IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 October 2024, accepted 4 November 2024, date of publication 11 November 2024,
date of current version 19 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3494877

RESEARCH ARTICLE

Designing a System to Recognize Main Arabic Dialects

DHEYAA ALHELAL[1,2] **AND TIMUR İNAN**[2,3]
[1]Computer Techniques Engineering, Northern Technical University, Mosul 41002, Iraq
[2]Electrical and Computer Engineering, Timur İnan-Computer Engineering, Altınbaş University, 34217 Istanbul, Türkiye
[3]Computer Engineering, Marmara University, 34722 Istanbul, Türkiye

Corresponding author: Dheyaa Alhelal (dalhelal@ntu.edu.iq)

ABSTRACT Identifying dialects can be considered as one of the recently attracted researchers. This study concentrates on recognizing between famous Arabic dialects based on speeches or talks. Three types of Arabic talks are considered here namely Arabian Peninsula Dialect (APD), Egyptian Dialect (ED), and Levantine Dialect (LD). We propose three artificial neural network models to identify between the three types of talking. These models are based on the Multi-Layer Perceptron (MLP) network, Convolutional Neural Network (CNN), and Deep Recurrent Neural Network (DRNN). Furthermore, comparisons between the three proposed models are provided. So, a comprehensive study is presented in this paper. Spoken Arabic Regional Archive (SARA) dataset is employed. It is prepared and divided into three groups. These are the Original SARA (OSARA), Filtered SARA (FSARA), and Mixed SARA (MSARA), which consists of both the OSARA and FSARA. According to the results, it has been found that the best accuracy of 90.70% is for the proposed DRNN model with the MSARA group of the employed dataset.

INDEX TERMS Arabic dialects, convolutional neural network, deep recurrent neural network, multi-layer perceptron.

I. INTRODUCTION

Automatic identification of dialects from people's speech has recently attained significant attention [1], [2], [3], [4]. Such identification refers to the process that reveals the speech type within a specific language. This subject is challenging because of the similarity between dialects especially for those which belong to the same original or native language [5]. Moreover, identifying dialects has a big advantage in predicting speakers' nationalities.

Like other languages, Arabic consists of multiple parts such as the written and speaking parts. The written part usually utilizes the Arabic formal language, which has grammar and unique pronunciation challenges [6]. Arabic people commonly use formal language in official government communications, media, and education. On the other hand, the speaking part usually utilizes an Arabic vernacular dialect, which does not follow the original language rules, and this makes it more challenging. Dialects differ according to the geographical locations as they have been acquired and influenced by residences. They are also affected by external factors such as colonialism, politics, and others.

Scientists classified Arabic dialects into four or five types according to the geographical locations [7]:

- Arabian Peninsula dialect: It is spoken by the Gulf states like Saudi Arabia, Kuwait, Emirate, Qatar, Bahrain, and Oman.
- Iraqi dialect: It is spoken by Iraqi people, however, it is sometimes combined with the Gulf.
- Levantine dialect: It is the dialect used by the people of the Levant, such as Syria, Jordan, Palestine, and Lebanon.
- Egyptian dialect: It is spoken by the Egyptians and part of the Libyan regions bordering Egypt.
- Maghrebi dialect: It is mainly spoken by Maghreb countries such as Morocco, Algeria, and Tunisia.

In view of the difficulties of distinguishing between Arab dialects, it seems advisable to achieve a system that uses

The associate editor coordinating the review of this manuscript and approving it for publication was Ghulam Muhammad.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

artificial intelligence to address such issues. Like too many works in various fields consider employing artificial intelligence models as [8], [9], [10], and [11]. Therefore, this paper aims to design a machine learning system to identify or distinguish between the three main Arabic dialects Arabian Peninsula Dialect (APD), Egyptian Dialect (ED), and Levantine Dialect (LD). Multiple contributions are provided in this study, they can be highlighted as follows:

- Collecting and preparing the related dataset of the Spoken Arabic Regional Archive (SARA) [12].
- Proposing three artificial neural network models based on the Multi-Layer Perceptron (MLP) network, Convolutional Neural Network (CNN), and Deep Recurrent Neural Network (DRNN) to distinguish between the three main Arabic dialects.
- Providing comprehensive study here including establishing comparisons between the three proposed models.

After this introduction, the remaining paper's sections will be ordered as follows: Section II reviews related work, Section III describes the theoretical part of this paper, Section IV demonstrates and discusses the results, and Section V provides the conclusion.

II. LITERATURE REVIEW

A comprehensive and up-to-date literature review of related work in the field of Arabic dialect recognition is offered as follows:

In 2015, Ali et al. applied more than one approach to dialect identification of Arabic speech. These approaches depend on phonetic and lexical features that were extracted from bottleneck features. The proposed speech recognition system depends on the i-vector framework. After that, two classifiers (generative and discriminative) were studied. A multi-class support vector machine was used to combine the two classifiers. When the features were evaluated by using a binary classifier to distinguish between Modern Standard Arabic MSA) and dialect Arabic, the accuracy was 100%. The proposed methods were utilized to discriminate among the Arabic dialects of (Egyptian, Arabian Peninsula, Levantine, North African, and MSA), the resulting accuracy was recorded as 59.2% [13].

In 2017, Shon designed an Arabic dialect identification system to differentiate among the significant Arabic dialects. Domain mismatches and dialect variability between the training and testing domains were focused. To build an effective identification system, the Siamese neural network was employed to understand similarities and dissimilarities among Arabic dialects. i-vector post-processing was used to adjust domain mismatches. At the same time, both acoustic and linguistic features were exploited. The system achieved 75% accuracy for 10-hours testing set [14].

In 2018, Shon et al. proposed an end-to-end dialect identification system with a Siamese neural network to extract language embedding. Acoustic features and linguistic features were utilized. Three types of acoustic features were used to train the end-to-end system. These were the log Mel-scale filter bank energies, spectrogram energies, and Mel-Frequency Cepstral Coefficients (MFCCs). The linguistic feature research was concerned with learning similarities and dissimilarities among dialects by using the Siamese neural network. Therefore, the performance of the system is improved as well as feature dimensionality is reduced. The results showed that 73% accuracy was attained when the system used a single feature set, while 78% accuracy was obtained for multiple feature sets. The end-to-end system achieved a good result compared to the Siamese network with language embeddings [15].

In the same year, Najafian et al. investigated some dialect identification methods for the Arabic language, like n-gram and i-vector phonotactic features with the support vector machine classifier. The paper studies the benefit of using Convolutional Neural Networks (CNN) for direct mapping of phonotactic and acoustic features to one of the Arabic dialects. Also, a new phonotactic dialect identification method with a novel feature representation methodology was proposed. It acquired probability statistics, phone duration, and phone sequences. The proposed method achieved 73.27% accuracy after using a combination of multi-lingual phonotactics of five languages (English, Arabic, Russian, Hungarian, and Czech). A confusion matrix was also used to study dialect identification errors. The proposed phonotactic method achieved 24.7% and 19% relative error rate reduction as the final system fusion [16].

In 2020, Hanani and Naser adapted the X-vector framework for Arabic dialect identification. At the same time, the traditional i-vectors, phonetic features, bottleneck features, Gaussian mixture model (GMM)-tokens, and word transcriptions were also used. The X-vector reported a performance of 16.60% improvements. When it is fused with the i-vectors, bottleneck features, and word unigram the performance further improved to 69.70% [17].

In 2022, Alsayadi et al. proposed a hybrid acoustic model combining CNN and Long Short-Term Memory (LSTM). Attention-based encoder-decoder techniques were utilized to build the model. Two datasets of the Standard Arabic Single Speaker Corpus (SASSC) and Multi-Genre Broadcast-3rd edition (MGB-3) were used. The proposed model achieved a 57.02% word error rate [18].

In 2023, Nasr et al. suggested a new transcribed corpus for Jordanian, Yamani, and multi-dialectal Arabic. Several baseline deep neural models were designed as, DeepSpeech2 and Bidirectional Long Short-Term Memory (Bi-LSTM) with attention model. Yamani speech corpus achieved a word error rate of 59% by utilizing the Bi-LSTM with attention model. The Jordanian speech corpus achieved a word error rate of 83% by employing the same model. The multi-dialectal Yamani, Jordanian, and Arab speech corpus also achieved a word error rate of 53% by exploiting the same model. The performances of the DeepSpeech2 model were the best, where it achieved 31%, 68%, and 30% for Yamani, Jordanian, and multi-dialectal Arabic corpus, respectively [19].

To critically analyze how the current study contributes to existing knowledge it can be noticed from the literature that the subject of identifying or distinguishing between dialects is challenging, where low performances were reported. Also, the Dialects of APD, ED, and LD have not been comprehensively concentrated before. Thus, this study is presented to further enhance the results of identifying such dialects.

III. METHODOLOGY

As mentioned, the aim here is to achieve a machine-learning system that can distinguish or identify between the three main Arabic dialects of APD, ED, and LD. To accomplish this a schematic procedure is suggested. It consists of multiple stages starting from collecting the Mixed SARA (MSARA) dataset, which includes both the Original SARA (OSARA) and Filtered SARA (FSARA) sets. Then, pre-processing operations only for the OSARA are provided. Consequently, the MFCC feature extraction is applied. Subsequently, all data are partitioned into training and testing groups. For the training group, the three proposed machine learning models of MLP, CNN, and DRNN are trained, and their weights are acquired. For the testing group, the three models are tested after acquiring the resulting training weights. All testing outputs of the three models are recorded. Then, evaluations and comparisons are applied to all of them. The schematic overview for achieving the suggested system of identifying between the dialects of APD, ED, and LD is demonstrated in Figure 1.

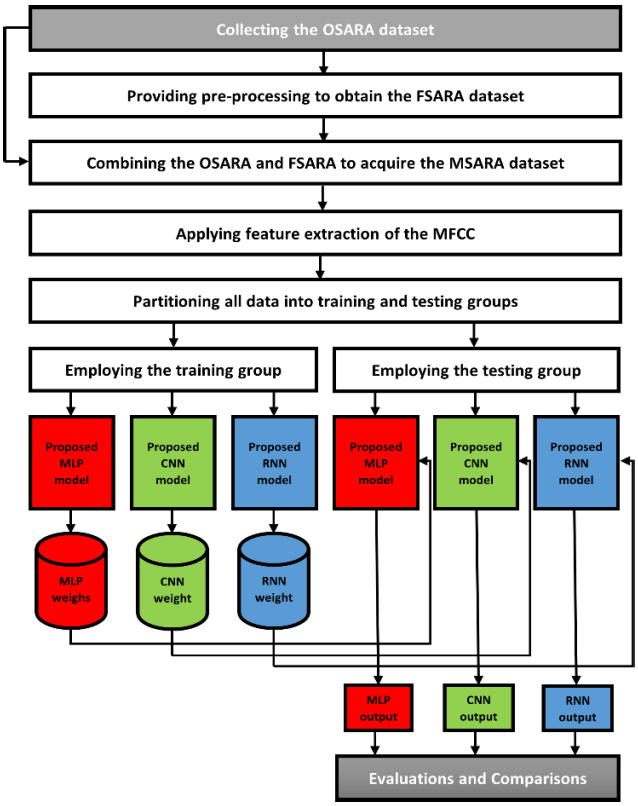


FIGURE 1. Schematic overview for achieving the suggested system of identifying the dialects of APD, ED, and LD.

A. COLLECTING THE OSARA DATASET

One of the most important reasons for choosing the SARA dataset is that it simulates real-world scenarios in terms of varying acoustic conditions, speaker accents, and the presence of background noise. The authors in [12] proposed the SARA dataset that contains spontaneous and undefined phrases (spontaneous conversations related to everyday life). This dataset has been acquired from series, films, and acting scenes in Arabic dialects (three different Arabic dialects). It is defined as the Spoken Arabic Regional Archive (SARA). It is not affected by live transmission Quality of Service (QoS) troubles such as unreliable packet delivery and delay. This is because the media for SARA is downloaded from YouTube and stored before the sampling process. The SARA dataset is useful for different machine learning fields such as automatic speech recognition, accent recognition, and speaker verification. To avoid incorrect pronunciations from children, which affects the detection process, the SARA dataset contains only adult speakers (both males and females) for three Arabic regional dialects: APD, ED, and LD. The dataset is sufficiently diverse and representative of the various Arabic dialects to ensure the models' generalizability, each group of dialects contains several different accents. In total, the SARA involves 12 different accents that represent 8 Arab countries as shown in table 3. Each statement (utterance) contains one speaker, while the number of the words and the speakers in the same dialect are unknown. To illustrate, SARA dataset specifications are explained in the following tables [12]:

TABLE 1. Number of participants in the SARA dataset.

Dialect	No. of Males	No. of Females	Total No. of Participants
ED	877	488	1365
APD	657	573	1230
LD	583	662	1245

TABLE 2. Recording times and number of samples in the SARA dataset.

Recording Time	No. of Samples in the ED	No. of Samples in the APD	No. of Samples in the LD	Total No. of Samples
3 sec	468	388	398	1254
4 sec	377	409	354	1140
5 sec	271	269	246	786
6 sec	166	117	158	441
7 sec	83	47	89	219
Total	1365	1230	1245	3840

These tables reveal the comprehensiveness of the SARA dataset. So, it is worth to be exploited in this paper. However, it has been extensively considered and prepared. It has been termed here the Original SARA (OSARA) dataset. Our models are trained by using this dataset so the performance in real-world scenarios could be estimated as highly acceptable.

TABLE 3. Considered accents with the number of speakers in the SARA dataset.

Dialect	Accent	No. of Male Speakers	No. of Female Speakers
APD	Saudi	143	207
	Hejazi	29	0
	Najdi	8	4
	Bahrani	335	282
	Gulf	64	46
	Omani	78	38
LD	Syrian	46	261
	Lebanese	76	341
	Jordanian	362	42
	Palestinian	99	18
ED	Egyptian	623	221
	Sa'idi	254	267

B. DATASET PREPARATION

In practical reality (actual reality), the audio signal is not pure. It is, in fact, a mixture between the desirable signal and undesirable (noise, background music, babble speech, ..., etc.) signals. Therefore, filtering such a mixture of signals in a processing stage has become necessary. Generally, Deep Neural Network (DNN) was used to purify and enhance the audio signal [20], [21]. It is worth noting that deep filtering was first proposed for enhancing and cleaning audio signals. One of the advantages of the deep filter is its ability to recover signal degradation, such as time frame zeroing or notch-filters, as the filter is applied to multiple Time-Frequency (TF) bins [22], [23]. To produce a speech enhancement framework with convincing performance the DeepFilterNet was approached. It has two-stage speech enhancement frameworks. In the first stage, since both speech and noise have a smooth spectral envelope, so, this advantage was considered. In the second stage, the enhancement focused on only the lower frequencies, because in the periodic speech components, these frequencies have the most energy [24]. Then, DeepFilterNet2 has been suggested as an expanded work by improving the network structure, data augmentation, and training procedure [25]. Figure 2 shows a schematic overview of the DeepFilterNet2 speech enhancement process.

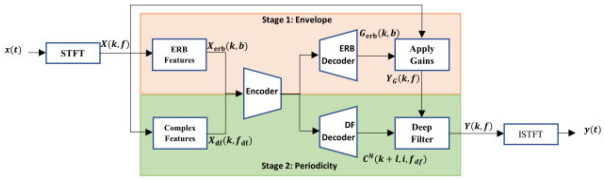


FIGURE 2. Schematic overview of the DeepFilterNet2 speech enhancement process [25].

From this figure, it can be observed that an original signal $x(t)$ inputs to the DeepFilterNet2. It is processed by the Short-Time Fourier Transform (STFT) where $X(k,f)$ is produced. Then, the produced $X(k,f)$ passes to the Equivalent Rectangular Bandwidth (ERB) Features in stage 1 (Envelope), where $X_{\text{erb}}(k,b)$ is generated. It is also passed to the Complex Features in stage 2 (Periodicity), where $X_{\text{df}}(k,f_{\text{df}})$ is generated. Both $X_{\text{erb}}(k,b)$ and $X_{\text{df}}(k,f_{\text{df}})$ go through Encoder. The outcome moves to the ERB Decoder in stage 1, where $G_{\text{erb}}(k,b)$ is produced, and DF Decoder in stage 2, where $C^N(k+l,i,f_{\text{df}})$ is generated. Both $X(k,f)$ and $G_{\text{erb}}(k,b)$ pass to Apply Gains in stage 1, where $Y_{\text{G}}(k,f)$ is provided. Then, $Y_{\text{G}}(k,f)$ and $C^N(k+l,i,f_{\text{df}})$ go through Deep Filter in stage 2, where $Y(k,f)$ is produced. Finally, $Y(k,f)$ is processed by the Inverse Short-Time Fourier Transform (ISTFT) and generate $y(t)$.

DeepFilterNet2 is used in this study, where the Filtered SARA (FSARA) dataset is derived from the OSARA dataset. Nevertheless, both two groups of the FSARA and OSARA datasets are concatenated and exploited here as one group called the Mixed SARA (MSARA) dataset. The relationship between the three groups of the SARA dataset is demonstrated as given in Figure 3.

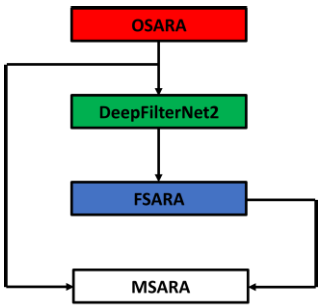


FIGURE 3. Preparing the MSARA dataset, which has been employed in this study.

C. FEATURE EXTRACTION

Although Dialect Identification (DID) or Language Identification (LID) began decades ago, it still holds a challenge to researchers. This is due to the similarities of phonotactics, acoustic phonetics, and prosodic cues among different dialects of languages. In addition, in many practical application cases, there is no control over the channel characteristics, available speech duration, and noise level [26]. To the best of the obtained knowledge, most of the research on detecting between Arabic dialects has not exceeded 78% accuracy. So, our goal here is to overcome the challenges and develop a system with high performance.

Usually in any system, extracting significant features from an input is important. There are many methods used for feature extraction such as Linear Prediction Coding (LPC), Perceptual Linear Prediction (PLP), Linear Predictive Cepstral Coefficients (LPCC), Neural Predictive Coding (NPC), Power Normalized Cepstral coefficient (PNCC) and Mel-Frequency Cepstral Coefficients (MFCC). The MFCC is considered the most common method utilized for audio recognition because it depends on known variations within critical frequency bandwidth for human ears [27]. Where researchers [28] demonstrated the superiority of MFCC when compared to PNCC. A block diagram of the total operations

that are employed for the MFCC feature extraction is given in Figure 4 [29]:

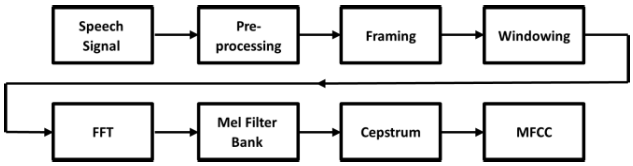


FIGURE 4. Block diagram of the total operations that are employed for the MFCC feature extraction.

The first operation is pre-processing. Naturally, the entered audio signal is analog, so it must be converted to digital. So, a set of samples $s(k)$ is obtained based on the sampling rate. After that, the $s(k)$ signal is passed through a first-order Finite Impulse Response (FIR) filter to increase its energy (magnitudes within the spectrum of the frequencies).

Then, the framing operation is applied since the analog signal is converted into N number of samples per frame. The adjacent frames are disjointed by M number of samples per other frame and ($M < N$), where the overlapping between the first frame and second frame is ($N - M$) number of samples. Each frame is individually analyzed and synthesized to keep the information from losing (preventing discontinuity in audio samples).

Consequently, the windowing operation is performed. This step aims to achieve continuity between a single frame on the one hand and all frames on the other hand. So that continuity is achieved as much as possible from the first point in the first frame to the last point in the last frame. Mathematically, this step is defined as $s[n] * w[n]$, where $s[n]$ is the signal itself and $w[n]$ can be represented as follows [30]:

$$w[n, \alpha] = (1 - \alpha) - \alpha cos(\frac{2\pi n}{N - 1}), \quad 0 \leq N \leq N - 1 \quad (1)$$

where: $w[n, \alpha]$ is the hamming window, n is the number of frames, α is a fixed value and N is the number of samples per frame.

Subsequently, the Fast Fourier Transform (FFT) operation is used to convert from the time domain to the frequency domain. The frequency spectrum for each frame (frequency domain) is significant as spectral analysis shows that different energy distributions at frequencies are proportional to different timbres in speech signals.

Then, Mel filter bank operation is considered. A human ear does not support linear frequency resolution because of its decreased sensitivity at high frequencies. In other words, human understanding of frequencies is non-linear. To overcome the limitation of understanding frequencies by a human's ear, the Mel scale is used. The Mel for any frequency is expressed as follows [31]:

$$Mel(f) = 2595 * log10\left(1 + \frac{f}{700}\right) \quad (2)$$

where: $Mel(f)$ is the Mel for any frequency, f is a given frequency in Hz.

Finally, Cepstrum operation is provided. Since the output of the previous step (Mel filter bank) is highly correlated, which is an issue for classification applications and machine learning algorithms to not deal with, it is necessary to de-correlate that output (features) and arrange them (in descending order of information) by using the Discrete Cosine Transform (DCT) [32]. DCT is like the Discrete Fourier Transform (DFT), except that it only utilizes real numbers, and its length is twice that in the DFT. The MFCC results, provide a valuable representation of the local spectral properties of a speech signal in analyzing a specific frame. This is because Mel spectrum coefficients and its logarithm are real numbers [33].

D. PROPOSED MACHINE LEARNING MODELS

As mentioned, three models are proposed in this paper. The first one is based on the MLP, the second one is based on the CNN, and the third one is based on the DRNN. They all are designed after extensive experiments regarding the number of layers, the number of neurons in each layer, and values of parameters. These are for providing appropriate adaptations to the task of distinguishing between the employed Arab dialects

1) PROPOSED MLP

The MLP is a type of machine learning that employs a classifier. So, it is here adapted to accept the extracted features of MFCC as inputs. On the other hand, its outputs are adapted for recognizing the three Arabic dialects of APD, ED, and LD. The output essentially consists of three nodes (or neurons), each node represents a class for an Arabic dialect. The main architecture of the proposed MLP model can be described as shown in Figure 5.

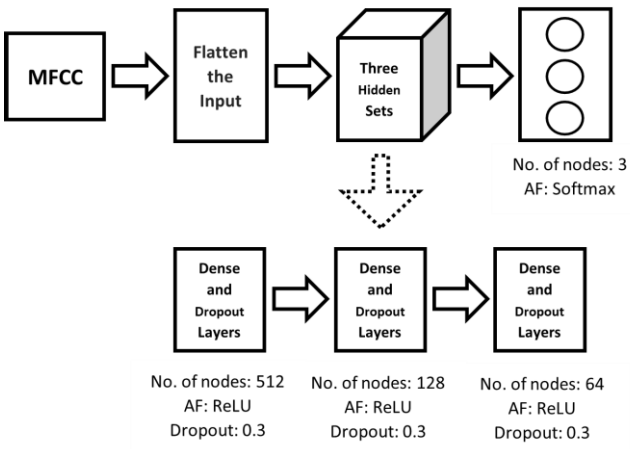


FIGURE 5. The main architecture of the proposed MLP model.

From this figure, each MFCC feature is prepared as a One-Dimensional (1D) vector. Consequently, this vector is applied as flatten in the input. Three hidden sets are experimentally chosen. They are three dense layers, and each one of them uses a Rectified Linear Unit (ReLU) Activation Function

(AF) which is fast and easy to implement. Also, each dense layer is followed by the dropout layer, which reduces the data size. The numbers of neurons in the dense layers are as follows: 512 nodes for the first dense layer, 256 nodes for the second dense layer, and 64 nodes for the third dense layer. Each dropout layer here reduces the data size by 0.3 (or 30%). In the output layer, three neurons are used, as mentioned, and SoftMax AF is applied. The SoftMax helps in choosing the node of the highest value to obtain a certain class of the Arabic dialect.

The problem of overfitting is addressed in this paper in four methods. First, keep the architecture of our models as simple as possible where the number of layers and neurons are as few as possible, and good network performances are maintained. Second, consider audio data augmentation, in other words, artificially increase the number of training samples by using the MSARA dataset which includes both the Original SARA (OSARA) and Filtered SARA (FSARA) sets. As can be seen in the (***E. TESTING RESULTS***) subsection, the overfitting issue is reduced when the MSARA dataset is used. Third, the dropout technique is used for all three models. Fourth, the L2 regularization technique is utilized in the MLP model.

2) PROPOSED CNN

Fundamentally, CNN is a type of deep learning that can be utilized for image processing. Therefore, it is here adapted to accept the MFCC as an input image. Likewise, the proposed MLP and its outputs are adapted for distinguishing between the three Arabic dialects. That is, in the output there are also three nodes, each node refers to a class for an Arabic dialect. The main architecture of the proposed CNN model can be demonstrated as given in Figure 6.

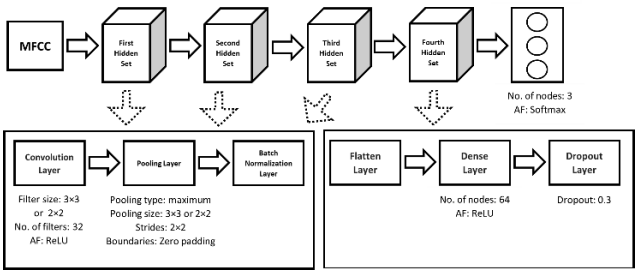


FIGURE 6. The main architecture of the proposed CNN model.

From this figure, it can be observed that each MFCC feature is prepared as a Two-Dimensional (2D) vector. This vector is applied as input. The hidden layers are also empirically chosen. They are three sets of convolution, pooling, and normalization layers which are sequentially arranged. The first convolution layer has 32 kernels (filters), each of size 3×3 pixels. The first pooling layer is of type maximum. It downsized 2D vectors by using the pooling size of 3×3 pixels, vertically and horizontally stride of 2×2 pixels, and boundaries of zero padding. The second and third convolution and pooling layers utilize the exact parameters that have been applied for the first convolution and pooling layers. However, the third convolution layer considers the filter size of 2×2 pixels and the third pooling layer exploits the pooling size of 2×2 pixels. Batch normalization is used for all normalization layers. In the last hidden set, it is necessary to flatten the output to a 1D vector and feed it into a dense layer with 64 neurons and (ReLU) AF followed by the dropout layer of 0.3 (or 30%), which reduces the data size to address the overfitting problem. In the output layer, the SoftMax AF is also applied for the three nodes. Again, it helps in determining the node of the highest value to attain a certain class of the Arabic dialect.

3) PROPOSED DRNN

Principally, the DRNN is a type of deep recurrent learning that can be exploited for pattern association. It is here also adapted to accept the MFCC as appropriate values. Similarly, the proposed MLP and CNN, outputs are adapted for classifying between the three Arabic dialects. So, there are also three nodes in the output, each node leads to a class for an Arabic dialect. The main architecture of the proposed DRNN model can be provided as demonstrated in Figure 7.

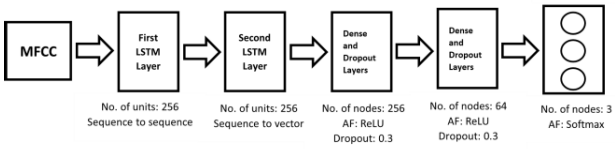


FIGURE 7. The main architecture of the proposed DRNN model.

From this figure, it can be investigated that each MFCC feature is prepared as a 2D vector. Subsequently, this vector is exploited for the input. The hidden layers are also experimentally determined. They consist of two Long Short-Term Memory (LSTM) layers followed by two dense layers. The first LSTM layer is a sequence-to-sequence, and it has 256 processing units. The second LSTM layer is a sequence-to-vector, and it has 256 processing units. The first dense layer includes the ReLU AF followed by a dropout of 0.3 (or 0.3%), to avoid overfitting, and it has 256 neurons. The second dense layer also includes ReLU AF followed by a dropout of 0.3%, again to avoid over-fitting, but it has 64 neurons. In the output layer, the SoftMax AF is also used for all three nodes. As mentioned, this can help in assigning the node of the highest value to reach a certain class of the Arabic dialect.

IV. RESULTS AND DISCUSSIONS

First, all the experiments are implemented in a Personal Computer (PC) with the following specifications: Asus Laptop, Intel core i7 processor with 2.20 GHz speed, 8 embedded Central Processing Units (CPUs), 8 GB main memory, NVIDIA external graphics card, Graphical Processing Unit (GPU) type GT 630M and 2 GB display memory. Furthermore, Python version 3.10.9 (64-bits) is used as a software tool and Anaconda version 3 is utilized as an editor.

A. EMPLOYED DATASET

As mentioned, the dataset of SARA [12] is exploited in this study. It involves 3840 samples for the three Arabic dialects of ARP, ED, and LD, whereas described in Table 1 they have 1230, 1365, and 1245 samples, respectively. It is also divided into five groups according to the recording times of 3, 4, 5, 6, and 7 seconds, where they include 1254, 1140, 786, 441, and 219 samples, respectively. Each group has samples of the three Arabic dialects as described in Table 2.

The SARA has been considered here as the OSARA, which is filtered as the FSARA. Then, both the OSARA and FSARA are concatenated and used as the MSARA.

B. PRE-PROCESSING RESULTS

Here, the DEEPFILTERNET2 is utilized to purify the OSARA dataset. This filter is applied to each sample individually, one by one. This can protect the data from losing some information. According to the nature of the dataset, which has been acquired from series, films, and acting scenes that consider the three Arabic dialects, it is classified into four groups according to the type of unwanted sounds in an original sample and the effect of the DEEPFILTERNET2 on that sample:

1) STANDARD SAMPLES

This represents most samples, as it contains slight noise without affecting the original speaker's voice. The DEEPFILTERNET2 has confirmed its effectiveness in this type by enhancing the purifications of original sounds. Figure 8 shows demonstrations of such achieved enhancement.

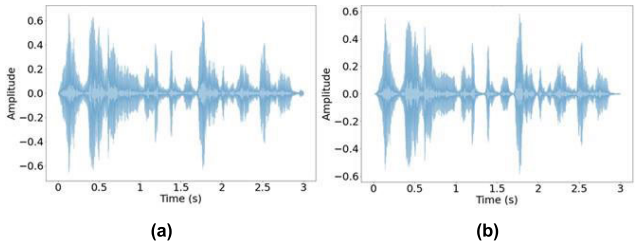


FIGURE 8. Demonstration of the DEEPFILTERNET2 effectiveness in purifying original samples: (a) an original sample and (b) its filtered sample.

2) SAMPLES WITH BACKGROUND SOUNDS

Depending on the nature of artistic environments, background sounds are combined within the standard samples. These expectedly affect the original signals. The DEEPFILTERNET2 has also confirmed its usefulness in removing background sounds such as music without influencing the speaker's voice, as given in Figure 9.

3) SAMPLES WITH EXISTING NOISE

As mentioned previously, the OSARA dataset was collected from series and movies. Therefore, it is natural for original samples to contain noise that may affect speakers' voices, some talks may need to be entirely understood by listeners. Once again, the DEEPFILTERNET2 also has confirmed its valuableness in dealing with this type as shown in Figure 10.

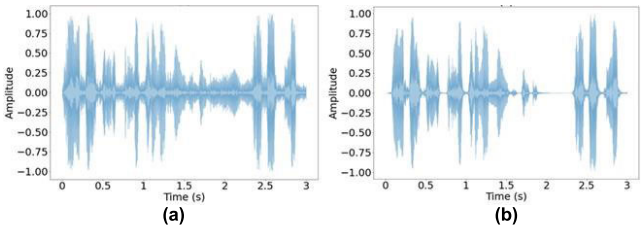


FIGURE 9. The usefulness of the DEEPFILTERNET2 in removing background sounds: (a) an original sample and (b) its filtered sample.

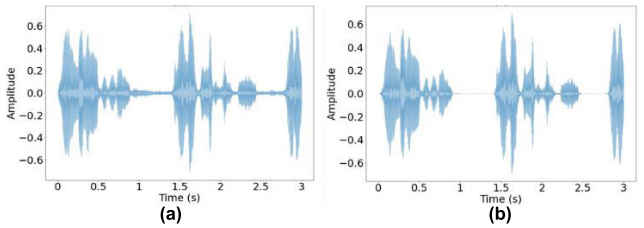


FIGURE 10. The usefulness of the DEEPFILTERNET2 in removing background sounds: (a) an original sample and (b) its filtered sample.

4) SAMPLES WITH MISSING INFORMATION

Like any other work, there is no ideality, especially in practical applications. There is a loss of information in some samples, even if of a small amount, as it is at most 1%. This has been dealt with by keeping the original samples and preserving the dataset as much as possible. Figure 11 illustrates this case.

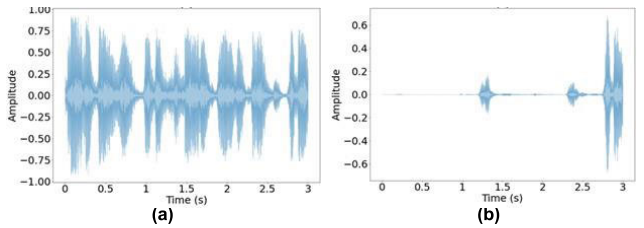


FIGURE 11. Illustration of the DEEPFILTERNET2 effectiveness in samples with missing information: (a) an original sample and (b) its filtered sample.

C. FEATURE EXTRACTION RESULTS

Practical experiments are considered for changing MFCC parameters to obtain the best feature extraction. The considered parameters here are as follows: Sample rate is 22050, Number of coefficients to extract is 40, Number of samples for FFT is 2048, Hop length (Sliding window for FFT) is 512, and number of segments which we want to divide sample tracks into is 5. Each MFCC and its label is saved in a file type JSON to be used later as input for a proposed model. Figure 12 shows examples of MFCC feature extractions.

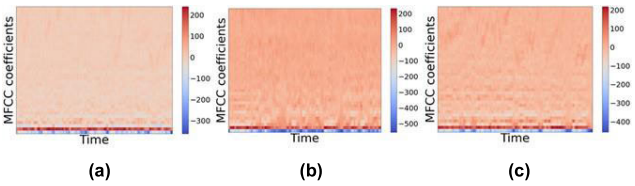


FIGURE 12. Examples of MFCC feature extractions: (a) MFCC for an APD sample (b) MFCC for an ED sample (c) MFCC for an LD sample.

From this figure, it can be investigated that the MFCC can provide extracted features that are useful in distinguishing between the three dialects of the APD, ED, and LD.

D. TRAINING RESULTS

As mentioned, three machine learning models are proposed in this paper (MLP, CNN, and DRNN models). All experiments here use the following training parameters: optimizer of type Adam, the fixed learning rate of 0.0001, Loss function of type sparse categorical cross-entropy, and mini-batch size of 32. Any dataset group is randomly partitioned into 70% for the training phase and 30% for the testing phase in the MLP model, while in the CNN and DRNN models, any dataset group (OSARA, FSARA, and MSARA) is randomly partitioned into 60% for the training phase, 15% for validation phase and 25% for testing phase. So, the partitioned testing parts are approximately equivalent.

Extensive training experiments are taken into account for each proposed model and all three groups of the employed dataset. In addition to investigating the five lengths of samples: 3, 4, 5, 6, and 7 seconds.

1) MLP TRAINING RESULTS

Figure 13 provides the training curves of the proposed MLP model for the OSARA dataset group and all five lengths of samples (3, 4, 5, 6, and 7 seconds). Similarly, figure 14 shows the training curves of the MLP model for the FSARA dataset group and all five lengths. Likewise, figure 15 displays the training curves of the proposed MLP for the MSARA dataset group and all five lengths too.

From these figures, all trains of the proposed MLP model have successfully been accomplished. That is, in all cases the training accuracy curves are acceptably increased until attaining very high values. The testing accuracy performances are also considered during the training phase at each epoch as they are demonstrated in the same figures. Their curves are increased along training epochs relative to the training curves. However, their performances vary between different cases. The best accuracies can be observed for the MSARA dataset group.

2) CNN TRAINING RESULTS

Figure 16 provides the training curves of the proposed CNN model for the OSARA dataset group and all five lengths of samples (3, 4, 5, 6, and 7 seconds). Similarly, figure 17 shows the training curves of the CNN model for the FSARA dataset group and all five lengths. Likewise, figure 18 displays the training curves of the proposed CNN for the MSARA dataset group and all five lengths too.

From these figures, it can be discovered that all trains of the proposed CNN model have positively been reported

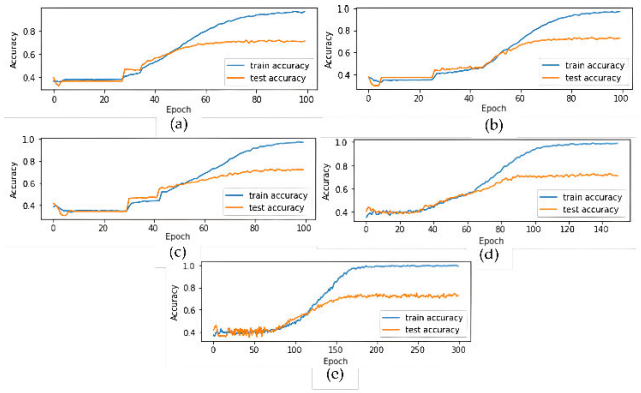


FIGURE 13. MLP training curves for the OSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

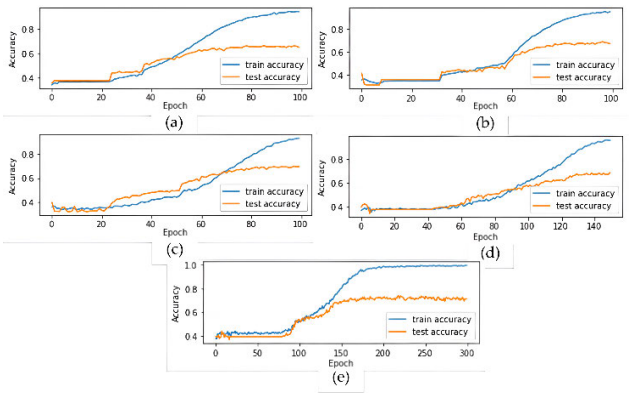


FIGURE 14. MLP training curves for the FSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

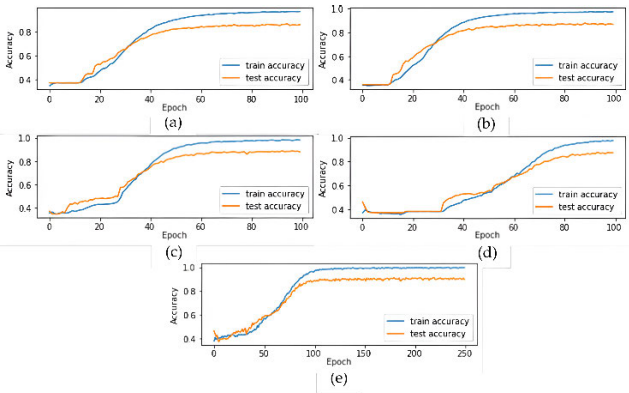


FIGURE 15. MLP training curves for the MSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

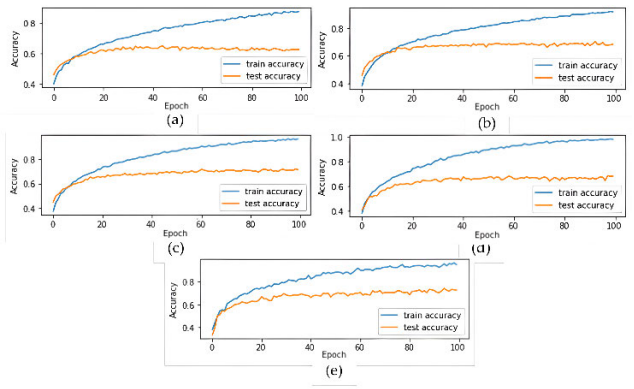


FIGURE 16. CNN training curves for the OSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

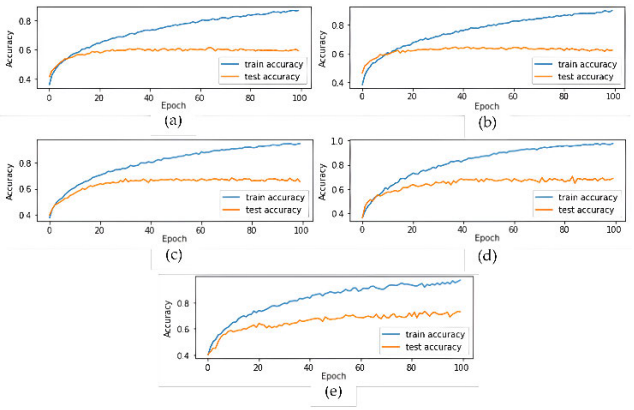


FIGURE 17. CNN training curves for the FSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

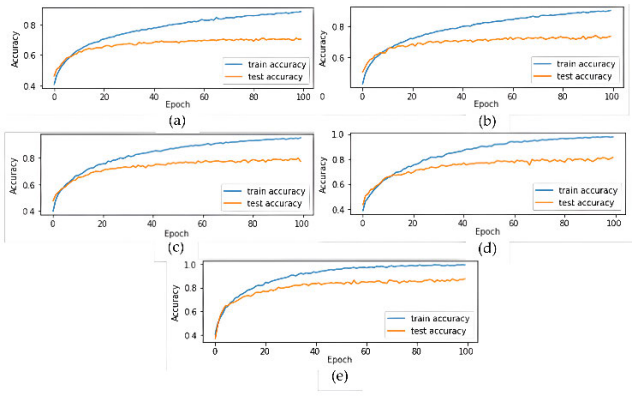


FIGURE 18. CNN training curves for the MSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

too. That is, in all cases the training accuracy curves are reasonably enhanced until reaching very high values. The testing accuracies are also carried out during the training phase at each epoch as they are provided in the same figures. Their curves are raised along training epochs consistently to the training curves. Nevertheless, their performances differ in various cases. The MSARA dataset group again confirms its ability to reach high accuracies.

3) DRNN TRAINING RESULTS

Figure 19 provides the training curves of the proposed DRNN model for the OSARA dataset group and all five lengths of samples (3, 4, 5, 6, and 7 seconds). Similarly, figure 20 shows the training curves of the DRNN model for the FSARA dataset group and all five lengths. Likewise, figure 21 displays the training curves of the proposed DRNN for the MSARA dataset group and all five lengths too.

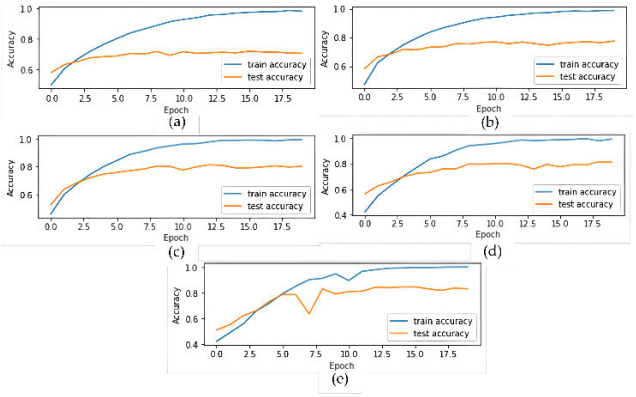


FIGURE 19. DRNN training curves for the OSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

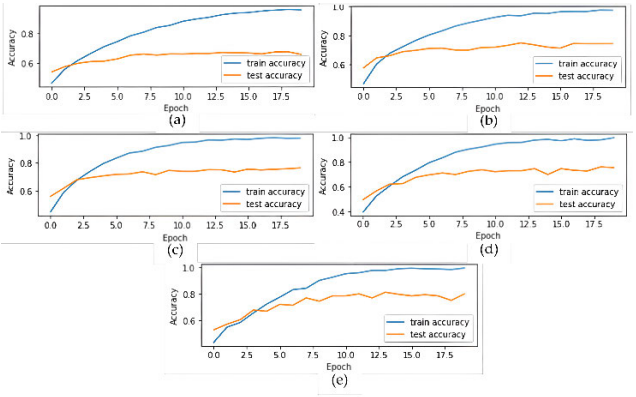


FIGURE 20. DRNN training curves for the FSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

From these figures, it can be explored that all trains of the proposed DRNN model have truly been achieved too. That is, in all cases the training accuracy curves are successfully raised until obtaining very high values. The testing performances are also taken into account during the training phase at each epoch as they are shown in the same figures. Their curves are enhanced along training epochs correspondingly to the training curves. Nonetheless, their performances vary

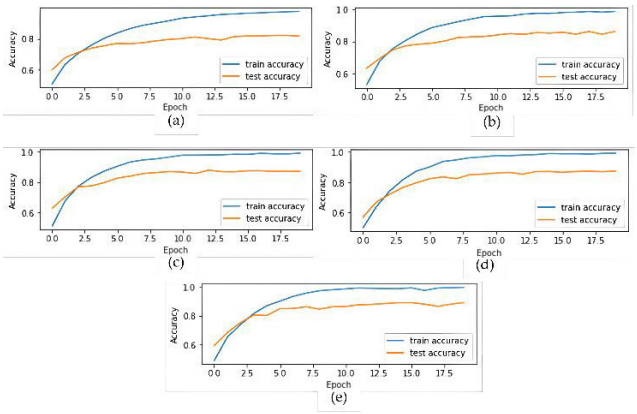


FIGURE 21. DRNN training curves for the MSARA dataset group and the lengths of (a) 3-second samples, (b) 4-second samples, (c) 5-second samples, (d) 6-second samples, and (e) 7-second samples.

between various cases. Again, the MSARA dataset group approves its capability of reporting high accuracies.

Moreover, the proposed DRNN model has generally recorded the best performances. Its accuracies in both training and testing are remarkable compared to other proposed models. This encourages selecting this model as the best proposed one in this study.

E. TESTING RESULTS

1) MLP TESTING RESULTS

The proposed MLP model has been tested for the three dataset groups of OSARA, FSARA, and MSARA. In addition to considering all the five lengths of samples: 3, 4, 5, 6, and 7 seconds. Table 4 shows the testing performances of this model for all the cases.

From this table, it can be noticed that the OSARA group of the employed dataset has obtained slight differences between almost all the performances of the five lengths of samples. That is, the range of differences for the accuracy is [70.7%, 72.8%] and the range of differences for the loss is [1.36, 2.35]. Such ranges are increased for the FSARA group where the accuracy range is being [65%, 71.2%] and the loss range is reached [1.38, 2.68]. The best performances can be found for the MSARA dataset group in all five lengths of samples. High and slight values for the variances between the accuracy outcomes are attained as the accuracy range values are raised to [85.6%, 89.5%]. Whereas low and slight values for the variances of almost all the loss outcomes are achieved where the range values are reduced to [0.67, 0.91]. These results are due to the specifications of each dataset group. To illustrate, the OSARA has the original or raw data that cannot be well analyzed, the FSARA has filtered data which may show big differences in evaluations, and the MSARA contains a combination between the OSARA and FSARA where this makes it more data with specifications to efficiently consider.

It is worth mentioning that the performances of the 7-second sample length for any dataset group are higher than the performances of the 3-second sample length for any dataset group. For example, the accuracy of the 7-second sample length for the MSARA dataset group is 89.5% and the loss is 0.91, whilst the accuracy of the 3-second sample length for the same dataset group is 85.6% and the loss is 0.67. This is because the length of samples is more effective if it is high as it involves lots of information that can be valuable in analyzing and providing significant outcomes. Although, the certain length contains a smaller number of samples such as the 7 seconds includes 438 samples for the MSARA dataset group, whereas, the 3 seconds involves 2508 samples for the same dataset group. Therefore, it can be yield that the longer length of samples, the higher the performances, even if the number of samples is smaller.

2) CNN TESTING RESULTS

The proposed CNN model has also been tested for the three dataset groups. In addition of takes into account all the five lengths of samples. Table 5 provides the testing performances of this model for all the cases.

From this table, it can be noticed that the OSARA group of the employed dataset has attained fluctuated variances between almost all the performances of the five lengths of samples. That is, the range of variances for the accuracy is [62.6%, 72.6%] and the range of differences for the loss is [0.74, 1.15]. Ranges are spread for the FSARA group where the accuracy range is found [59.1%, 72.9%] and the loss range is reported [0.75, 1.22]. Again, the best performances are recorded for the MSARA dataset group in all five lengths of samples. Dramatically increased values for the variances between the accuracy outcomes are achieved, and the accuracy range values are increased to [70.1%, 87.3%]. Whilst low decreased values for the variances of all the loss outcomes are obtained, the range values are decreased to [0.43, 0.92]. These results are consistent with the achievements in the MLP model. So, the same explanations can be considered here. That is, the provided performances are due to the specifications of each dataset group, where the OSARA has the original signals that cannot be well analyzed, the FSARA has filtered signals that may demonstrate big differences in assessments, and the MSARA includes fusion between the OSARA and FSARA where this makes it involves more signals with specifications to efficiently analyzed.

Also, here, the performances of the 7-second sample length for each dataset group are greater than the performances of the 3-second sample length for each dataset group. For instance, the accuracy of the 7-second sample length for the MSARA dataset group is 87.3% and the loss is 0.43, whilst the accuracy of the 3-second sample length for the same dataset group is 70.1% and the loss is 0.92. This is also can be explained as the length of samples being more valuable if it is high as it contains lots of information that can be effective in obtaining interesting outcomes. Thus, it can also be revealed that the longer the length of samples, the better the outcomes, despite the number of samples being smaller.

TABLE 4. Testing performances of the proposed MLP model for the three dataset groups (OSARA, FSARA, and MSARA) and all the lengths of samples (3, 4, 5, 6, and 7).

Sample Length		3 seconds		4 seconds		5 seconds		6 seconds		7 seconds	
Accuracy & Loss		Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
Dataset Group	OSARA	70.7	1.47	72.8	1.36	72.1	1.6	71.2	1.99	72.7	2.35
	FSARA	65	1.64	66.9	1.64	69.3	1.38	68.3	1.82	71.2	2.68
	MSARA	85.6	0.67	86.5	0.71	88	0.71	87.2	0.79	89.5	0.91

TABLE 5. Testing performances of the proposed CNN model for the three dataset groups (OSARA, FSARA, and MSARA) and all the lengths of samples (3, 4, 5, 6, and 7).

Sample Length		3 seconds		4 seconds		5 seconds		6 seconds		7 seconds	
Accuracy & Loss		Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
Dataset Group	OSARA	62.6	1.11	68.1	1.09	71.2	0.99	68.1	1.15	72.6	0.74
	FSARA	59.1	1.17	62.3	1.19	65.2	1.22	68.4	1.09	72.9	0.75
	MSARA	70.1	0.92	73.6	0.81	79.8	0.71	81.2	0.69	87.3	0.43

TABLE 6. Testing performances of the proposed DRNN model for the three dataset groups (OSARA, FSARA, and MSARA) and all the lengths of samples (3, 4, 5, 6, and 7).

Sample Length		3 seconds		4 seconds		5 seconds		6 seconds		7 seconds	
Accuracy & Loss		Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
Dataset Group	OSARA	70.3	1.71	77.4	1.34	79.9	1.05	81.1	0.78	82.9	0.95
	FSARA	65.9	1.98	74.3	1.37	76.3	1.25	75.7	1.27	79.9	0.87
	MSARA	81.6	0.95	85.9	0.68	87.1	0.59	88.7	0.60	90.7	0.46

3) DRNN TESTING RESULTS

The proposed DRNN model has been tested for the three dataset groups too. In addition to carrying out all the five lengths of samples. Table 6 demonstrates the testing performances of this model for all the cases.

From this table, the OSARA group of the employed dataset has achieved dramatic differences between almost all the performances of the five lengths of samples. That is, differences in the accuracies are dramatically increased between the range of [70.3%, 82.9%], and differences in the loss significantly fluctuate between the range of [1.36, 2.35]. Such ranges and variances are also expanded for the FSARA group, the accuracy range is reached [65.9%, 79.9%] and the loss range is [0.87, 1.98]. The best performances can also be observed for the MSARA dataset group in all five lengths of samples. Significant values for the outcomes are bench-marked as the accuracy range values are dramatically raised between the range of [81.6%, 90.7%], and the loss range values fluctuate between the range of [0.46, 0.95]. These results are also consistent with the attainments in the MLP and CNN models. Again, the same illustrations can be provided here. To illustrate, the provided performances depend on the specifications of each dataset group, where the OSARA has the raw data that cannot be well considered, the FSARA has filtered data which may give big variances in examinations, and the MSARA has combined information between the OSARA and FSARA, it involves more data with specifications to effectively consider.

Moreover, the performances of the 7-second sample length for all dataset groups are also higher than the performances of the 3-second sample length for all dataset groups. For example, the accuracy of the 7-second sample length for the MSARA dataset group is 90.7% and the loss is 0.46, whilst the accuracy of the 3-second sample length for the same dataset group is 81.6% and the loss is 0.95. This is also due to the length of samples which is being more influenced if it is high as it includes a lot of information that can be important in analyzing and yielding valuable outcomes. So, it can be emphasized that the bigger the length of samples, the greater the performances, even though the number of samples is smaller.

To justify the choice of a successful model, various aspects have been considered as the architecture and training time. Regarding the architecture, each of the three models is composed of a special and useful architecture. To illustrate, the MLP uses a feedforward architecture with considers suitable adaptation for its inputs and adopts a small number of hidden layers, the CNN utilizes a comparable architecture that is especially designed for the inputs of images but with a big number of hidden layers, and the DRNN exploits more effective architecture which involves valuable feedback added to multiple hidden layers. Regarding the training time,

more computations are carried out to the employed MLP, CNN, and DRNN models. They yield that the MLP consumes 605.7 seconds, the CNN takes 404.5 seconds and the DRNN spends 485.8 seconds. Such results show near-consuming times for training phases, especially for the CNN and DRNN. Reasonably, the testing times take further smaller times than the training times as they only require very short times for straightforward analyses. To this end, accuracy is used here as the most acceptable justification for choosing the best model, where the best results between the three proposed machine learning models are benchmarked for the DRNN. Specifically, the best performances are found for the DRNN with the MSARA dataset group and 7 seconds sample length, where the highest accuracy is recorded as 90.7% and the lowest loss is reported as 0.46.

V. CONCLUSION AND FUTURE WORK

In this work, a system was designed to distinguish between the three main Arabic dialects of APD, ED, and LD. The system went through several stages which began with collecting the SARA dataset. Followed by the preparation stage, where the SARA was prepared and derived into three groups (OSARA, FSARA, and MSARA). The next stage was for the feature extraction, here the MFCC was considered. After that, three artificial neural network models were proposed, these are MLP, CNN, and DRNN. The three models were trained and tested comprehensively by applying extensive experiments considering the three groups of the dataset, multiple lengths of samples, and different evaluations.

In conclusion, it has been found that the highest accuracy of 90.70% and lowest loss of 0.46 are achieved for the proposed DRNN model with the 7-second sample length and MSARA group. The MSARA group has confirmed its ability to obtain the highest testing performances for all three proposed models. It could also be investigated that the longer the length of samples used, the higher the performances attained, despite the smaller number of samples.

In future work, it is possible to use this study to detect the nationality of speakers by using a part of the daily speech. In addition to acquiring a database that includes all Arab countries' dialects. It is crucial to include dialects from all of the main locations, including North Africa, especially in light of the growing interaction between speakers of different Arabic dialects.

REFERENCES

[1] N. B. Chittaragi and S. G. Koolagudi, "Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms," *Lang. Resour. Eval.*, vol. 54, no. 2, pp. 553–585, Jun. 2020, doi: 10.1007/s10579-019-09481-5.

[2] S. Shivaprasad and M. Sadanandam, "Identification of regional dialects of Telugu language using text independent speech processing models," *Int. J. Speech Technol.*, vol. 23, no. 2, pp. 251–258, Jun. 2020, doi: 10.1007/s10772-020-09678-y.

[3] D. Wang, S. Ye, X. Hu, S. Li, and X. Xu, "An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model," in *Proc. Interspeech*, Aug. 2021, pp. 3266–3270, doi: 10.21437/interspeech.2021-374.

[4] A. B. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: Automatic speech recognition and ethnicity-related dialects," *Speech Commun.*, vol. 140, pp. 50–70, May 2022, doi: 10.1016/j.specom.2022.03.009.

[5] F. Biadsy, J. B. Hirschberg, and N. Y. Habash. (2009). *Spoken Arabic Dialect Identification Using Phonotactic Modeling*. [Online]. Available: http://biadsy_al_09b(1).pdf

[6] N. Y. Habash, *Introduction to Arabic Natural Language Processing*. Switzerland: Springer, 2010, doi: 10.2200/S00277ED1V01Y201008HLT010.

[7] M. Elaraby and M. Abdul-Mageed, "Deep models for Arabic dialect identification on benchmarked data," in *Proc. 27th Int. Conf. Comput. Linguistics (COLING), 5th Workshop NLP Similar Languages, Varieties Dialects*, 2018, pp. 263–274.

[8] R. R. O. Al-Nima, F. H. Abdulraheem, and M. Y. Al-Ridha, "Using hand-dorsal images to reproduce face images by applying back propagation and cascade-forward neural networks," in *Proc. 2nd Int. Conf. Electr., Commun., Comput., Power Control Eng. (ICECCPCE)*, Feb. 2019, pp. 97–102.

[9] A. A. Ismael, S. J. Suleiman, R. R. O. Al-Nima, and N. Al-Ansari, "Predicting the discharge coefficient of oblique cylindrical weir using neural network techniques," *Arabian J. Geosci.*, vol. 14, no. 16, pp. 1–8, Aug. 2021.

[10] M. Y. Al-Ridha, A. S. Anaz, and R. R. O. Al-Nima, "Expecting confirmed and death cases of COVID-19 in Iraq by utilizing backpropagation neural network," *Bull. Electr. Eng. Informat.*, vol. 10, no. 4, pp. 2137–2143, Aug. 2021.

[11] A. H. S. Hamdany, R. R. Omar-Nima, and L. H. Albak, "Translating cuneiform symbols using artificial neural network," *TELKOMNIKA*, vol. 19, no. 2, pp. 438–443, Apr. 2021.

[12] R. R. Ziedan, M. N. Micheal, A. K. Alsammak, M. F. M. Mursi, and A. S. Elmaghraby, "A unified approach for Arabic language dialect detection," in *Proc. 29th Int. Conf. Comput. Appl. Ind. Eng.*, 2016, pp. 165–170.

[13] A. Ali, N. Dehak, P. Cardinal, S. Khurana, S. H. Yella, J. Glass, P. Bell, and S. Renals, "Automatic dialect detection in Arabic broadcast speech," 2015, *arXiv:1509.06928*.

[14] S. Shon, A. Ali, and J. Glass, "MIT-QCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 374–380.

[15] S. Shon, A. Ali, and J. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," 2018, *arXiv:1803.04567*.

[16] M. Najafian, S. Khurana, S. Shan, A. Ali, and J. Glass, "Exploiting convolutional neural networks for phonotactic based dialect identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5174–5178.

[17] A. Hanani and R. Naser, "Spoken Arabic dialect recognition using X-vectors," *Natural Lang. Eng.*, vol. 26, no. 6, pp. 691–700, Nov. 2020, doi: 10.1017/s1351324920000091.

[18] H. A. Alsayadi, S. Al-Hagree, F. A. Alqasemi, and A. A. Abdelhamid, "Dialectal Arabic speech recognition using CNN-LSTM based on end-to-end deep learning," in *Proc. 2nd Int. Conf. Emerg. Smart Technol. Appl. (eSmarTA)*, Oct. 2022, pp. 1–8, doi: 10.1109/eSmarTA56775.2022.9935427.

[19] S. Nasr, R. Duwairi, and M. Quwaider, "End-to-end speech recognition for Arabic dialects," *Arabian J. Sci. Eng.*, vol. 48, no. 8, pp. 10617–10633, Aug. 2023, doi: 10.1007/s13369-023-07670-7.

[20] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016, doi: 10.1109/TASLP.2015.2512042.

[21] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 241–245, doi: 10.1109/ICASSP.2017.7952154.

[22] W. Mack and E. A. P. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Process. Lett.*, vol. 27, pp. 61–65, 2020.

[23] H. Schröter, T. Rosenkranz, A. N. Escalante-B, M. Aubreville, and A. Maier, "CLCNET: Deep learning-based noise reduction for hearing aids using complex linear coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6949–6953, doi: 10.1109/ICASSP40776.2020.9053563.

[24] H. Schroter, A. N. Escalante, T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7407–7411.

[25] H. Schröter, A. Maier, A. N. Escalante, and T. Rosenkranz, "DeepFilterNet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2022, pp. 1–5.

[26] M. H. Bahari, N. Dehak, H. Van hamme, L. Burget, A. M. Ali, and J. Glass, "Non-negative factor analysis of Gaussian mixture model weight adaptation for language and dialect recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 7, pp. 1117–1129, Jul. 2014, doi: 10.1109/TASLP.2014.2319159.

[27] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *Proc. 4th Int. Conf. Signal Process. Commun. Syst.*, Dec. 2010, pp. 1–5.

[28] M. T. S. Al-Kaltakchi, R. R. O. Al-Nima, and M. A. M. Abdullah, "Comparisons of extreme learning machine and backpropagation-based I-vector approach for speaker identification," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 28, no. 3, pp. 1236–1245, May 2020.

[29] A. S. Haq, M. Nasrun, C. Setianingsih, and M. A. Murti, "Speech recognition implementation using MFCC and DTW algorithm for home automation," *Proc. Electr. Eng. Comput. Sci. Informat.*, vol. 7, no. 2, pp. 78–85, Oct. 2020.

[30] R. Ranjan and A. Thakur, "Analysis of feature extraction techniques for speech recognition system," *Int. J. Innov. Technol. Exploring Eng.*, vol. 8, pp. 197–200, May 2019.

[31] M. Limkar, R. Rao, and V. Sagvekar, "Isolated digit recognition using MFCC and DTW," *Int. J. Adv. Electr. Electron. Eng., (IJAEEE)*, vol. 1, no. 1, 2012.

[32] M. E. Rahaman, S. M. S. Alam, H. S. Mondal, A. S. Muntaseer, R. Mandal, and M. Raihan, "Performance analysis of isolated speech recognition technique using MFCC and cross-correlation," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–4, doi: 10.1109/ICCCNT45670.2019.8944534.

[33] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database," in *Proc. IEEE 4th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Feb. 2019, pp. 130–133.



DHEYAA ALHELAL was born in Mosul, Iraq, in 1983. He received the B.S. degree in computer engineering from Northern Technical University, Mosul, in 2006, and the M.S. degree in computer engineering from Bridgeport University, Bridgeport, CT, USA, in 2014. He is currently pursuing the Ph.D. degree in computer engineering with Altınbaş University, Istanbul, Türkiye.

Since 2007, he has been working as a Lecturer with the Department of Computer Engineering, Technical Engineering College Mosul, Northern Technical University. His research interests include embedded systems, signal processing, image processing, and artificial intelligence.

Mr. Alhelal has memberships in the Iraqi Academics Syndicate other than the Iraqi Engineers Union.



TIMUR İNAN was born in Istanbul, Türkiye, in 1979. He received the combined B.S. and M.S. degree in electronics and computer education and the Ph.D. degree in electrics-electronics engineering from the University of Marmara, in 2013 and 2019, respectively.

From 2011 to 2019, he was a Lecturer at the Department of Computer Programming, İstanbul Arel University. From 2021 to 2024, he worked as an Assistant Professor at the Department of Software Engineering, Altınbaş University, Türkiye. Since 2024, he has been working as an Assistant Professor with the Department of Computer Engineering, Marmara University. His research interests include artificial intelligence, optimization algorithms, fuzzy logic, and embedded systems.