Hindawi
Advances in Multimedia
Volume 2023, Article ID 2642558, 14 pages
https://doi.org/10.1155/2023/2642558

Research Article

The Recognition of Holy Qur'an Reciters Using the MFCCs' Technique and Deep Learning

Ghassan Samara,[1] **Essam Al-Daoud**,[1] **Nael Swerki**,[1] **and Dalia Alzu'bi**[2]

[1]*Department of Computer Science, Zarqa University, Zarqa, Jordan*
[2]*Department of Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan*

Correspondence should be addressed to Ghassan Samara; gsamara@zu.edu.jo

Received 22 November 2022; Revised 16 January 2023; Accepted 1 March 2023; Published 21 March 2023

Academic Editor: Zhongxu Hu

The Holy Qur'an has recently gained recognition in the field of speech-processing research. It is the central book of Islam, from which Muslims derive their religious teachings. The Qur'an is the primary source and highest authority for all Islamic beliefs and legislation. It is also one of the most widely memorized and recited texts around the world. Listening to and reciting the Qur'an is one of the most important daily practices for Muslims. In this study, we propose a deep learning model using convolutional neural networks (CNNs) and a dataset consisting of seven well-known reciters. We utilize mel frequency cepstral coefficients (MFCCs) to extract and evaluate information from audio sources. We compare our proposed model to different deep learning and machine learning methodologies. Our proposed model outperformed the competing models with an accuracy of 99.66%, compared to the support vector machine's accuracy of 99%.

1. Introduction

The Qur'an is the book that was revealed to Prophet Muhammad. There are nearly two billion Muslims worldwide, who represent approximately 25% of the global population [1]. The Qur'an is the primary text for Muslims [2]. It is divided into thirty chapters and contains 6,236 verses organized into 114 groups called "Surahs." The book was revealed in Arabic, one of the world's most extensive and challenging languages [3, 4]. Muslims recite the Qur'an daily as a religious act that brings them closer to their Lord. They follow the Ahkam Al-Tajweed rules for recitation, which include Al-Edgham, Al-Aqlab, and Endowment [5].

Although the reciters of the Qur'an recite the same verses, there are differences in recitation due to the characteristics of the text, the differences in reading from one reader to another, and the presence of dialects in Arab tribes, which resulted in multiple readings that reached ten [6, 7]. This adds a challenge to recognizing a specific reciter.

The melodies used to recite the Qur'an enhance its beauty. These tunes, known as "Maqams" in Arabic, are of eight types: Hijaz, Seka, Nahawand, Bayat, Rast, Ajam, Kurd, and Saba [8]. One of the many applications of audio classification is the recognition of Qur'an reciters.

Automatic audio classification is based on matching a sound wave to a set of its samples, typically basic linguistic units known as compounds of phonemes [9–11]. It includes a variety of applications, such as speaker age, gender, speaker identification, speech emotion, and audio archiving management [12]. Despite more research being conducted on images and video than on audio, the recent availability of audio databases has resulted in a resurgence of research in this field, which has become one of the most popular and significant topics being studied. Today, there are more audio datasets, and researchers are increasingly interested in managing them.

There are numerous studies on speech processing, but the majority are limited to English and a few other languages [13–16]. More research needs to be conducted on recognizing Arabic, particularly in reciting the Qur'an. The lack of standardized datasets for all famous reciters of the Qur'an is one of the primary reasons for the delay in recognizing the

recitation of the Qur'an compared to speech recognition [17–20]. Recognizing the reciters of the Qur'an is challenging because each reciter has distinctive phonetic characteristics, which vary over time in terms of letter and word pronunciation and recitation style [21–23]. Significant parallels exist between recognizing Qur'an reciters and speaker recognition [24–27]. The global spread of Islam has led to increased demand for audio recordings of several famous reciters of the Qur'an and an increase in demand for Qur'an applications. This book was published on the Internet due to its significance to Muslims. Although the publication of the Qur'an has made it more accessible to Muslims and non-Muslims worldwide, there were multiple attempts to alter and falsify the recitation of Qur'anic verses to harm Islam [10, 11, 19, 20, 24]. Therefore, there is a need for a system that distinguishes trustworthy reciters of the Qur'an from those who are not. The contribution of this paper is to determine the relationship between performance and data volume for deep learning and machine learning approaches, as well as the point at which deep learning outperforms machine learning on the dataset used in this study. The document is divided into five sections. Section 2 presents the available literature review on recognizing Qur'an reciters. Section 3 provides an overview of the proposed system and methodology. Section 4 describes the data and experiments conducted and thoroughly analyzes the results. Section 5 concludes the paper with observations on possible directions for future research.

2. Literature Review

In [28], the researcher presented a computer system that automatically applies some rules of recitation that must be considered when reading the Qur'an and identifies the correct use of the recitation rules throughout the entire Qur'an. This paper focuses on eight fundamental rules of intonation for reading. They employed some feature extraction and classification techniques, including mel-frequency cepstral coefficients (MFCCs), wavelet packet decomposition (WPD), and linear predictive coefficient (LPC). In addition, they used classification techniques, including SVM, random forest (RF), bagging, K-nearest neighbors (KNN), and artificial neural networks (ANN). Using bagging, they were able to achieve a maximum accuracy of 94 percent, but the achieved accuracy can be improved.

In [17], the researcher presented an improved computer system that contains an automatic method for applying certain rules of recitation that must be considered when reading the Qur'an and identifies the correct usage of the recitation rules throughout the entire Qur'an. This paper also focuses on eight fundamental rules of intonation for reading. They employed some feature extraction and classification techniques, including MFCCs, WPD, and LPC. In addition, they used classification techniques, including SVM, RF, bagging, KNN, and ANN. The highest degree of accuracy they achieved was 94%. They also utilized a convolutional deep belief network (CDBN), a deep learning method. The CDBN result exceeded previous results, reaching 97.7 percent, but the accuracy achieved can still be improved, and the rules used for recitation were strict and cannot be applied regularly.

In [29], the authors introduced a machine learning technique for identifying recitations of the Qur'an for ten notable reciters. They utilized MFCCs for audio processing and feature extraction. They extracted twenty features and incorporated them into the proposed system. In addition, they utilized KNN and ANN classification methods. Using KNN, the highest accuracy obtained was 97.03. The ANN achieves a maximum accuracy of 97.6%, but the accuracy achieved can still be improved.

In [30], the researchers presented a method for identifying the Qur'an reciter's voice. They utilized a dataset consisting of five reciters and multiple Surahs of the Qur'an. As a method for audio processing and feature extraction from the dataset, they selected the MFCCs' method and used the Gaussian mixture model (GMM) as a classification technique. They claimed to achieve a 100 percent accuracy rate during the training and testing phases. When tested, the system also claimed the ability to reject unknown samples.

The authors of [31] proposed a system that employs deep learning to identify Qur'an reciters. The dataset utilized by the proposed model consists of five reciters. They utilized MFCCs to extract thirteen features from the audio file. They used recurrent neural networks (RNNs) as a classification technique with bidirectional long short-term memory (BLSTM). The BLSTM is suitable for identifying Qur'an reciters and yields accurate results. The proposed system demonstrated greater accuracy than conventional ANNs and is computationally inexpensive. The proposed system achieved 99.89 percent accuracy.

In [18], the authors presented a machine learning-based system for recognizing Qur'an reciters. Twelve reciters recited ten surahs of the Qur'an using the dataset. They utilized the frequency domain and the spectrogram forms of audio representation. The model's features were extracted in the frequency domain using MFCCs and pitch. Autocorrelograms were used to extract system features from the spectrogram. Their system utilized a conventional machine learning model. They employed naive Bayes (NB), random forest (RF), and Java 48 (J48). With the use of MFCCs and pitch features, the classifiers performed well. With NB and RF, they achieved an accuracy of 88 percent, and with J48, the accuracy reached 78 percent. However, the accuracy achieved is low compared with other systems deployed.

In [32], the authors proposed a system for identifying the Qira'ah of the Qur'an. Qira'ah is a form of recitation of the Qur'an. There are ten distinct Qira'ah of the Qur'an. Some of the Qur'an reciters who read the Qur'an in numerous Qira'ah were included in the dataset. Utilizing MFCCs, features were extracted from the audio file. Several machine learning models were adopted for the proposed system, and a comparison between these models was conducted. The support vector machine (SVM) result was superior to those of the other models. SVM achieved a 96 percent accuracy, whereas ANN and RF achieved 62 and 68 percent accuracy, respectively. However, the accuracy achieved is low compared with other systems deployed.

In [33], the authors presented an automatic model for recognizing TajweedShould we change "Tajweed" to "Tajwid" here and other instances. Please confirm, the Qur'an recitation rules. They focused on four rules for reciting the Qur'an. The utilized dataset includes 657 audio files for the four rules. The model used filter banks for audio processing and feature extraction. They utilized seventy filter banks, and the audio signal passed through them to extract features. They employed SVM as classifiers for machine learning. They achieved a 99 percent accuracy in precision.

The authors of [34] present a system for recognizing the identity of Qur'an reciters. The authors selected two surahs for the dataset: Al-Baqraa and Al-Kahaf. The dataset contains 15 individuals who recite the Qur'an. The dataset contains 1,650 audio files. They utilized MFCCs for audio processing and extraction of features. They extracted twenty characteristics from each audio file and incorporated them into the system. As classification strategies, SVM and ANNs are employed. Experiments revealed that the results obtained by SVM were superior to those obtained by ANNs. SVM achieved a precision of 96.59 percent, while ANN achieved a precision of 86.17 percent. However, the accuracy achieved can be improved.

The authors of [35] present a comparative study for a supervised classification system to identify Qur'an reciters. They created a dataset containing 2,134 wave audio files for seven renowned reciters. A collection of perceptual features, including short-time energy features, tempo-based features, and pitch fractures, were chosen to achieve more favorable results. Wave audio files had twenty-one features extracted, including spectral entropy, frequency centroid, skewness, and spectral flatness. Several classification models, including eXtreme gradient boosting (XGBoost), logistic regression (LR), SVM, ensemble adaptive boosting (Ensemble AdaBoost), RF, SVM-linear, SVM-radial basis function (SVM-RBF), and decision tree (DT), were utilized in the comparative study. The XGBoost classifier provided the most accurate results, achieving an accuracy rate greater than 93%. However, the achieved accuracy can be improved.

The authors of [36] proposed a system for the Qur'an memorization test to verify the Qur'an recitation. Five reciters who recited the last ten surahs were used to compile the dataset. The data used for testing came from other reciters. MFCCs, mel-frequency spectral coefficient (MFSC), and delta were tested as feature extraction techniques for the wave audio file. They used two classifiers to evaluate the system: the Siamese and the Manhattan long short-term memory (MaLSTM). Both of these entities are neural networks. Siamese measures the symmetry between pairs of samples, while MaLSTM is used for sequential data. MaLSTM was the most accurate classifier with the highest precision. Utilizing MFCCs and delta, the $F1$ score was increased to 77.35 percent. However, the achieved accuracy can be improved.

In [8], the authors proposed a deep learning system for categorizing the melodies used to recite the Qur'an. They were limited to only eight melodies (Hijaz, Seka, Nahawand, Bayat, Rast, Ajam, Kurd, and Saba). The dataset included 874 wave audio files from two famous reciters who used all the melodies when they recited. The research utilized MFCCs, zero-crossing rate, chroma feature, root mean square (RMS) energy, spectral centroid, spectral bandwidth, and spectral roll-off. In addition, the research utilized various deep learning models, including CNN, LSTM, and deep ANN. After using 26 input features, the highest accuracy achieved was 95.7%. Five layers of deep ANN were utilized to achieve this level of precision. However, the accuracy achieved can still be improved.

In [7], the authors present a speech recognition system for categorizing the Qur'an recitations into ten different types, including Warsh and Hafss. The technique of MFCC was used to extract characteristics from the recitation of Quranic verses. Hidden Markov models (HMM) were used for training and recognition as a classification model.

In [37], the authors presented a method for estimating the speaker's age using three distinct datasets. The first dataset is TIMID, which contains recordings of 630 users. The second dataset is Switchboard-1, which contains 2,400 phone conversations. The third corpus is the CMU Kids corpus, which contains 5,180 phrases. Several characteristics were extracted and incorporated into the model, including cepstral coefficients, shimmer, centroid, spectral entropy, spectral decrease, jitter, F0DIFF, and flatness. The authors utilized three classifiers based on machine learning: SVM, GMM, and a combination of GMM and SVM. The classifiers with the highest accuracy were SVM at 79 percent, GMM at 73 percent, and GMM-SVM at 83 percent. Figure 1 depicts the CNN architecture.

3. Proposed System Stages

This section proposes a machine learning system for recognizing Qur'an reciters. The proposed system consists of four fundamental phases: data acquisition, data preprocessing, feature extraction, and classification. Figure 2 illustrates the overall system architecture, which shows the flow of data through each phase of the system.

3.1. Data Acquisitions. A dataset of seven famous Qur'an reciters was constructed. Mp3 files were obtained for each reciter and wave format conversion (sampling frequency: 22.05 kHz). Each reciter recited 80 minutes of Quran surahs on an audio file. Table 1 lists the names of Qur'an reciters.

3.2. Data Preprocessing

3.2.1. Sampling. The sound wave is first converted from analog to digital form through a process called sampling, which generates an array [38]. The sample rate is the number of samples obtained per second. In Figure 3 [39], the red line represents the sound wave, while the blue dots represent the samples. Reducing the distance between each point increases the number of samples taken, resulting in a higher sample rate. This means that the higher the sample rate, the closer

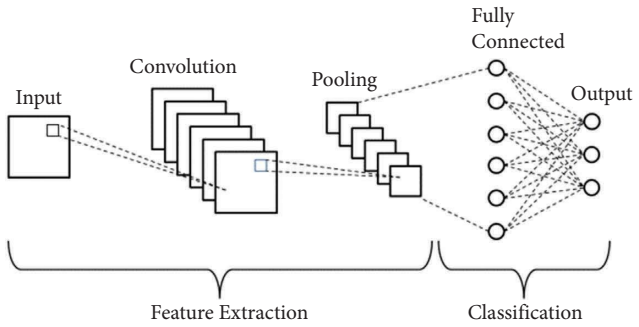

Figure 1: CNN architecture.

the reproduction is to the original sound wave. Before features can be extracted, the frequency is reduced from 22.05 kHz to 16 kHz.

3.2.2. Librosa Library. This library is frequently used for audio processing because it facilitates the following [40]:

(i) It converges the signal into a single carrier (mono) channel

(ii) represents an audio signal relative to a normalized pattern between −1 and 1 so that a consistent pattern is observed throughout the entire audio

(iii) It can also detect the sample rate, which converts to 22,050 Hz by default

3.3. Feature Extraction. The extraction of features is a crucial step in speech recognition systems. It involves computing sequences of distinct feature vectors to represent an audio signal [28]. The primary purpose of the feature extraction process is to identify the most important components of the audio signal and remove any unnecessary data. When dealing with large databases, it is necessary to employ techniques for feature extraction. This technique reduces the size of the data while preserving all essential information. It aids in accelerating the machine learning system's learning rate and delivering accurate results with less time, effort, and resource consumption [41].

3.3.1. Mel-Frequency Cepstral Coefficients (MFCCs). The audio processing technique used for feature extraction will be MFCCs. This technique is the most widely used in speech recognition systems and has achieved spectacular results [42]. The frequency will be resampled from 22.05 kHz to 16 kHz. The audio signal will be divided into segments of varying lengths (2, 3, or 4 seconds) with a frame size of 2,048 and a hop length of 512. Sixty features will then be extracted from each segment. Figure 4 shows an example of an audio signal, while Figure 5 displays thirteen MFFC features for the audio signal shown in Figure 4. The characteristics of MFCCs will be extracted using the Librosa library.

The MFCC technique for extracting features consists of three main phases [30, 43]. Figure 6 demonstrates

(i) Short-time processing: the audio signal is divided into overlapping frames, and a Hamming window is applied to obtain the magnitude of the discrete Fourier transform (DFT). The signal's sound is transformed from the time domain to the frequency domain because it is easier to analyze in the frequency domain than in the time domain.

(ii) Mel-frequency wrapping and filter bank [44]: each tone has a specific pitch with a frequency weighted on a scale known as the Mel scale. Using equation (1), we can compute an estimated mel for each frequency.

$$\mathrm{mel}(f) = 117 \ln\left(1 + \frac{f_{\mathrm{Hz}}}{700}\right), \tag{1}$$

where *fmel* is the frequency in mels and fHZ is the frequency in Hz.

After that, a coefficient output is obtained by applying equation (2) to a single triangular filter bank of M for each mel frequency.

$$H_m[k] = \begin{cases} 0, & k < f[m-1], \\ \dfrac{k - f[m-1]}{f[m] - f[m-1]}, & f[m-1] < k \le f[m], \\ \dfrac{f[m+1] - k}{f[m+1] - f[m]}, & f[m] \le k < f[m+1], \\ 0, & k > f[m+1], \end{cases} \tag{2}$$

where $m = 0, 1, \ldots, M-1$.

In the final step of this phase, the log-energy mel spectrum is computed by applying

$$[m] = \ln\left[\sum_{k=0}^{N-1} |X[k]|^2 H_{m[k]}\right], \tag{3}$$

where $X[k]$ is the DFT of the input audio $x[n]$.

(iii) Discrete cosine transform (DCT): using the DCT, we convert the log mel spectrum back to the time domain to obtain the final output, which we refer to as the MFCCs. Equation (4) demonstrates

$$\hat{x}[n] = \sum_{m=0}^{M-1} S[m] \cos\left[\left(m + \frac{1}{2}\right)\frac{\pi n}{M}\right]. \tag{4}$$

3.4. Classification. In the proposed system, multiple classifiers can be used. We give priority to classifiers that have achieved a high level of speech recognition accuracy. Several factors are taken into consideration when selecting the appropriate classifier. First, the size of the dataset, followed by computation cost, and finally, usability [45]. According to [46], classifiers can be categorized as either shallow learning

Figure 2: Proposed system stages.

Table 1: The Qur'an reciter name.

#	Reciter name in English	Reciter name in Arabic
1	Mishari Al-Afasi	مشاري العفاسي
2	Ahmad Al-Ajmy	احمد العجمي
3	Sa'ad Al-Ghamidi	سعد الغامدي
4	Khaled Al-Jaleel	خالد الجليل
5	Maher Al-Muaiqly	ماهر المعيقلي
6	So'oud Al-Shuraim	سعود الشريم
7	Abed Al-Rahman Al-Sudaes	عبد الرحمن السديس

models or deep learning models. In the proposed system, we will use classifiers from both categories. In conclusion, the performance of the classifiers will be compared.

4. The Experimental Results and Analysis

This section presents the experimental results of the proposed CNN model and other machine learning models. Several performance metrics, including accuracy, precision, recall, *F*-measure, and root mean square error (RMSE), were used to evaluate the models.

For reciter recognition, the proposed CNN model and various machine learning models were evaluated using a dataset that contained 60 features of varying segment lengths. The models were trained on 80% of the dataset and tested on the remaining 20%. The dataset is presented in Table 2.

4.1. Full Dataset with Different Segment Lengths and Different Feature Numbers Using the Proposed Model (CNN). The proposed model for reciter recognition uses different segment lengths and feature counts. The objective of evaluating the model is to determine the optimal segment length and the optimal number of features that produce the highest level of accuracy.

4.1.1. Two Seconds Segment Length. Table 3 presents the results of the proposed model for reciter recognition based on different feature counts. The results are evaluated using several performance metrics, including accuracy, *F*1-measure, recall, precision, and root mean square error (RMSE).

Based on the results, it can be observed that as the number of features increases, the accuracy and other performance metrics also increase, except for RMSE, which decreases. The proposed system achieves its highest level of accuracy, 0.989143, and its lowest level of RMSE, 0.406086, when sixty features are utilized.

It is also interesting to note that the proposed system achieves a high level of accuracy (0.988262) with only 30 features while using 20 feature results in the lowest accuracy (0.984741) and the highest RMSE (0.504091). This indicates that increasing the number of features can significantly improve the model's performance, but the improvement is not necessarily linear.

In conclusion, the optimal number of features for maximizing accuracy and minimizing RMSE is sixty, but using 30 or more features can also result in a high level of accuracy.

It has been observed that the accuracy increases as the number of features increases. The proposed system achieves its highest level of accuracy, 0.989143, and its lowest level of RMSE, 0.406086, when there are sixty features. When 20 features were utilized, the proposed system achieved the lowest accuracy of 0.984741 and the highest RMSE of 0.504091. With a segment length of 2 seconds, the optimal number of features for maximizing accuracy and minimizing RMSE is sixty.

4.1.2. Three Seconds Segment Length. The results show the performance of the proposed model on a 3-second segment length dataset. Table 4 compares the results for different numbers of features used in the model, including accuracy, *F*1-measure, recall, precision, and RMSE.

For 20 features, the accuracy achieved is 0.987676, and the model has an *F*1-measure of 0.987639, recall of 0.987676, precision of 0.987661, and an RMSE of 0.470053. When 30 features are used, the accuracy and other performance metrics are very similar to the 20-feature model.

The model with 40 features achieves the highest accuracy of 0.992957, with an *F*1-measure of 0.992972, recall of 0.992957, precision of 0.99301, and an RMSE of 0.275144. The accuracy decreases slightly with 50 and 60 features, but the other performance metrics are still relatively high, with an RMSE of 0.402003 for 50 features and 0.401362 for 60 features.

Overall, the results show that the proposed model performs well on the 3-second segment length dataset, achieving high accuracy and other performance metrics, especially with 40 features.

4.1.3. Four Seconds Segment Length. Table 5 presents the testing results for the proposed model using a dataset with 4-second segment lengths and varying numbers of features. The performance of the model is evaluated based on several metrics, including accuracy, *F*1-measure, recall, precision, and root mean square error (RMSE).

The results show that the proposed model achieved high levels of accuracy for all feature numbers, ranging from 0.993279 to 0.995406. The highest accuracy is achieved with 30 features, with an accuracy of 0.995406, followed closely by 60 features with an accuracy of 0.994601.

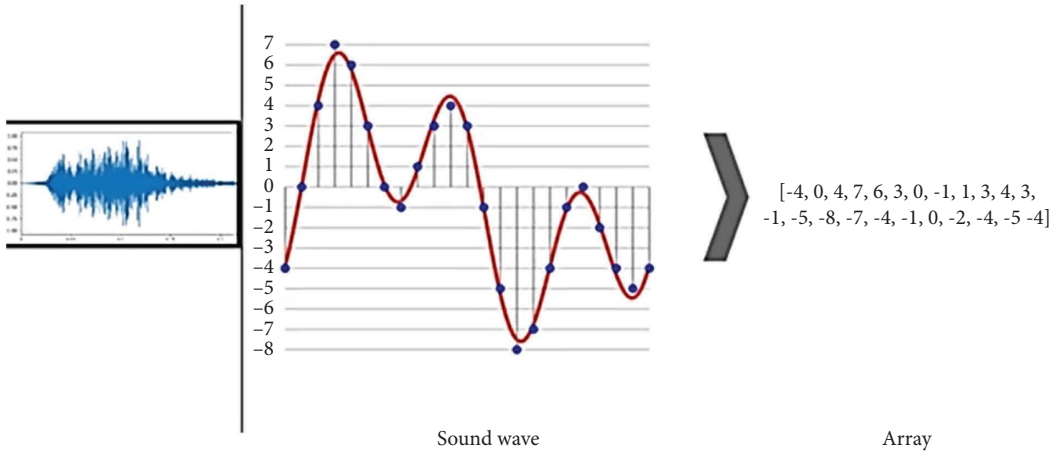

Figure 3: Sampling.

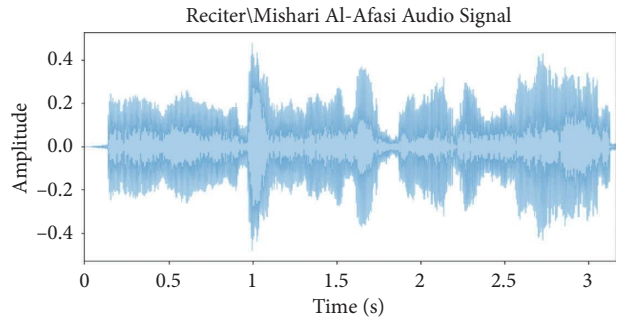

Figure 4: Sample of audio signal.

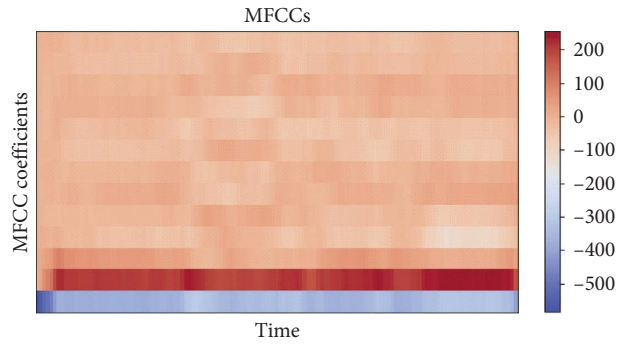

Figure 5: Sample of MFCC file.

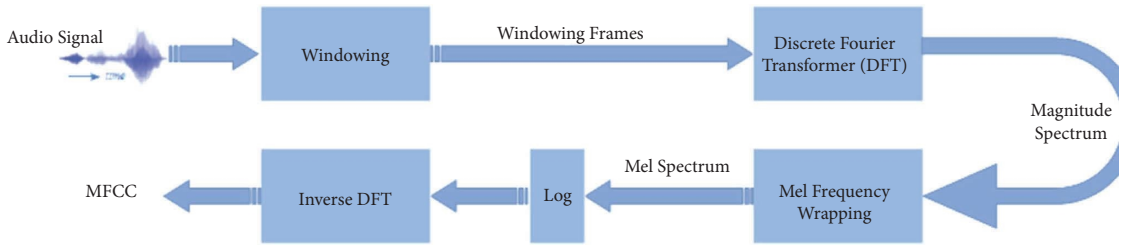

Figure 6: MFCC technique phases.

Table 2: Datasets.

Datasets	Segment length (seconds)	Features' number	Number of records
Dataset 1	2	60	17037
Dataset 2	3	60	11358
Dataset 3	4	60	8517

Table 3: Proposed model testing results for 2 sec segment length dataset.

Features' number	Accuracy	*F*1-measure	Recall	Precision	RMSE
20	0.984741	0.984692	0.984741	0.984699	0.504091
30	0.988262	0.988229	0.988262	0.988323	0.434703
40	0.986208	0.986201	0.986208	0.986224	0.459319
50	0.987676	0.987678	0.987676	0.987731	0.459319
60	0.989143	0.989128	0.989143	0.989187	0.406086

Table 4: Proposed model testing results for 3 sec segment length dataset.

Features' number	Accuracy	*F*1-measure	Recall	Precision	RMSE
20	0.987676	0.987639	0.987676	0.987661	0.470053
30	0.987676	0.987652	0.987676	0.987684	0.423765
40	0.992957	0.992972	0.992957	0.99301	0.275144
50	0.989836	0.989834	0.989836	0.989896	0.402003
60	0.989436	0.989448	0.989436	0.989499	0.401362

Table 5: Proposed model testing results for 4 sec segment length dataset.

Features' number	Accuracy	*F*1-measure	Recall	Precision	RMSE
20	0.993279	0.99331	0.993279	0.993457	0.334797
30	0.995406	0.995403	0.995406	0.995431	0.287566
40	0.993794	0.993802	0.993794	0.993844	0.312118
50	0.994305	0.994312	0.994305	0.994324	0.305915
60	0.994601	0.994605	0.994601	0.994621	0.280163

All the feature numbers also show high levels of *F*1-measure, recall, and precision, indicating that the model is effective in correctly identifying the reciter. The RMSE values are relatively low, ranging from 0.280163 to 0.334797, indicating that the model's predictions are close to the actual values.

Overall, the results suggest that the proposed model is effective in recognizing the reciter using a dataset with 4-second segment lengths and varying numbers of features. The model achieved high levels of accuracy, and the optimal number of features is 30, as it achieved the highest accuracy among all the feature numbers.

Based on the analysis of the results in Tables 3–5, it can be concluded that the combination of 30 features and a 4-second segment length yields the highest accuracy rate and the lowest error rate.

4.2. Feature Importance of Using Machine Learning. An ensemble learning model, ExtraTreesClassifier, has been used to determine the significance of the features. Three datasets with 60 features and varying segment lengths were utilized in the experiments.

4.2.1. Dataset Based on 60 Features and 2 Seconds Segment Length. In this experiment, 60 features with 2-second segment lengths were used. Figure 7 demonstrates that characteristics between 30 and 59 are less significant than the initial 30 characteristics. However, comparing this result to Table 3 reveals that using sixty features yielded the highest accuracy.

4.2.2. Dataset Based on 60 Features and 3 Seconds Segment Length. This experiment utilized 60 features with a segment length of 3 seconds. Figure 8 demonstrates that features between 30 and 59 are less significant than the first 30 features. Comparing this result with the result from Table 4 reveals that the highest accuracy was achieved by using the first forty features.

4.2.3. Dataset Based on 60 Features and 4 Seconds Segment Length. In this experiment, 60 features with a segment length of 4 seconds were utilized. Figure 9 demonstrates that the features between 30 and 59 are less significant than the first 30 features. Comparing this result to the one from Table 5 reveals that the highest accuracy was achieved by utilizing the first 30 features.

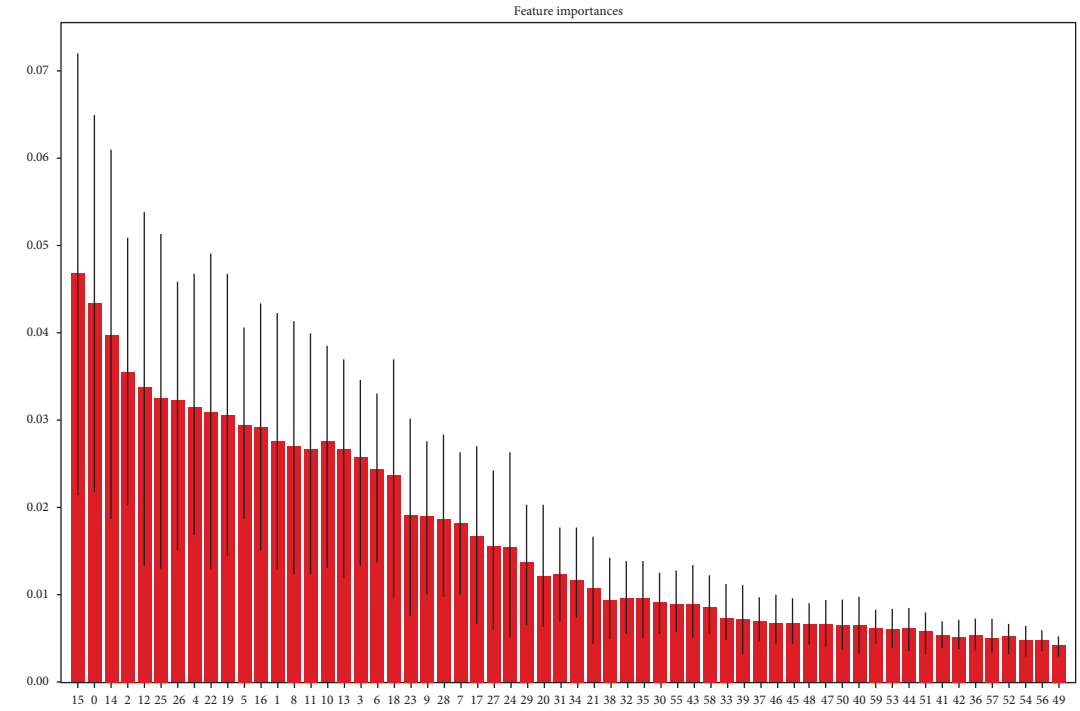

Figure 7: Feature importance based on 2 sec segment length.

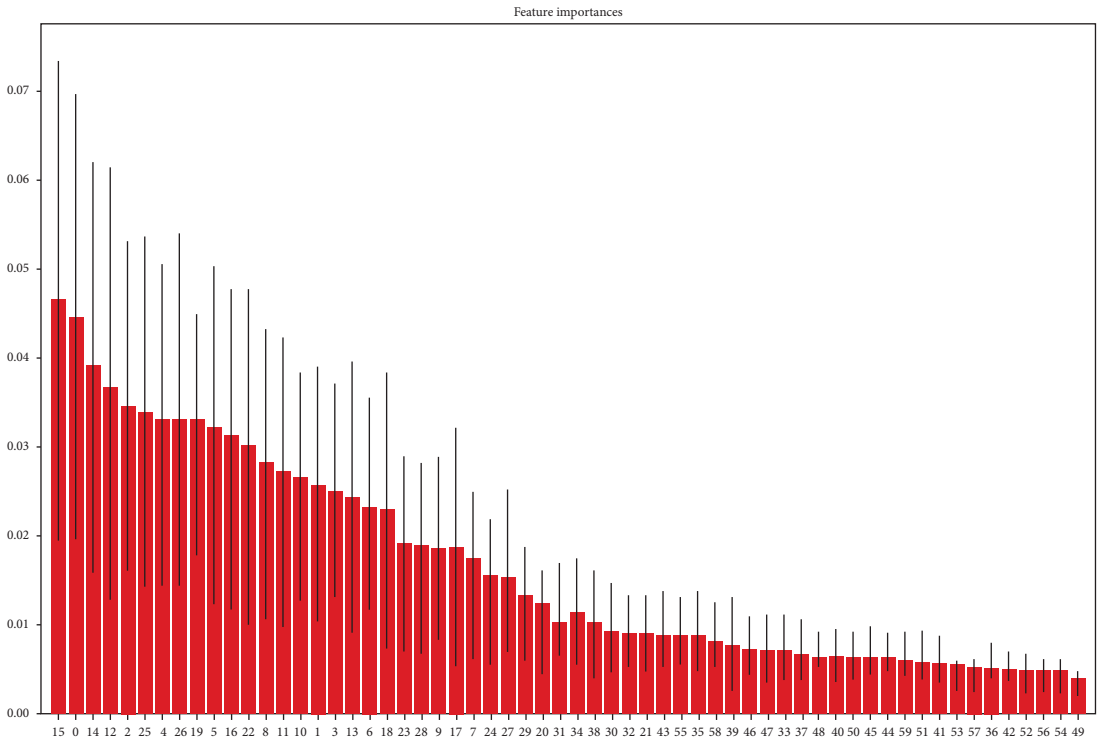

Figure 8: Feature importance based on 3 sec segment length.

According to previous results, consensus exists regarding the importance of the first 30 features, whereas the importance of the remaining 30 features is less clear.

4.3. Full Dataset with 30 Features and Different Segment Lengths. Three datasets with 30 features and varying segment lengths were utilized in the experiments.

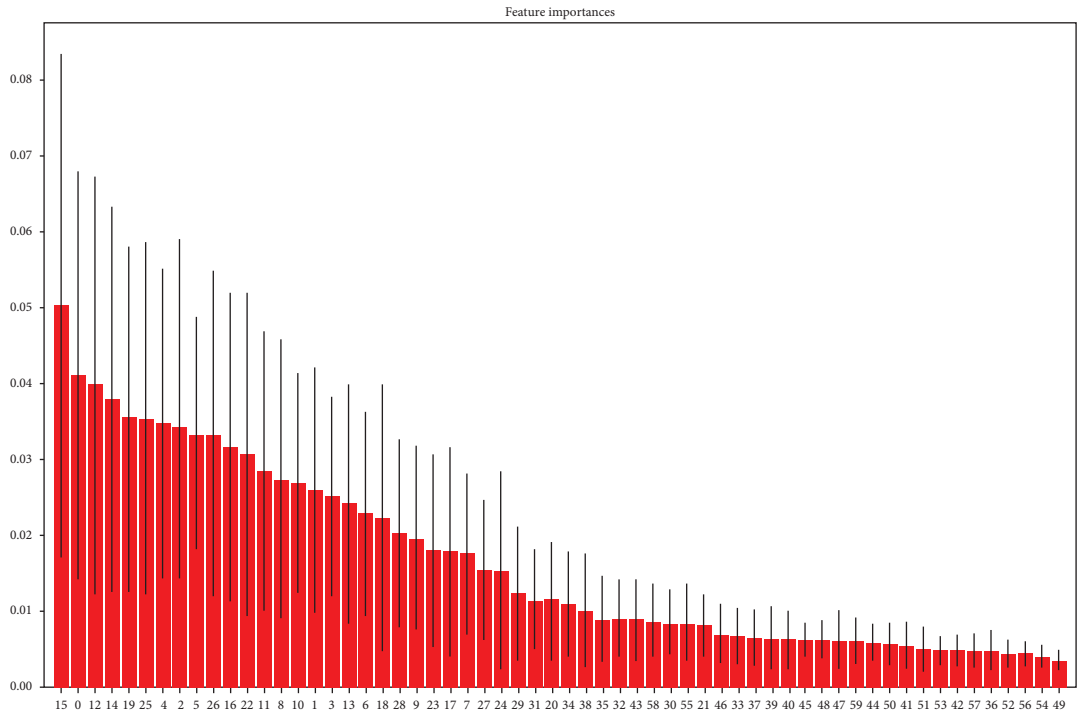

Figure 9: Feature importance based on 4 sec segment length.

4.3.1. Dataset Based on 30 Features and 2 Seconds Segment Length. Table 6 and Figure 10 show the testing results for various classification models using a dataset with a 2-second segment length and 60 features. The proposed model achieved the highest accuracy rate of 0.988262, as well as the highest *F*1-measure, recall, and precision scores. The root mean squared error (RMSE) for the proposed model is 0.434703, which is the lowest among all the models. The support vector machine (SVM) model achieved the second-highest accuracy rate of 0.980861, followed by the K-nearest neighbor (KNN) model with an accuracy rate of 0.956279. The decision tree (DT) and AdaBoost models had the lowest accuracy rates of 0.682805 and 0.724471, respectively. The other models, including logistic regression (LR), random forest (RF), naive Bayes (NB), and artificial neural network (ANN), achieved moderate accuracy rates ranging from 0.886737 to 0.971244. Overall, the proposed model outperformed all the other models in terms of accuracy, precision, recall, *F*1-measure, and RMSE.

4.3.2. Dataset Based on 30 Features and 3 Seconds Segment Length. Table 7 and Figure 11 show the testing phase results based on a 3-second segment length and the number of features. The results are shown for various models, including SVM, LR, DT, RF, NB, KNN, ANN, AdaBoost, and the proposed model.

For each model, the table displays the accuracy, *F*1-measure, recall, precision, and RMSE. The proposed model achieves the highest accuracy (0.987676) and *F*1-measure (0.987652), as well as the highest recall and precision. The SVM model also achieves high accuracy (0.987235) and *F*1-measure (0.987206), but the proposed model outperforms it in terms of recall and precision.

The results for the other models vary, with some achieving moderate accuracy (LR, RF, KNN, and ANN), while others have lower accuracy (DT, NB, and AdaBoost). The RMSE values also vary significantly, with the proposed model achieving the lowest value (0.423765), indicating that it has the best predictive performance.

Overall, the results suggest that the proposed model performs well compared to other models and that a 3-second segment length may be more effective than a 2-second segment length.

4.3.3. Dataset Based on 30 Features and 4 Seconds Segment Length. In Table 8 and Figure 12, the testing phase results are presented for a 4-second segment length. The table shows the accuracy, *F*1-measure, recall, precision, and root mean square error (RMSE) for each of the tested algorithms. The algorithms used in this experiment include SVM, LR, DT, RF, NB, KNN, ANN, AdaBoost, and the proposed model.

The results show that the proposed model achieved the highest accuracy (0.995406) and *F*1-measure (0.995403) among all the tested algorithms. It also achieved the highest recall (0.995406) and precision (0.995431) scores. The RMSE value for the proposed model was the lowest (0.287566), indicating a good fit of the model to the data.

The SVM algorithm achieved the second-highest accuracy (0.98761) and *F*1-measure (0.987617) scores, followed by LR with an accuracy of 0.981807 and an *F*1-measure of 0.981838. The DT algorithm performed the worst among all the tested algorithms, with an accuracy of only 0.668427, an *F*1-measure of 0.669207, and a high RMSE value of 1.978908.

Table 6: Testing phase result based on 2-second segment length.

Features' number	Accuracy	*F*1-measure	Recall	Precision	RMSE
SVM	0.980861	0.980858	0.980861	0.980868	0.516454
LR	0.968896	0.968941	0.968896	0.968998	0.634262
DT	0.682805	0.68334	0.682805	0.684155	1.859028
RF	0.940434	0.940081	0.940434	0.940039	0.839006
NB	0.886737	0.885253	0.886737	0.886315	1.135998
KNN	0.956279	0.95667	0.956279	0.957806	0.789826
ANN	0.971244	0.971063	0.971244	0.971251	0.614763
AdaBoost	0.724471	0.724527	0.724471	0.731606	1.820349
Proposed model	0.988262	0.988229	0.988262	0.988323	0.434703

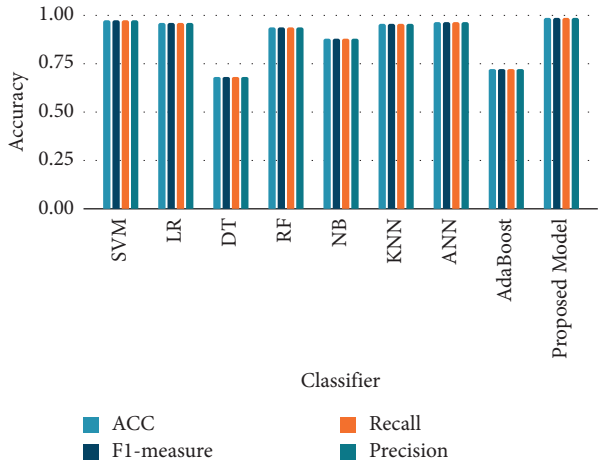

Figure 10: Testing phase result based on 2-second segment length.

Table 7: Testing phase result based on 3-second segment length.

Features' number	Accuracy	*F*1-measure	Recall	Precision	RMSE
SVM	0.987235	0.987206	0.987235	0.987243	0.407348
LR	0.974471	0.974466	0.974471	0.974542	0.576841
DT	0.677816	0.67824	0.677816	0.679116	1.94624
RF	0.953785	0.953264	0.953785	0.953817	0.723717
NB	0.923415	0.922244	0.923415	0.923427	0.919277
KNN	0.968309	0.968487	0.968309	0.969337	0.618451
ANN	0.977112	0.977062	0.977112	0.977154	0.566835
AdaBoost	0.700704	0.700658	0.700704	0.701963	1.763653
Proposed model	0.987676	0.987652	0.987676	0.987684	0.423765

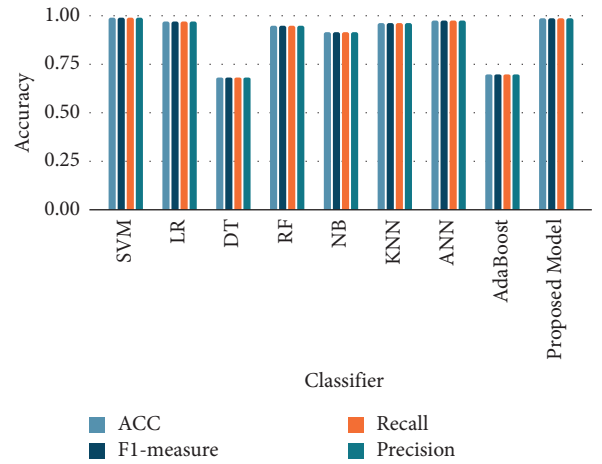

Figure 11: Testing phase result based on 3-second segment length.

Table 8: Testing phase result based on 4-second segment length.

Features' number	Accuracy	*F*1-measure	Recall	Precision	RMSE
SVM	0.98761	0.987617	0.98761	0.987648	0.390015
LR	0.981807	0.981838	0.981807	0.981945	0.481463
DT	0.668427	0.669207	0.668427	0.672091	1.978908
RF	0.96068	0.960466	0.96068	0.960494	0.667839
NB	0.927816	0.926662	0.927816	0.927617	0.933215
KNN	0.969483	0.96975	0.969483	0.970703	0.649575
ANN	0.969483	0.96946	0.969483	0.969577	0.608044
AdaBoost	0.682511	0.683807	0.682511	0.689531	1.920362
Proposed model	0.995406	0.995403	0.995406	0.995431	0.287566

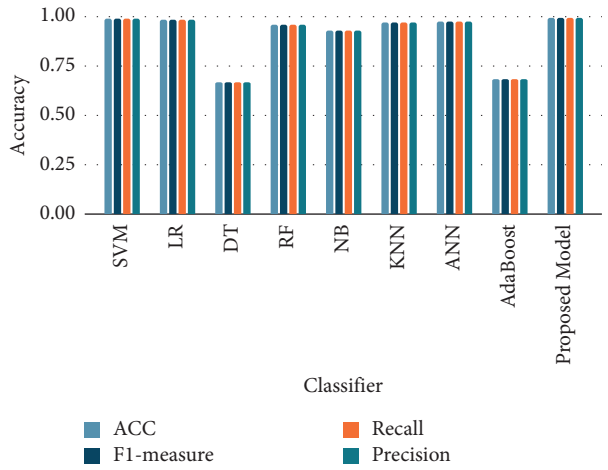

Figure 12: Testing phase result based on 4-second segment length.

Overall, the results suggest that the proposed model is highly effective in recognizing Qur'an reciters, and its performance is superior to the other tested algorithms.

The previous results indicate that the proposed model outperformed the other models and achieved higher accuracy with the various segment lengths used in the experiments. It is also observed that the SVM model follows the proposed model with a slightly smaller accuracy gap. However, some models, such as LR, RF, ANN, and KNN, had low accuracy, while others, such as DT and AdaBoost, had higher accuracy.

Based on the results of this experiment and experiment *B*, it can be concluded that the proposed model with an optimal feature number of 30 and an optimal segment length of 4 seconds achieved the highest accuracy of 0.995406 with an RMSE of 0.287566.

4.4. Performance Comparison between the Proposed Model (CNN) and SVM Model Using GridSearchCV and RandomizedSearchCV. In this study, two models were selected, the proposed model and the SVM model, which exhibited the highest degree of precision. Dataset 3, with 30 features and a 4-second segment length, was used for this comparison. Table 9 and Figure 13 compare the performance of the proposed and SVM models before and after applying GridSearchCV and RandomizedSearchCV.

Both models utilized several parameters. For the proposed model, various parameters, such as kernel size, pool size, learning rate, dropout, and activation function, were used, which significantly impacted the results of the experiments in general. As for the SVM model, the linear and radial basis function (RBF) kernel hyperparameters were utilized. Both models employed ten folds as a cross-validation parameter.

The table shows the comparison of the accuracy, time in minutes, and the results of GridSearchCV and RandomizedSearchCV for the SVM and proposed models. Both search methods aim to optimize the hyperparameters of the models.

For the SVM model, GridSearchCV achieved an accuracy of 0.990002 with a running time of 71 minutes, while RandomizedSearchCV achieved an accuracy of 0.989998 with a shorter running time of 23 minutes. The difference in accuracy between the two search methods is very small.

For the proposed model, GridSearchCV achieved an accuracy of 0.996578 with a running time of 124 minutes, while RandomizedSearchCV achieved an accuracy of 0.996511 with a running time of 49 minutes. The proposed model with the optimized hyperparameters achieved a higher accuracy than the SVM model with the optimized hyperparameters. However, the running time for the proposed model was longer than the SVM model for both search methods.

Overall, the proposed model outperformed the SVM model in terms of accuracy but required more time to optimize the hyperparameters. The GridSearchCV method took more time than the RandomizedSearchCV method for both models but achieved slightly better results for the proposed model.

4.5. Performance Comparison between the Proposed Model (CNN) and BLSTM. Table 10 compares the performance of two classifiers, the proposed model and BLSTM, using five classes. The evaluation metrics used are accuracy, *F*1-measure, recall, and precision. Both models achieved very

Table 9: GridSearchCV and RandomizedSearchCV comparison.

Classifier	Accuracy	GridSearchCV Accuracy	Time (minute)	RandomizedSearchCV Accuracy	Time (minute)
SVM	0.98761	0.990002	71	0.989998	23
Proposed model	0.995406	0.996578	124	0.996511	49

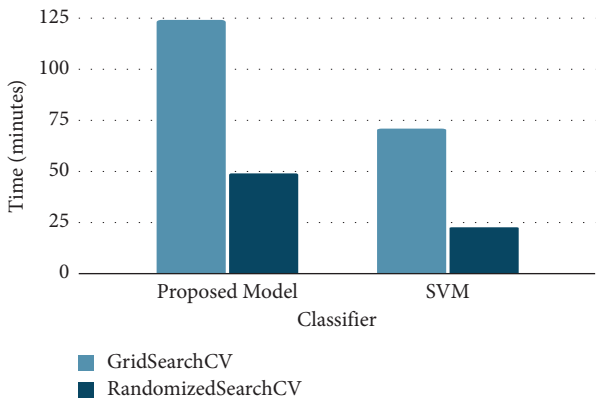

Figure 13: GridSearchCV and RandomizedSearchCV comparison.

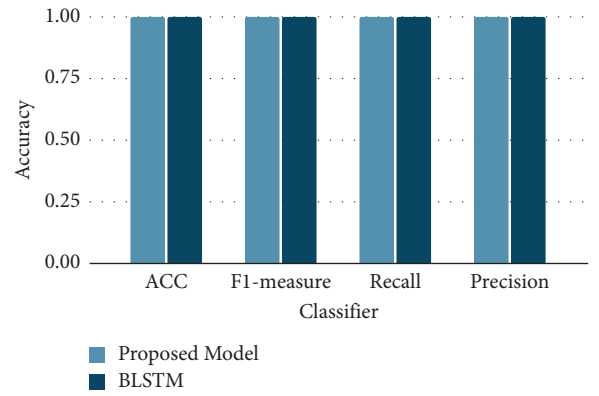

Figure 14: Comparison between (proposed model and BLSTM) using five classes.

Table 10: Comparison between (proposed model and BLSTM) using five classes.

Classifier	Accuracy	*F*1-measure	Recall	Precision
Proposed model	0.998803	0.998811	0.998803	0.998809
BLSTM	0.998992	0.998992	0.998992	0.998992

high accuracy, with the BLSTM model performing slightly better than the proposed model. The *F*1-measure, recall, and precision values were the same for both models, indicating that they both had good overall performance. However, it is important to note that the difference in accuracy between the two models is very small, so other factors, such as the complexity and computational cost of the models, should also be considered when choosing between them. Also, Figure 14 depicts the performance comparison.

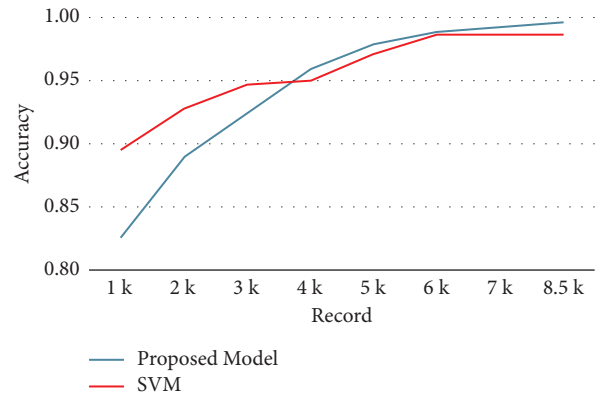

Figure 15: Comparison between the proposed model and SVM on different numbers of record.

4.6. Performance Comparison between the Proposed Model (CNN) and the SVM Model. Figure 15 and Table 11 depict a performance comparison between the proposed model and SVM. This comparison utilizes Dataset 3 with 30 features, 4-second segment length, and a different number of records.

Table 11 presents a comparison between the proposed model and SVM on different numbers of records. The accuracy of both models increases as the number of records increases. However, the proposed model outperforms the SVM model in terms of accuracy for all numbers of records. For instance, for 1 K records, the proposed model achieved an accuracy of 0.825339, while the SVM model achieved 0.89501. For 8.5 K records, the proposed model achieved an accuracy of 0.996578, which is significantly higher than the accuracy achieved by the SVM model (0.98761). These results indicate that the proposed model can handle a larger number of records and achieve higher accuracy compared to the SVM model, and it is worth noting that the proposed model continues to improve when increasing records, reaching an accuracy of 0.996578 when using the entire dataset. Proving that in deep learning methods, performance is directly proportional to the data used.

Table 11: Comparison between the proposed model and SVM on different numbers of record.

Number of records (K)	Proposed model	SVM
1	0.825339	0.89501
2	0.889017	0.92818
3	0.925482	0.947259
4	0.959456	0.950123
5	0.978885	0.971691
6	0.9891	0.987166
7	0.991787	0.987593
8.5	0.996578	0.98761

5. Conclusion

This paper proposes a deep learning-based model for recognizing Qur'an reciters using MFCCs. The aim of this study was to distinguish between trustworthy and fraudulent reciters of the Qur'an. The second objective was to determine whether the deep learning approach is superior to the machine learning approach in the current dataset.

Three datasets with varying segment lengths and feature counts were utilized. Using the proposed model, numerous segment length and feature number comparisons were conducted to determine the optimal segment length and number of features for achieving the highest possible accuracy. The experiment demonstrated that 4 seconds is the optimal segment length and 30 features are the optimal number.

ExtraTreesClassifier was utilized to determine the significance of features for each dataset. In each of the three datasets, the first 30 features were deemed to be the most significant. Then, the first 30 features from each dataset were evaluated using machine learning models and compared to the outcomes of the proposed system with varying segment lengths. Results indicate that the proposed system outperformed all other models with an accuracy of 0.995406 at a segment length of 3 seconds, while SVM and LR achieved an accuracy of 0.98761 and 0.981807, respectively.

GridSearchCV and RandomizedSearchCV were utilized with the proposed model and SVM to find the optimal hyperparameters that improve the model's accuracy. The accuracy of the models improved relatively in both models, from 0.995406 to 0.996578 for the proposed model and from 0.98761 to 0.990002 for the SVM.

Five classes were used to compare the performance of the proposed model and BLSTM. The BLSTM outperformed the proposed model by a small margin, with the proposed model achieving an accuracy of 0.998803 and the BLSTM achieving an accuracy of 0.998992.

Finally, a performance comparison between the proposed and SVM models was conducted to determine the point at which deep learning outperforms machine learning. At 4 k records, the proposed model, which represents deep learning via CNN, outperformed SVM, which represents machine learning.

In the future, a new dataset will be constructed that contains the entirety of the Quran to eliminate the inconsistency of research findings in this field. However, creating such a dataset requires collecting and processing samples. To identify the correct use of the recitation rules (Tajweed-rules) in the entirety of the Qur'an, the study aims to extend its work and employ a deep learning approach (CNN) and MFCCs' technique. Additionally, an application will be developed for the correct application of recitation rules (Tajweed-rules) throughout the Qur'an.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Deanship of Research in Zarqa University/Jordan.

References

[1] Wikipedia, "Muslims – Wikipedia," 2022, https://en.wikipedia.org/wiki/Muslims.

[2] B. Lawrence, *The Qur'an: A Biography*, Atlantic Books Ltd, London, UK, 2014.

[3] Fluentarabic, "3 Reasons why starting to learn arabic is difficult," 2022, https://www.fluentarabic.net/why-learning-arabic-is-difficult/.

[4] A. S. Sadi, T. Anam, M. Abdirazak et al., "Applying ontological modeling on Quranic nature domain," in *Proceedings of the 2016 7th International Conference on Information And Communication Systems (ICICS)*, pp. 151–155, IEEE, Irbid, Jordan, April 2016.

[5] I. Alsmadi and M. Zarour, "Online integrity and authentication checking for Quran electronic versions," *Applied Computing and Informatics*, vol. 13, no. 1, pp. 38–46, 2017.

[6] Z. Razak, N. J. Ibrahim, M. I. Idris et al., "Quranic verse recitation recognition module for support in j-QAF learning: a review," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 8, no. 8, pp. 207–216, 2008.

[7] B. Yousfi and A. M. Zeki, "Holy Qur'an speech recognition system distinguishing the type of recitation," in *Proceedings of the 2016 7th International Conference on Computer Science and Information Technology (CSIT)*, pp. 1–6, IEEE, Amman, Jordan, July 2016.

[8] S. Shahriar and U. Tariq, "Classifying maqams of qur'anic recitations using deep learning," *IEEE Access*, vol. 9, pp. 117271–117281, 2021.

[9] M. O. Khelifa, M. Belkasmi, Y. Abdellah, and Y. Abdellah, "Strategies for implementing an optimal ASR system for quranic recitation recognition," *International Journal of Computer Application*, vol. 172, no. 9, pp. 35–41, 2017.

[10] G. Samara and K. M. Blaou, "Wireless sensor networks hierarchical protocols," in *Proceedings of the 2017 8th International Conference on Information Technology (ICIT)*, pp. 998–1001, IEEE, Amman, Jordan, May 2017.

[11] A. O. A. Salem, T. Alhmiedat, and G. Samara, "Cache discovery policies of MANET," *World of Computer Science and Information Technology Journal*, vol. 3, no. 8, 2013.

[12] T. Zhang and C. C. J. Kuo, "Hierarchical system for content-based audio classification and retrieval," in *Multimedia Storage and Archiving Systems 3*, vol. 3527, pp. 398–409, International Society for Optics and Photonics, 1998.

[13] D. Ebrahimi, S. Sharafeddine, P. H. Ho, and C. Assi, "UAV-aided projection-based compressive data gathering in wireless sensor networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1893–1905, 2019.

[14] P. Ray, R. Kaluri, T. Reddy, and K. Lakshmanna, "Contemporary developments and technologies in deep learning–based IoT," in *Deep Learning for Internet of Things Infrastructure*, pp. 61–82, CRC Press, Boca Raton, FL, USA, 2021.

[15] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin lesion

classification of dermoscopic images using machine learning and convolutional neural network," *Scientific Reports*, vol. 12, no. 1, pp. 18134–18211, 2022.

[16] B. B. Gupta, A. Gaurav, E. C. Marín, and W. Alhalabi, "Novel graph-based machine learning technique to secure smart vehicles in intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 188, pp. 1–9, 2022.

[17] M. Al-Ayyoub, N. A. Damer, and I. Hmeidi, "Using deep learning for automatically determining correct application of basic quranic recitation rules," *The International Arab Journal of Information Technology*, vol. 15, no. 3, pp. 620–625, 2018.

[18] R. U. Khan, A. M. Qamar, and M. Hadwan, "Quranic reciter recognition: a machine learning approach," *Advances in Science, Technology and Engineering Systems Journal*, vol. 4, no. 6, pp. 173–176, 2019.

[19] G. Samara and M. Al-okour, "Optimal number of cluster heads in wireless sensors networks based on LEACH," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 891–895, 2020.

[20] G. Samara, "An intelligent routing protocol in VANET," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 29, no. 1/2, pp. 77–84, 2018.

[21] A. Mahmood, M. Alsulaiman, G. Muhammad, and S. Akram, "Artificially intelligent recognition of Arabic speaker using voice print-based local features," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28, no. 6, pp. 1009–1020, 2016.

[22] G. Samara, G. Al Besani, M. Alauthman, and M. Al Khaldy, "Energy-Efficiency routing algorithms in wireless sensor networks: a survey," *International Journal of Scientific & Technology Research*, vol. 9, no. 1, 2020.

[23] G. Samara and W. A. Ali Alsali, "Message broadcasting protocols in VANET," *Information Technology Journal*, vol. 11, no. 9, pp. 1235–1242, 2012.

[24] T. M. Hasan Asda, T. Surya Gunawan, M. Kartiwi, and H. Mansor, "Development of Quran reciter identification system using MFCC and neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 1, no. 1, pp. 168–175, 2016.

[25] G. Samara and M. Aljaidi, "Aware-routing protocol using best first search algorithm in wireless sensor," *The International Arab Journal of Information Technology*, vol. 15, no. 3, pp. 592–598, 2018.

[26] G. Samara, "Wireless sensor network MAC energy-efficiency protocols: a survey," in *Proceedings of the 2020 21st International Arab Conference on Information Technology (ACIT)*, pp. 1–5, IEEE, Giza, Egypt, November 2020.

[27] Z. Touati-Hamad, M. Ridda Laouar, I. Bendib, and S. Hakak, "Arabic quran verses authentication using deep learning and word embeddings," *The International Arab Journal of Information Technology*, vol. 19, no. 4, pp. 681–688, 2022.

[28] N. A. Damer, M. Al-Ayyoub, and I. Hmeidi, "Automatically determining correct application of basic quranic recitation rules," in *Proceedings of the International Arab Conference on Information Technology*, Muscat, Oman, April 2017.

[29] J. H. Alkhateeb, "A machine learning approach for recognizing the holy quran reciter," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020.

[30] T. S. Gunawan, N. A. Muhamat Saleh, and M. Kartiwi, "Development of quranic reciter identification system using MFCC and GMM classifier," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 1, pp. 372–8708, 2018.

[31] A. Qayyum, S. Latif, and J. Qadir, "Quran reciter identification: a deep learning approach," in *Proceedings of the 2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, pp. 492–497, IEEE, IIUM Gombak, Malaysia, September 2018.

[32] K. M. Nahar, R. khatib, M. Shannaq, and M. Barhoush, "An efficient holy quran recitation recognizer based on svm learning model," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 6, no. 0, p. 1, 2020.

[33] A. M. Alagrami and M. M. Eljazzar, "Smartajweed automatic recognition of Arabic quranic recitation rules," 2020, https://arxiv.org/abs/2101.04200.

[34] K. M. O. Nahar, M. Al-Shannaq, A. Manasrah, R. Alshorman, and I. Alazzam, "A holy quran reader/reciter identification system using support vector machine," *International Journal of Machine Learning and Computing*, vol. 9, no. 4, pp. 458–464, 2019.

[35] A. Elnagar and M. Lataifeh, "Predicting quranic audio clips reciters using classical machine learning algorithms: a comparative study," in *Recent Advances in NLP: The Case of Arabic Language*, pp. 187–209, Springer, Cham, Switzerland, 2020.

[36] R. A. Rajagede and R. P. Hastuti, "Al-Quran recitation verification for memorization test using Siamese LSTM network," *Communications in Science and Technology*, vol. 6, no. 1, pp. 35–40, 2021.

[37] K. Kuppusamy and C. Eswaran, *Speaker Recognition System Based on Age-Related Features Using Convolutional and Deep Neural Networks*, Research Square, Durham, NC, USA, 2020.

[38] Y. J. Park and H. S. Cho, "An experiment of sound recognition using machine learning," in *Proceedings of the 2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pp. 1–3, IEEE, Seoul, South Korea, November, 2020.

[39] P. M. Djuric, Y. Huang, and T. Ghirmai, "Perfect sampling: a review and applications to signal processing," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 345–356, 2002.

[40] B. McFee, C. Raffel, D. Liang et al., "librosa: audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, pp. 18–25, Austin TX, USA, July 2015.

[41] D. Gupta, P. Bansal, and K. Choudhary, "The state of the art of feature extraction techniques in speech recognition," *Speech and language processing for human-machine communications*, pp. 195–207, Springer, Berlin, Germany, 2018.

[42] V. Tiwari, "MFCC and its applications in speaker recognition," *International Journal on Emerging Technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[43] D. S. Kumar, "Feature normalisation for robust speech recognition," 2015, https://arxiv.org/abs/1507.04019.

[44] Haythamfayek, "Speech processing for machine learning: filter banks," 2022, https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.

[45] A. Elnagar, R. Ismail, B. Alattas, and A. Alfalasi, "Automatic classification of reciters of quranic audio clips," in *Proceedings of the 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–6, IEEE, Aqaba, Jordan, October, 2018.

[46] M. Lataifeh, A. Elnagar, I. Shahin, and A. B. Nassif, "Arabic audio clips: identification and discrimination of authentic Cantillations from imitations," *Neurocomputing*, vol. 418, pp. 162–177, 2020.