.
[36] W. Bouchelligua, R. Al-Dayil, and A. Algaith, "Effective data augmentation techniques for Arabic speech emotion recognition using convolutional neural networks," *Applied Sciences*, vol. 15, no. 4, 2114, 2025.
[37] R. Sujatha, J. M. Chatterjee, B. Pathy, and Y. C. Hu, "Automatic emotion recognition using deep neural network," *Multimedia Tools and Applications*, vol. 84, pp. 33633−33662, Nov. 2025.

Sarmad Hamad Ibrahim Alfarag received his B.Sc. degree in electrical engineering and his M.Sc. degree in electronic engineering from Kogakuin University, Japan. His research interests include electronic circuits, embedded systems, artificial intelligence applications in engineering, and simulation-based design approaches.