Emotion Recognition in Arabic Speech from Saudi Dialect Corpus Using Machine Learning and Deep Learning Algorithms

Hanaa Alamri† **and Hanan S. Alshanbari**†,

† Department of Computer Science, College of Computer Science and Information System, Umm Al-Qura University,
Makkah, Saudi Arabi

Abstract
Speech can actively elicit feelings and attitudes by using words. It is important for researchers to identify the emotional content contained in speech signals as well as the sort of emotion that resulted from the speech that was made. In this study, we studied the emotion recognition system using a database in Arabic, especially in the Saudi dialect, the database is from a YouTube channel called Telfaz11, The four emotions that were examined were anger, happiness, sadness, and neutral. In our experiments, we extracted features from audio signals, such as Mel Frequency Cepstral Coefficient (MFCC) and Zero-Crossing Rate (ZCR), then we classified emotions using many classification algorithms such as machine learning algorithms (Support Vector Machine (SVM) and K-Nearest Neighbor (KNN)) and deep learning algorithms such as (Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM)). Our Experiments showed that the MFCC feature extraction method and CNN model obtained the best accuracy result with 95%, proving the effectiveness of this classification system in recognizing Arabic spoken emotions.

Keywords:
Speech Emotion Recognition, Arabic speech, Saudi dialect, KNN, SVM, CNN, LSTM.

1. Introduction

Human-Computer Interaction (HCI) is vital for better understanding and communicating each other's goals. Because humans have a natural ability to discern emotions, researchers aimed to develop algorithms that could detect human emotions using various approaches. However, the machine must have a lot of information about human emotions to understand the emotions in speech. While speech recognition has improved significantly, computer-human interaction is still far from being a common occurrence, as computers are unable to comprehend the state of human emotions [1].

Speech Emotions Recognition (SER) can be defined as a method of detecting and recognizing emotions through human expression [2]. When building artificial intelligence services, Speech Emotion Recognition is critical for understanding user emotions [3].

For years, speech recognition had been a hallmark of science fiction, but in 1976, real-world capabilities paled in comparison to the fictitious realm's far-fetched capabilities. In 1971, Allen Newell, the chair of a speech recognition study committee, advised that many additional sources of knowledge be brought to bear on the problem [4]. Six levels of knowledge were examined in the report: acoustic, parametric, phonemic, lexical, phrase, and semantic.

By 1976, Reddy was leading a team at Carnegie Mellon University that created the Hearsay, Dragon, Harpy, and Sphinx I/II speech recognition systems [4].

The approaches for automatically recognizing emotions from voice signals have lately matured greatly, particularly for real-world settings such as contact centers, conversation analyses used to improve service quality, analysis of incoming calls to emergency services, additional disease analysis, and distance education [5].

The speaker has a huge impact on the speech characteristic values. Individual differences in these parameters make categorization difficult. Multidimensional complicated data cannot be properly manipulated by a typical classification method. This justification motivates the usage of deep learning algorithms in the SER system[6].

For SER systems, many researchers present learning methods such as ANN (Artificial neural network), LSTM (Long short term memory), CNN (Convolutional neural network), and SVM (support vector machine).

Convolutional Neural Networks (CNNs), for example, are deep learning systems that may infer a hierarchical representation of input data to aid in categorization [7]. CNN is a multi-layer perceptron that was specifically created to recognize two-dimensional shapes. CNN effectively uses the dimensions information stored in waveform points as a result [6]. CNN model is used for voice emotion

Manuscript received August 5, 2023
Manuscript revised August 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.8.2

recognition and image recognition because of its adaptive feature extraction properties.

In this paper, section II represents Literature reviews related to the Arabic speech emotion recognition system, section III describes Dataset, section IV describes implemented system methodology section V describes the results observed in the implemented system.

Classification of the learning fields with mixed certificates according to their level of acceptance by job market is important to universities and decision makers.

Studying the effect of including some specific online courses in the traditional curriculum is important to scientific departments such as computer science.

Our findings and recommendations could benefit the job market as well as the universities.

This paper is organized as follows. The next section describes the methodology we followed int his paper. Section 3 discusses the companies' survey results. In section 4, we discuss the results of the students' survey.

Section 5 summarizes the results and gives some recommendations. Finally, in section 6, we give some concluding remarks.

2. Literature Reviews

Many studies have focused on speech emotion recognition, where researchers studied the subject in various languages by using many artificial intelligence algorithms.

In comparison to other languages, the literature on emotional speech recognition for the Arabic language is lacking. So, The King Saud University Emotions (KSUEmotions) corpus was created and verified by the researchers in this study[3].

The emotional speech of twenty-three speakers from Saudi Arabia, Syria, and Yemen is included in KSUEmotions, and it comprises the following emotions: neutral, happiness, sadness, surprise, and anger. The researchers used two methods to verify the content of the data. The first is human verification in this step they used nine human listeners to categorize the sounds. The second method is by automatic classification of feelings by using artificial intelligence algorithms. In addition, by using the Emotional Prosody Speech and Transcripts Corpus, this study also tests emotion recognition for the English language (EPST).

There are several researchers interested in investigating and classifying emotions by using the King Saud University Emotions (KSUEmotions) corpus, for example, Meftah et al [8] presented the results of cross-corpus emotion recognition in the Arabic and English languages. For Arabic speech, the King Saud University Emotions (KSUEmotions) corpus is utilized, and for English speech, the Emotional Prosody Speech and Transcripts (EPST) corpus is used. Deep Belief Networks (DBN) and Multi-Layer Perceptron (MLP) classifiers were utilized in the study.

Another study that applied SER to detect emotions by using (KSUEmotions) was by Hifny et al [9] they were able to work on developing the CNN model result in the essential study by using deep neural networks. They created an attention-based CNN-LSTM-DNN emotion classifier. The convolutional layers (CNN) extract significant characteristics in their classifier, while the bidirectional long short-term memory (BLSTM) layers handle the speech signal's sequential occurrences. Their new strategy can lead to significant increases (2.2 percent absolute improvements) over a powerful deep CNN baseline system on an Arabic speech emotion recognition test, according to the results.

The researchers were also able to delve into the complexities of the Arabic language, working on Speech Emotions Recognition in a variety of Saudi dialects. AL Juhani et al [10], studied Arabic Speech Emotion Recognition from Saudi Dialect Corpus by using different algorithms such as support vector machine (SVM), k-nearest neighbor (KNN), and multi-layer perceptron (MLP). Their dataset was created from YouTube videos and labeled using four perceived emotions: neutral, happiness, anger, and sadness. Various spectral features such as the Mel-frequency cepstral coefficient (MFCC) and Mel spectrogram, and Spectral contrast. were extracted, and then the classification methods were applied. They got the highest accuracy rate 77% by using (SVM).

When a person becomes emotional, his voice is modified depending on the state of emotion in emotion detection. As acoustic characteristics such as pressure, intensity, and loudness change from one emotional state to the next. The most difficult challenge is to extract the best features that describe the emotional statistics. Mohammad et al [1] concentrated on the feature extraction phase in this paper, employing a mix of Linear Predictive Codes (LPC) and 10-degree polynomial Curve fitting Coefficients over periodogram power spectral density function. Finally, they used different Machine learning algorithms for classification and then they evaluate the accuracy results of various machine learning algorithms (Decision Tree, ANN, SVM Logistic Regression, KNN) to find the best accuracy.

Furthermore, a group of researchers studied speech emotion recognition using REGIM TES an emotional speech database, that was constructed and evaluated to provide all practical extraction experiences. Pitch of voice, Energy, MFCCs, Formant, LPC, and spectrogram are the descriptors used in the study. The impact of the Arabic language on physiological events and the influence of culture on emotional behavior was demonstrated by descriptors [11].

The proposed work aims to use signal processing and pattern recognition techniques to choose the most efficient features for Arabic speech to construct an autonomous emotion recognition system. We focus to improve Arabic language research, so we use an Arabic dataset, especially from the Saudi dialect corpus. In the corpus data, the system will be trained and evaluated to extract the features and classify the emotions with different machine learning and deep learning algorithms.

3 Dataset

In this paper, we used A semi-natural emotion speech dataset in Saudi dialect that was constructed from YouTube recordings from the prominent Saudi YouTube channel Telfaz11 [10] [12].

Telfaz11 is an entertainment station that broadcasts a variety of programming, including comedy shows, clips, and short films. Their work is mostly concerned with presenting Saudi culture and regionally relevant information.

The dataset contained four main emotions: anger, happiness, sadness, and neutral. These emotions were used to categorize the audio clips. Male and female performers were represented in the final collection's 175 records, which had 113 male actors and 62 female actors. The duration of the dataset was roughly 11 minutes.

3. Methodology

This study applied different algorithms for Arabic speech emotion recognition. At first, MFCC and ZCR features were used to extract features from speech signals and then machine learning and deep learning algorithms were applied to recognize emotions.

3.1 Feature Etraction

Feature extraction is used in preliminary SER. It primarily uses voice signals to derive the number of features. The collected features are then provided as a set of inputs to the classifier, which is used to detect various moods.

In this work, two types of spectral features were extracted from the dataset.

1) MEL frequency cepstral coefficient (MFCC)

MFCC is one of the most extensively used spectral characteristics. It has several advantages, including ease of calculation, improved differentiation ability, and great noise resilience. Because of its high-frequency resolution and good sidelobe suppression properties, the hamming window is generally used [13].

We applied Mel Frequency Cepstral Coefficient (MFCC) by Librosa library, which is a Python library used for audio analysis and feature extraction techniques.

2) zero crossing rate (zcr)

The rate at which a signal transitions from positive to zero to negative or from negative to zero to positive is known as the zero-crossing rate (ZCR). Its importance has been extensively recognized in voice recognition and music information retrieval, and it is a vital element in classifying percussive sounds [14]. We applied Zero-Crossing Rate (ZCR) by the Librosa library.

3.2 CLASSIFICATION

All Algorithms were implemented on a PC having MS-Windows OS and an Intel Core i7-10510U processor at 1.80GHz. We used the Python programming language on Jupyter Notebook [15] which is a non-profit open-source software that has grown to support collaborative data science. In addition, we used different libraries such as Scikit-learn library [16] version 0.24.1 to create the emotion classification models and Keras [17] is an open-source Python library for developing and evaluating deep learning emotion classification models.

For audio analysis, we used the Librosa library[18]. Librosa aids in the visualization of audio signals as well as feature extractions using various signal processing techniques.

Moreover, we used Matplotlib [19] for visualizations which is a Python package that allows you to create static, animated, and interactive visualizations.

For classification, the data was split 90 percent for the training set and 10 percent for the testing set.

1) classification by using machine learning algorithms

We applied speech emotions classification by using SVM and KNN algorithms.

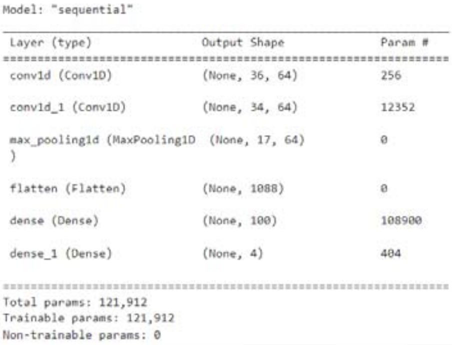
SVM is a binary classification method that identifies the optimum linear decision surface by employing the concept of structural risk reduction. The decision surface is a weighted combination of the properties of the training set. The MFCC and ZCR features were extracted separately, then the classification of emotions was applied to each feature independently

KNN algorithm assumes that the new case/data and existing cases are similar and places the new case in the category that is most like the existing categories. It maintains all of the available data and classifies a new data point based on its similarity to the existing data. This means that new data can be quickly sorted into a suitable category using the KNN method. We used 5 neighbors for kneighbors queries. MFCC and ZCR features were extracted separately, then the classification of emotions was applied to each feature independently.

2) classification by using deep learning algorithms

We applied speech emotions classification by using two deep learning models CNN and LSTM. figure 1 shows the CNN model architecture.

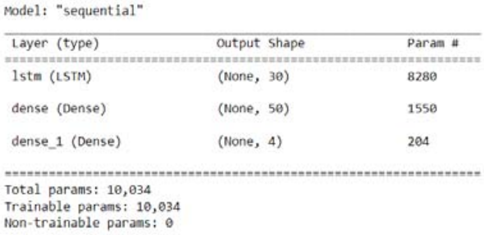
Figure 1: CNN model architecture.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 36, 64)            256

 conv1d_1 (Conv1D)           (None, 34, 64)            12352

 max_pooling1d (MaxPooling1D  (None, 17, 64)           0
 )

 flatten (Flatten)           (None, 1088)              0

 dense (Dense)               (None, 100)               108900

 dense_1 (Dense)             (None, 4)                 404

=================================================================
Total params: 121,912
Trainable params: 121,912
Non-trainable params: 0
```

We complied our model with the Adam optimizer, which is a deep neural network training algorithm that uses an adaptive learning rate optimization mechanism. To train our CNN model we used 30 epochs and 32 batch sizes. For another deep learning algorithm, we used LSTM. Figure 2 presents LSTM model architects.

figure2: LSTM model architecture



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 30)                8280

 dense (Dense)               (None, 50)                1550

 dense_1 (Dense)             (None, 4)                 204

=================================================================
Total params: 10,034
Trainable params: 10,034
Non-trainable params: 0
```

We trained our LSTM model with 30 epochs and 32 batch sizes.

Finally, MFCC and ZCR features were extracted separately, then the algorithms for the classification of emotions were applied to each feature independently.

4 Result

In this study, two feature extraction methods and four classification algorithms for Arabic speech emotion recognition were applied. Below we review the most important results of Arabic Speech Emotion Recognition from the Saudi Dialect Corpus.

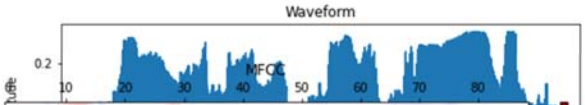
4.1 EXPERIMENTAL RESULTS

1) Feature Extraction

The introduced dataset was used to extract spectral characteristics, and then the frequency-based features were generated by converting the time-based signals.

Figure 3 shows the waveform extracted from an audio sample of happy emotions. In Figures 4 and 5 we can see the MFCC and ZCR features extracted from the audio sample of happy emotions.

Figure 3: Sound sample of happy emotion as a waveform



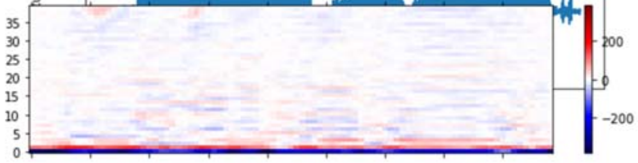


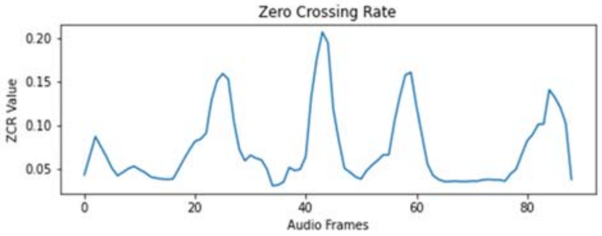
Figure 4: MFCC feature extraction from a happy emotion sound sample.

Figure 4: ZCR feature extraction from a happy emotion sound sample.

2) *CLASSIFICATION*

The goal of this study was to examine different machine learning and deep learning algorithms for predicting emotions.

We performed the different experiments using MFCC and ZCR features extracted separately, then the classification of emotions was applied to each feature independently.



For machine learning experiments we used two algorithms SVM and KNN. Figure 19 shows the result of the confusion matrix for the MFCC feature and SVM algorithm. The accuracy of the SVM algorithm with the MFCC feature is 87%.

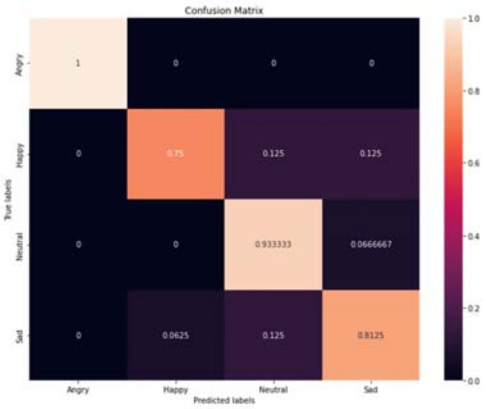


Figure 6: Confusion matrix for SVM algorithm with MFCC feature.

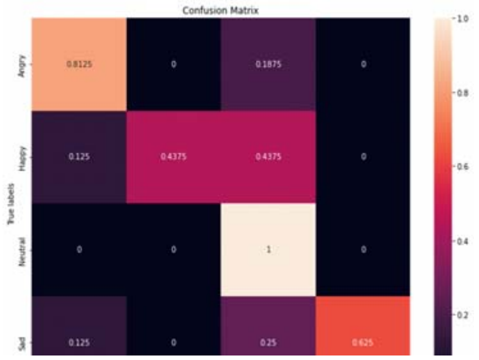


Figure 7: Confusion matrix for SVM algorithm with ZCR feature.

Figure 7 presents the result of the confusion matrix for the ZCR feature and SVM algorithm. The result of training SVM with the ZCR feature is 71%.

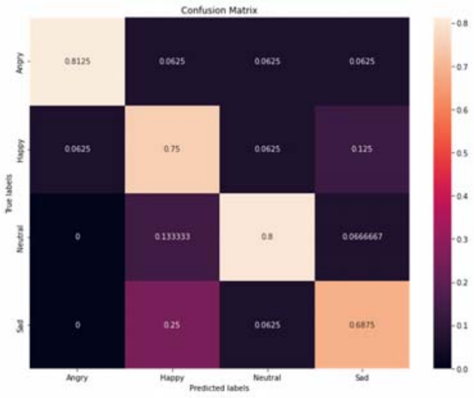


Figure 8: Confusion matrix for KNN algorithm with MFCC feature.

The second machine learning model we applied for classification is KNN. Figure 8 presents the result of the confusion matrix for the MFCC feature and KNN algorithm. The result of training KNN with the MFCC feature is 76%.

Figure 9 shows the confusion matrix for KNN with the ZCR feature. The accuracy result of training KNN with the ZCR feature is 49%.

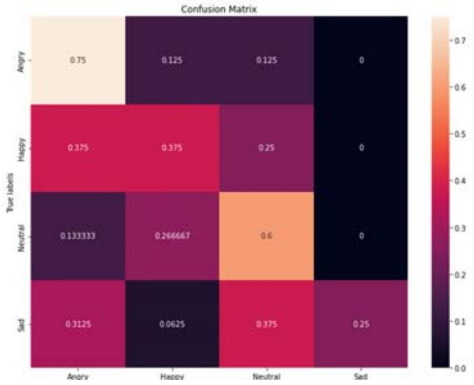


Figure 9: Confusion matrix for KNN algorithm with ZCR feature.

For deep learning experiments we used two algorithms CNN and LSTM. Figure 10 shows the result of the confusion matrix for the MFCC feature and CNN algorithm. The accuracy of the CNN algorithm with the MFCC feature is 95%.

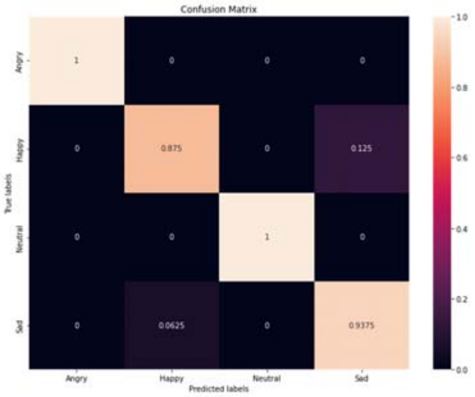


Figure 10: Confusion matrix for CNN algorithm with MFCC feature.

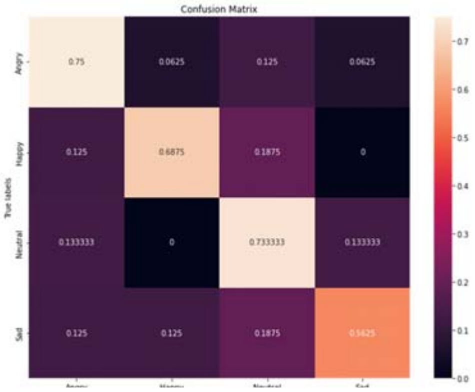


Figure 11: Confusion matrix for CNN algorithm with ZCR feature.

Figure 11 shows the confusion matrix for CNN model with the ZCR feature. The accuracy result of training CNN with the ZCR feature is 68%.

The last model we applied is LSTM with two features MFCC and ZCR. In Figure 12 we see the confusion matrix for LSTM model with MFCC feature the accuracy result for this model is 90%.

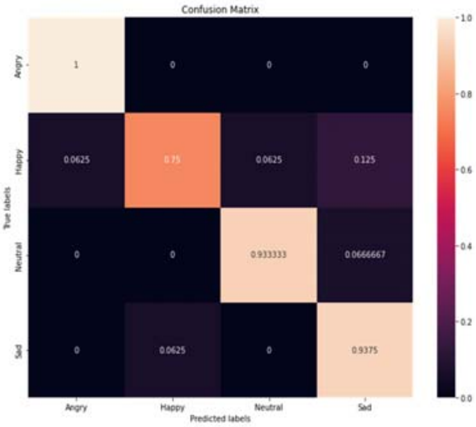


Figure 12: Confusion matrix for LSTM algorithm with MFCC feature.

Figure 13 presents the confusion matrix for LSTM model with ZCR feature. The accuracy result of the training this model is 34%.

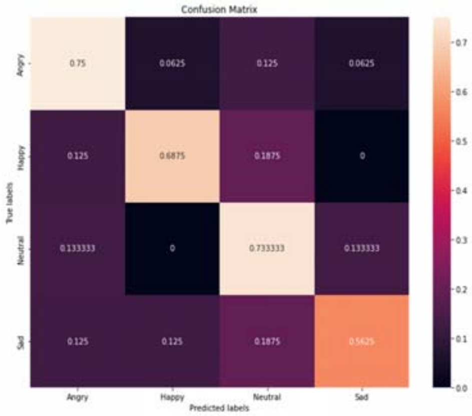


Figure 13: Confusion matrix for LSTM algorithm with ZCR feature.

Overall, all classifiers accurately predicted anger, with the greatest prediction rate for this emotion. MFCC-SVM, MFCC-CNN, MFCC- LSTM all these models returned the highest prediction rate for anger emotion compared with all other classifiers with 100%. With prediction rates of 93%, MFCC-CNN and MFCC-LSTM outperformed other classifiers in their ability to predict negative emotion (sadness). ZCR-SVM and MFCC- CNN models accurately predicted the (neutral) emotions with a prediction rate of 100%.

All classifiers performed badly when predicting happiness compared to the other emotions. The MFCC feature extraction approach with the CNN model beat the other classifiers in emotions prediction.

In general, based on all the previous results, we concluded that when we used Mel Frequency Cepstral Coefficient (MFCC) feature, we were able to obtain a high predication rate for all emotions, as it reached 95% when we used MFCC feature with deep learning model (CNN). But on the other hand, when we used Zero Crossing Rate feature extraction method the results of the prediction were much lower as it reached 34% in ZCR- LSTM model.

4.2 *Comparison with Previous Works*

By comparing our results to those of the previous study, we discovered that utilizing MFCC feature extraction with the CNN model, we were able to get a high accuracy rate of 95%, whereas using the machine learning model, the researchers achieved the maximum accuracy rate of 77% by using the combination of features extraction MFCC + Mel spectrogram+ spectral contrast with SVM.

Table I shows the result comparison for Arabic speech emotion recognition of the current and the previous study.

Table 1: Comparison results (%) between the prior study and the proposed study.

Paper	Features extraction methods	Classification Algorithms	Result (%)
	MFCC	KNN	60.00%
		MLP	57.14%
		SVM	54.29%
	MFCC+ Mel spectrogram	KNN	54.29%
		MLP	65.71%
		SVM	65.71%
	MFCC+ spectral contrast	KNN	68.57%
		MLP	60.00%

Aljuhani et al. [10]		SVM	54.29%
	Mel spectrogram	KNN	42.86%
		MLP	51.43%
		SVM	40.00%
	MFCC + Mel spectrogram + spectral contrast	KNN	57.14%
		MLP	71.43%
		SVM	77.14%
Current Study	MFCC	KNN	76.00%
		SVM	87.00%
		CNN	95.00%
		LSTM	90.00%
	ZCR	KNN	49.00%
		SVM	71.00%
		CNN	68.00%
		LSTM	34.00%

5. Conclusion and Future Work

There are extremely few studies that focus on the Arabic language when it comes to recognizing emotions through speech.

In this study, we worked on comparing the capacity of various deep learning and machine learning classifiers to forecast emotions in a speech in the Saudi dialect.

The four emotions that were examined were anger, happiness, sadness, and neutral. For classification, spectral features were used such as MFCC and ZCR features extraction with different machine learning and deep learning algorithms. Deep learning algorithms had the best accuracy in our Arabic speech emotion recognition. MFCC with the CNN model had the best accuracy of 95% and the result of MFCC with the LSTM model is 90%. All the classifiers agreed that anger was the best predictor of emotion.

In future work, we will support this data with other Arabic data from various Arabic dialects, as we will work on developing Arabic language research in this field. To achieve better outcomes, we will also make use of numerous different classification algorithms.

References

[1] O. A. Mohammad and M. Elhadef, “Arabic speech emotion recognition method based on LPC and PPSD,” in 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), 2021.

[2] B. H. Su and C. C. Lee, “Unsupervised Cross-Corpus Speech Emotion Recognition Using a Multi-Source Cycle-GAN,” IEEE Transactions on Affective Computing, 2022.

[3] 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.

[4] X. Huang, J. Baker, and R. Reddy, “A historical perspective of speech recognition, ” Commun,” Commun. ACM, vol. 57, no. 1, pp. 94–103, 2014.

[5] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, “A review on emotion recognition using speech,” in 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017.

[6] B. Zhang, C. Quan, and F. Ren, “Study on CNN in the recognition of emotion in audio and images,” in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016.

[7] Y. Hifny and A. Ali, “Efficient Arabic emotion recognition using deep neural networks,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

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

[9] M. Meddeb, H. Karray, and A. M. Alimi, “Building and analysing emotion corpus of the Arabic speech,” in 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), 2017.

[10] Telfaz11. Youtube Channel. [Online]. Available:https://www.youtube.com/user/telfaz11

[11] A.Milton, S. Sharmy Roy, and S. Tamil Selvi, “SVM scheme for speech emotion recognition using MFCC feature,” Int. J. Comput. Appl., vol. 69, no. 9, pp. 34–39, 2013.

[12] A. Torres Garcia, C. A. Reyes Garcia, L. Villasenor-Pineda, and O. Mendoza-Montoya, Eds., Biosignal processing and classification using computational learning and intelligence: Principles, algorithms, and applications. San Diego, CA: Academic Press, 2021.

[13] “Project jupyter,” Jupyter.org. [Online]. Available: https://jupyter.org/. [Accessed: 25-Jul-2022].

[14] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, “Scikit-learn: Machine learning without learning the machinery, ” GetMob,” GetMob. Mob. Comput. Commun, vol. 19, no. 1, pp. 29–33, 2015.

[15] Keras.io. [Online]. Available: https://keras.io/. [Accessed: 25-Jul-2022].

[16] Mcfee, ““librosa: Audio and Music Signal Analysis in Python,” in Proceedings of the 14th Python in Science Conference, 2015.

[17] “Matplotlib — visualization with python,” Matplotlib.org. [Online]. Available: https://matplotlib.org/. [Accessed: 25-Jul-2022]