

# Recognizing AlMuezzin and his Maqam using deep learning approach

Nahlah Mohammad Shatnawi<sup>1</sup>, Khalid M. O. Nahar<sup>1</sup>, Suhad Al-Issa<sup>2</sup>, Enas Ahmad Alikhashashneh<sup>3</sup>

<sup>1</sup>Department of Computer Science, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, Jordan

<sup>2</sup>Department of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK

<sup>3</sup>Department of Information Systems, Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, Jordan

## Article Info

### Article history:

Received Jul 14, 2024

Revised Mar 27, 2025

Accepted Jul 2, 2025

### Keywords:

Aladhan

AlMuezzin

Arabic language

Deep learning

Maqam

Speech recognition

VGG-16

## Abstract

Speech recognition is an important topic in deep learning, especially to Arabic language in an attempt to recognize Arabic speech, due to the difficulty of applying it because of the nature of the Arabic language, its frequent overlap, and the lack of available sources, and some other limitations related to the programming matters. This paper attempts to reduce the gap that exists between speech recognition and the Arabic language and attempts to address it through deep learning. In this paper, the focus is on Call for Prayer (Aladhan: الأذان) as one of the most famous Arabic words, where its form is stable, but it differs in the notes and shape of its sound, which is known as the phonetic Maqam (Maqam: المقام الصوتي). In this paper, a solution to identify the voice of AlMuezzin (المؤذن), recognize AlMuezzin, and determine the form of the Maqam through VGG-16 model presented. The VGG-16 model examined with 4 extracted features: Chroma feature, LogFbank feature, MFCC feature, and spectral centroids. The best result obtained was with chroma features, where the accuracy of Aladhan recognition reached 96%. On the other hand, the classification of Maqam with the highest accuracy reached of 95% using spectral centroids feature.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Nahlah Shatnawi

Department of Computer Science, Faculty of Information Technology and Computer Sciences

Yarmouk University

21163, Irbid, Jordan

Email: nahlah.s@yu.edu.jo

## 1. INTRODUCTION

Speech recognition is one of the most active research areas that aims to identify the speaker based on the characteristics of their voice [1]. Speech recognition contributes to improving several disciplines, such as health care and security. Several state-of-the-art works have recently explored the use of feature extraction techniques to describe a massive amount of data using different feature vectors that represent different physical and acoustic meanings. Selecting a good feature will help to improve the accuracy of the recognition. Thus, choosing the feature extraction technique is considered a critical step in the speaker recognition process. Currently, the most used speech characteristics are the linear reduction spectrum coefficient (LPCC) [2] and the MEL spectrum coefficient (MFCC) [3]. These features have achieved good recognition effects in speech recognition [4], [5].

Traditional automatic speech recognition (ASR) systems still employ an architecture consisting of numerous components, including but not limited to lexicon building, language models, and acoustic models. Various techniques are employed to construct and process these components, including traditional machine

learning (ML) techniques, Gaussian mixture models, hidden Markov models, a deep neural network, and a hybrid HMM-DNN [6].

Aladhan is a call to prayer for Muslims, where the AlMuezzin pronounces the call to prayer every day at the beginning of the time for each of the five obligatory prayers. In the past, the AlMuezzin used to give the call to prayer from a high place, or from the lighthouse, or from the roof of the mosque, but now the AlMuezzin gives the call to prayer through amplification devices, which makes the matter much easier for him.

Aladhan is an announcement of the time of prayer with specific and customized words, through which the AlMuezzin informs people of the times of prayer and invites them to prayer. Aladhan is announced through specific words in the following format, as in Figure 1.

<p>Allah is Great, Allah is Great          I bear witness that there is no God except Allah          I bear witness that Mohammad is the Messenger of Allah          Come to the prayer          Come to the prayer          Come to the success          Come to the success          Allah is Great, Allah is Great          There is no God except Allah</p>	<p>الله اكبر، الله اكبر          اشهد ان لا اله الا الله          اشهد ان محمدا رسول الله          حي على الصلاة          حي على الصلاة          حي على الفلاح          حي على الفلاح          الله اكبر، الله اكبر          لا اله الا الله</p>
---	--

Figure 1. Aladhan words format

AlMaqamat is a complex tonal system used in traditional Arabic music. Phonetic Maqams are characterized by a set of specific pitches and their own rules of performance. AlMaqamat can be used to create a wide range of musical emotions and effects. Phonetic Maqams are usually classified into two main groups: basic Maqams and sub-Maqams. The basic Maqams are the Maqams that form the basis of the Arabic musical system. There are nine basic shrines, which are: AlRust, AlNahawand, AlHijaz, AlSiyka, AlBayat, AlSaba, AlEajam. Sub-Maqams are Maqams that are derived from the basic Maqams. There are many sub-Maqams, but some are more common than others.

Various Maqams with distinctive melodies accompanying them; these are the features of the call to prayer in most different Islamic countries, as it is a call to prayer with sweet maqams appropriate to the time of the obligatory prayer and the psychological state of the people. The Muzzins in all mosques and mosques in most different Islamic countries are keen to perform this call in the best way, in a manner that suits the maqams adopted hundreds of years ago.

Hence, the importance of this work, as it links deep learning to the Arabic language, especially classical Arabic music. And because of the importance of Arabic music, its chosen and selected Aladhan as an actual application for speech recognition because it is known among people, especially Muslims, and because it contains a specific and fixed set of words, its form is stable, and it also has a specific and fixed tone in musical Maqamat.

To achieve the objective of this paper, the authors present a VGG-16 model [7] to identify AlMuezzin (المؤذن) and classify his Maqam (المقام الصوتي). This performed by extracting a set of features from the collected dataset. Then, speaker-independent speech recognition performed through the VGG-16 model because no two speakers have the same voice and the organs of the sound differ. So, the VGG-16 model can distinguish different Azzan collected from different Muezzin under several Maqamat: Al-Hejaz, Al-Sika, Al-Rust, Al-Saba, Al-Ashaq, and Nahawand.

This paper describes an interesting idea for using sound features for Muezzin identification. The proposed approach in this research is hot, uncommon, and could be improved to be applied in many other fields. The work in this study can be the first milestone to show the effectiveness of deep learning (DL) for classification and the dependence on sound features to identify AlMuezzin.

The rest of this paper is organized as follows: in section 2, the related works to the proposed approach shown. Next, the collected dataset information presented. In section 4, introduce the work method. Section 5, include experimental results. Finally, the conclusion is determined.

## 2. RELATED WORK

This section presents a few works that utilize DL, ML, feature extraction, feature reduction, and view how they relate to this work. The researchers of the paper [8] collected a modern arabic dataset to assess the performance of a few of the DL strategies in human speech recognition (HSR).

In this work, the accuracy of the modular hidden Markov model-deep neural network (HMM-DNN) frameworks was compared to the native speaker performance. The comparison makes it appear that human performance within the Arabic dialect is still significantly better than that of machines, with an absolute word error rate (WER) gap of 3.5% on average.

On the other hand, in paper [9], there's an endeavor to construct a strong, robust diacritized Arabic ASR utilizing Deep learning approaches. They utilized the standard arabic single speaker corpus (SASSC), which contains seven hours of cutting-edge standard Arabic discourse, to prepare and test a modern CTC-based ASR, convolutional neural network (CNN)-long short-term memory (LSTM), and an attention-based end-to-end approach to make strides in diacritized Arabic ASR. From the exploratory results, the researchers conclude that the CNN-LSTM with an attention framework outperforms conventional ASR and the Joint CTC-attention ASR framework within the task of Arabic speech recognition.

In work [10], the researchers used a deep feed-forward neural network (DFFNN) to the Arabic natural audio dataset (ANAD), which is designed for Arabic automatic speech recognition. The ANAD dataset contains three discrete feelings: angry (A), surprised (S), and happy (H). The researchers also utilized eight videos of live calls between an anchor and a human outside the studio that were downloaded from online Arabic talk shows to test and evaluate the proposed approach. The target was to recognize human feelings from the sounds. They proposed an automated Arabic speech emotion recognition system using feature extraction to extract the foremost imperative features from the dataset, which was at that point utilized to train the DFFNN. In this investigation, it appears that the DFFNN achieves the highest accuracy when applying PCA to the extracted features, with an accuracy of 98.56%.

Moreover, the work [11] a speech emotion recognition system based on deep neural network hidden markov models (DNN-HMM) by extricating MFCC and epoch-based features. The researchers concluded that the accuracy when utilizing MFCC features was 60.86% whereas when utilizing epoch-based features, it was 54.52%. Also, the the recognition performance to 64.2% when MFCC and epoch features are combined.

Fahad *et al.* [12], they presented a convolutional neural network for Arabic speech recognition. In this investigation, they centered on single-word Arabic automatic speech recognition (AASR). They utilized log-frequency spectral coefficients (MFSC) and Gammatone-frequency cepstral coefficients (GFCC) with their first and second-order derivatives. They found that the greatest accuracy gotten when utilizing GFCC with CNN is 99.77%, and it appeared that the CNN accomplished way better performance in AASR. A traditional ML approaches such as the random forest (RF) were used by [13] to distinguish different speakers by extraction mel-frequency cepstral coefficients (MFCC) and reconstructed phase space (RPS) features. The researchers of this investigation observed that the accuracy in MFCC is higher than in RPS, where the accuracy obtained from utilizing RPS features was 71% and the accuracy obtained from utilizing MFCC features was 97%.

Another speaker-identification framework was proposed by the researchers [14] to recognize spoken sounds by utilizing particular words. The researchers extracted the MFCC features and then utilized them as input for the recurrent neural network (RNN) and LSTM. They found that the accuracy in different RNNs is 87.74%, and the accuracy that showed up in a single RNN is 80.58%. On the other hand, Utomo *et al.* [15] proposed automatic speaker recognition by artificial neural network (ANN).

They extracted the Another speaker-identification framework by utilizing particular words. The researchers extracted the MFCC features and then utilized them as input for the RNN and LSTM. They found that the accuracy in different RNNs is 87.74%, and the accuracy that showed up in a single RNN is 80.58%.

Moreover, the work in [16] proposed text-speaker recognition to recognize what the speaker said. They utilized MFCC, spectrum, and log-spectrum to extract the features from the speaker sound wave the extracted features were at that point utilized to to train and evaluate the LSTM and RNN models. The accuracy by utilizing MFCC was 95.33%, whereas by utilizing spectrum and log-spectrum, it was 98.7%. Analysts in [17] proposed speaker-identification in a noisy environment. They utilized CNN to classify 60 speakers and divided 4 voice samples for each speaker. The researchers utilized MFCC to extract the features from the speaker signal and found an accuracy of 87.5%.

In addition, ion, ion, analysts in [18] propose a speaker-identification framework by utilizing the gaussian mixture model (GMM), and MFCC to extract the features. The researchers extracted features and compared

them with all the features they had saved. In DNN-HMM, numerous sorts of inquiry have been made about comparing two strategies for this point.

In Chowdary *et al.* [19], analysts compared extraction and normalization strategies for speakers; they utilized MFCC and PNCC for feature extraction. This framework was connected to six men and two women, and the performance of the speaker's identification was 100%. PNCC had far better, much better, higher, stronger, and improved performance than MFCC in recognizing women. Al-Kaltakchi *et al.* [20], analysts compared extraction strategies by utilizing FFANN and SVM and utilized numerous extraction methods such as MFCC, PLP, and LPC. They connected these feature extractions to two calculations, ANN and SVM, to discover the finest include extractions that seemed to apply to these two calculations. The best accuracy they obtained was when using MFCC, PLP, and LPC with SVM and ANN; it was 100%.

The work in [21] examined the efficient of extricating and coordinating feature strategies utilizing vector quantization and MFCC together in speaker-identification and speech recognition applications. They found that utilizing vector quantization reduced the time of comparison between the input speech and the testing speech. Several papers have used speech recognition system as part of smart home automation systems. For example, the work in [22] proposed a speech recognition framework without having to utilize the web to assist individuals with disabilities by utilizing Ivona. converted to text and received and gotten by GSM modems.

This framework permits individuals with inabilities to deliver house voice commands to carry out a particular command. Another study done by [23] proposed a speech recognition framework for individuals with disabilities to control their wheelchairs and other devices. The researchers used MATLAB to create 8 commands (GO, HELP, START, REPEAT, ERASE, NO, ENTER, YES) by extricating MFCC features; all extracted features are put away in K-mean cluster forms. The average accuracy of 8 commands is 73.54%, and the average accuracy of listener is 82.25%.

In the work of [24], [25], the researchers constructed speaker-independent NSR systems using the DeepSpeech model, then assessed them using the WER. The DeepSpeech model is one of the most well-known open-source ASR models from Mozilla for Quranic recitations for male and female reciters. Where the target is to be utilized efficiently by anyone, regardless of gender or age, and it obtained intriguing results.

Shareef and Al-Irhayim, [26] they do speech sound errors classification impairments children when are incorrectly pronounced in Arabic. They employ Mel frequency spectral coefficients for feature extraction, and deep LSTM network. They gain classification accuracy reaches 97.99% and loss 0.18%.

To the best of our knowledge, deep learning approach has not been used in the automatic speech recognition of AlMuezzin. The call to prayer is a special type of speech that announces the call to prayer in a mosque through a standardized set of words with a Maqam. Recognize the AlMuezzin with the Maqam he follows will contribute to identifying the speaker, just as identifying the Maqam will contribute to the problem of speech synthesis. Based on the investigation and literature review conducted, this paper has been prepared.

### 3. METHOD

The proposed method for recognizing AlMuezzin and his maqam using deep learning approach consists of two phases. The first phase is for AlMuezzin identification, and the second phase is for Acoustic stand classification of Azan (Maqam classification). Where for each phase, different dataset collected and used.

In this work, the VGG16 model in [7] used. VGG16 model reaches a test accuracy of 92.7% with almost 14 million training photos from 1000 item classes in ImageNet, and was one of the best models from the ILSVRC-2014 competition. VGG one of the most widely used deep learning models for image recognition. As the name implies, VGG16 is a 16-layer deep neural network. With 138 million parameters overall, VGG16 is a relatively large network—huge even by today's standards. Nevertheless, the key selling point of the VGGNet16 architecture is its simplicity, as it incorporates the most significant convolution neural network features.

Since the problem is to achieve both AlMuezzin identification, and acoustic stand of Azzan, the pictorial design of the methodology is stated into two phases as shown in Figure 2. Figure 2(a): AlMuezzin identification. Figure 2(b): Aladhan Maqam classification. After collecting the data, it needs to be processed, then the audio file spectrum is extracted using different features to be trained on a VGG16 model. The detailed steps of the methodology are explained in the following subsections.

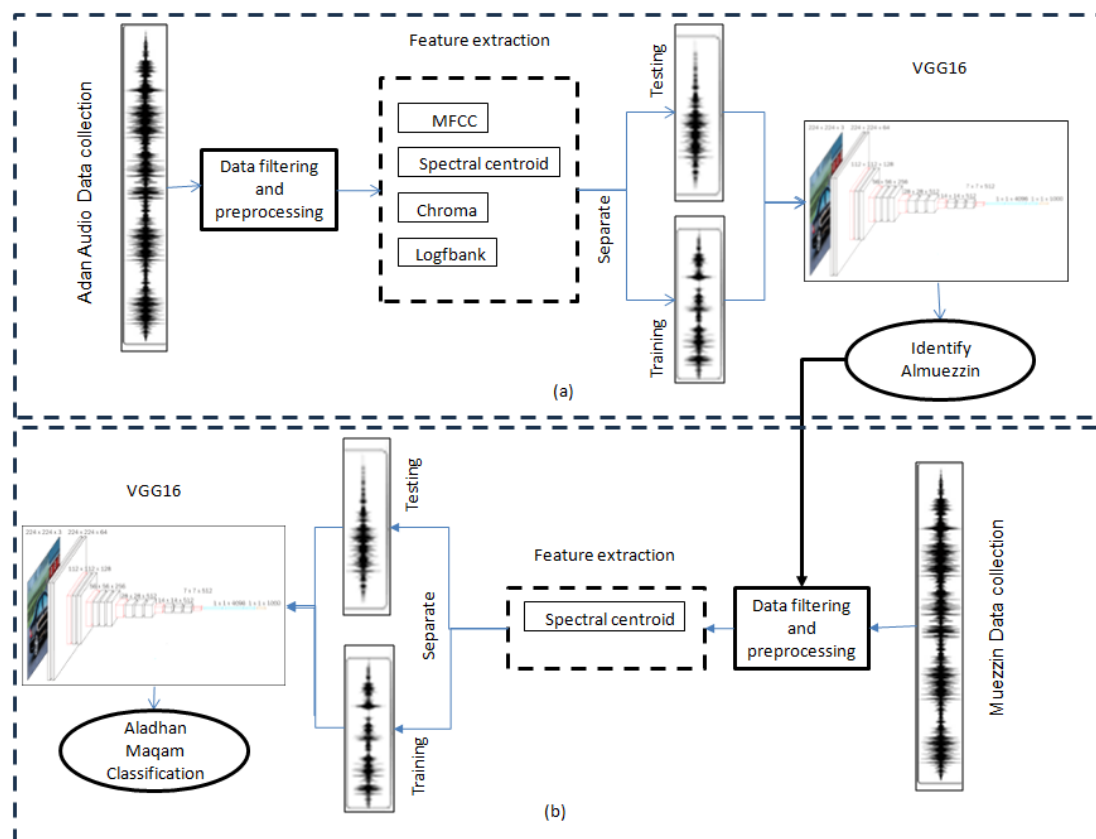


Figure 2. The methodology phases: (a) AlMuezzin identification and (b) Aladhan Maqam classification

To ensure full reproducibility of the experimental setup, the methodology is structured into two clearly defined phases: AlMuezzin identification and Maqam classification, each with distinct datasets and preprocessing pipelines. The experimental workflow begins with a carefully curated dataset of audio recordings collected from YouTube, comprising well-known Muezzins and multiple Maqam styles. These recordings were converted to WAV format (16-bit stereo, 44.1 kHz) and segmented into 20-second clips to manage memory efficiency and enhance model performance. The feature extraction was then performed using four widely validated acoustic features: MFCC, LogFBank, Spectral Centroid, and Chroma. Each feature set was independently used to train a pre-trained VGG-16 model, allowing a comparative evaluation of their effectiveness. For AlMuezzin identification, the model was trained on 1,211 samples and tested on 295, while for Maqam classification, 287 training samples and 71 testing samples were used. The models were validated using standard performance metrics (accuracy and loss) over multiple epochs, and visualizations of training behavior (e.g., accuracy/loss curves) were included in the Results section. A detailed pictorial representation of the pipeline shown in Figure 3, and structured equations for each feature extraction method Table 1 are provided to ensure transparency and reproducibility of our experimental design.

### 3.1. Dataset

The dataset was collected manually and carefully from YouTube in two stages, you can find the dataset in the link in [27]. In the first stage, 19 different audio records of the Aladhan for 19 famous male Muezzins collected, a total of 105 audio records were collected. Then, transferring each audio file in the datasets into WAV audio files with a 16-bit stereo and 44.1 kHz sample rate so a VGG-16 model's can handle [7]. After that, audio files divided by reciters and created different folders with the names of the reciters. The reader's audio files are stored in the folder with the reader's name. Thus, the intersection of audio files between readers to perform speaker-independent identification avoided. The audio files divided into 80% for the training group, 20% for testing. The largest training ratio is to ensure sufficient and good training of the system.

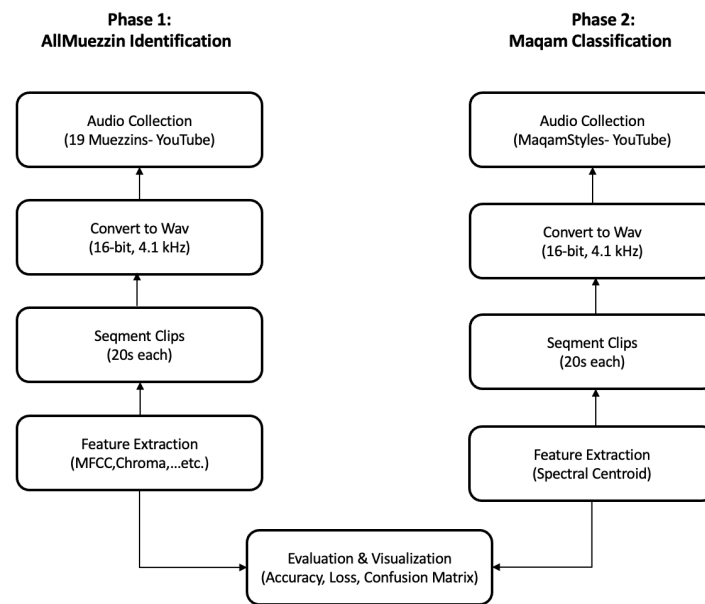


Figure 3. Graphical overview of simulation and experimental setup

To increase the performance of the prediction, each audio of Aladhan divided into several audio tracks of 20 seconds each because the memory resources are limited and to avoid out of memory issue, so the total amount of data in the training becomes 1506 audio. But after deleting the empty audio files and the corrupted audio files, the remaining 1506 audio files separated so 1211 audio records for training and 295 audio records for testing. Finally, noise removed from data to ensure the data is clean. The target of this stage is to perform AlMuezzin identification.

In the second stage, records collected of different calls to prayer from different audio maqams for different Muezzins, such as Hijaz, Sikka, Al-Sada', Al-Saba', Al-Ashaq, and Al-Nahawand, as in Table 1. The total number of audio recordings collected was 36. Also, as in the first stage, the data divided in each records here into several audio tracks of 20 seconds each, bringing the size of the data to be trained to 358 in this phase. A splitting ratio of 80% for training and 20% for testing was used. So, 287 audios used for training and 71 audios for testing. The target of this stage is to perform Al-Maqam identification.

### 3.2. AlMuezzin identification

For AlMuezzin identification the collected data is first preprocessed as mentioned in dataset section, after that features extraction using four distinct feature types to train them into pre-trained VGG16 model. The features are extracted from the speech signal for analysis are: MFCC, Spectral - Centroid, Chroma, and LogFBank as shown in Figure 4.

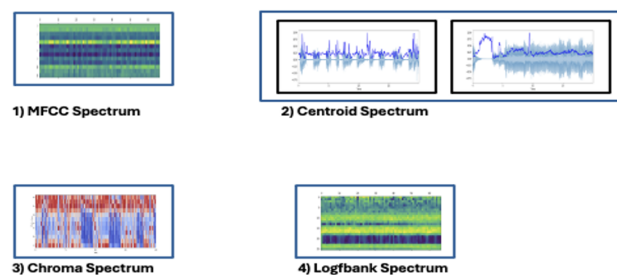


Figure 4. Features extraction from four different feature types

MFCC is so well-liked because its foundation is a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. In spectral centroid every frequency band's spectrum has a center of gravity.

When a piece of music has pitches that can be meaningfully categorized and is tuned close to the equal-tempered scale, it is said to have chroma. While LogFBank idly used in robust speech recognition community. Table 1 showed the computation of the four features.

Table 1. Features computations of MFCC, Spectral Centroid, Chroma, and LogFBank

Features	Description
MFCC	$x'(n) = x(n) - \alpha * x(n-1)$ $X(f) = FFT(w(n))$ <p>Apply a Mel filter bank to the power spectrum of the signal</p> $y_k = \log(\sum  y_k(n) ^2)$
Spectral - Centroid	<p>Compute the discrete cosine transform (DCT) for the log-energy output</p> $centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$
Chroma	<p>Ccalculated as the weighted mean of the frequencies present in the signal Using short-time Fourier transforms in combination with binning strategies</p> $f_g f[n, k] = \sum_{m=0}^{M-1} f[n-m]g[m] \in k[m]$ <p>Where</p> $\in k[m] = e^{-2\pi m \frac{k}{N}}$
LogFBank	<p>M is the window length of g and N is the number of samples in f</p> $x'(n) = x(n) - \alpha * x(n-1)$ $w(n) = x'(n) * h(n)$ $X(f) = FFT(w(n))$ <p>Apply a logarithmically spaced filter bank to the power spectrum</p> $y_k = \log(\sum  y_k(f) ^2)$

After applying the features on the dataset, VGG16 used to identify AlMuezzin. Tiny convolution filters make up a VGG network. Thirteen convolutional layers and three fully linked layers make up VGG16. An overview of the VGG architecture is provided below:


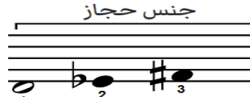
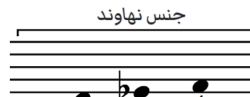
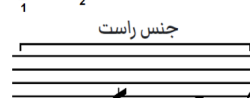
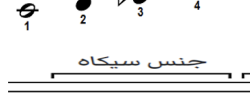
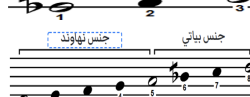
- Input: VGGNet is fed a 224 by 224 picture input.
- Convolutional layers: VGG's convolutional filters use a 3x3 receptive field, the smallest available. In addition, VGG performs a linear transformation on the input using a 1x1 convolution filter.
- ReLu activation: AlexNet's primary innovation for cutting training time is the rectified linear unit activation function (ReLU) component. ReLU is a linear function that yields zero for negative inputs and a matching output for positive inputs. To maintain the spatial resolution following convolution, VGG uses a fixed convolution stride of 1 pixel (the stride value shows how many pixels the filter “moves” to cover the complete space of the picture).
- Hidden layers: unlike AlexNet, which uses local response normalization, all of the VGG network's hidden layers employ ReLU. The latter adds little to the accuracy overall but lengthens the training period and memory usage.

- Pooling layers: a pooling layer is placed after a series of convolutional layers, which serves to lower the feature maps produced by each convolution step's dimensions and parameter count. Given the quick increase in the number of filters accessible from 64 to 128 to 256 and finally 512 in the final layers, pooling is essential.

### 3.3. Aladhan Maqam classification

Aladhan Maqam classification is the second phase where the resulted data from phase 1 is used beside another dataset collected as mentioned in the dataset section. In this phase, the data is also preprocessed and filtered, and then apply spectral centroid for feature extraction using the equations in Table 1. So, the VGG-16 model can distinguish different Azan collected from different Muezzin under several Maqamat: Al-Hejaz, Al-Sika, Al-Rust, Al-Saba, Al-Ashaq, and Nahawand, details of each Maqam mentioned in Table 2.

Table 2. Different Muezzins under several Maqams

Acoustic stand	Description	Musical scale
Al-Saba	Saba is a very common maqam in Arabic music. The ladder of this maqam begins on the decision, and the Hijaz on the third degree overlaps with it.	جنس صبا 
Al-Hejaz	Maqam Hijaz is the main maqam in the Maqam al-Hijaz family. The scale of the place of the Hijaz begins with the genus Hijaz on the decision, followed by the Rust.	جنس حجاز 
Nahawand	Maqam Nahawand is the main maqam in the Maqam Nahawand family. The ladder of this maqam begins with the genus al-Nahawand on the first degree (Qarar), followed by the genus al-Hejaz.	جنس نهاوند 
Al-Rust	The Rust Maqam is the main maqam in the Rust Maqam family. The scale of this maqam begins with the genus of the Rust on the first degree (Qarar), followed by any of the genus Nahawand or the genus of the higher Rust.	جنس راست 
Al-Sika	It is the main maqam in the Sika Maqam family, but it is rarely used as an independent maqam.	جنس سیکاه 
Al-Ashaq	The Maqam Ashaq Egyptian is a sub-maqam in the Maqam Nahawand family.	جنس بیاتی 

## 4. RESULTS AND DISCUSSION

At the beginning, a neural network model built, where the identification problem was resolved using a pre-trained VGG-16 model. Ultimately, the necessary model for the identification process was generated by using 80% of the gathered data in the training phase, which was then sent as a sample to VGG16. In this section the experiments conducted and the results that's obtained presented.

In the first phase, four different experiments are conducted to perform AlMuezzin identification. In the first experiment, the proposed model trained using MFCC features got a 93% accuracy. In the second experiment, the proposed model trained using Logfbank features got a 96% accuracy. In the third experiment, the proposed model trained using spectral centroid features got a 94% accuracy. In the fourth experiment, the proposed model trained using Chroma features got a 96% accuracy. All accuracy results from the conducted experiments are listed in Table 3.

Table 3. Classification accuracy by VGG16 for AlMuezzin identification using different features

Feature	MFCC	Logfbank	Spectral centroid	Chroma
Accuracy	93%	96%	94%	96%



For Adan identification, Logfbank and Chroma performed better than the other criteria in terms of accuracy. The accuracy and loss for each model in relation to the epoch number (100) are shown in Figures 5-8 (each figure shows the number of epochs for training and validation (a) accuracy and (b) loss).

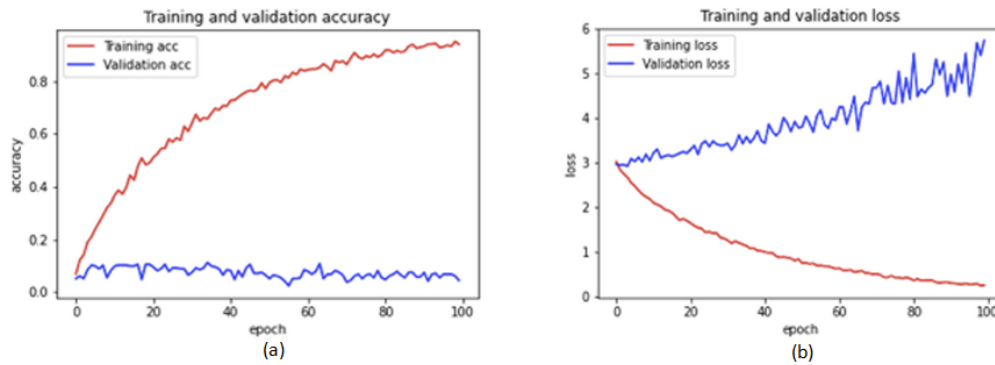


Figure 5. The number of epochs for (a) accuracy trends during training and validation using MFCC features and (b) loss function behavior showing the model's adaptation over epochs

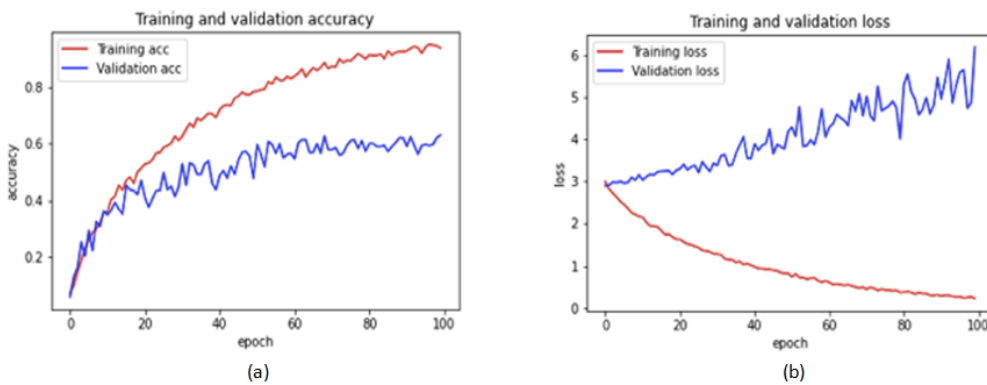


Figure 6. The number of epochs for (a) accuracy progression using Chroma features and (b) loss function graph highlighting convergence and overfitting tendencies

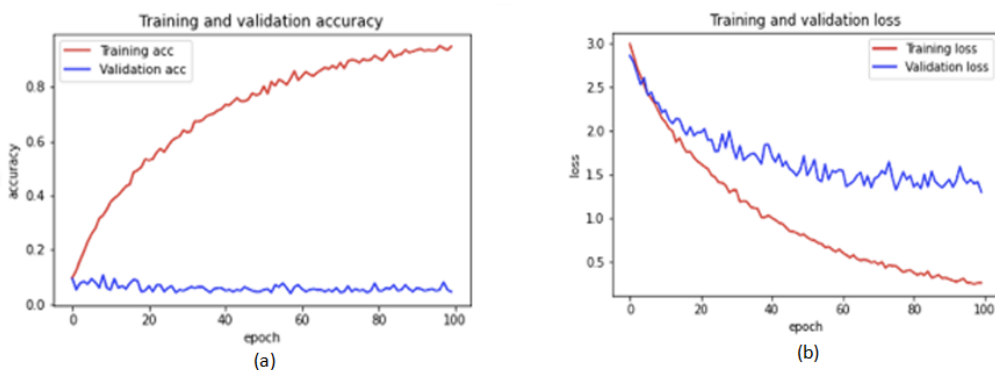


Figure 7. The number of epochs for (a) accuracy progression during training and validation using Logfbank features and (b) loss function behavior indicating model convergence

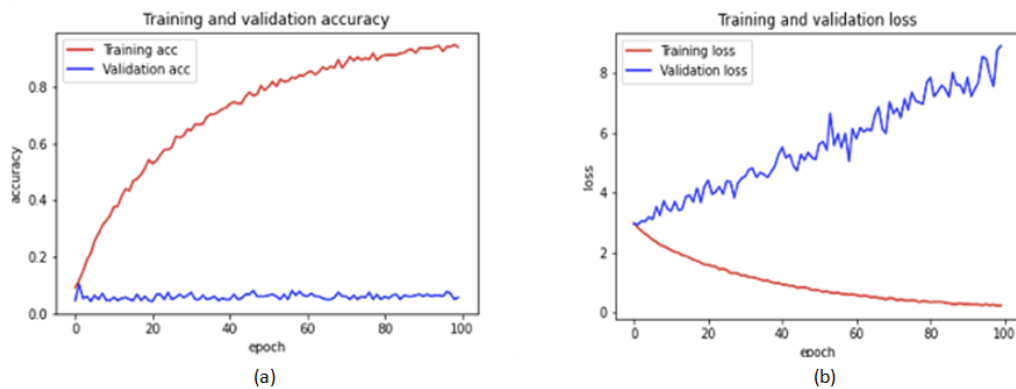


Figure 8. The number of epochs for (a) validation accuracy trends across epochs using spectral centroid features and (b) loss graph demonstrating overfitting tendencies in spectral centroid-based classification

Because validation data are a collection of fresh data points that the model is unfamiliar with, the validation accuracy is typically lower than the training accuracy. Training data is data that the model is already familiar with, where this is noticed in Figures 5-8. Therefore, it stands to reason that the accuracy is lower when using validation data than when using training data. But in Figure 6, noticed in the first epochs (approximately the first 15 epochs) the validation data's accuracy exceeds that of the training data, can interpret this by saying that the proposed model is a highly accurate predictor that takes into account a wide range of boundary situations. However, considering that some of the data points in the validation data present some challenge to the model, the model can be considered good if its accuracy (validation data) is approximately 80% of the training data.

When the accuracy of the validation data is higher than the accuracy of the rainfall data, this can be interpreted as a good indicator that the hyperparameters in the training data were properly adjusted, leading to a superior prediction in the validation.

Found that the validation loss is significantly higher than the training loss, as shown in Figures 5-8. And this is because of the overfitting of this model, in this instance, the validation loss is significantly higher than the training loss. While the validation loss is not continuously lowering, the training loss is. This indicates that the complexity of presented model is sufficient for it to "memories" the patterns found in the training set. In these kinds of cases, the proposed model needs to be regularized, and that is what are attempting to do in the upcoming work.

Also, an experiment to identify Al-Maqam, where we got a 95% as training accuracy, and 74% as validation accuracy, as shown in Figure 9 in relation to the epochs number (60).

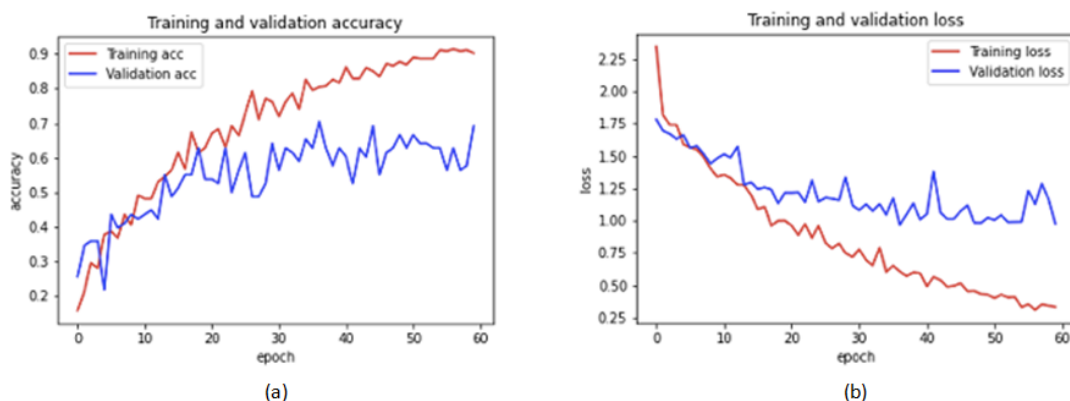


Figure 9. The number of epochs for (a) accuracy trends for Al-Maqam classification using spectral centroid and (b) corresponding loss function analysis highlighting performance variation across training phase using spectral centroid

In summary, this work clearly demonstrated the ability of deep learning techniques to differentiate AlMuezzin voices and their related Maqam. Using the VGG-16 model, high classification accuracy was obtained; spectral centroid features for Maqam classification (95%) and chroma features for AlMuezzin recognition (96%), produced the best results. These findings show that deep learning is promising in the recognition of Arabic speech in a musical and religious context.

The results of this research are comparable with prior studies in deep learning and speech recognition and support the application of feature extracting methods to raise classification accuracy. While MFCC and spectrum analysis methods have long been employed in speech-related activities, this work is the first to apply these methods to AlMuezzin and Maqam categorization recognition. Especially for complex tone fluctuations prevalent in Arabic Maqams, deep learning models such as VGG-16 offer a more dependable and scalable solution than traditional ASR algorithms depending on HMM and gaussian models.

The research ends by demonstrating that in religious Arabic environments, deep learning approaches could significantly raise the accuracy of speech recognition. By precisely identifying AlMuezzin voices and Maqams, historical audio preservation, automated recitation analysis, and adaptive prayer call systems have new chances. These advances mark a turning point in the way artificial intelligence might be combined with preservation of religious and cultural legacy.

## 5. CONCLUSION

In this work, AlMuezzin (المؤذن) voice identification, and determine the form of the Maqam which is a complex tonal system used in traditional Arabic music through VGG-16 model presented. The VGG-16 model examined using dataset collected manually and carefully from YouTube, with 4 extracted features Chroma feature, LogFbank feature, MFCC feature, and spectral centroids.

As a conclusion that MFCC, Logfbank, and Spectral Centroid are not suitable for solving problem for AlMuezzin identification since saw overfitting because of the validation loss being much larger than the training loss. Furthermore, determine that, among the four suggested features, the Chroma feature offers the best validation and training accuracy with accuracy reach 96%. For Aladhan Maqam Classification found that spectral centroid feature is a valuable tool that offers good training and validation accuracy for this kind of challenge with accuracy reach 95%. The results of the proposed work demonstrate the applicability for identification of different Arabic sounds.

Future studies should look at enhancing the efficacy of the model by leveraging transformer-based architectures or attention processes so increasing generality across many reciters and environmental situations. Expanding the dataset to include more varied audio samples from other geographical places could also help to improve model robustness. Moreover, adding real-time identification technologies into IoT or mobile apps could help Islamic researchers, academics, and audio restoration initiatives to realize their ideas.

## FUNDING INFORMATION

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nahlah Shatnawi	✓	✓	✓		✓	✓		✓	✓	✓		✓	✓	
Khalid M.O. Nahar	✓		✓		✓	✓	✓		✓	✓	✓		✓	
Suhad Al-Issa		✓	✓	✓		✓	✓		✓	✓	✓			
AEnas Alikhashashneh	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		

C	: Conceptualization	I	: Investigation	Vi	: Visualization
M	: Methodology	R	: Resources	Su	: Supervision
So	: Software	D	: Data Curation	P	: Project administration
Va	: Validation	O	: Writing - Original Draft	Fu	: Funding acquisition
Fo	: Formal analysis	E	: Writing - Review & Editing		

## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

The supporting data for this study are openly available at [https://yarmouk university2014-my.sharepoint.com/:f:/r/personal/khalids\\_yu\\_edu\\_jo/Documents/Moazzine% 20Dataset?csf=1& web=1& e=dPOzI1](https://yarmouk university2014-my.sharepoint.com/:f:/r/personal/khalids_yu_edu_jo/Documents/Moazzine%20Dataset?csf=1&web=1&e=dPOzI1), reference number [27].




## REFERENCES

- [1] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 2, pp. 91–101, Mar. 2018, doi: 10.1049/IET-BMT.2017.0065.
- [2] H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in speech recognition system," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan. 2016, pp. 498–502, doi: 10.1109/CONFLU-ENCE.2016.7508171.
- [3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in Speech Recognition*, Elsevier, 1990, pp. 65–74.
- [4] C. Hao, M. Xin, and Y. Xu, "A study of speech feature extraction based on manifold learning," *Journal of Physics: Conference Series*, vol. 1187, no. 5, p. 052021, Apr. 2019, doi: 10.1088/1742-6596/1187/5/052021.
- [5] I. Guyon and A. Elisseeff, "An introduction to feature extraction," in *Feature Extraction*, vol. 207, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–25.
- [6] A. Vellido, J. D. Mart'ín-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable," in *ESANN 2012 Proceedings, 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2012, pp. 163–172.
- [7] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, p. 14, 2015.
- [8] M. Picheny textslet al., "Trends and advances in speech recognition," *IBM Journal of Research and Development*, vol. 55, no. 5, pp. 447–464, Sep. 2011, doi: 10.1147/JRD.2011.2163277.
- [9] A. Hussein, S. Watanabe, and A. Ali, "Arabic speech recognition by end-to-end, modular systems and human," *Computer Speech & Language*, vol. 71, p. 101272, Jan. 2022, doi: 10.1016/j.csl.2021.101272.
- [10] H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, and Z. T. Fayed, "Arabic speech recognition using end-to-end deep learning," *IET Signal Processing*, vol. 15, no. 8, pp. 521–534, Oct. 2021, doi: 10.1049/sil2.12057.
- [11] E. Abdelmaksoud, "Arabic automatic speech recognition based on emotion detection," *The Egyptian Journal of Language Engineering*, vol. 8, no. 1, pp. 17–26, Apr. 2021, doi: 10.21608/ejle.2020.49690.1016.
- [12] M. S. Fahad, A. Deepak, G. Pradhan, and J. Yadav, "DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features," *Circuits, Systems, and Signal Processing*, vol. 40, no. 1, pp. 466–489, 2021, doi: 10.1007/s00034-020-01486-8.
- [13] E. Abdelmaksoud, A. Hassen, N. Hassan, and M. Hesham, "Convolutional neural network for arabic speech recognition," *The Egyptian Journal of Language Engineering*, vol. 8, no. 1, pp. 27–38, Apr. 2021, doi: 10.21608/ejle.2020.47685.1015.
- [14] K. K. Nawas, M. K. Barik, and A. N. Khan, "Speaker recognition using random forest," *ITM Web of Conferences*, vol. 37, p. 01022, 2021, doi: 10.1051/itmconf/20213701022.
- [15] Y. F. Utomo, E. C. Djamal, F. Nugraha, and F. Renaldi, "Spoken word and speaker recognition using MFCC and multiple recurrent neural networks," in *2020 7th International Conference on Electrical Engineering, Computer Sciences and Informatics (EECSI)*, Oct. 2020, vol. 2020-Octob, pp. 192–197, doi: 10.23919/EECSI50503.2020.9251870.
- [16] K. J. Devi, N. H. Singh, and K. Thongam, "Automatic speaker recognition from speech signals using self organizing feature map and hybrid neural network," *Microprocessors and Microsystems*, vol. 79, p. 103264, Nov. 2020, doi: 10.1016/j.micpro.2020.103264.
- [17] S. A. El-Moneim, M. A. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. Abd El-Samie, "Text-independent speaker recognition using LSTM-RNN and speech enhancement," *Multimedia Tools and Applications*, vol. 79, no. 33–34, pp. 24013–24028, Sep. 2020, doi: 10.1007/s11042-019-08293-7.
- [18] A. Ashar, M. S. Bhatti, and U. Mushtaq, "Speaker identification using a hybrid CNN-MFCC approach," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, Mar. 2020, pp. 1–4, doi: 10.1109/ICETST49965.2020.9080730.
- [19] P. N. Chowdary, K. Joshi, L. Kranthikiran, Y. Deepthi, K. Radhika, and A. Harika, "Speaker identification using GMM with MFCC in python," *Journal of critical reviews*, vol. 7, no. 14, pp. 586–590, Jul. 2020, doi: 10.31838/jcr.07.14.103.
- [20] M. T. S. Al-Kaltakchi, H. A. A. Taha, M. A. Shehab, and M. A. M. Abdullah, "Comparison of feature extraction and normalization methods for speaker recognition using grid-audiovisual database," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 18, no. 2, pp. 782–789, 2020, doi: 10.11591/ijeecs.v18.i2.pp782-789.
- [21] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using fusion of features with feedforward artificial neural network and support vector machine," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, Jun. 2020, pp. 170–176, doi: 10.1109/ICIEM48762.2020.9160269.
- [22] B. Alkhatib and M. M. W. K. Eddin, "Voice identification using MFCC and vector quantization," *Baghdad Science Journal*, vol. 17, no. 3, pp. 1019–1028, Sep. 2020, doi: 10.21123/bsj.2020.17.3(Suppl.).1019.
- [23] M. Kamala, G. Bhanusree, P. Thakkar, and R. Patel, "Offline voice recognition with low cost implementation based intelligent home automation system," *Journal of Engineering Sciences*, vol. 11, no. 11, 2020, [Online]. Available: [www.jespublication.com](http://www.jespublication.com).
- [24] A. Revathi, C. Ravichandran, P. Saisiddarth, and G. S. R. Prasad, "Isolated command recognition using MFCC and clustering algorithm," *SN Computer Science*, vol. 1, no. 2, p. 82, Mar. 2020, doi: 10.1007/s42979-020-0093-x.




- [25] S. Al-Issa, M. Al-Ayyoub, O. Al-Khaleel, and N. Elmitwally, "Building a neural speech recognizer for quranic recitations," *International Journal of Speech Technology*, vol. 26, no. 4, pp. 1131–1151, 2023, doi: 10.1007/s10772-022-09988-3.
- [26] S. R. Shareef and Y. F. Al-Irhayim, "Towards developing impairments arabic speech dataset using deep learning," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 25, no. 3, p. 1400, Mar. 2022, doi: 10.11591/ijeecs.v25.i3.pp1400-1405.
- [27] N. Shatnawi, K. Nahar, M. O., S. Al-Issa, and E. Alikhashashneh, "Moazzine dataset," *Youtube*, 2014.

## BIOGRAPHIES OF AUTHORS






**Nahlah Mohammad Shatnawi**    is an assistant professor in the Department of Computer Sciences- College of Information Technology and Computer Sciences, Yarmouk University, Jordan. She received a B.Sc. and M.Sc. degrees in computer science from the Jordan University of science and Technology, Jordan, and a Ph.D. degree in Computer Science from UKM, Malaysia. Her research areas are pattern recognition, optimization, machine learning, and deep learning. She can be contacted at email: [nahlah.s@yu.edu.jo](mailto:nahlah.s@yu.edu.jo).






**Khalid M. O. Nahar**    is a full professor in the Department of Computer Sciences-Faculty of IT, Yarmouk University, Irbid-Jordan. He received his BS and MS degrees in computer sciences from Yarmouk University in Jordan, in 1992 and 2005 respectively. He was awarded a full scholarship to continue his PhD in Computer Science and Engineering from King Fahd University of Petroleum and Minerals (KFUPM), KSA. In 2013 he completed his PhD and started his job as an assistant professor at Tabuk University, KSA for 2 years. In 2015 he backs to Yarmouk University as an assistant professor, he was awarded the upgrade promotion degree as an associate professor in 2020 then to full professor in 2024. Prof. Khalid was the assistant dean for quality control in the Faculty of IT-Yarmouk University for the year 2019. He was assigned the chairman of the training and development department in the accreditation and quality center at Yarmouk University from the year 2020 to the year 2022. Prof. Khalid's research interests include but are not limited to, Continuous Speech Recognition, Arabic Computing, Natural Language Processing, Multimedia Computing, Content-Based Retrieval, Artificial Intelligence, Machine Learning, Pattern Recognition, IoT, and Data Science. He can be contacted at email: [khalids@yu.edu.jo](mailto:khalids@yu.edu.jo).



**Suhad Al-Issa**    has a B.Sc. in computer engineering from Yarmouk University. She has earned an M.Sc. in computer engineering from JUST University. Also, she was a research assistant at the CEIP center at JUST University. Her role involves continuing the latest work done on the master's thesis. Also, she was an lecturer at many universities in Jordan. Currently, she is a Ph.D. student in Robotics at Queen's University in the UK. Her interest is in working on topics related to artificial intelligence, deep learning, machine learning, and NLP. She can be contacted at email: [saleassa01@qub.ac.uk](mailto:saleassa01@qub.ac.uk).



**Enas Ahmad Alikhashashneh**    is an Assistant Professor at the College of Information Technology and Computer Science, Yarmouk University, Jordan. She received a B.Sc. degree in computer science from the AlBalqa Applied University, Jordan, an M.Sc. degree in Computer and Information Systems from the Yarmouk University, Jordan, another M.Sc. degree in Science from the Purdue University/ IUPUI, USA, and a Ph.D. degree in Computer Science from Purdue University/ IUPUI, USA. Her research areas are Machine Learning, Deep learning, NLP, and pattern recognition. She is Assistant Dean of the College of Information Technology and Computer Science at Yarmouk University from 2023 to now. She can be contacted at email: [enas.a@yu.edu.jo](mailto:enas.a@yu.edu.jo).