# Effective Data Augmentation Techniques for Arabic Speech Emotion Recognition Using Convolutional Neural Networks

Wided Bouchelligua * , Reham Al-Dayil , Areej Algaith

Posted Date: 3 January 2025

doi: 10.20944/preprints202501.0126.v1

Keywords: Arabic Speech Emotion Recognition; Data augmentations; Convolutional Neural Networks

*Article*

# Effective Data Augmentation Techniques for Arabic Speech Emotion Recognition Using Convolutional Neural Networks

**Wided Bouchelligua *, Reham Al-Dayil and Areej Algaith**

Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 11432, Saudi Arabia
* Correspondence: wabouchelligua@imamu.edu.sa

**Abstract:** This paper investigates the effectiveness of various data augmentation techniques for enhancing Arabic Speech Emotion Recognition (SER) using Convolutional Neural Networks (CNNs). Utilizing the Saudi Dialect and BAVED datasets, we address the challenges of limited and imbalanced data commonly found in Arabic SER. To improve model performance, we apply augmentation techniques such as noise addition, time shifting, increasing volume, and reducing volume. Additionally, we examine the optimal number of augmentations required to achieve the best results. Our experiments reveal that these augmentations significantly enhance the CNN's ability to recognize emotions, with certain techniques proving more effective than others. Furthermore, the number of augmentations plays a critical role in balancing model accuracy. The Saudi Dialect dataset achieved its best results with two augmentations (increasing volume and decreasing volume), reaching an accuracy of 96.81%. Similarly, the BAVED dataset demonstrated optimal performance with a combination of three augmentations (noise addition, increasing volume, and reducing volume), achieving an accuracy of 92.60%. These findings indicate that carefully selected augmentation strategies can greatly improve the performance of CNN-based SER systems, particularly in the context of Arabic speech. This research underscores the importance of tailored augmentation techniques to enhance SER performance and sets a foundation for future advancements in this field.

**Keywords:** arabic speech emotion recognition; data augmentations; convolutional neural networks

## 1. Introduction

In recent years, Speech Emotion Recognition (SER) has garnered significant attention due to its applications in human-computer interaction [1], psychological assessment [2], and entertainment technologies such as robotics [3] and computer games [4]. The ability of machines to detect and interpret human emotions from speech enhances user experience, enabling more empathetic and natural interactions in various domains, such as virtual assistants and automated customer service. The advent of machine learning and deep learning approaches has led to notable advances in SER systems. However, one of the critical challenges in this area is the limited availability of large, balanced datasets that capture a wide range of emotional expressions in speech [5].

This challenge is particularly pronounced in Arabic Speech Emotion Recognition, where the complexity of dialects and regional variations poses significant difficulties for effective emotion recognition. Arabic, one of the most widely spoken languages globally, includes distinct dialects that differ in pronunciation, intonation, and emotional expression [6]. Despite its importance, Arabic SER has received less attention compared to other languages such as English, German, and French [7]. This gap is compounded by the scarcity of labeled emotional datasets for Arabic speech, limiting the performance of machine learning models in this area.

 doi:10.20944/preprints202501.0126.v1



A promising solution to address the problem of small and imbalanced datasets is data augmentation, which involves artificially increasing the size and variability of training data. Techniques such as noise addition, volume adjustments, and time shifting are commonly used in speech processing to simulate real-world variations, improving model robustness and generalization [8,9]. Although these techniques have been applied effectively in other languages, their impact on Arabic speech emotion recognition remains largely unexplored.

In this study, we aim to address the challenges of limited data in Arabic SER by investigating the effect of multi-technique data augmentation on the performance of Convolutional Neural Networks (CNNs). Instead of proposing a novel SER architecture, we focus on enhancing the training process using augmentation techniques to improve the model's ability to generalize across different dialects and emotional expressions. Specifically, we employ two datasets: the Saudi dialect dataset [10], which captures regional variations in Arabic, and the BAVED (Basic Arabic Vocal Emotions Dataset), which represents standard Arabic emotions [11]. By applying various augmentation techniques, we aim to evaluate their effectiveness in improving emotion recognition accuracy and determine the optimal number of augmentations to enhance model performance.

This paper is structured as follows: Section 2 presents a review of related work in SER, with a particular focus on Arabic speech emotion recognition and data augmentation techniques. Section 3 then provides an explanation of the proposed approach and system architecture. Section 4 presents and discusses the results, while Section 5 concludes with a summary of the key findings and an outline of future perspectives.

## 2. Related Work

In recent years, neural network-based modelling approaches, including variants such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid architectures, have made notable strides in the realm of speech emotion recognition. This has led to a significant enhancement in the capacity to accurately detect and classify emotions from spoken language. One of the key advantages of deep learning models is their ability to automatically learn hierarchical representations of data. RNNs are particularly good at capturing temporal relationships in sequential data, like speech, they have been utilized extensively for speech emotion recognition. RNNs can represent the dynamics and context of emotional expressions by processing speech signals across time. This allows for more precise emotion recognition. CNNs, however, are good at picking up spectral characteristics and local patterns in voice signals. They can examine the spectral content of brief speech segments, essential for differentiating between emotional states. Furthermore, advancements in deep learning architectures, such as long short-term memory (LSTM) [12] networks, gated recurrent units (GRUs) [13], and attention mechanisms [14], have further improved the performance of deep learning models for speech emotion recognition.

While SER has seen significant progress in languages like English, Spanish, and German, the research focus on Arabic speech emotion recognition is relatively limited. This is notable given Arabic's status as one of the most widely spoken languages globally. Recent studies have attempted to fill this gap by developing models and datasets for Arabic SER. For instance, Meftah et al. [15] explored the application of CNN models for Arabic speech emotion recognition using the KSUEmotions corpus. The authors demonstrated that CNN models outperformed traditional machine learning approaches, effectively capturing local features in speech signals. Furthermore, the study also compared the performance of CNNs with other architectures like ResNet and CRNN, highlighting CNNs' advantages in multilingual emotion recognition tasks. This contribution enriches the understanding of CNN's potential for SER, particularly in the context of Arabic language datasets. In [16], the authors proposed two models for emotion recognition using the KSUEmotions dataset. The first approach integrates an attention mechanism with a convolutional neural network (CNN), while the second employs a deep CNN. The attention-based model achieved a 2.2% improvement in performance over the CNN baseline, whereas the deep CNN demonstrated faster training and classification times.

 doi:10.20944/preprints202501.0126.v1



The aforementioned studies have established significant foundations; nevertheless, challenges persist due to the paucity of labelled Arabic datasets and the variability between Arabic dialects. To address these challenges, researchers such as Shahin et al. [17] and Abdel-Hamid et al. [18] have presented datasets focusing on specific dialects, including Emirati Arabic and Egyptian. This work underscores the growing recognition of the significance of dialect variation in Arabic SER.

One critical area for improvement in Speech Emotion Recognition (SER) models, particularly those for languages like Arabic where data availability is limited, is the implementation of data augmentation techniques. The utilization of techniques such as noise injection, pitch shifting, time warping, and volume adjustments is a common practice with the objective of artificially increasing the diversity of training datasets. The rationale behind this approach is to enhance model robustness and generalization.

Recent studies have demonstrated the efficacy of these data augmentation techniques in SER. Padi et al. [19] employed random time-frequency masks on log-mel spectrograms to generate supplementary training samples, effectively mitigating overfitting and enhancing generalization, as demonstrated on the IEMOCAP dataset. Similarly, Paraskevopoulou et al. [20] introduced techniques such as noise addition, audio shifting, and adjustments in pitch and speed to enhance CNN-based emotion recognition using the EMOVO dataset. This resulted in superior performance compared to previous methods that did not incorporate data augmentation. Furthermore, the authors of [21] investigated the impact of data augmentation on Japanese Twitter-based emotional speech and IEMOCAP datasets, identifying glottal source extraction and silence removal as particularly effective for speaker-independent data. The results demonstrated that while augmenting the number of instances can enhance performance for text-independent data, selecting appropriate augmentation methods based on specific conditions is crucial. This indicates a trade-off between the number of augmentations and recognition performance.

This research contributes to the existing literature by examining the impact of multi-technique data augmentation on Arabic speech emotion recognition using Convolutional Neural Networks (CNNs). This study employs both the Saudi dialect dataset and the BAVED dataset, which encompass disparate dialectal variations in Arabic speech, with the objective of addressing the existing gap in the literature regarding effective augmentation strategies for Arabic emotion recognition. The study examines the efficacy of combining noise addition, volume adjustments and time shifting as a means of enhancing the robustness and generalization of Arabic speech emotion recognition systems.

## 3. Methodology

### *3.1. Datasets*

#### 3.1.1. Saudi Dialect Dataset

This study utilized the semi-natural emotion speech dataset in the Saudi dialect [10], which was curated from YouTube videos sourced from the Saudi YouTube channel, Telfaz11 [22]. Telfaz11 is an entertainment platform known for producing diverse content, including short films, comedy clips, and shows aimed at showcasing Saudi culture and region-specific information. The dataset consists of 175 records, with 113 segments featuring male actors and 62 segments featuring females. It encapsulates four primary emotions: anger, happiness, sadness, and neutral. The distribution of emotions in the dataset is as follows: 69 records for anger, 31 for happiness, 37 for neutral, and 38 for sadness. In Figure 1, we present the distribution of emotions in the Saudi dialect dataset.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2025 doi:10.20944/preprints202501.0126.v1



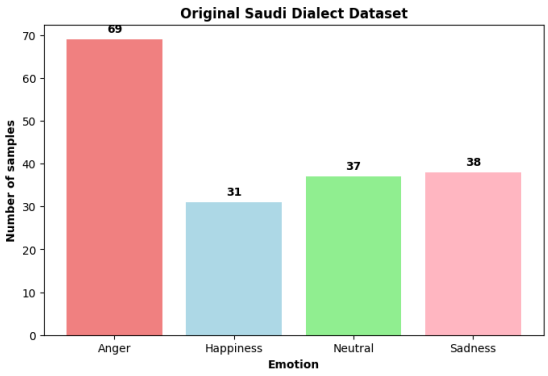


**Figure 1.** Original Saudi dialect Dataset distribution.

3.1.2. BAVED Dataset

Basic Arabic Vocal-Emotions Dataset (BAVED) is a collection of audio/wav recorded Arabic words spoken in various expressed emotions [11]. The BAVED dataset includes 7 words given as 0- أعجبني = like, 1- لم يعجبني = unlike, 2- هذا = this, 3- الفيلم = film, 4- رائع = good, 5- مقبول = neutral, and 6- سيئ = bad. The BAVED dataset comprises 1935 recordings, made by 61 speakers (45 males and 16 females), in which the word is pronounced at three levels, corresponding to the speaker's emotional state. The levels are as follows: 0 for low emotion (tired or exhausted), 1 for neutral emotion, and 2 for high emotion (positive or negative emotions, e.g. happiness, joy, sadness, anger). Figure 2 illustrates the distribution of recorders across each emotion level.

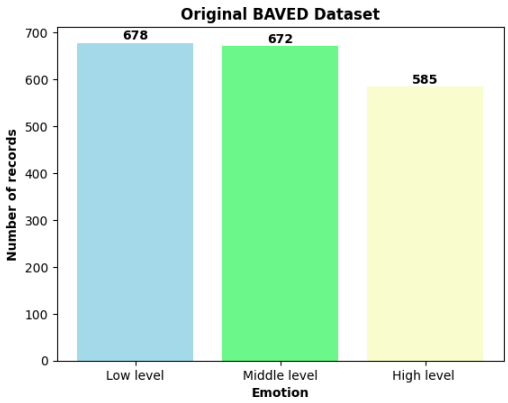


**Figure 2.** Original BAVED dataset distribution.

*3.2. Data Augmentation*

Due to the limited availability of Saudi dialogue datasets, we intend to increase the dataset's size by utilizing various data augmentation techniques. This is an essential step to ensure the success of the project. Data augmentation is a simple yet effective method for expanding the training dataset's size and diversity, especially in cases where obtaining a large, annotated dataset is not feasible. Data augmentation employs diverse processing techniques to generate new training samples from the original dataset while preserving the accuracy of the original class labels. This approach addresses overfitting concerns and enhances the model's resilience and capacity for generalization through training with augmented data. In the field of audio signal processing, this entails precisely modifying and transforming existing audio samples to generate new samples that maintain the essential features of the original audio, while introducing deliberate changes or perturbations. Figure 3 illustrates the process of data augmentation in SER, showcasing the application of various techniques such as noise addition, increasing volume, reducing volume, and time shifting. The figure also highlights the impact of using different numbers of augmentations, ranging from one to four, applied individually or in combination to enhance the diversity and robustness of the dataset.

 doi:10.20944/preprints202501.0126.v1



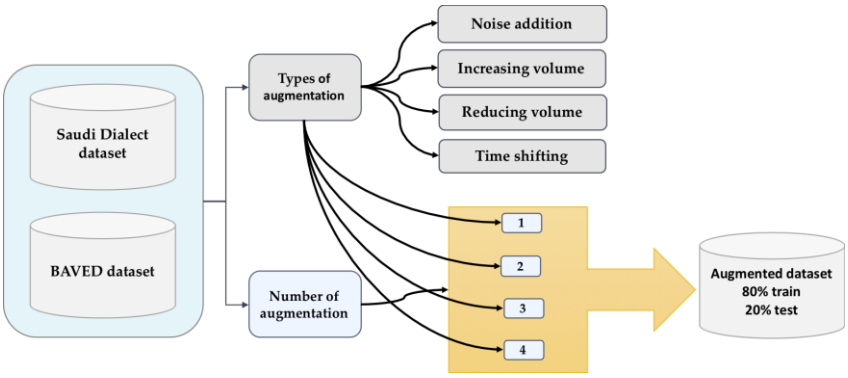


**Figure 3.** The flow of data preparation for the SER.

3.2.1. Noise Addition

The addition of noise is a fundamental data augmentation technique used in audio signal processing, where simulated background noise is added to the original speech signal. This technique is essential for enhancing the model's resilience by exposing it to variations in acoustic environments, enabling the system to generalize effectively in real-world scenarios where ambient noise is unavoidable. The pseudocode for noise addition is given in Algorithm 1. This algorithm uses a normal distribution to generate realistic background noise, enhancing the model's ability to generalize in noisy environments.

**Algorithm 1** Data augmentation: Noise addition

**Input:**
audio_data: Numpy array representing the original audio time series data.
**Output:**
augmented_audio: Numpy array representing the augmented audio data.
1: noise_ratio ← 0.05
// Calculate Noise Standard Deviation
2: noise_std ← noise_ratio * np.max(np.abs(audio_data))
// Generate Random Noise from Normal Distribution
3: noise ← np.random.normal(0, noise_std, len(audio_data))
// Add Noise to Audio Data:
4: augmented_audio ← np.clip(audio_data + noise, -1.0, 1.0)
5: **return** augmented_audio

3.2.2. Time Shifting

Time shifting refers to the manipulation of the entire audio signal, whereby it is moved forward or backward in time. This technique simulates variations in the timing of audio events, thereby assisting the model in recognizing signals that may not always commence at the same point in time. In our work, we shift the spectrogram in the right direction. The pseudocode for time shifting is presented in Algorithm 2. In this study, we shift the spectrogram to the right.

**Algorithm 2** Data augmentation: Time shifting

**Input:**
audio_data: Numpy array representing the original audio time series data.
**Output:**
augmented_audio: Numpy array representing the augmented audio data.
1: shift_amounts ← 0.1
// Calculate Shift in Samples
2: shift_samples ← int(shift_amount * len(audio_data))
// Apply Time shifting

 doi:10.20944/preprints202501.0126.v1



3: augmented_audio ← np.roll(audio_data, shift_samples)
// Fill with zeros at the start
4: augmented_audio[shift_samples:] ← 0
5: **return** augmented_audio

3.2.3. Increasing Volume

Increasing volume is a data augmentation technique used in machine learning, particularly in audio signal processing, to artificially enhance the training dataset by adjusting the amplitude or loudness of the original audio signals. The process involves scaling the amplitude of the audio waveform, effectively amplifying the volume. Algorithm 3 outlines the pseudocode for implementing the volume-increasing augmentation technique.

**Algorithm 3** Data augmentation: Increasing volume
**Input:**
audio_data: Numpy array representing the original audio time series data.
**Output:**
augmented_audio: Numpy array representing the augmented audio data.
1: gain_factor ← 2
// Apply Volume Increase:
2: augmented_audio ← audio_data * gain_factor
// Clip Values to Ensure Valid Audio Range:
3: augmented_audio ← np.clip(augmented_audio, -1.0, 1.0)
4: **return** augmented_audio

3.2.4. Reducing Volume

Reducing volume is a data augmentation technique that has found extensive application in machine learning, particularly in the domain of audio signal processing. This technique aims to expand the diversity of training datasets by manipulating the amplitude or loudness of audio signals. This technique involves decreasing the amplitude of the audio waveform, thereby reducing the volume. The pseudocode for the volume-reducing augmentation technique is presented in Algorithm 4.

**Algorithm 4** Data augmentation: Reducing volume
**Input:**
audio_data: Numpy array representing the original audio time series data.
**Output:**
augmented_audio: Numpy array representing the augmented audio data.
1: reduction_factor ← 0.5
// Apply Volume Reduction
2: augmented_audio ← audio_data * reduction_factor
// Clip Values to Ensure Valid Audio Range:
3: augmented_audio ← np.clip(augmented_audio, -1.0, 1.0)
4: **return** augmented_audio

Figure 4 shows examples of data augmentations applied to the Saudi dialect dataset using the techniques described above. The sample used corresponds to an angry emotion (01). The leftmost example represents the original sample (a) from the dataset, while the remaining examples illustrate augmented data generated through noise addition (b), time shifting (c), increasing volume (d), and reducing volume (e). Clear differences can be observed among the augmented data. For instance, in the noise addition example, random perturbations are introduced into the waveform, making it appear more irregular. In the time shifting example, the waveform is shifted slightly along the time axis, altering its alignment without affecting its overall structure. Similarly, increasing the volume

 doi:10.20944/preprints202501.0126.v1



amplifies the waveform's amplitude, making it more pronounced, while decreasing the volume reduces the amplitude, resulting in a less prominent waveform. These variations demonstrate how each augmentation technique uniquely transforms the original data to improve the model's generalization and robustness.

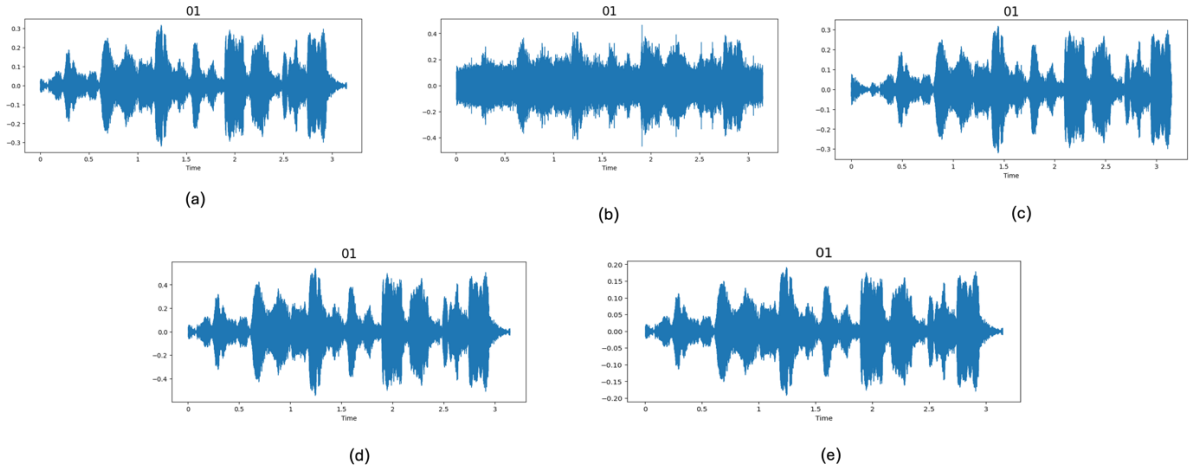


**Figure 4.** Examples of the audio files with data augmentation, (a) original audio for an angry emotion (01), (b) noise addition, (c) time shift, (d) increasing volume and (e) reducing volume.

In addition to evaluating the efficacy of the four individual data augmentation techniques, we also assessed the performance of various combinations of these techniques. Specifically, the combinations included six pairwise arrangements, four configurations involving three augmentations, and one configuration incorporating all four augmentations. Table 1 provides an overview of the number of samples generated for each data augmentation configuration.

**Table 1.** Number of data in each number of data augmentation.

| Number of data augmentations | Number of data after augmentation | |
|---|---|---|
| | Saudi dialect dataset | BAVED dataset |
| Without augmentation | 175 | 1935 |
| With one augmentation | 350 | 3870 |
| With two augmentations | 525 | 5805 |
| With three augmentations | 700 | 7740 |
| With four augmentations | 875 | 9675 |

*3.3. Feature Extraction*

Mel Frequency Cepstral Coefficient (MFCC) is the most common feature extraction technique used for processing the human voice to calculate the cepstral coefficients with the consideration of human hearing. The block diagram in Figure 5 shows the steps to compute the MFCC.

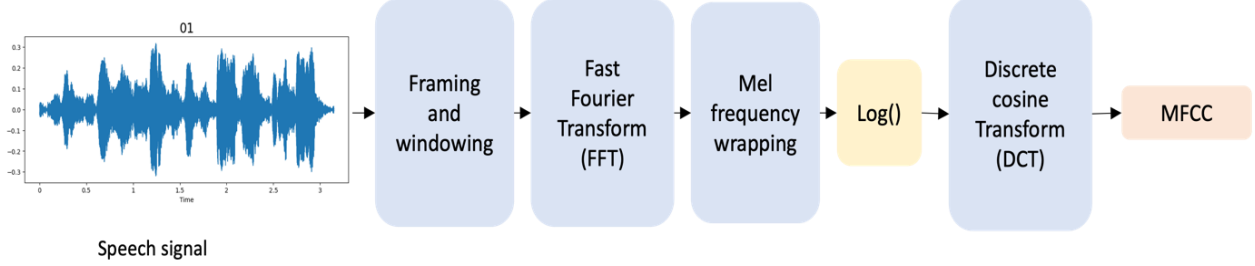


**Figure 5.** The block diagram of the MFCC computation.

 doi:10.20944/preprints202501.0126.v1



*3.4. CNN Architecture*

Convolutional neural networks (CNNs) are a prevalent tool in the field of speech emotion recognition, due to their capacity to automatically extract and analyze intricate emotional patterns in speech. In a CNN-based SER system, the input speech signal is converted into a visual representation, such as a spectrogram or MFCC, which enables the network to recognize the spatial and temporal features associated with the emotional content of the speech. Convolutional layers apply filters to identify fundamental cues such as pitch variations and energy shifts, while pooling layers downsample in order to retain key features in an efficient manner. Dense layers map the aforementioned features to emotion labels, thereby enabling CNNs to effectively classify emotions such as happiness, sadness, or anger based on input speech. Figure 6 presents our proposed architecture.

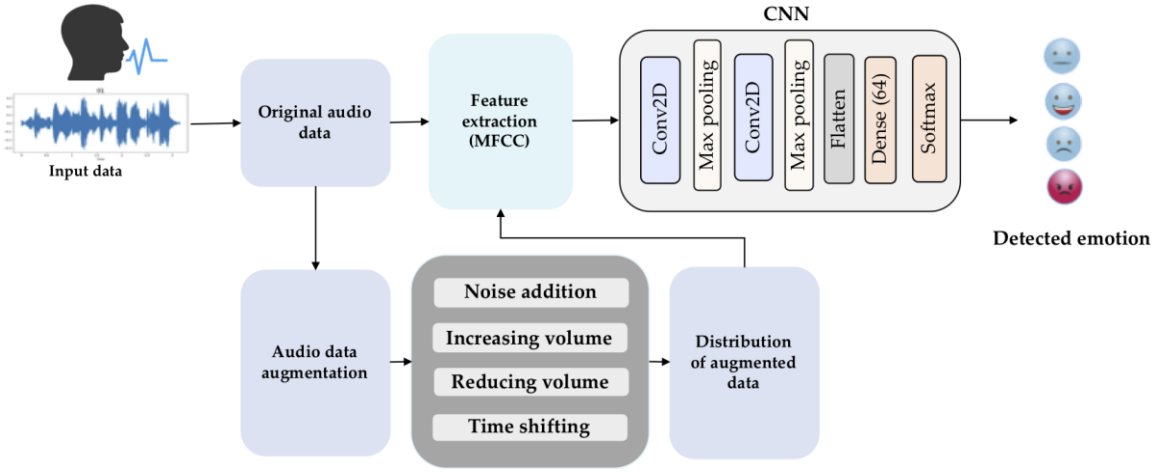


**Figure 6.** The proposed SER architecture

*3.4. Evaluation Metrics:*

In order to evaluate the performance of the model, four key metrics (accuracy, precision, recall, and F1-score) were calculated during both the training and testing phases. The mathematical equations of these metrics can be expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

where TP, FP, TN, and FN respectively represent the number of true positives, false positive, true negative, and false negative samples.

## 4. Results and Discussion

The results of our experiments, presented in Tables 2 and 3, demonstrate the substantial impact of various data augmentation techniques and their combinations on the performance of the SER model for both the Saudi Dialect and BAVED datasets. Performance metrics, including accuracy, recall, precision, and F1-score, were evaluated on a 0–100% scale to ensure clarity and facilitate comparison. The findings highlight the effectiveness of augmentation techniques in enhancing model performance across diverse datasets.

 doi:10.20944/preprints202501.0126.v1



**Table 2.** Classification report using various data augmentation techniques and different numbers of augmentations on the Saudi dialect dataset (OD = Original Dataset). The highest scores are in bold. .

| Data | Accuracy (%) | Recall (%) | Precision (%) | F1_Score (%) |
|---|---|---|---|---|
| **Without augmentation** | | | | |
| Original Dataset (OD) | 42.85 | 42.85 | 46.79 | 43.49 |
| **With one augmentation** | | | | |
| OD + noise(0.05) | 74.60 | 74.60 | 77.76 | 72.46 |
| OD + volumeUp(2) | 88.89 | 88.89 | 89.83 | 88.61 |
| OD + volumeDown(0.5) | 87.30 | 87.30 | 87.81 | 86.57 |
| OD + timeShift(0.1) | 68.25 | 68.25 | 75.09 | 66.32 |
| **With two augmentations** | | | | |
| OD + noise(0.05) + volumeUp(2) | 89.36 | 89.36 | 90.14 | 89.41 |
| OD + noise(0.05) + volumeDown(0.5) | 90.43 | 90.43 | 91.05 | 90.33 |
| OD + noise(0.05) + timeShift(0.1) | 84.04 | 84.04 | 86.33 | 84.51 |
| OD + volumeUp (2) + volumeDown(0.5) | **96.81** | **96.81** | **97.01** | **96.76** |
| OD + volumeUp(2) + timeShift(0.1) | 89.36 | 89.36 | 90.02 | 89.18 |
| OD + volumeDown(0.8) + timeShift(0.1) | 88.30 | 88.30 | 89.68 | 88.00 |
| **With three augmentations** | | | | |
| OD + noise(0.05) + volumeUp(2) + volumeDown(0.5) | 95.24 | 95.24 | 95.77 | 95.26 |
| OD + noise(0.05) + volumeDown(0.5) + timeShift(0.1) | 85.71 | 85.71 | 86.59 | 85.86 |
| OD + noise(0.02) + volumeUp(2) + timeShift(0.1) | 92.86 | 92.86 | 93.47 | 92.72 |
| OD + volumeUp(2) + volumeDown(0.5) + timeShift(0.1) | 95.20 | 95.20 | 95.36 | 95.16 |
| **With four augmentations** | | | | |
| OD + noise(0.05) + volumeUp(2) + volumeDown(0.5) + timeShift(0.1) | 96.13 | 96.13 | 96.26 | 96.11 |

**Table 3.** Classification report using various data augmentation techniques and different numbers of augmentations on the BAVED dataset (OD = Original Dataset). The highest scores are in bold. .

| Data | Accuracy (%) | Recall (%) | Precision (%) | F1_Score (%) |
|---|---|---|---|---|
| **Without augmentation** | | | | |
| Original Dataset (OD) | 65.89 | 65.89 | 66.36 | 65.85 |
| **With one augmentation** | | | | |
| OD + noise(0.05) | 72.28 | 72.28 | 72.71 | 72.35 |
| OD + volumeUp(2) | 75.32 | 75.32 | 78.79 | 75.30 |
| OD + volumeDown(0.5) | 78.14 | 78.14 | 78.18 | 78.08 |
| OD + timeShift(0.1) | 76.44 | 76.44 | 78.67 | 76.15 |
| **With two augmentations** | | | | |
| OD + noise(0.05) + volumeUp(2) | 82.31 | 82.31 | 82.69 | 82.22 |
| OD + noise(0.05) + volumeDown(0.5) | 86.54 | 86.54 | 86.80 | 86.59 |
| OD + noise(0.05) + timeShift(0.1) | 80.86 | 80.86 | 81.62 | 80.96 |
| OD + volumeUp (2) + volumeDown(0.5) | 88.45 | 88.45 | 89.00 | 88.47 |
| OD + volumeUp(2) + timeShift(0.1) | 83.72 | 83.72 | 84.03 | 83.79 |
| OD + volumeDown(0.8) + timeShift(0.1) | 83.79 | 83.79 | 84.23 | 83.88 |
| **With three augmentations** | | | | |

 doi:10.20944/preprints202501.0126.v1



| | | | | |
|---|---|---|---|---|
| OD + noise(0.05) + volumeUp(2) + volumeDown(0.5) | **92.60** | **92.60** | **92.73** | **92.63** |
| OD + noise(0.05) + volumeDown(0.5) + timeShift(0.1) | 88.16 | 88.16 | 88.18 | 88.13 |
| OD + noise(0.02) + volumeUp(2) + timeShift(0.1) | 89.39 | 89.39 | 89.41 | 89.39 |
| OD + volumeUp(2) + volumeDown(0.5) + timeShift(0.1) | 89.14 | 89.14 | 89.13 | 89.07 |
| **With four augmentations** | | | | |
| OD + noise(0.05) + volumeUp(2) + volumeDown(0.5) + timeShift(0.1) | 89.4 | 89.4 | 89.6 | 89.2 |

*4.1. Saudi Dialect Dataset*

For the Saudi Dialect dataset, the baseline model without augmentation achieved an accuracy of 42.85%, highlighting the challenges associated with this dataset. The application of individual augmentations significantly improved performance, with volume adjustments yielding the best results. Specifically, decreasing the volume (OD + VolumeUp) achieved the highest accuracy of 88.89% and an F1-score of 88.61%.

When two augmentation techniques were combined, the best results were achieved with the combination of volume increase and volume decrease (OD + VolumeUp + VolumeDown), resulting in an accuracy of 96.81% and an F1-score of 96.76%. These findings align with the type of data augmentation techniques (one augmentation), where augmentations involving volume increase and volume reduction outperformed other types of data augmentation. Adding a third augmentation technique, such as noise, slightly reduced performance compared to the best two-augmentation setup, but it still resulted in high scores. The application of all four augmentations (OD + Noise + VolumeUp + VolumeDown + TimeShift) also led to a slight reduction in performance.

*4.2. BAVED Dataset*

The BAVED dataset also showed notable improvements with data augmentation, though the baseline accuracy of 65.89% was higher compared to the Saudi Dialect dataset. Among individual augmentations, reducing the volume (OD + volumeDown) achieved the best performance, with an accuracy of 78.14% and an F1-score of 78.08%. Combining two augmentations, specifically volume increase and volume decrease (OD + volumeUp + volumeDown), resulted in the highest accuracy of 88.45% and an F1-score of 88.47%. Adding a third augmentation, such as noise, yielded the highest overall performance, with an accuracy of 92.60% and an F1-score of 92.63%. However, incorporating all four augmentation techniques resulted in similar performance (accuracy: 89.40%, F1-score: 89.20%), indicating diminishing returns with the addition of more augmentations.

A comparison of the results across the two datasets reveals that the Saudi Dialect dataset exhibited a more substantial enhancement following data augmentation, with an improvement of over 54 percentage points in accuracy from the baseline to the best-performing model. Conversely, the BAVED dataset demonstrated a more modest enhancement of 26 percentage points, likely attributable to its superior initial quality and equilibrium. This finding suggests that the impact of augmentation techniques is more pronounced in datasets that present greater challenges in terms of balance or representation.

*4.3. Comparative Study*

The results of this study demonstrate substantial improvements over previous works that utilized the BAVED and Saudi Dialect datasets for Speech Emotion Recognition (SER). For the Saudi Dialect dataset, where prior studies [10] faced challenges due to the dataset's limited size and imbalance, baseline models without augmentation often achieved accuracies below 77.14%. Our findings demonstrate a substantial improvement, with the best results obtained using the

 doi:10.20944/preprints202501.0126.v1



combination of volume increase and volume decrease, yielding an accuracy of 96.81% and an F1-score of 96.76%. Compared to previous works that utilized traditional feature extraction methods or simpler classifiers, our results showcase the superior performance of CNNs when combined with appropriate augmentation techniques.

For the BAVED dataset, prior studies such as [23] reported accuracies ranging between 84% and 89%, depending on the features and models used. In contrast, our approach, which incorporates three data augmentation techniques—noise addition, volume increase, and volume decrease—achieved a significantly higher accuracy of 92.60% and an F1-score of 92.63%, demonstrating the effectiveness of our augmentation strategy in enhancing CNN performance. These findings validate the importance of augmentation in addressing dataset limitations and emphasize the potential of our approach for advancing Arabic SER systems. As shown in Table 4, a comparison with the state-of-the-art methods demonstrates that our model achieved superior performance over existing research.

**Table 4.** Comparing Our Model with State-of-the-Art Models.

| Model | Accuracy (%) |
|---|---|
| Saudi dialect SER [10] | 77.14 |
| BAVED SER [23] | 89 |
| Our Model SER with Saudi dialect dataset | 96.81 |
| Our Model SER with BAVED dataset | 92.63 |

## 5. Conclusions

This study investigated the impact of data augmentation techniques on the performance of a speech emotion recognition (SER) model for the Saudi dialect and BAVED datasets. A systematic approach was employed to evaluate the impact of single and combined augmentations, including noise addition, volume increase, volume decrease, and time shifting, on classification performance. The results demonstrated a notable enhancement in accuracy, recall, precision, and F1-score for both datasets, with the Saudi Dialect dataset exhibiting a more pronounced benefit from augmentation due to its inherent complexity.

In the case of the Saudi Dialect dataset, the highest level of performance was achieved by combining two augmentation techniques, resulting in an accuracy of 96.81% and an F1-score of 96.76%. Similarly, the BAVED dataset demonstrated considerable improvement, with a combination of three augmentations producing the most favorable results, achieving an accuracy of 92.60% and an F1-score of 92.63%. These findings emphasize the significance of data augmentation in addressing challenges specific to a given dataset, particularly for underrepresented dialects or less balanced datasets.

The comparative analysis also demonstrated that the selection and combination of augmentation techniques are of critical importance in improving SER performance. While individual augmentations provided substantial improvements, the combination of complementary techniques yielded the most favorable results, thereby illustrating the potential of augmentation to enhance model generalization and robustness.

Further work could investigate additional augmentation methods, such as pitch shifting or advanced signal processing techniques, to enhance model performance further. Furthermore, extending the scope to encompass real-world datasets and cross-dialect assessment could furnish invaluable insights into the generalizability and scalability of the proposed methodologies. These findings provide researchers and practitioners with practical guidance for the development of more robust and accurate SER systems, particularly for dialect-specific applications.

**Author Contributions:** Conceptualization, W.B. and R.A.; methodology, W.B. and R.A.; software, W.B. and R.A.; validation, W.B. and R.A.; formal analysis, W.B. and R.A.; investigation, W.B. and R.A.; resources, W.B. and R.A.;

 doi:10.20944/preprints202501.0126.v1



data curation, W.B. and R.A.; writing—original draft preparation, W.B. and R.A.; writing—review and editing, W.B., R.A. and A.A.; visualization, W.B., R.A. and A.A.; supervision, A.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Cowie, R.; Douglas-Cowie, E.; Savvidou, S.; McMahon, E.; Sawey, M.; Schröder, M. FEELTRACE: An Instrument for Recording Perceived Emotion in Real Time; **2000**.
2. Low, L.-S. A.; Maddage, N. C.; Lech, M.; Sheeber, L.; Allen, N. Influence of Acoustic Low-Level Descriptors in the Detection of Clinical Depression in Adolescents. In **2010** IEEE International Conference on Acoustics, Speech and Signal Processing; 2010; pp 5154–5157. https://doi.org/10.1109/ICASSP.2010.5495018.
3. Spezialetti, M.; Placidi, G.; Rossi, S. Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. Front. Robot. AI **2020**, 7. https://doi.org/10.3389/frobt.2020.532279.
4. Kozlov, P.; Akram, A.; Shamoi, P. Fuzzy Approach for Audio-Video Emotion Recognition in Computer Games for Children. arXiv August 31, **2023**. https://doi.org/10.48550/arXiv.2309.00138.
5. El Ayadi, M.; Kamel, M. S.; Karray, F. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. Pattern Recognit. **2011**, 44 (3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020.
6. Barakat, A.; Al Hammadi, O.; Aldhaheri, A.; Elnagar, A. Arabic Dialect Identification from Speech. In 2024 15th Annual Undergraduate Research Conference on Applied Computing (URC); **2024**; pp 1–6. https://doi.org/10.1109/URC62276.2024.10604557.
7. Meftah, A. H.; Qamhan, M.; Alotaibi, Y. A.; Zakariah, M. Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus. Int. J. Simul. Syst. Sci. Technol. **2020**. https://doi.org/10.5013/IJSSST.a.21.02.21.
8. Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; Le, Q. V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. arXiv December 3, **2019**. https://doi.org/10.48550/arXiv.1904.08779.
9. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio Augmentation for Speech Recognition. In Interspeech 2015; ISCA, **2015**; pp 3586–3589. https://doi.org/10.21437/Interspeech.2015-711.
10. Aljuhani, R. H.; Alshutayri, A.; Alahdal, S. Arabic Speech Emotion Recognition From Saudi Dialect Corpus. IEEE Access **2021**, 9, 127081–127085. https://doi.org/10.1109/ACCESS.2021.3110992.
11. Aouf, A. Basic-Arabic-Vocal-Emotions-Dataset, **2024**. https://github.com/40uf411/Basic-Arabic-Vocal-Emotions-Dataset (accessed 2024-6-15).
12. Zhao, J.; Mao, X.; Chen, L. Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks. Biomed. Signal Process. Control **2019**, 47, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035.
13. Rayhan Ahmed, Md.; Islam, S.; Muzahidul Islam, A. K. M.; Shatabda, S. An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition. Expert Syst. Appl. **2023**, 218, 119633. https://doi.org/10.1016/j.eswa.2023.119633.
14. Chen, S.; Zhang, M.; Yang, X.; Zhao, Z.; Zou, T.; Sun, X. The Impact of Attention Mechanisms on Speech Emotion Recognition. Sensors **2021**, 21 (22), 7530. https://doi.org/10.3390/s21227530.
15. Meftah, A. H; Qamhan, M. A.; Seddiq, Y.; Alotaibi, Y. A.; Selouani, S. A. King Saud University Emotions Corpus: Construction, Analysis, Evaluation, and Comparison. *IEEE Access* **2021**, vol. 9, pp. 54201-54219. doi: 10.1109/ACCESS.2021.3070751.
16. Hifny, Y.; Ali, A. Efficient Arabic Emotion Recognition Using Deep Neural Networks. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); **2019**; pp 6710–6714. https://doi.org/10.1109/ICASSP.2019.8683632.

17. Shahin, I.; Nassif, A. B.; Hamsa, S. Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network. IEEE Access **2019**, 7, 26777–26787. https://doi.org/10.1109/ACCESS.2019.2901352.
18. Abdel-Hamid, L. Egyptian Arabic Speech Emotion Recognition Using Prosodic, Spectral and Wavelet Features. Speech Commun. **2020**, 122, 19–30. https://doi.org/10.1016/j.specom.2020.04.005.
19. Padi, S.; Sadjadi, S. O.; Manocha, D.; Sriram, R. D. Improved Speech Emotion Recognition Using Transfer Learning and Spectrogram Augmentation. arXiv August 16, **2021**. https://doi.org/10.48550/arXiv.2108.02510.
20. Paraskevopoulou, G.; Spyrou, E.; Perantonis, S. A Data Augmentation Approach for Improving the Performance of Speech Emotion Recognition; **2024**; pp 61–69.
21. Atmaja, B. T.; Sasou, A. Effects of Data Augmentations on Speech Emotion Recognition. Sensors **2022**, 22 (16), 5941. https://doi.org/10.3390/s22165941.
22. telfaz11. telfaz11. https://telfaz11.com/ (accessed 2024-6-21).
23. Mohamed, O.; Aly, S. A. Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset. arXiv October 9, **2021**. http://arxiv.org/abs/2110.04425.