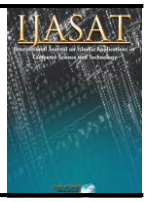

Acoustic Modeling for Indexing and Retrieval of Quranic Verses

Muhammad Aleem Shakeel[1,a], Hasan Ali Khattak[1,2,b], Numan Khurshid[1,c] and Kamran Zeb[1,d]

[1]National University of Sciences and Technology (NUST), Islamabad, Pakistan
[2]Faculty of Computing and Information Technology, Sohar University, Sohar Oman
[[a] mshakeel.msee21seecs, [b] hasan.alikhattak, [c] numan.khurshid, [d] kamran.zeb]@seecs.edu.pk

ABSTRACT

The Holy Quran, revered as the singular scripture of the universe and preserved in its original entirety since its divine revelation, holds profound significance within the Muslim community. Originally shown in the Arabic language, practitioners must understand and adhere to the prescribed methods of recitation and memorization as defined by native Arabic speakers. Despite the advancement of AI technology in acoustic modeling, the intricate nature of Arabic, with its diverse accents and dialects, poses a formidable challenge for developing a resilient model for Quranic recitation. Our research addresses this challenge by introducing a deep learning model that withstands linguistic variations and stays unaffected by diverse recitation styles and the nuances of the Tajweed. In this paper, the deep features extracted from this model prove exceptional performance, achieving a remarkable accuracy of approximately 96.30% in classification tasks. To underscore the significance of our deep learning network as an acoustic model, we developed a content-based verse retrieval system (CBVeRse). Utilizing the previously trained model, this system exhibited an impressive performance with a mean Average Precision (mAP) of 96.52%. This underscores the efficiency and importance of our approach in enhancing the understanding and application of the Holy Quran's acoustic attributes.

Keywords: Acoustic Modeling, Quran, Deep Learning, Content-Based Verse Retrieval System (CBVeRse)

1 Introduction

The Quran is the most essential book for millions of Muslims. (Irmi et al., 2023). "Tilawah," or the poetry recital in Quranic verses, is a profound art form that captivates listeners with its rhythmic flow and serves as a vehicle for fostering spiritual awareness. Maintaining precise recitation has always been the most important for the Muslim world. Because Quranic recitation has always been passed down orally, the art form has been preserved in its most basic form. (Samara, 2021) However, new opportunities have emerged to use deep learning and machine intelligence to enhance our understanding of Quranic recitation and make it more accessible to a broader audience.

One area that has received much attention lately is deep modeling techniques for Quran recitation. Deep learning, natural language processing, and speech recognition have significantly progressed. Deep acoustic modeling processes and analyzes acoustic data using advanced neural networks in speech recognition and voice analysis. By employing these innovative methods for Quranic recitation, researchers hope to gain a deeper understanding of the art of "Tilawah." (Samara & Al-okour, 2020). Additionally, they aim to make the existing Quranic recitation recognition systems more reliable and accurate.

Numerous researchers have diligently explored various facets of the Quran, employing advanced deep-learning techniques. For classifying Arabic Speech Genres, Devin et al. (Stewart, 2022) offered multiple methods for identifying distinct speech genres in the Quran. They highlight the kinds of evidence scholars should focus on when examining genres, provide fundamental criteria for interpreting Qur'anic verses, and discuss mistakes and difficulties that should be considered in follow-up research. They advocated closely examining unique words, phrases, and structures. For text summarization using deep learning models, Wahdan et al. (Wahdan et al., 2020) proposed a text categorization for the Arabic language. They focused on deep learning-based text categorization methods such as CNN, RNN, LSTM, etc. They provided a detailed analysis of the system models, accuracy, and outcomes of twelve relevant research publications before recommending the model that would be most useful. In the end, they offer recommendations on models to employ to enhance text classification.

For Reciters Classification, Khan et al. (R. U. Khan et al., 2019) presented a machine-learning method for identifying the Holy Quran reciter. Twelve different reciters in the dataset recited the last ten surahs of the Quran, so the model has 12 classes to categorize. First, features are extracted using the sound pitch and MFCC. Audio spectrogram auto-correlograms are the second. Next, apply J48, Random Forest, and Naive-Bayes for classification. The highest accuracy of 88% was attained with Naive-Bayes and Random Forest. For Tajweed and Short-Vowels, Alqadheeb et al. (Alqadheeb et al., 2021) utilized an audio collection of Arabic words that included short vowels. There are 84 classes and 2892 Arabic short vowels in the entire dataset. They put the preprocessing methods into practice and use CNN for testing and classification. With the word "ALIF" as a test word, 312 Arabic phonemes were used to test the model, and 100% accuracy was attained.

Despite the substantial progress made in the field of Quranic research, encompassing areas such as Quranic semantics, Natural Language Processing (NLP) for text summarization, reciter classification, Tajweed and Makhraj correction, and Automatic Speech Recognition (ASR), it is evident that certain holistic aspects remain unaddressed. These unmet challenges are covered in this paper and count as a contributing factor to the research gap.

- Building a comprehensive and efficient dataset for Quranic verses and developing efficient data processing pipelines for modeling.
- Building a state-of-the-art deep acoustic model for Quranic recitation modeling efficient enough to cover Quranic scripture recitation.
- An efficient recommendation system for a Content-Based Verse Retrieval System that can help retrieve and index Quranic verses.

2 System Model

We aim to create a system model to develop state-of-the-art deep-learning approaches that correctly categorize Quranic Surahs based on their acoustic feature and help find indexing and localization of verses. The system model is broken down into key stages, as shown in Figure 1.

2.1 Data Acquisition

Data acquisition involves gathering audio recordings of individuals reciting verses of the Holy Quran. We took different considerations for precise data, including selecting Reciters with well-known knowledge of Quranic grammar and phonemes, acoustic environments including closed-room recordings, and different dialectal recitation styles and speeds. (Alajmi, 2023) The reciters we used for modeling include:

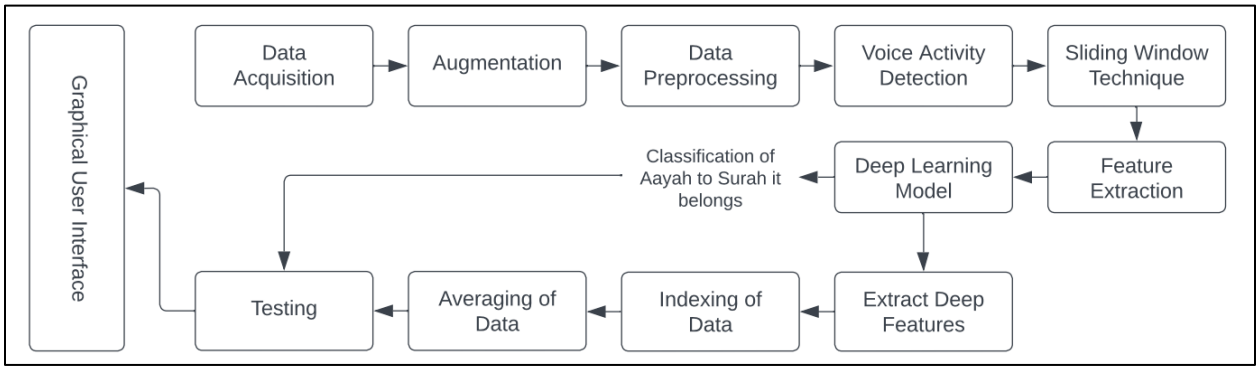

Figure 1: Detailed system model of the proposed method and data preprocessing pipeline

- Sheikh Saad Al Ghamdi
- Sheikh Ali Al Huhaifi
- Sheikh Mishary Rashid Al Afasy
- Sheikh Salah Al Budair

We consider the Quranic recitation of the last 10 Surahs, from Al-Fil (105) to An-Nas (114), each having a random length of every Surah, depending on their recitation style. The maximum size of a recitation is 27 sec. We have collected these audio files from various sources on the internet in MP3 format and converted them to. WAV format for further processing.

2.2 Data Preprocessing and Augmentation

The data preparation process employs various techniques and steps to improve the audio data's effectiveness. These techniques include:

1. Each recitation consists of random audio lengths, so we implemented a force-alignment technique to fix every audio to the same size manually.
2. Conversion of audio samples from stereo-channel to mono-channel.
3. Adjust the sampling rate from 44.1KHz and 22.05KHz to 16KHz to generate an array. (Park & Cho, 2020)

For data augmentation, we use Librosa's built-in augmentation features (Gambhir et al., 2023) for different lengths, as shown in Table 1.

Table 1: Data Augmentation transformation and rates to produce robustness.

Transformations	Rate
Time Stretch	0.1x – 0.5x
Pitch Scale	1x – 8x
Random Gain	Min Factor: 1x-3x, Max Factor: 2x-4x
Adding White Noise	0.1x – 0.8x
Polarity Inversion	-1

These augmentation techniques help the model to learn diversity and train on different environmental conditions.

2.3 Voice Activity Detection

Voice activity detection (VAD) finds speech in audio segments and helps differentiate speech from background noise or silence. (Andersen et al., 2023) VAD must separate distinct "ayahs" (verses) from a more extensive surah audio recording (chapter). We implemented Silero-VAD technology, which can distinguish between areas of an audio sample with speech activity and others without. Silero-ASR (Automatic Speech Recognition) systems, on which the Silero-VAD model works, handle noises and artifacts that might interfere with the identification process.

2.4 Feature Extraction

The feature extraction technique creates a more manageable and valuable representation of raw audio data. (Arpitha et al., 2022) We implemented a Sliding window technique to break the audio signal further into (1,2, and 3 seconds) to get more features and create robustness in the model. Since we had Arabic speech audio signals, we prefer the Mel-Frequency Cepstral Coefficient (MFCC) for feature extraction of the audio signals. (Abbaskhah et al., 2023) MFCC computation requires the following steps, shown in Figure 2.

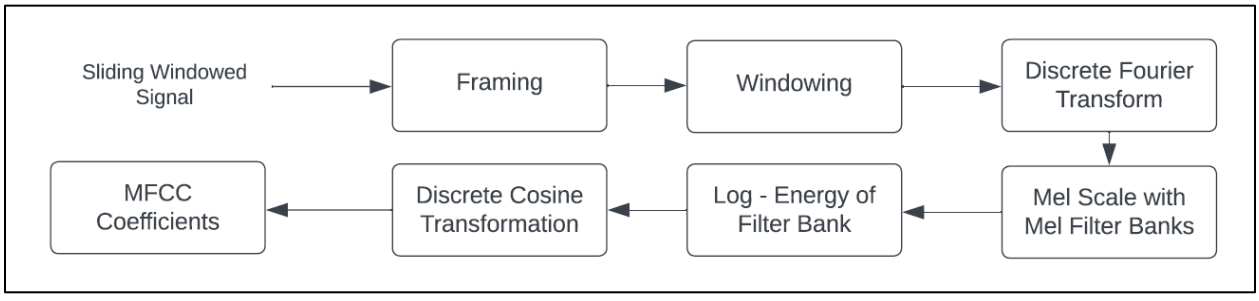

Figure 2: Detailed Methodology of MFCC process

2.4.1 Short-Time Processing - Framing & Windowing

A 50% common overlap between the start of the next frame and the conclusion of the preceding frame for short-time processing. We used the Hamming Window for the windowing function because it controls and balances the frequency domain's main lobe width and side lobe attenuation.

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) \quad (2.1)$$

2.4.2 Mel-Frequency Wrapping and Filter banks

Mel filter banks simulate the sensitivity of the human auditory system to various frequency ranges while extracting relevant spectral information from the power spectrum of the stream. (Bhangale & Mohanaprasad, 2021). The perceptual frequency scale is a nonlinear scale that more closely reflects how humans hear pitch. The following eq. (2.2) is widely used to approximate frequency in Hertz (f) and frequency in Mel (m):

$$M(f) = 2595.\log_{10}(1+\frac{f}{700}) \quad (2.2)$$

A group of triangle filters positioned along the Mel scale make up Mel filter banks. Each filter's breadth and center frequency, commonly represented in Mel units, characterizes it. The center frequencies of the filter bank are linearly spaced in Mel scale before being transformed back to Hertz. The response of each filter is computed using the following equations:

$$H(m,k) = \begin{cases} 0 & \text{if } f(k) < f(m-1) \\ \dfrac{f(k) - f(m-1)}{f(m) - f(m-1)} & \text{if } f(m-1) \le f(k) \le f(m) \\ \dfrac{f(m+1) - f(k)}{f(m+1) - f(m)} & \text{if } f(m) \le f(k) \le f(m+1) \\ 0 & \text{if } f(k) > f(m+1) \end{cases} \quad (2.3)$$

Where $m = \{0,1,2,\ldots,M-1\}$, $H(m,k)$ is the response of the m^{th} filter at the frequency bin k, and $f(k)$ is the frequency corresponding to bin k in eq. (2.3)

2.4.3 Discrete Cosine Transforms

To return the log mel spectrums, also known as MFCCs, to the time domain, the DCT is applied to the logarithmically scaled filterbank energies. The DCT's equation is:

$$C(i) = \sum_{k=1}^{K}(\log(|X(k)|^2)\cos(\frac{i(k-0.5)\pi}{K})) \quad (2.4)$$

Where $C(i)$ is the i-th MFCC coefficient, and K is the total frequency bins in eq. (2.4)

2.5 Deep Learning Model

In the proposed approach, we used a multi-class classifier to build the state-of-the-art deep acoustic model for the Quran that classifies each Ayah according to the Surah to which it belongs. (Bano et al., 2023) We prioritize speech recognition systems' classifiers that obtain high accuracy. (Lataifeh et al., 2020). Model details are shown in Table 2.

Table 2: Deep Learning Model architecture of the proposed system

Layers	No. of Filters	Filter Size	Padding	Activation
Conv-2D	64	3x3	Same	Tanh
MaxPool-2D	-	2x2	-	-
Conv-2D	64	3x3	Same	Tanh
MaxPool-2D	-	2x2	-	-
Conv-2D	64	3x3	Same	Tanh
MaxPool-2D	-	2x2	-	-
Dropout	Rate – 0.1	-	-	-
Dense	1024	-	-	Tanh
Dense	10	-	-	-

To help guard against overfitting, we applied the L2 regularizer to the first layer.

2.6 Extraction of Audio Deep Features

To extract deep features, we defined the second-last layer, i.e., the 'dense' layer of the model, which has 1024 deep features for each verse of the Quran. To show its use, we find this layer's index in the list of layers for the model, which means the model uses the same input as earlier but outputs the activations of the 'dense' layer and supplies the deep 1024 features of each audio. (Lataifeh et al., 2020)

With our deep feature model, we began by passing an extensive collection of Quranic verse audio recordings used in training earlier. (Amiriparian et al., 2018). Using the prediction function with the deep feature model, we systematically collected the output from the 'dense' layer for each audio sample and called the deep features of the audio.

2.7 Indexing & Retrieval of Verses

To organize the Quranic verses into a database so they would be simple to search, they are arranged so that the first verse's audio goes with the first ayah, the second verse accompanies the second ayah, and so on. In this manner, we obtained deep features that matched the Quranic verses' chronological order. (H. Khan et al., 2023)

Figure 3: Example of Indexing of the Quranic verses

To deal with many reciters who might speak in various dialects, we use a method to decide the average (or mean) of the identical verses performed by many reciters shown in eq (2.5). This means we cut any dialect-related variances and are left with a more consistent and robust rendition of the verses, making it more straightforward.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \tag{2.5}$$

3 Experimentation and Results

The experimental outcomes of the proposed CNN model are shown in this section. The models were assessed using several performance measures, including accuracy, precision, F1-score, and mAP. For every audio segment of length (1, 2, or 3 seconds), all 40 MFCC features are retrieved, and the model is assessed based on the previously specified parameters for each segment. The dataset has been divided into 90-10, 80-20, and 70-30 data categories for training and testing.

The best result of the proposed system shows that the model performs very well when tested on a smaller fraction of the dataset (10%) and trained on a relatively significant percentage of the dataset (90%). With the help of such a large training dataset, the model could find underlying patterns in the data, which allowed it to predict extremely well on test data. The data has been split into consecutive segments, each lasting one second, based on using a sliding window size of one 1-second. Figure 4 displays the ideal training outcomes, including training accuracy and loss per epoch.

The model's high accuracy and dependability are proven by achieving a normalized confusion matrix with highly correct predictions, as shown in Figure 5, suggesting that it is more reliable for practical applications.

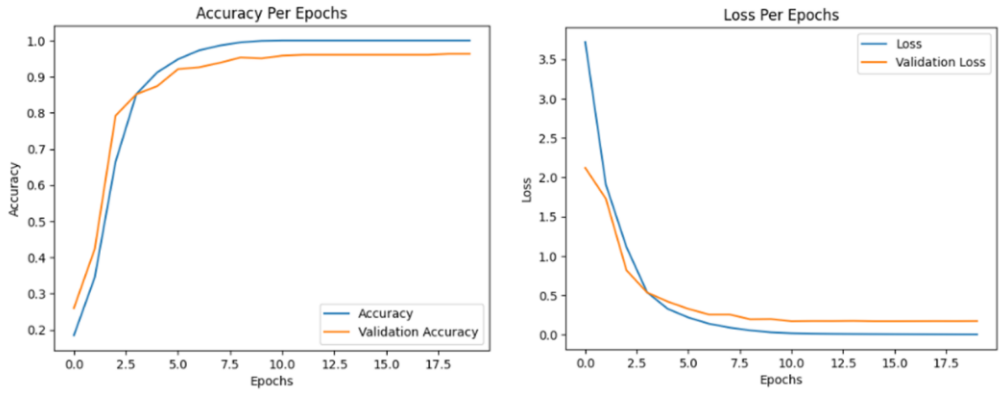

Figure 4: Accuracy and Loss Per Epoch of our Proposed System

Figure 5: Normalized Confusion Matrix of the trained model

Performance parameters were slightly decreased due to the model's performance being affected by the change in sliding window size from 1 to 3 seconds. This suggests that the 1-second window may have been more helpful for the specific task than the 3-second. The findings show that accuracy and other performance measures increase with increased features. The maximum accuracy reached by the proposed system is 96.30%. The detailed results are shown in Table 3.

Table 3: Performance parameter for different SW sizes with best Train-Test Split and accuracies.

SW - Size	Train-Test Split	Accuracy (%)	Precision	Recall	F1- Score
1 sec	90-10	96.30	0.9641	0.9628	0.9627
2 sec	90-10	96.29	0.9643	0.9628	0.9626
3 sec	90-10	95.05	0.9530	0.9504	0.9504

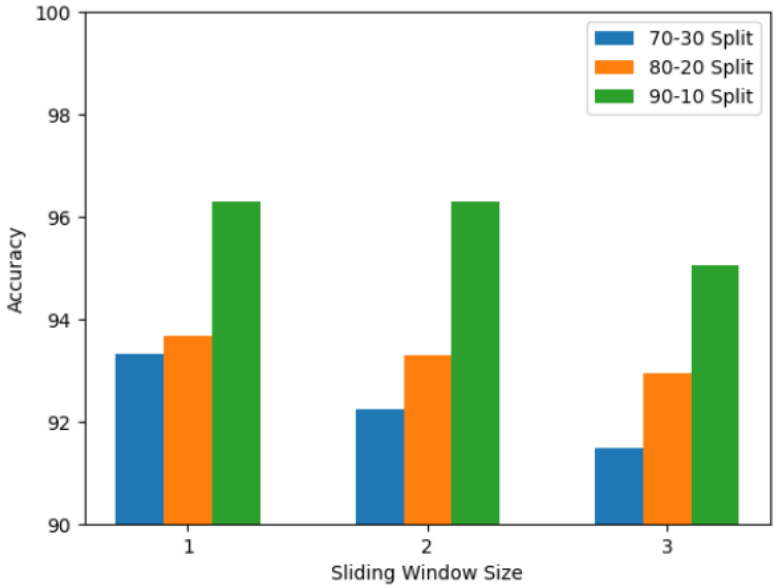

Figure 6: Accuracies achieved for different SW-Size and different Train-Test-Split

3.1 Testing: Automated Surah Classification and Localization via Mean Representations

We use audio samples recited by a different reciter for testing except for our training data, which may have a different dialect. Just like with the training data, we implement preprocessing steps on the testing data to ensure it's in a suitable format for analysis. Once the testing audio data has been preprocessed, we pass it through the same Dense Layer with 1024 neurons that we used during the training phase. This Dense Layer helps extract relevant features and representations from the audio data. Despite this change in audio samples with different dialect and recitation styles, our proposed deep learning based CNN model correctly predicted the name of the Surah to which the ayah belongs. Secondly, our database, which holds the average representations of each ayah in chronological sequence, uses the principle of Euclidean distance, to find the closest match, when trying to find the testing audio sample index within our database. We can find the closest match with accuracy by calculating the difference between the test ayah's deep feature representation, and each representation kept in the database.

3.2 Graphical User Interface (GUI)

Our user-friendly Graphical User Interface (GUI), shown in Figure 7, makes Quranic audio analysis simple. The GUI supplies the following functions:

- Simple input section where users can easily upload testing audio samples. WAV format.
- Integration of an audio playback capability: make sure the user has chosen the correct audio file for analysis by listening to it.
- For output, our trained model predicts the name of the Surah and presents the index or locale to find the Surah within the Quran corresponding to the tested audio. This data is presented in a standard style, like "(Surah Name: Indexing).

The mean Average Precision (mAP) value reached by our recommendation system in the context of a content-based verse retrieval system (CBVeRse) stands impressively at 96.52%. This metric underscores the performance of our system in accurately localizing and predicting Quranic verses.

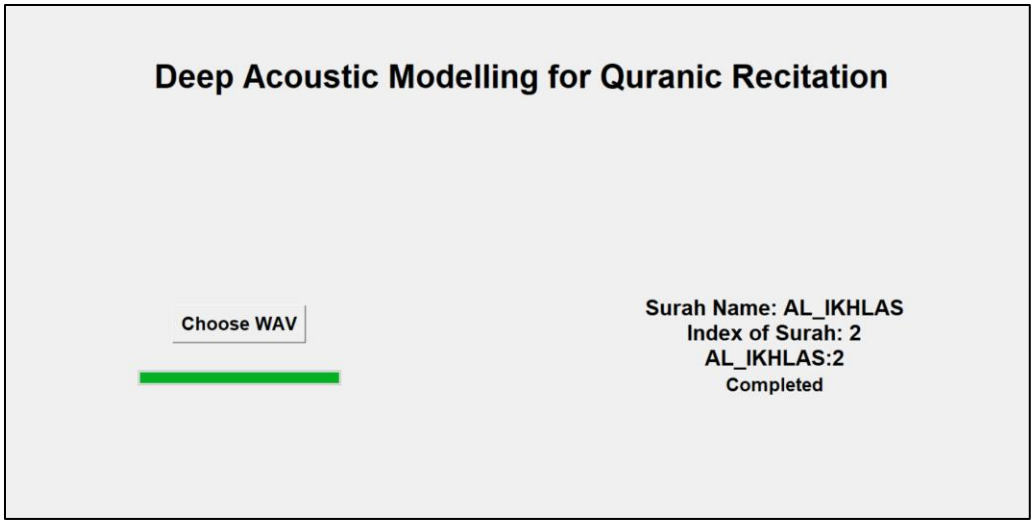

Figure 7: Graphical User Interface (GUI) of the proposed system with outputs of Indexing

4 Conclusion & Future Work

In conclusion, deep acoustic modeling for Quranic recitation has significantly enhanced the understanding and analysis of Quranic recitations. Initially, we gathered many datasets from different reciters and dialects to extract individual verses from Quranic chapters. The data was then ready for training by performing feature extraction using the MFCC. We trained a state-of-the-art deep acoustic model, which produced an incredible 96.30% accuracy. After deep feature extraction, we tested and used the same preprocessing methods to guarantee accuracy and consistency during testing and compared to deep featured data in our database using Euclidian distance to find the closest match. Through this procedure, we were able to get valuable insights about the recitation of the Quran with a mAP value of 96.52%.

For future work, this acoustic modeling can be used for the Quranic Recitation Tajweed recognition and mispronunciation detection on the verse level. Also, the approach could be extended to complete Quranic recitation modeling, including all Quranic Surahs. Most importantly, Quranic Recitation can also be used as the Arabic repository for building large Arabic Language Modeling irrespective of the dialects.

References

Abbaskhah, A., Sedighi, H., & Marvi, H. (2023). Infant cry classification by MFCC feature extraction with MLP and CNN structures. *Biomedical Signal Processing and Control*, *86*, 105261.

Alajmi, N. M. (2023). Regional and Sociolinguistic Variation of Personal Pronouns in Dialects of Najdi Arabic. *Journal of Language Teaching and Research*, *14*(5), 1313–1319.

Alqadheeb, F., Asif, A., & Ahmad, H. F. (2021). Correct pronunciation detection for classical Arabic phonemes using deep learning. *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, 1–6.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Pugachevskiy, S., & Schuller, B. (2018). Bag-of-deep-features: Noise-robust deep feature representations for audio analysis. *2018*

International Joint Conference on Neural Networks (IJCNN), 1–7.

Andersen, L. R., Jacobsen, L. J., & Campos, D. (2023). Compressed, Real-Time Voice Activity Detection with Open Source Implementation for Small Devices. *Proceedings of the 8th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*, 1–10.

Arpitha, Y., Madhumathi, G. L., & Balaji, N. (2022). Spectrogram analysis of ECG signal and classification efficiency using MFCC feature extraction technique. *Journal of Ambient Intelligence and Humanized Computing*, 1–11.

Bano, S., Khalid, S., Tairan, N. M., Shah, H., & Khattak, H. A. (2023). Summarization of scholarly articles using BERT and BiGRU: Deep learning-based extractive approach. *Journal of King Saud University-Computer and Information Sciences*, *35*(9), 101739.

Bhangale, K. B., & Mohanaprasad, K. (2021). A review on speech processing using machine learning paradigm. *International Journal of Speech Technology*, *24*, 367–388.

Gambhir, P., Dev, A., Bansal, P., & Sharma, D. K. (2023). End-to-End Multi-modal Low-Resourced Speech keywords Recognition using Sequential Conv2D Nets. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Irmi, R. F., Al Farabi, M., & Darlis, A. (2023). Technology Education in the Quran. *Solo Universal Journal of Islamic Education and Multiculturalism*, *1*(01), 1–9.

Khan, H., Saqib, M., Khattak, H. A., Ali, S. I., & Lee, S. (2023). Ontology Alignment for Accurate Ontology Matching: A Survey. *International Conference on Smart Homes and Health Telematics*, 338–349.

Khan, R. U., Qamar, A. M., & Hadwan, M. (2019). Quranic reciter recognition: a machine learning approach. *Advances in Science, Technology and Engineering Systems Journal*, *4*(6), 173–176.

Lataifeh, M., Elnagar, A., Shahin, I., & Nassif, A. B. (2020). Arabic audio clips: Identification and discrimination of authentic Cantillations from imitations. *Neurocomputing*, *418*, 162–177.

Park, Y.-J., & Cho, H.-S. (2020). An experiment of sound recognition using machine learning. *2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 1–3.

Samara, G. (2021). Lane prediction optimization in VANET. *Egyptian Informatics Journal*, *22*(4), 411–416.

Samara, G., & Al-okour, M. (2020). Optimal number of cluster heads in wireless sensors networks based on LEACH. *ArXiv Preprint ArXiv:2003.13765*.

Stewart, D. J. (2022). Approaches to the Investigation of Speech Genres in the Qur'an. *Journal of Qur'anic Studies*, *24*(1), 1–45.

Wahdan, A., Hantoobi, S., Salloum, S. A., & Shaalan, K. (2020). A systematic review of text classification research based on deep learning models in Arabic language. *Int. J. Electr. Comput. Eng*, *10*(6), 6629–6643.