



OPEN ACCESS

EDITED BY

Ashwani Kumar,
SRM Institute of Science and Technology, India

REVIEWED BY

Mohamed Nabih,
Bruno Kessler Foundation (FBK), Italy
Qasem Abu Al-Haija,
Jordan University of Science and Technology,
Jordan

*CORRESPONDENCE

Izzeldin I. Mohamed,
✉ iabdelaziz@su.edu.om

RECEIVED 27 November 2024

ACCEPTED 30 May 2025

PUBLISHED 13 October 2025

CITATION

Al-Shidi A, Nordin FHB, Mohamed II, Younis MA,
Elmutasim IE and Abusham EE (2025) A
literature review on natural language
processing techniques for Qur'anic studies:
challenges and insights.
Front. Signal Process. 5:1535166.
doi: 10.3389/frsip.2025.1535166

COPYRIGHT

© 2025 Al-Shidi, Nordin, Mohamed, Younis,
Elmutasim and Abusham. This is an open-
access article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

A literature review on natural language processing techniques for Qur'anic studies: challenges and insights

Alia Al-Shidi¹, Farah Hani Bt Nordin², Izzeldin I. Mohamed^{1*},
M. A. Younis¹, Imadeldin Elsayed Elmutasim³ and
Eimad Eldin Abusham⁴

¹Electrical and Computer Engineering Department, Faculty of Engineering, Sohar University, Sohar, Oman, ²Electrical & Electronic Engineering, Institute of Power Engineering, Universiti Tenaga Nasional, Kajang, Selangor, Malaysia, ³Faculty of Electrical and Electronics Engineering Technology Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia, ⁴Faculty of Computing and IT Sohar University, Sohar, Oman

Studies related to the Qur'an have unearthed a multitude of unique challenges and opportunities within the Natural Language Processing (NLP) research domain. These challenges tend to stem from the distinctive orthography of the Qur'an, the need for technical and linguistic support, as well as the requirement for advanced NLP technologies to address stumbling intricacies. As such, this paper delves into the significance of NLP in analyzing the holy Qur'an across ten notable dimensions; Grammatical NPL analysis, speech Recognition, Semantic Work, Ontology Base Work, Translation, Search system, Classifier system, Topic Extraction, knowledge Extraction, and Maqams. In particular, this paper looks into the following aspects - linguistic structure, positive discourse, and the use of conjunctions in translation. The probabilistic topic modeling was also deployed, along with paragraph vectors and K-Mean clustering techniques, to explore the semantically rich subjects identified across the Qur'an. Besides enabling researchers to apply deep learning techniques when analyzing Qur'anic passages and recitations, the listed methods map out the influence of ancient Arab philosophers on the development of both the Maqam theory and practice. Having said the above, NLP methods are indeed applicable for Qur'anic research purposes to unveil rich information waiting to be discovered.

KEYWORDS

Qur'anic, NLP, review, studies, challenges

1 Introduction

The Qur'an is a timeless compass with Divine significance. The verses outlined in the Qur'an has echoed for centuries; touching hearts, inspiring minds, and forming civilizations. The Qur'an was revealed to Prophet Muhammad (pbuh) for over 23 years until 632 BC. The Holy Qur'an is composed of 157,935 words, 5,277 unique terms, 6,236 verses, and 114 chapters. Its language is renowned for being poetic and succinct. Roughly 1.8 billion Muslims use this Holy Book for liturgical purposes, with 73% of them being non-Arabs. Studying the Qur'anic scriptures has been a lifelong passion for hundreds of academics upon gaining prominence more than 1,400 years ago. The topic of Quranic

Studies is a vast scholarly investigation hidden behind a celestial curtain. Fundamentally, the goal of Quranic Studies is to dissect the complex structure of the Qur'an. It goes beyond simple recitation and requires one to examine the following contexts:

Textual Criticism: The foundation is the textual history of the Qur'an, which includes its revelation, compilation, and transmission. The aim here is to get a glimpse into the earliest revelations by closely examining old manuscripts and tracking down alternate readings (qira'at).

Linguistic Marvels: The eloquence of the Qur'an surpass human language. Linguists and poets alike are enthralled by its grammatical subtleties, poetic rhythms, and rhetorical tactics. Its word choices, similes, and metaphors are analyzed while marveling at its heavenly eloquence.

Exegesis (Tafsir): In tafsir, academics dissect meaning in layers; it is the core of Quranic studies. Every verse turns into a universe, a call to consider, inquire, and pursue knowledge. Traditional analyses provide valuable perspectives, whilst contemporary methods integrate customs and the current environment.

Ethics and Law: Social norms and individual behavior are governed by the precepts of the Quran. Islamic jurisprudence is shaped by legal scholars' extraction of legal principles (usul al-fiqh) from its passages. In addition to the nature of God, predestination, free choice, and eschatology are among the theological issues addressed in the Qur'an. Academics participate in theological discussions by utilizing the boundless resources found in the Qur'an.

Natural language processing (NLP) has promoted advances in computer approaches to make Qur'anic research and studies easier, opening new possibilities for those interested in studying and comprehending the Qur'an. However, Arabic NLP is still at least 10 years behind English NLP in terms of development. This is because, Arabic is a rich language with intricate syntactic and grammatical features. One of the biggest obstacles to NLP tasks involving Modern Standard Arabic (MSA), the most common and literary version of Arabic, refers to the absence of capitalization, which makes Named Entity Recognition (NER) more intricate and difficult to identify proper names. In addition, the writing style in MSA excludes diacritical marking, perhaps because readers may infer meaning from context. Diacritic give words additional meaning. Given that Arabic has a distinct sentence structure that demands careful alignment for accurate meaning, translating from Arabic can be a laborious task. For example, [Altammami and Atwell \(2022\)](#) disclosed some challenges found in both the Qur'an and the Hadith. They reported that the unique orthography and distinct meanings of words in both Books make it challenging for tokenization, segmentation, and morphological analysis. The Quran and the Hadith, particularly referring to their related texts, are distinct in structure, style, and orthography, thus making it difficult for readers to identify semantic similarities. The Quran and the Hadith have been noted for their complex texts embedded with polysemy, sarcasm, hyperbole, and other linguistic features that the current NLP models may dismiss or overlook.

The Quran, which is considered to be God's exact word revealed to Prophet Muhammad, is distinguished by its highly organized, rhythmic, and eloquent language, as well as a unique orthography with a unique orthography that has been preserved meticulously since its compilation. Thus, the paper provides a clear sound to go

into two sections: NLP works with the Qur'an and Qur'anic NLP techniques.

2 NLP works for Qur'an

Even though studies within the Qur'anic NLP domain are not in the limelight at the present moment, much research work has been published in the past 2 decades revolving around the application of linguistic computation to the Qur'an. As illustrated in [Figure 1](#), NLP is composed of numerous categories, including Grammatical Analysis, Semantic Work, Ontology-Based Work, Translations, Search Systems, Classifier Systems, Topic Extraction and Knowledge Extraction, Speech Processing, and Maqams. Each subsection delineates additional information.

2.1 Grammatical NLP analysis

Arabic grammar can be examined by using both morphology and syntactical analyses; syntactic analysis concentrates on sentence structure, word form, proper noun, verb, and particle relations as shown in [Table 1](#). In an instance, [Dukes and Buckwalter \(2010\)](#) established the Qur'anic Arabic Dependency Treebank (QADT) to explore Qur'anic grammar. The researchers manually annotated text collection that encompassed all 77,430 words found in the Qur'an. Morphological analysis involves breaking down words into morphemes and assigning grammatical roles. The QADT process is composed of three phases: automated tagging, manual verification, and collaborative online annotation. The automated system recorded an impressive 87% coverage of the Qur'an with 77% accuracy and an 83% recall rate.

In the attempt to annotate Qur'anic verses, [Dukes et al. \(2010\)](#) developed several guidelines based on the QADT. Their approach is inclusive of a rule-based dependency parser with a robust F-measure of 78%. The manual checks adhere to standard grammar rules as a guide. Some examples of guidelines cover prepositional phrases and subject-verb-object sentence structures. Expert supervisors ensure the quality of decisions by drawing on their deep subject knowledge ([Dukes et al., 2013](#)).

In another study, [Bentrcia et al. \(2018\)](#) applied the conjunction "AND" to examine conjunctive phrases found in the Qur'an. Three cases were identified: words that occur in a particular order just once, words that occur in a particular order more than once, and words that occur in a different order either once or more. The study mined the necessary conjunctive phrases using the QAC corpus and a combination of statistical and grammatical methods. The findings revealed that words that occur in a particular order appear to be logical. A similar study by [Fadlun \(2016\)](#) assessed the functions of conjunctions in translation, their impact on textual meaning, and methods to generate concordances and frequency lists using both corpus analysis and AntConc software.

The QADT contains graphical dependency graphs that are useful for scholars studying the Qur'an and Arabic linguistics. These visual representations facilitate a deeper understanding of the traditional Arabic grammar. By utilizing color-coded annotations, scholars can meticulously examine each morpheme while also conducting extensive morphological analyses of verses.

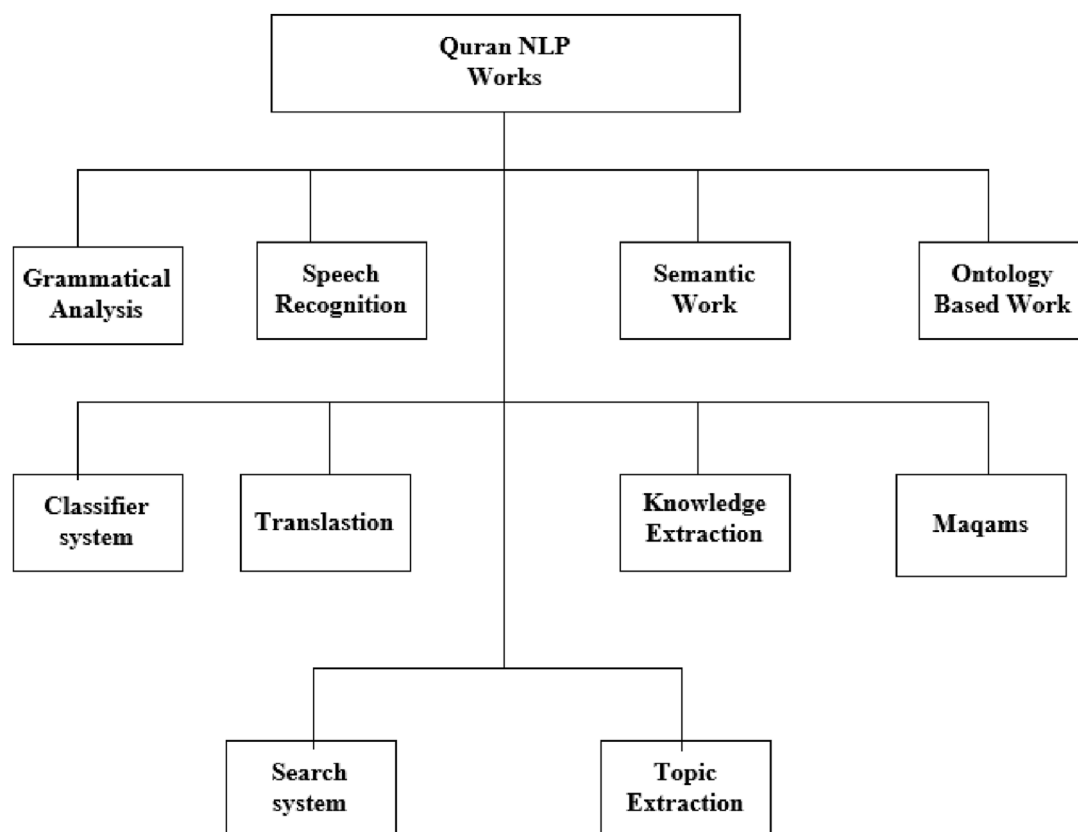


FIGURE 1
Categorization of NLP work for the Qur'an.

The interconnectedness of words within phrases and across sentences becomes evident through these graphical structures. ElAffendi et al. (2021) explored morphological patterns and part-of-speech (POS) tagging for Arabic using a neural network-based approach. In order to achieve real-time embeddings for words, sentences, and characters simultaneously, they leveraged the Galois Power-of-Two (GPOW2) technique. Their research work benefited from rich morphological and syntactic annotations available in the QAC dataset. The model calculated contextual embeddings for each word during the input phase and sent them to the neural network along with GPOW2 representations. Remarkably, their POS tag prediction achieved an impressive accuracy of 98.8%, underscoring the intricacies of Arabic morphology. Proper syntax parsing is paramount in Arabic because an erroneous i'raab (grammatical inflection) can lead to misinterpretation of the intended meaning. In a study, Mannaa et al. (2022) developed an enhanced context-free grammar (eCFG) technique that covers all rules taught in grammar textbooks and recorded an accuracy of 88.33% on 300 sentences. Khadhim and Chiad (2022) explained that morphological analysis is indeed a crucial technique in Qur'anic NLP that enables information extraction, summarization, and translation. However, due to factors including the complexity of the Arabic language, the sensitivity of religious writings, the diversity of genres, and the paucity of huge corpora, NLP approaches appear to be challenging. In addition, ethical and legal concerns such as protecting intellectual

property rights, maintaining the purity of the Qur'an, and avoiding the exploitation and abuse of NLP tools must be addressed. Recently, Abouelnour et al. (2023) investigated the idea of Qur'anic.

2.2 Speech processing and Qur'anic NLP

The entire Qur'an has been memorized by Muslims worldwide and many find pleasure in reciting it out loud. Modern speech-based learning tools are crucial to ensure accurate recitation in the age of technological innovation, thus encouraging Muslims in their spiritual path. The advancement in speaker-independent automatic speech recognition (ASR) technology has become essential in combining tradition and innovation to improve the holy practice of Qur'anic recitation. The Arabic voice recognition was studied by Satori et al. (2007) by using the CMUSphinx toolset. The researchers utilized 300 spoken words from six Moroccan male speakers and reported an impressive mean recognition ratio of 86.66%. They showcased a demonstration of an Arabic voice recognition program that included greetings expressed in Arabic digits. A similar speech recognition technique was proposed by Tabbal et al. (2006) to differentiate between Qur'anic verses. Utilizing the CMUSphinx toolset, an automatic verse-level cutting system was developed. The study revealed that the recognition ratios for reciters in both Tajweed and Tarteel styles were 90% and 92%, respectively, while the ratio for regular Arabic speakers was 85%.

TABLE 1 Grammatical NLP work for the Qur'an.

| Sr | Authors | Paper title | Limitation/objectives | Techniques/outcome |
|----|-----------------------------|--|--|--|
| 1 | Dukes et al. (2013) | Supervised collaboration for syntactic annotation of Qur'anic Arabic | Insufficient oversight of the Qur'anic syntactic tasks | Supervisors kept an eye on the annotation |
| 2 | Fadlun (2016) | The Analysis of Conjunctions as Cohesive Device in the Qur'anic Translation of Surah Al- Baqarah | Reviews the literature on cohesion, conjunction, translation, and corpus linguistics | The AntConc software was used for corpus analysis to generate concordances, collocates, and frequency lists of Conjunctions in the translation |
| 3 | Bentrcia et al. (2018) | An analytical study of the holy Qur'an based on the order of words in Arabic and conjunction | Recognizing the reasoning behind the Qur'anic word placement around conjunctions | Assessed the arrangement of adjectives, pronouns, and nouns that occur with "AND" |
| 4 | Dukes and Buckwalter (2010) | A dependency treebank of the Qur'an using traditional Arabic grammar | Absence of a manually checked treebank and Qur'anic corpus | Introduced the Qur'anic Corpus and the QADT |
| 5 | Dukes et al. (2010) | Online visualization of traditional Qur'anic grammar using dependency graphs | The Qur'an has no syntactic annotation | Methodical, rule- based, cooperative approach to the syntactic annotation of the Qur'an |
| 6 | Dukes and Habash (2010) | Morphological annotation of Qur'anic Arabic | The Qur'an lacks morphological annotation | Methodical, cooperative approach to the Qur'anic morphological annotation |
| 7 | ELAffendi et al. (2021) | A simple Galois power- of-two (GPOW2) real-time embedding scheme for performing Arabic morphology deep learning tasks | Forecast of POS for the Qur'an | Applied GPOW2 to neural network- based context-based POS prediction |
| 8 | Mannaa et al. (2022) | Computer- assisted i'raab of Arabic sentences for teaching grammar to students | Arabic sentence structure (I'raab) | Improved syntax without context |
| 9 | Khadhim and Chiad (2022) | Parsing Facets of Letters Inflicted by the Three Diacritics in the Holy Qur'an | Large, annotated corpora, as well as the precision and tact needed when working with sacred texts | Examined the different efforts of Qur'anic NLP, as well as their strengths and weaknesses |
| 10 | Abouelnour et al. (2023) | Linguistic structure and its impact on diversifying positive discourse in the Holy Qur'an (A Semantic Grammatical Study) | The textual complexity, diverse interpretations, as well as ethical and religious sensitivities, make it challenging to deal with due to the lack of standardized datasets | Extended some positive linguistic discourse models from the Qur'an and its compositional techniques to interpret their meanings and linguistic ramifications |

Razak et al. (2008) outlined the four primary phases of speech recognition, namely, pre-processing, feature extraction, training, and testing. Pre-processing employs methods such as Endpoint Detection, Noise filtering, Smoothing, and Channel Normalization to enhance the readability of audio data. Employing spectrographic analysis, Mel-Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction, and Linear Predictive Coding, features are extracted. Training and testing employ Hidden Markov Model (HMM), Artificial Neural Network (ANN), and Vector Quantization (VQ). The authors pointed out that HMM and MFCC emerged as the most effective training approach and feature extraction method, respectively. In a similar vein, Mohammed et al. (2015) described a four-step procedure for using voice recognition techniques to authenticate Qur'anic recitations: input preparation, feature extraction, audio match, and training. In particular, the matching step is crucial to ascertain validity.

In another study, Ahmed and Abdo (2017) proposed a method for Qur'anic verse verification by leveraging audio recordings from reputable academics. They highlighted the approaches prescribed by Razak et al. (2008). Their approach involved analyzing audio recordings from reputable scholars and aligning the audio signals with the corresponding Qur'anic text. To improve Qur'anic learning, Jamaliah Ibrahim et al. (2013) introduced the Tajweed checking system that applies the speech recognition technology;

HMM for classification and MFCC for feature extraction. It demonstrated efficacy for teaching when tested on Surah Al Fatiha by achieving recognition rates of 86.41% at the phoneme level and 91.95% at the verse level. A tajweed verification system was initiated by Ahsiah et al. (2013) to support the learning of Qur'anic recitation. The algorithm detects discrepancies and suggests errors in recitations by comparing them with those of experts. The HMM is used for feature categorization and MFCC for feature extraction. Consonants in the Qur'an that produce the prosodic effect of Qalqalah (vibration) have been the subject of many recent studies. Words with Qalqalah effect can be found by using NLP techniques via a tool called Semantic Pathway developed by Brierley et al. (2014).

In a study, Al-Bakeri and Basuhail (2017) developed an ASR and integrated it into a self-study environment for tajweed testing. To train the acoustic model, they employed HMM, MFCC for feature extraction, and Gaussian mixture density for state emission probabilities. They achieved a phoneme accuracy of 89.47% by using Surah Al-Ikhlās and Surah Al-Rahman as input. Next, Mohammed et al. (2018) assessed the length of phonemes in Gunna and Madd letter characters included in Tajweed guidelines for Qur'anic recitation. To detect error in the rules, data were gathered via expert recitations. Ten recitations yielded 600 words for the study, from which the mean time of each Madd and Gunna was computed.

Meanwhile, e-Hafiz is a clever technique designed by [Muhammad et al. \(2012\)](#) for memorizing and reciting the Qur'an. To quantify accuracy, the system makes use of expert corrections and MFCC feature extraction. The accuracy rates of the system were 92% for men, 90% for children, and 86% for women. [Abro et al. \(2012\)](#) used speech recognition algorithms to automate the memorizing of the Qur'an. They employed ANN for pattern recognition and acoustic modeling, as well as MFCC for feature extraction. Their trials, however, failed to distinguish errors from accurate answers, signifying the insufficiency of this straightforward method.

[Yekache et al. \(2012\)](#) developed a method to control the Qur'an reader via spoken commands by employing speech data from 114 surah and reciter names. To enhance Qur'anic learning, [Putra et al. \(2012\)](#) combined speech recognition software with learning tools. With a limited dataset, their performance was subpar despite achieving 90% accuracy for reciting law and 70% accuracy for pronunciation. Next, [El Amrani et al. \(2016b\)](#) explored ASR for Qur'anic recitation using the CMUSphinx toolkit. They applied a simplified set of phonemes to develop an ASR system for the entire Qur'an by utilizing 40-min audio data derived from male speakers. In addition, they developed a list of phonetic vocabulary for a speech recognition system that focuses on simplified Arabic phonemes. Their approach involved using audio recordings of renowned recitations of the first and last three surahs of the Qur'an. The outcomes resulted in a phone list, transcription file, and phonetic dictionary. Notably, their experiments achieved a low word error rate of 50.0% and 55.7% when utilizing 90% and 80% of the audio data, respectively.

A computer-aided training method for Qur'anic recitation was initiated by [Tabbaa and Soudan \(2015\)](#). The system recognizes recitation phones using HMM-based ASR and it trains two classifiers for stressed and unstressed letter pronunciation. The system recorded 97.6% word-level accuracy, but after adding a classifier, the performance improved to 91.2% due to reduced false positives and false negatives. The 60-minute audio data were generated from 18 female and 14 male reciters. [Qayyum et al. \(2018\)](#) used a Bidirectional Long Short-Term Memory (BLSTM) model in conjunction with a deep learning technique to identify Qur'an reciter. With five individual reciters, the model was able to extract MFCC features with an accuracy of up to 99.89% for 3 seconds of recitation. This outperformed other robust models, such as Support Vector Machine (SVM), which remained below 90%. The study amplified its significance to reciting the Qur'an because even a slight modification can affect its meaning. Similarly, [Mahmudin and Akbar \(2023\)](#) used Deep Speech, a deep learning-based speech recognition model, to introduce a Qur'an recitation correction system and assess its efficacy using several metrics. The authors trained the Deep Speech model on 172,895 voice samples of the Qur'an using MFCC features. They also applied MaLSTM (Manhattan Long Short-Term Memory) and Siamese-Classifiers models to compare the similarity between the two audio verses. Additionally, [Harere and Jallad \(2023\)](#) presented a novel end-to-end deep learning model for the difficult task of identifying Qur'an recitation characterized by its rules and variations. The authors used a character-based decoder (a beam search decoder), a CNN-Bidirectional, and a Gated Recurrent Unit (GRU) encoder that uses connectionist temporal classification (CTC) as an objective

function. The prescribed model was embedded with the recently released Ar-DAD public dataset that includes 37 chapters of the Qur'an read aloud by 30 reciters with various accents and pacing. On the other hand, [Hadwan et al. \(2023\)](#) introduced a novel end-to-end transformer-based model with Arabic discretized text as the output for Qur'anic ASR from reciters. The transformer-based model, which comprises of an encoder, a decoder, and an attention mechanism, was implemented by using the Espresso toolkit. Multi-head attention, positional encoding, and Convolutional Neural Network (CNN) layers were used to extract and represent features. As for language modeling, both RNN and LSTM were used to decode shallow fusion. The study gathered a dataset of 10-h Qur'anic recitations, along with transcripts for training and assessment. The same classification was used in [Hadiyansah and Andamira \(2023\)](#), who developed a CNN model to classify and identify readings of the Al-Qur'an that can enable a Hafiz to memorize the Qur'an independently. Data from a webpage were preprocessed using Librosa library and the model was trained using 80% of the data.

Meanwhile, [Kuppusamy and Eswaran \(2020\)](#) developed a speaker recognition system by embedding CNN and deep neural network (DNN) based on age-related characteristics. The study categorized speech signals into distinct age groups by extracting robust and discriminative characteristics and applying a variety of classifiers. The researchers introduced a four-step approach: CNN and DNN for feature extraction; weighted feature fusion; Gaussian Mixture Model (GMM), SVM, and GMM-SVM classification; and score-level fusion using a linear combination of probabilities. The study took advantage of upgraded bottleneck features to enhance the performance of the DNN model.

In [Samara et al. \(2023\)](#), the dataset of seven reciters was tested. Using an 80-minute recording file, data from each reciter were compared with various deep learning and machine learning (ML) techniques, including SVM, random forests (RF), and ANN. The authors found a positive association between the number of features and accuracy, while the opposite relationship for root mean square error (RMSE). In addition, [Alkhateeb \(2020\)](#) prescribed an ML method based on audio analysis to identify Qur'an reciter. Features from 10 reciters were extracted by using MFCC, as well as ANN and K-nearest neighbor (KNN) classifiers. The developed system outperformed KNN and scored high recognition rates. A similar strategy was undertaken by [Al Anazi and Shahin \(2022\)](#) by introducing an ML model embedded with ANN and KNN classifiers to identify Qur'an reciters. The efficacy of the system was contrasted with SVM. The results showed that ANN outperformed SVM for voice recognition.

[Nahar et al. \(2020\)](#) discussed the difficulties faced when distinguishing among various regional accents in Arabic communities. They clarified that the pronunciation and meaning of the words found in the Qur'an are affected by the type of narration or Qira'ah that a Qur'an reciter follows. [Larabi-Marie-Sainte et al. \(2022\)](#) developed a novel Arabic recitation system named Samee'a by embedding Jaro Winkler algorithm and speech recognition. The system is meant to ease its users to learn and memorize Arabic texts, including speeches, poetry, and the Qur'an. The system measures the similarity between the original and converted texts using the Jaro Winkler Distance algorithm and later converts Arabic speech to text using the Google Cloud Speech Recognition API. To reduce noise

and improve the outcomes, the system performs a few preprocessing operations on the text files. The system offers a tool for reciting texts in Arabic, not just the Qur'an. Regardless of age or proficiency level, Arabic speakers can use the system as a self-learning tool.

Moving on, [Al-Jarrah et al. \(2022\)](#) introduced a method that uses a codebook matching procedure and an enhanced VQ technique to identify Qur'an reciter. The feature vectors of the speech signals were extracted in the study via MFCC. A modified version of the Linde–Buzo–Gray (LBG) algorithm was used to cluster the feature vectors into a finite number of centroids. Next, the squared Euclidean distance (SED) was applied in the study to determine which reciter scored the lowest matching error by comparing the centroids of several codebooks. Four verses of Surah Al-Kawthar and 14 expert reciters were the sole input data for the study. Apparently, the prescribed model performed the others for effective speaker recognition for Arabic and Qur'an recitation. [AlTalmas et al. \(2022\)](#) developed a classification-based method using ML techniques and a variety of audio characteristics to identify mispronounced Sifaat with opposites in Qur'anic letters. The study employed the Relieff algorithm for feature selection, while MFCC, PLP, and a combination of the two as feature extraction techniques. In addition, the study evaluated several classification models for every group of Sifaat with opposites, including KNN, ensemble random under-sampling (RUSBoosted), and SVM. The study presented the top-performing models, as determined by certain criteria: accuracy, recall, precision, and F1-score, for each Sifaat group. The authors recommended that future work should use deep learning approaches and increase the number of samples for underrepresented classes to improve performance. The prescribed method can be included in an automated real-time assessment system for pronouncing Qur'anic letters, as well as aiding both the students and instructors with Qur'anic recitation instruction.

[Balula et al. \(2021\)](#) presented an overview of the literature on ASR systems for Al-Qur'an recitation and Arabic language learning. The study tapped into the difficulties, uses, methods, and earlier research work in this area. It suggests an approach based on MFCC and HMM to confirm Qur'an recitation. The primary phases of ASR systems, including preprocessing, feature extraction, training, testing, and feature classification, are covered in the study. It also describes how to pronounce Arabic words and sounds, along with Tajweed principles, for reciting the Qur'an. Although ASR systems are helpful and promising for teaching Arabic and memorizing the Qur'an, they still require more work and development to address their obstacles. The authors listed some recommendations for future work, including the use of hybrid models, deep learning, big and varied datasets, more Tajweed rules, and effective feedback systems.

In another study, [Aljohani and Jaha \(2023\)](#) introduced a new Arabic lip-reading dataset that is based on the book written by Al-Qaida Al-Noorania. Here, a deep learning model is trained to recognize Arabic words and letters visually. Three digital cameras were used to capture films of 22 reciters uttering 10 Qur'anic words, 14 separated letters, and 29 single letters from three distinct viewpoints. Preprocessing techniques such as frame scaling, data augmentation, grayscale conversion, and face- and-mouth detection were used. Their method involved applying transfer learning by using a three-layer GRU as the backend and a ResNet-18 model as the front end. The model was comparable to the English and Mandarin lip-reading models, and it performed better than the

previous lip-reading models for Arabic. The subsection is summarized in [Table 2](#).

2.3 Semantic technologies

In terms of Islamic law, the Qur'an is the most important religious text written in Arabic. Numerous Qur'anic search engines have been developed and are in widespread usage over the last 20 years. Using semantic information, the primary goal of a semantic search engine for the Qur'an is to improve recall and precision. These search engines, however, are often plagued by drawbacks, such as only using keywords or root searches while omitting the semantic relationships between the words. [Sherif and Ngonga Ngomo \(2015\)](#) used Tanzil (a web-based resource) and QAC (provided morphological data) to create a database containing semantic datasets for the Qur'an in 43 languages. For information aggregation, morphological queries, and data retrieval, the dataset is connected to numerous web platforms. Meanwhile, [Recherche \(Al-Khalifa et al., 2009\)](#) introduced "SemQ," a web-based framework for semantic opposition analysis that makes use of semantic web technologies and NLP. Qur'anic English WordNet (QEWN), a semantic and lexical-based search system for the Qur'an, was developed by [Afzal and Mukhtar \(2019\)](#). The Vocabulary of Qur'anic Search (VQC) performs a unique search of Islamic concepts, including the "pillars of Islam," by storing a variety of concepts found in the Qur'an. By using the WordNet database for semantic search, [Shoaib et al. \(2009\)](#) tackled the issue of keyword-based search and achieved 80% accuracy.

[Fakhrezi et al. \(2021\)](#) summarized relevant documents while maintaining the original context. They applied a text summarization technique based on the TextRank method. Next, [Alargami and Eljazzar \(2020\)](#) introduced a novel approach by incorporating words from the Qur'an and Hadith into the word2vec technique, specifically tailored to Islam-specific knowledge. Meanwhile, [Alshammeri et al. \(2021b\)](#) utilized a simpler method by employing the Doc2Vec approach to identify comparable verses in the Qur'an with a high 76% accuracy rate. This technique facilitates semantic-based searches by establishing a degree of similarity between query questions and relevant verses. Using the query expansion technique, [Beirade et al. \(2021\)](#) achieved a 70% precision rate for semantic Qur'anic search. To illustrate the meaning or reasoning behind each verse, [Khazani et al. \(2021\)](#) presented a framework for semantic graphs for the Qur'an that makes use of word dependency within verses and POS tagging. To improve query results, this system creates semantic graph rules and POS tags for words. The AyaTEC is a series of Qur'anic tests created by [Malhas and Elsayed \(2020\)](#) to assess possible QA systems. It offers 207 verse-based questions on 11 themes to satisfy both skeptics and inquisitive users. The subsection is summarized in [Table 3](#).

2.4 Ontology work for the Qur'an

An ontology is a hierarchical concept dictionary that makes it easier for machines to comprehend and interpret concepts. It presents a methodical framework for knowledge organization and

TABLE 2 Speech recognition work for Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|--------------------------------|---|---|---|
| 1 | Tabbal et al. (2006) | Analysis and implementation of Qur'anic verses delimitation system in audio files using speech recognition techniques | Verse definition using speech recognition | Verses from different reciters are recognized by voice using the CMUs- Phinx program |
| 2 | Satori et al. (2007) | Introduction to Arabic speech recognition using the CMUSphinx system | The use of open-source CMU Sphinx-4 to develop an Arabic ASR System | A discrete HMM-based large-vocabulary speaker-independent continuous voice recognition system |
| 3 | Razak et al. (2008) | Qur'anic verse recitation recognition module for support in j-QAF learning: A review | The development of an ASR system based on the Sphinx framework | A comprehensive examination of Qur'anic verse recitation recognition methods, while addressing the benefits and limitations of each method |
| 4 | Muhammad et al. (2012) | E-hafiz: Intelligent system to help Muslims in recitation and memorization of Qur'an | The efficacy of E-Hafiz in correctly identifying the recitation of the Qur'an | The system compares recorded voices with expert voices using MFCC to detect word-level mismatches that need to be corrected |
| 5 | Abro et al. (2012) | Qur'an recognition for memorization using the Speech Recognition technique | The webpage suggests an intelligent system as a replacement for Hafiz | Provide learners with direction and feedback in the age of information technology |
| 6 | Yekache et al. (2012) | A Qur'anic reader is controlled by speech | The processes involved in creating an Arabi task-oriented continuous speech recognition system c | Data collection (corpus) based on CMU Sphinx4 was required to develop the acoustic model to be utilized in the Qur'anic reader's voice interface |
| 7 | Putra et al. (2012) | Developing speech recognition system for Qur'anic verse recitation learning software | Designed and created a prototype software for learning Qur'an verse recitation | Analyzing and identifying pronunciation errors for correction with the use of signal processing techniques (e.g., signal energy, delta MFCC, delta-delta MFCC, and MFCC coefficient extraction) |
| 8 | Jamaliah Ibrahim et al. (2013) | Automated tajweed checking rules engine for Qur'anic learning | The development of a software prototype that includes an automated Tajweed checking rules engine for Qur'anic learning | A recognition rate of 86.41% for phonemes and 91.95% for yates with HMM algorithm for recognition while MFCC for feature extraction |
| 9 | Ahsiah et al. (2013) | Tajweed checking system to support recitation | Accurate Tajweed verification | ASR for identifying errors during recitation via MFCC and HMM |
| 10 | Brierley et al. (2014) | Tools for Arabic NLP: a case study in Qalqalah prosody | focuses on finding Qalqalah events in the Qur'an dataset by using regular expressions to identify and gather them | Meccan and Medinan chapters of the Qur'an were substantially distinguished between Qalqalah content terms, indicating that Qalqalah prosody can emphasize important textual components |
| 11 | Mohammed et al. (2015) | Qur'anic verse verification using speech recognition techniques | Developed a mechanism that can prevent corruption and distortion of the Qur'an's integrity, especially in the internet media | After reciting 200 Qur'anic verses by 10 reciters, the algorithm detected mistakes (e.g., word order, punctuation, and pronunciation) with an accuracy of 95% |
| 12 | Tabbaa and Soudan (2015) | Computer-aided training for Qur'anic recitation | The system distinguishes between emphasized and non- emphasized utterances of the letter R and closely similar letters with the same articulation point using an HMM-based speech recognition engine with two classifiers | The algorithm achieved 91.2% word-level accuracy in correctly identifying and correcting recitation errors |
| 13 | El Amrani et al. (2016a) | Building a CMU Sphinx language model for the Holy Qur'an using simplified Arabic phonemes | Instead of using Romanized phonemes to indicate how Arabic words are pronounced in the Qur'an, diacritical Arabic characters were used | Using the Sphinx tools to train and test the language model, one can vary the number of Senones and the dimension of Gaussians, with 1.5% WER |
| 14 | El Amrani et al. (2016b) | Towards using CMU sphinx tools for Qur'an recitation verification | Reduce the complexity of developing a strong acoustic model for Qur'an recitation verification by considering Tajweed and Arabic rules | A phonetic vocabulary, a set of phonemes, transcription files, and audio files from the Qur'an chapters were included in the training data |
| 15 | Ahmed and Abdo (2017) | Verification system for Qur'an recitation recordings | An overview of voice recognition methods | An explanation of the many techniques used in speech recognition for Qur'anic recitation |
| 16 | Al-Bakeri and Basuhail (2017) | ASR for Tajweed rules: integrated with self-learning environments | MVC design and ASR to improve students' understanding of Tajweed and promote self-learning | ASR performance is measured by word error rate (WER). Both phonemes and syllables are the preferable training data due to their higher accuracy |

(Continued on following page)

TABLE 2 (Continued) Speech recognition work for Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|-----------------------------------|--|--|--|
| 17 | Mohammed et al. (2018) | Recognition of Qur'an recitation rules using phoneme duration | Corrects mistakes in Al-Medd and Ghunnah rules | Based on phoneme duration, the model determines the Medd and Ghunnah rules for reciting the Qur'an |
| 18 | Qayyum et al. (2018) | Qur'an reciter identification: A deep learning approach | The Qur'anic reciter's identity | 99.89% accuracy in voice identification and recognition using deep learning |
| 19 | Nahar et al. (2020) | An efficient Holy Qur'an recitation recognizer based on the SVM learning model | Pronunciation, punctuation, and word length are the primary areas where the 10 primary reading styles, or "Qira'at" diverge | Using ML methods to recognize recitations: To identify a voice or a reading style, various methods apply MFCC and learning algorithms (HMM, SVM, and ANN). SVM is the most effective method. |
| 20 | Alkhateeb (2020) | A machine learning approach for recognizing the Holy Qur'an reciter | Only two Qur'an chapters were used for system validation. The model is not contrasted with other current techniques or datasets. The analysis of mistakes or the variables influencing the recognition performance is absent | Excellent accuracy is achieved by the model when both classifiers were used. For chapters 18 and 36, the ANN scored 97.62% and 96.7% accuracy, respectively. For chapters 18 and 36, the KNN classifier scored 97.03% and 96.08% accuracy, respectively |
| 21 | Kuppusamy and Eswaran (2020) | Speaker Recognition System based on Age-related Features using Convolutional and Deep Neural Networks | The lack of generalization to other languages and dialects, the reliance on the quantity and quality of speech data, and the challenge of clearly establishing age boundaries | Significant improvements over the current systems were achieved by the new model, which had an accuracy of 82.5%, a recall of 63.0%, and WER of 1.0503. More use of bottleneck features and the combination of CNN-DNN features improved speaker age recognition |
| 22 | Balula et al. (2021) | Automatic Speech Recognition (ASR) Systems for Learning Arabic Language and Al-Qur'an Recitation: A Review | Faced difficulties with the Arabic language's morphology, phonology, discretization, and variety in pronunciation | Several feature extraction strategies (spectrographic analysis, PLP, and MFCC) and classification strategies (HMM, VQ, ANN, and SVM) |
| 23 | Al Anazi and Shahin (2022) | A machine learning model for the identification of the Holy Qur'an reciter utilizing KNN and ANN | The model analyzed two chapters and ten reciters for assessment. The approach omitted the reciters' varying reading preferences and emotional characteristics | The model used KNN (97.02% and 96.07% for chapters 7 and 32) and ANN (98.5% and 97.2% for two chapters) to achieve excellent accuracy. In recognizing the voice of reciter, the model performed better than SVM. |
| 24 | Larabi-Marie-Sainte et al. (2022) | A new framework for Arabic recitation using speech recognition and Jaro Winkler algorithm | The system dismissed Arabic pronunciation variances or the Tajweed criteria | Seventy files and some chapters from the Qur'an were used to test the system. The average similarity for files was 83.33% and 69.0% for chapters. After the preprocessing techniques, these scores improved to 91.33% and 95.66% |
| 25 | Al-Jarrah et al. (2022) | Accurate Reader Identification for the Arabic Holy Qur'an Recitations Based on an Enhanced VQ Algorithm | The input data consisted of four Surah Al-Kawthar verses and 14 expert reciters. The suggested model is not contrasted with other approaches currently in use, such as ANN or HMM. | The new model had a recognition accuracy rate of 85.11% (8 centroids) and 98.21% (16 centroids) for the codebook. It outperformed the k-means clustering algorithm in terms of precision and recall |
| 26 | AlTalmas et al. (2022) | Characteristics with opposite of Qur'anic letters mispronunciation detection: a classifier-based approach | Lack of a database with Tajweed-compliant pronunciations for the letters in the Qur'an | The approach based on categorization effectively identified Sifaat mistakes, except for groups whose data distribution was unbalanced |
| 27 | Mahmudin and Akbar (2023) | Qur'an Recitation Correction System Using Deep speech | The model did not give good generalization to other sections of the Qur'an because it was trained and evaluated on a small dataset of Juz 30. The model relied on human error and bias in the manual labeling of verse pairs | The best model performed well in Qur'an recitation correction, with an accuracy of 0.91 on the test set and 0.88 on the inference set |
| 28 | Hadiyansah and Andamira (2023) | Convolutional Neural Network (CNN) for Detecting Al- Qur'an Reciting and Memorizing | The data source was limited to a single surah. Besides omitting differences in reader dialect, intonation, and pronunciation, the model did not offer criticism or rectification of reading mistakes | The algorithm found 57 verses out of 64 data points with an accuracy of 89.06% The concept can be included in a program to aid with memorization of the Qur'an in the absence of the internet. |
| 29 | Harere and Jallad (2023) | Qur'an Recitation Recognition Using End-to-End Deep Learning | The model omits error detection or correction but focuses on recognition task Should refrain from using word- or sub-word-based decoders and instead use character-based ones. The model should not | The performances of the model on the Ar-DAD dataset were 8.34% WER and 2.42% CER, indicating promise in the field of Qur'an recitation recognition |

(Continued on following page)

TABLE 2 (Continued) Speech recognition work for Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|--------------------------|--|--|---|
| | | | be compared to other cutting-edge models using the same dataset. | |
| 30 | Hadwan et al. (2023) | An End-to-End Transformer- Based Automatic Speech Recognition for Qur'an Reciters | Lack of accessible datasets, difficulty in pronouncing words in Arabic and their diacritical marks, and the variety of Qur'anic recitation conventions and styles | Surpassed both earlier techniques and cutting-edge models for Qur'an reciter recognition, as the new model achieved a low character error rate (CER) of 1.98% and a WER of 6.16% |
| 31 | Aljohani and Jaha (2023) | Visual Lip- Reading for Qur'anic Arabic Alphabets and Words Using Deep Learning | The dataset covered only the content and context of the Al-Qaida Al-Noorania book; data from other sources would be helpful | For Qur'anic words, the new model had an accuracy of 83.3%, while 80.5% and 77.5% for separated and single letters, respectively |
| 32 | Samara et al. (2023) | The Recognition of Holy Qur'an Reciters Using the MFCC Technique and Deep Learning | Provide a deep learning model to identify the reciters of the Qur'an by employing CNN and MFCC. The study assessed how performance relates to data volume for various strategies | Using a dataset of seven renowned reciters, the developed model was contrasted with other ML and deep learning techniques (RF, SVM, and ANN). The developed model had 99.66% accuracy and SVM 99.0%; signifying superior performance of the former. |

TABLE 3 Semantic work of the Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|--------------------------------|---|--|--|
| 1 | Shoaib et al. (2009) | Relational WordNet model for semantic search in the Holy Qur'an | Semantics are lacking in Qur'anic keyword searches | English verses from the Al-Baqarah chapter were retrieved using different levels of synonymy |
| 2 | Al-Khalifa et al. (2009) | SemQ: A framework for representing semantic opposition in the Holy Qur'an using Semantic Web technologies | An examination of the Qur'an using semantic opposition | Creation of "SemQ" to conduct Qur'anic verse semantic opposition analysis |
| 3 | Sherif and Ngonga Ngomo (2015) | Semantic Qur'an: Semantic Web | Word meanings in the Qur'an | Creation of a semantic database for the Qur'an and its interaction with internet platforms |
| 4 | Afzal and Mukhtar (2019) | Semantically enhanced concept search of the Holy Qur'an: Qur'anic English WordNet | Semantics missing from Qur'anic keyword-based searches | Creation of the QEWN database and the VQC (Qur'anic vocabulary) search engine for the Qur'an using word synonyms |
| 5 | Alargami and Eljazzar (2020) | Imam: Word Embedding Model for Islamic Arabic NLP | Islamic knowledge in a semantic framework | The word2vec embedded vocabulary unique to Islam |
| 6 | Malhas and Elsayed (2020) | Ayatec: building a reusable verse- based test collection for Arabic question answering on the Holy Qur'an | A Benchmark dataset for the Qur'anic QA system is absent | The AyaTEC dataset and assessment of a quality assurance system that yielded Qur'anic versus |
| 7 | Fakhrezi et al. (2021) | Implementation of automatic text summarization with the TextRank method in the development of the Al-Qur'an vocabulary encyclopedia | Absence of short responses to subjects covered in the Qur'an | Applied semantic cosine search in conjunction with the text rank algorithm |
| 8 | Beirade et al. (2021) | Semantic query for Qur'anic ontology | Qur'anic semantic exploration | Using ontologies and the query expansion technique |
| 9 | Khazani et al. (2021) | Semantic graph knowledge representation for Al-Qur'an verses based on word dependencies | Search the Qur'an semantically | Creation of a framework for semantic graphs |
| 10 | Alshammeri et al. (2021b) | Detecting semantic-based similarity between verses of the Qur'an and Doc2vec | Improved semantic outcomes | Vectorization of verses with Doc2Vec |

representation to enable more efficient machine-human communication (Stevens et al., 2010). In a study, Al-Yahya et al. (2010) developed an ontology for the Qur'an that categorized words in a hierarchy with 18 conceptual classes defined for time field, which contributed to the Arabic computational lexicon. To provide semantic results for queries based on Qur'anic words and their meanings, Beirade et al. (2021) introduced an ontology for the Qur'an. Next, Iqbal et al. (2013) initiated an ontology utilizing a genuine Qur'anic corpus, offering verse context derived from Hadith and Tafsir. An ontological method for semantic Qur'anic search—specifically for the animal domain—was put forth by Khan et al. (2013). Al-Qur'an is a theme-based ontology, as

introduced by Ta'a et al. (2017), which includes three primary classes: Allah, Angels, and Unseen. The SQL queries can be used to retrieve Qur'anic data stored in a database. In their examination of Qur'anic ontologies and Qur'anic Search tools, Alqahtani and Atwell (2016) highlighted several challenges. These included the absence of effective solutions, the limited vocabulary found in Arabic WordNet, and the incomplete linkages between verses and their associated concepts. To address these issues, they proposed the Arabic Qur'anic Semantic Search Tool (AQSSST) by incorporating a comprehensive database comprising of Tafsir, eight English translations, the original Qur'anic text, and semantic tagging. Notably, it integrates four distinct ontologies by

TABLE 4 Ontologies work for the Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|-----------------------------|--|--|---|
| 1 | Al-Yahya et al. (2010) | An ontological model for representing semantic lexicons: an application on time nouns in the Holy Qur'an | Temporal domain | A hierarchical approach for locating Qur'anic synonyms and antonyms |
| 2 | Iqbal et al. (2013) | An experience of developing Qur'an ontology with contextual information support | The 30th Division in the Qur'an | Verses' genuine contextual meaning is provided by ontology; both English and Malay translations are offered |
| 3 | Khan et al. (2013) | Ontology-based semantic search in the Holy Qur'an | Animals | A framework to answer semantic questions about animals in the Qur'an |
| 4 | Ahmed and Atwell (2016) | Developing an ontology of concepts in the Qur'an | Incomplete Arabic language ontology and Holy Qur'an ontology | Ontology development with the combination of preexisting, Qur'anic original text extraction, and domain experts' assistance |
| 5 | Alqahtani and Atwell (2016) | Arabic Qur'anic search tool based on ontology | Combining current ontologies | Combining current ontologies and creating the AQSST search engine to increase the amount of knowledge to avail |
| 6 | Ta'a et al. (2017) | Al-Qur'an ontology based on knowledge themes | Allah, Angels, and Unseen | An interface-based Qur'anic theme search system |
| 7 | Beirade et al. (2021) | Semantic query for Qur'anic ontology | Using ontology to locate semantic outcomes | Construction of ontology using terms and definitions from the Qur'an |

TABLE 5 Search work of the Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|---------------------|--|--|---|
| 1 | Nassourou (2012) | Towards a Knowledge-Based Learning System for The Qur'anic Text | Unable to offer thorough responses to complex questions because it does not possess a thorough grasp of the Qur'anic context, chronology or analytical abilities | Knowledge-based translation developed with contemporary technology to offer concepts or phrases with exact meanings |
| 2 | Hanum et al. (2013) | Evaluation of Malay grammar on the translation of Al-Qur'an sentences using Earley algorithm | Several of the sentences did not follow the accepted Malay grammar | Proposed a parser system for Qur'anic translation into Malay by using Earley's method to verify that the grammar is correct based on the accepted Malay grammar norms |
| 3 | Putra et al. (2017) | Text mining for Indonesian translation of the Qur'an: A systematic review | Word ambiguity: a single word may have several meanings, hence the right answers may not be provided for a given inquiry | Analyze questions and deliver answers, the SQA component Discovering Source Answer (DSA) uses text processing, POS tagging, and sentence identification |

enhancing the search experience for users (Ahmed and Atwell, 2016) and investigating the usefulness of applying merged ontologies to conduct a semantic search for Qur'anic abstract concepts. The subsection is summarized in Table 4.

2.5 Translation

Translating the Qur'an from its original Arabic text into other languages is a tedious task because it is full of subtleties and obstacles. Having said that, Nassourou (2012) initiated a conceptual method to address both morphological and lexical challenges in Qur'anic translation. The existing body of research has identified the following gaps: Lack of Tools, Context and Chronology, and Sophisticated Responses. In an attempt to bridge these gaps, Nassourou proposed a translation system that leverages the Qur'an and genuine Hadith as the knowledge base. Meanwhile, Putra et al. (2017) investigated ideas about Indonesian translation (ITQ), text mining, as well as searching and question-answering (SQA) applications. The SQA is an ITQ-based search engine that manages text processing, POS tagging, and sentence

detection. In another study, Hanum et al. (2013) proposed a parser system for the Qur'anic translation into Malay by applying Earley's algorithm to verify, the accuracy of the grammar is based on the standard Malay grammar rules. In total, 115 rules are embedded into the new set, in comparison to 94 in the original standard Malay grammar. The subsection is summarized in Table 5.

2.6 Search systems and Qur'anic NLP

Both the Qur'anic NLP and the search systems play a crucial role in making the Qur'an accessible and comprehensible to a broader audience. By applying the query expansion techniques, Hammo et al. (2007) introduced a Qur'an search engine that stores words in three RDBMS indices: Vowelized-Word Index, Non-vowelized-Word Index, and Root Index. While the Non-Vowelized Index eliminates diacritical marks, the Vowelized Index retains unique Qur'anic words without preprocessing. By stemming every word to its base, root words that are created are kept in the Root Index. Synonyms for words found in the Qur'an are listed in a database thesaurus. The outcomes derived from three experiments

TABLE 6 Knowledge extraction of NPL of Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|-----------------------------|--|---|---|
| 1 | Hammo et al. (2007) | Effectiveness of query expansion in Searching the Holy Qur'an | Reduced precision in the Qur'anic verses search method using a word | Tokenization, query expansion, stemming, and word storage in three distinct indices are features of search engines |
| 2 | Yunus et al. (2010) | Visualizing Qur'an documents results by stemming semantic speech query | The ability to respond to inquiries in English, Arabic or Malay; preprocessing is necessary to translate the inquiry into an identical query in the target language | A Qur'anic search framework is proposed by applying stemming and semantic analysis of the query to accomplish CLIR |
| 3 | Khan et al. (2013) | Ontology-based semantic search in the Holy Qur'an | Semantic search engine built on an ontological framework | An ontological method is proposed for the animal domain |
| 4 | Alqahtani and Atwell (2016) | Arabic qur'anic search tool based on ontology | Lack of a detailed ontology, concept-based search, and NER for the Qur'an | The AQSST tool is proposed for ontology- based search supported by keyword search |
| 5 | Alqahtani and Atwell (2017) | Evaluation criteria for computational Qur'an search | Limitations of the existing Qur'anic search tools such as, failing to apply modern NLP techniques (e.g., parsing or spell check) | Use of both text-based and semantic search technologies to assess common standards, including search characteristics, result precision, database size, query and answer types, response speed, and contents of database |

TABLE 7 Classifier system of the Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Technique/outcome |
|----|-------------------------|---|---|---|
| 1 | Al-Kabi et al. (2005) | Statistical classifier of the Holy Qur'an verses (Fatiha and Yaseen chapters) | Automatically classifies text in the Qur'an | Microsoft Visual Basic was used in the tool development and the method is straightforward |
| 2 | Ta'a et al. (2013) | Al-Qur'an themes classification using ontology | Discusses two primary themes of the Al-Qur'an, faith (Iman), and deed (Akhlaq), along with their sub-themes. The ontology is constructed in Malay, a language some users might not comprehend. Practical application and assessment are omitted in the ontology | An ontology of Qur'anic knowledge is proposed based on verses, themes, and sub-themes. Domain experts validated the ontology after it was built with Protégé-OWL and the Malay language |
| 3 | Rostam and Malim (2021) | Text categorization in Qur'an and Hadith: Overcoming the interrelation challenges using ML and term weighting | Text categorization is used to group topics by analyzing the associations between Hadith and Qur'anic sources. Mere use of three categories might not fully represent the range of Islamic concepts | Classified texts from the Qur'an and Hadith into three categories (Hajj, Prayer, and Zakat) using supervised ML (SVM, NB, and KNN) with term weighting (TF-IDF). The accuracy and performance of each approach was later compared |

significantly improved after stemming and root index searching. Using semantic and stemming analyses, Yunus et al. (2010) introduced a Qur'anic search framework that can perform CLIR, translate user queries into equivalent target language sentences, and return results in Arabic, Malay or English. In another study, Alqahtani and Atwell (2017) assessed a range of Qur'anic search methods and resources, with an emphasis on text-based and semantic search. An ontologically-based semantic search tool covering Qur'anic references to animals and birds was proposed by Khan et al. (2013). The Arabic Qur'anic Semantic Search Tool was first presented by Alqahtani and Atwell (2016). The tool can resolve ambiguities in results and offer comprehensive outcomes that include the concepts of each word. The subsection is summarized in Table 6.

2.7 Classifier systems and Qur'anic NLP

An automatic text categorization tool for the Qur'an was initiated by Al-Kabi et al. (2005). It groups verses based on topics covered by Islamic scholars, such as faith, prayer, and pilgrimage. The tool recorded a 91% accuracy rate on Microsoft

Visual Basic. In another study, Ta'a et al. (2013) proposed an ontology to group Qur'anic verses according to themes, with sub-themes accessible within the classes. To determine the association between the Qur'an and the Hadith, as well as to classify texts using Naive Bayes (NB), KNN, and SVM, Rostam and Malim (2021) conducted ML experiments. As a result, the SVM outperformed other methods when classifying Qur'anic verses and identifying related topics. The subsection is summarized in Table 7.

2.8 Topic extraction for Qur'an

A statistical technique called linear discriminant analysis (LDA) has been deployed in many domains to determine a linear feature combination that distinguishes or describes two or more classes of objects or occurrences. This approach was employed by Atwell (2015) for topic modeling in the Qur'an corpus and the outcomes revealed it superior performance to K-means clustering in earlier research work. Most of the topics included a combination of multiple topics and the results with roots were unacceptable. Meanwhile, Hassan et al. (2015) trained the KNN algorithm on data and then labeled it based on similarity with neighboring data to

TABLE 8 Topic Extraction work in Qur'an.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|---|--|--|--|
| 1 | Yauri et al. (2012) | Qur'anic-based concepts: Verse relations extraction using Manchester OWL syntax | The subjects in Qur'an were categorized using ontology | A framework was developed by using the already-built Leeds University ontology that only discussed nouns found in the Qur'an |
| 2 | Atwell (2015) | A Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus | The LDA model was ineffective for the Qur'an | Based on the original words of the Surah, their stems, and their roots, several experiments were conducted using versions of LDA |
| 3 | Hassan et al. (2015) | Categorization of 'Holy Qur'an-Tafseer' using K- nearest neighbor algorithm | A supervised ML approach was deployed for subject classification | The KNN method was applied to classify the text of the Malay translations of passages from the Qur'an |
| 4 | Alshammeri et al. (2021a) | Qur'anic topic modeling using paragraph vectors | Identifies features for K-Mean clustering algorithm | The semantically-rich topics in the Qur'an were achieved with the aid of clusters based on document-level vectors |

TABLE 9 Knowledge extraction for Qur'anic NLP.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|---|--|--|--|
| 1 | Yauri et al. (2012) | Qur'anic- based concepts: Verse relations extraction using Manchester OWL syntax | The system lacks a user-friendly interface that would allow a query to be input in normal language, although complicated queries can be executed | Proposed an ontology-assisted semantic search by utilizing Manchester OWL syntax to model Qur'anic domain knowledge. The system achieved an impressive 95% accuracy level in semantically searching for verses related to user queries in the Qur'an |
| 2 | Saad et al. (2013) | Rules and Natural Language Patterns in Extracting Qur'anic Knowledge | Generating the ontology in the Web Ontology Language (OWL) format using Protégé, an ontology editor tool | Three levels were the basis of the ontology development <ul style="list-style-type: none"> - The Qur'anic meta concept - Salah's domain ontology layer - A last layer that connects the second and first levels |
| 3 | Siddiqui et al. (2013) | Discovering the thematic structure of the Qur'an using a probabilistic topic model | Identified the Qur'anic theme structure by applying an unsupervised ML technique called Latent Dirichlet allocation (LDA) | The Meccan and Medinan Chapters of the Qur'an were accurately categorized by the LDA model based on word differences, signifying its efficacy for topic-based knowledge extraction from the Qur'an |
| 4 | Mohamed et al. (2015) | Al-Bayan: A knowledge- based system for Arabic answer selection | The model includes a supervised ML approach (decision tree classifier), a Qur'an ontology, and a keyword matcher | The system uses a decision tree classifier to determine the semantic relevance of the response vectors to the question vector after converting questions and potential replies to the vectors |
| 5 | Ahmed and Atwell (2016) | Developing an ontology of concepts in the Qur'an | This new ontology blends the following three ontologies to cover more topics: ontology from the Qur'an, ontology from Tafsir, and ontology from reviews published by experts on various verses | The PROMPT tool facilitated the merger of small- and medium-sized ontologies. The results were further enhanced by up to 85% |
| 6 | Ta'a et al. (2017) | Al-Qur'an ontology based on knowledge themes | The Al-Qur'an ontology was developed with Protégé-OWL and the use of a semantic search technique using a relational database to focus on three themes: Allah, Angels, and Unseen | Test cases and expert reviews were used to assess the accuracy of the search strategy and the validity of the ontology specification |

perform ML for topic categorization. Next, [Alshammeri et al. \(2021a\)](#) proposed document vectors based on original Qur'anic verses by using embeddings for topic modeling. This method was fed as features for a K-Mean clustering algorithm, which appeared more effective at retrieving semantically similar verses for a given verse. [Yauri et al. \(2012\)](#), on the other hand, used ontology—the most popular technique—for comprehension and interpretation. The subsection is summarized in [Table 8](#).

2.9 Knowledge extraction for Qur'anic NLP

In a study, [Saad et al. \(2013\)](#) outlined a set of guidelines to derive knowledge from the Qur'an by using a three-layer ontology and

English translation. Meanwhile, with an accuracy of 74.53%, [Siddiqui et al. \(2013\)](#) used unsupervised ML techniques to detect Qur'anic thematic structure. [Mohamed et al. \(2015\)](#) found answers to questions about the Qur'an with 74.53% accuracy using supervised ML approach, a decision tree classifier, and a keyword matcher. Next, [Yauri et al. \(2012\)](#) and [Ta'a et al. \(2017\)](#) applied ontology for knowledge extraction. By combining two ontologies, [Ahmed and Atwell \(2016\)](#) enhanced the ontology and achieved an 82% accuracy rate. By combining small- and medium-sized ontologies, the PROMPT tool enhanced results to 85%. Research and development on ontology-based and ML techniques for knowledge extraction from the Qur'an appear to be ongoing in a vigorous manner. The subsection is summarized in [Table 9](#).

2.10 Maqams of Qur'anic recitation

Maqams are melodic modes used in Qur'anic recitation. Eight maqams are applied in Qur'anic recitation, each with its unique melody and character ('Unlock the Secrets of 8 Maqamat of Qur'an Recitation! - KQ', no date). The maqams are employed to heighten the beauty of Qur'an recitation and to foster a sense of heavenly favor that strengthens the listener's bond with the Words of the God (Saleh, 2021; Jabar et al., 2018). The eight maqams are listed as follows.

1. Bayati: Invokes patience and spiritual serenity. It is typically used while reciting longer, story-focused passages.
2. Rast: Often employed when reciting passages about the majesty of Allah and scenarios from the day of judgment.
3. Hijaz: Evokes sorrow and anguish, particularly when reciting passages about loss, hardship or the difficulties faced by prophets.
4. Saba: Topics about cautions, wisdom, and reflections that serve as a reminder of the greater significance of events.
5. Sikah: Strikes a balance between fun and reflection. It imparts hope and lightness.
6. Nahawand: This maqam uses dark yet lovely tunes to depict profound sadness and tragedy.
7. Kurd: This maqam is for poetry that highlights the wonders of the natural world and our surroundings.
8. Adjam: It is frequently employed to recite poetry that discusses the characteristics of the human spirit and the value of introspection and introspection.

Shahriar and Tariq (2021) proposed a novel approach that uses deep learning models and audio cues to categorize the eight maqams of Qur'anic recitations. It presents a brand-new dataset of 478 audio files to assess how classification performance can be affected by audio features. Four deep learning architectures were compared in the study, with deep ANN displaying superior performance (a score of the highest accuracy rate at 95.7%). Meanwhile, Nahar et al. (2020) introduced an acoustic wave corpus by extracting MFCC characteristics, training an SVM-based model, and contrasting the outcomes with those of other ML techniques (e.g., Least Square-SVM (LSSVM) and ANN). Maqamat refers to the melodic modes or scales that alter the pitch, rhythm, and embellishment during recitation. The formant frequencies of the maqamat signals or the spectral peaks that define the vowel sounds can be obtained by comparing two types of cepstral analysis: MFCC and Warped Discrete Fourier Transform (W-DFT). In a study, Al Bakri and Mallah (2016) identified numerous sorts of human voices suitable for the performance of the Qur'an, explored the background and evolution of the Qur'an musical rendition, and provided guidelines and precepts of tarteel (the art of vocalization) and tajweed (the art of recitation). The subsection is summarized in Table 10.

3 Qur'anic NLP techniques

Several techniques employed in NLP to extract meaningful information from input data will be discussed. Task-specific methods are utilized, while fundamental techniques are

applicable in nearly every application. Further details are provided in each subsection.

3.1 Text preprocessing

Text pre-processing for NLP involving the Qur'an requires several processes. First, words are converted into symbolic vectors using word representation methods, which expose linguistic and semantic commonalities (Alsubhi et al., 2021). Subsequently, a new preprocessing method is suggested for pronoun mapping and extraction in the English translations of the pronominal anaphora resolution mechanism for the Qur'an (Touati-Hamad et al., 2020). This method achieves excellent precision, recall, and accuracy in pronoun extraction and mapping by using morphologic, statistical, and anaphoric knowledge acquired from the Arabic corpus of the Qur'an (Tabrizi et al., 2016). The prosodic-syntactic markup of the Qur'an is examined, and the tajwid markup is seen as further text-based data for computer analysis (Sawalha et al., 2017). To facilitate the statistical study of sound and meaning correlations, a grapheme-phoneme mapping method and software have been constructed for each word in the Qur'an to produce a stressed and syllabified phonemic transcription (Elbarougy et al., 2020). Lastly, preprocessing methods such as tokenization, normalization, structural processing, and stop-word removal can improve the performance of Arabic text summarization.

3.2 Text matching

The text-matching technique has been extensively employed in Qur'anic search systems. Text matching has several applications. Some software programs match the exact phrases in a verse to a user's question. The inquiry words and verse words can be reverted to their original forms in other applications, thus allowing them to be matched to any form they have been used in. In this method, the search window is expanded. Semantic activities relating to the Qur'an can benefit from the application of advanced kinds of text matching, where matches are not only established based on words but also on the concepts behind those words. A simple search, for example, may focus solely on a few search phrases and provide results that are not precisely what was intended. Here, a synonym or antonym can be used to increase the size of search window. It is not possible to retrieve concepts from any text by only matching words. Ontology, which matches words at various hierarchical levels, is applicable to retrieve conceptual results. Even while text matching can appear to be a relatively simple technique, with the right application, it can assist achieve amazing outcomes.

3.3 Clustering

Using clustering, one can amalgamate similar types of data together by utilizing an unsupervised ML technique. The absence of labels is present in unsupervised learning. To facilitate the process of categorization and knowledge extraction, comparable verse types in the Qur'an are located. Of all the existing clustering algorithms for

TABLE 10 Maqams of Qur'anic recitation.

| Sr | Authors | Paper title | Objective/limitation | Techniques/outcome |
|----|----------------------------|---|---|---|
| 1 | Al Bakri and Mallah (2016) | Musical performance of the Holy Qur'an with the assistance of the Arabic Maqams | The study describes the academic and emotional qualities of the most often utilized maqams in reciting the Holy Qur'an | The study assessed the relationship between Arabic cult music and philosophical legacy of the Middle Ages as well as the impact of ancient Arab philosophers on the theory and application of maqams |
| 2 | Nahar et al. (2020) | An efficient holy Qur'an recitation recognizer based on the SVM learning model | A novel technique that embeds SVM and MFCC to detect types of recitation among 10 approved ways of reading the Qur'an | The SVM-based model outperformed other models (e.g., ANN, LSSVM, Decision Tree, RF, and Naive Bayes) to identify different recitations by achieving a high accuracy rate of 96% |
| 3 | Shahriar and Tariq (2021) | Classifying maqams of Qur'anic recitations using deep learning | A new dataset of 478 audio files was initiated by using a novel approach to categorize the eight maqams of Qur'anic recitations | The study assessed the outcomes of four distinct deep learning architectures—CNN, LSTM, CNN + LSTM, and deep ANN. The greatest accuracy of 95.7% was scored by deep ANN with 26 input characteristics |

topic modeling, the Latent Dirichlet Allocation (LDA) is arguably the most widely used (Alhawarat, 2015). The LDA assigns multiple themes to each document based on the occurrence frequency of words across a given subject. Through the use of clustering analysis to the Qur'anic verses, the topics can be listed. Next, the K-means method classifies the verses found in the Qur'an, in addition to LDA. The K-means clustering method creates a specified number of groups based on how different the words from each verse are from the centroids of each group.

3.4 Classification

In contrast to clustering, classification is a supervised ML method utilized for knowledge extraction and Qur'anic theme modeling. Data in classification have labels. The KNN is a widely used algorithm for classification. There is a collection of verses in KNN that is already classified with specific subjects. Next, a distance algorithm determines how similar the test verses are to labeled verses to assign each verse to a topic. The alternative categorization model is the decision tree. Mohamed et al. (2015) claimed that the approach is applied for knowledge extraction and Qur'anic question answering. Training data are classified using decision tree algorithm based on input features. The labels are assigned to the testing data based on the rules learned by decision tree during training.

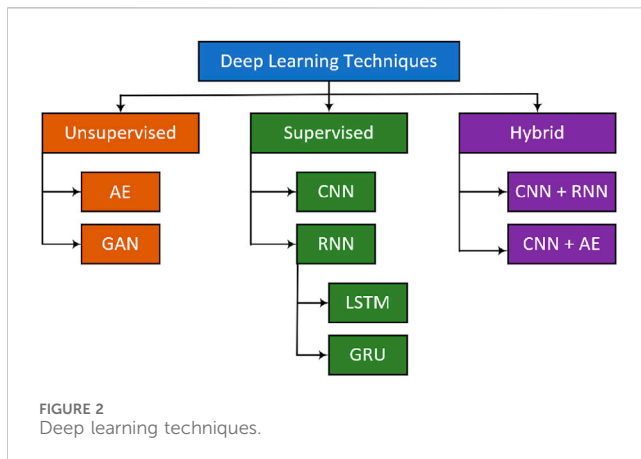
3.5 Speech processing

The Qur'anic NLP works section discusses the various applications of speech processing for accurate Qur'anic recitation. The natural feel of spoken languages is conveyed with the use of speech processing. It is crucial to recite the Qur'an based on specific guidelines. The MFCC is used in speech recognition to extract speech features as it can mimic the human auditory system (Chakraborty et al., 2008). Speech data are used to train and assess HMM. An HMM-based recognizer takes feature vectors built by MFCC as input and applies Bayesian probability to

detect words from phonemes trained over many pronunciations of each word (Gales and Young, 2008). The HMM is a state machine where each state in a certain period is connected with a distinct phoneme and a word is detected by observing the sequence of phonemes (Dimitrakakis and Bengio, 2011).

4 Critical analysis of existing NLP techniques in Qur'anic studies

The current landscape of natural language processing (NLP) techniques in Qur'anic studies shows significant progress and challenges in different areas. In grammatical analysis, rule-based parsing and real-time embedding improve syntactic understanding, but the lack of standardized data sets and the complexity of Arabic remains a major hurdle. Speech recognition benefits from high accuracy through models such as HMM and MFCC, although different recitation styles require large annotation files. Semantic work has progressed with tools such as WordNet and Doc2Vec, which help to remember and be accurate, but still faces complex relationships and limited resources. Ontology-based approaches effectively categorize Qur'anic concepts, but lack comprehensive data sets and face language problems. Translation systems deal with morphological complexity using algorithms like Earley's, but they require more sophisticated processing of the language. Search systems improve accessibility by using stemming and semantic analysis, but due to language complexity, deeper integration with NLP is necessary. Classifier systems use naive Bayes, KNN, and SVM to classify topics but require more robust data sets. The theme extraction is improved by LDA and clustering, but is still influenced by legacy complexity. Knowledge extraction combines ontology and ML techniques to produce meaningful insights, but it lacks robustness. Finally, the maqam of recitation is analysed using deep learning and cepstral features, although Arabic phonetics is a challenge. Overall, the trend shows an increasing use of deep learning, adaptive models, and semantic tools to overcome the Arabic language problems in Qur'anic NLP. When comparing the effectiveness of different NLP tasks for the Qur'an, the grammatical NLP and search systems are the most effective due to their comprehensive approach



and high accuracy in syntactic and morphological analysis, as well as improved search performance. Speech recognition also shows great promise in the accuracy of its recognition of recitation, although it faces problems related to the complexity of Arabic phonetics and the variability of recitation styles. Semantic works, ontologies and translation systems are effective in improving understanding and retrieving semantic information and translations, but they struggle with the complexity of Arabic structures and the need for advanced NLP techniques. Classifying systems, topic extraction, and knowledge extraction techniques are effective in categorizing and extracting meaningful information, but they are limited by the lack of comprehensive data sets and the sensitivity of religious texts. Finally, the analysis of maqam in Qur'anic recitation is moderately effective and improves the understanding of melodic modes, but it also faces difficulties due to the complexity of Arabic phonetics and the need for sophisticated techniques.

5 Deep learning

Deep learning (DL), a branch of machine learning (ML) and artificial intelligence (AI), is now considered to be the key technology of the fourth industrial revolution (4IR or Industry 4.0). Due to its ability to learn from data, Deep Learning (DL) technology, based on artificial neural networks (ANN), has become a hot topic in the computer world and is widely used in various applications, such as healthcare, visual recognition, text analysis, cybersecurity, and many others. This section discusses different types of deep learning approaches, which typically involve multiple layers of information processing steps in hierarchical structures to learn. The input and output layers are among the many hidden layers found in a typical deep learning algorithm. As shown in Figure 2, we categorize deep learning techniques into three main groups for taxonomy: deep learning broadly divided into three major categories: deep networks for supervised learning, deep networks for unsupervised learning and deep networks for hybrid learning.

5.1 Supervised or learning

Supervised learning is a machine learning approach defined by the use of labelled datasets to train algorithms to classify data and to

predict outcomes. A labeled dataset has outputs that match the input data so that the computer can understand what to look for in the invisible data. Supervised architectures include, in particular, convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These techniques are briefly discussed in the following sections.

5.1.1 Convolutional neural network (CNN)

Convolutional neural network (CNN), also known as ConvNet, is a specialised type of deep learning algorithm designed primarily for object recognition tasks, including image classification, image detection and segmentation (LeCun et al., 1998). Figure 3 shows an example of a CNN with multiple convolutions and pooling layer. As a result, CNN improves the design of traditional ANNs such as regularised ANNs. Each layer in CNN takes into account the optimal parameters for meaningful output, and reduces the complexity of the model. CNN also uses a sunset feature (Géron, 2019) to address the problem of overloading that may occur in a traditional network. CNNs are designed to handle a variety of 2D shapes and are therefore widely used in visual recognition, medical imaging, image segmentation, natural language processing and many other applications (Sarker, 2021). The capacity to automatically detect key features of a feed without any human intervention makes it more powerful than a traditional net. There are several variants of CNN, including visual geometry group (VGG), (He et al., 2015), AlexNet (Krizhevsky et al., 2012), Xception (Chollet, 2017), Inception (Szegedy et al., 2015), and ResNet (He et al., 2016), that used in various application domains according to their learning capabilities.

5.1.2 Recurrent neural network (RNN)

The Recurrent Neural Network (RNN) is a common class of neural network architecture that captures temporal dynamics in sequence or time-series data. RNNs are different from traditional feedforward networks or Convolutional Neural Networks (CNNs) in that they have an internal memory that allows the network to remember information about previous time steps. This is done through providing the output of a previous state as input of the current state in the same network, allowing the capturing of dependencies in sequences (Dupond, 2019). Standard Deep Neural Networks (DNNs) often assume that inputs/outputs are independent; however, the sequential nature of the RNN lends to appropriate modeling where the current output depends on elements that preceded it. A key drawback of traditional RNNs, however, is the vanishing gradient problem, making them less effective at learning long-range dependencies. Various enhanced architectures-like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) provide better performance in many real-world scenarios.

5.1.2.1 Long short-term memory (LSTM)

LSTMs are a popular variant of the RNN architecture that were proposed specifically to solve the vanishing gradient problem, which prevents standard RNNs from properly learning long term dependencies. Long Short-Term Memory (LSTM) networks were introduced by Hochreiter and Schmidhuber (1997), they include special type of memory units, also called memory cells, that allow information to be stored for long time

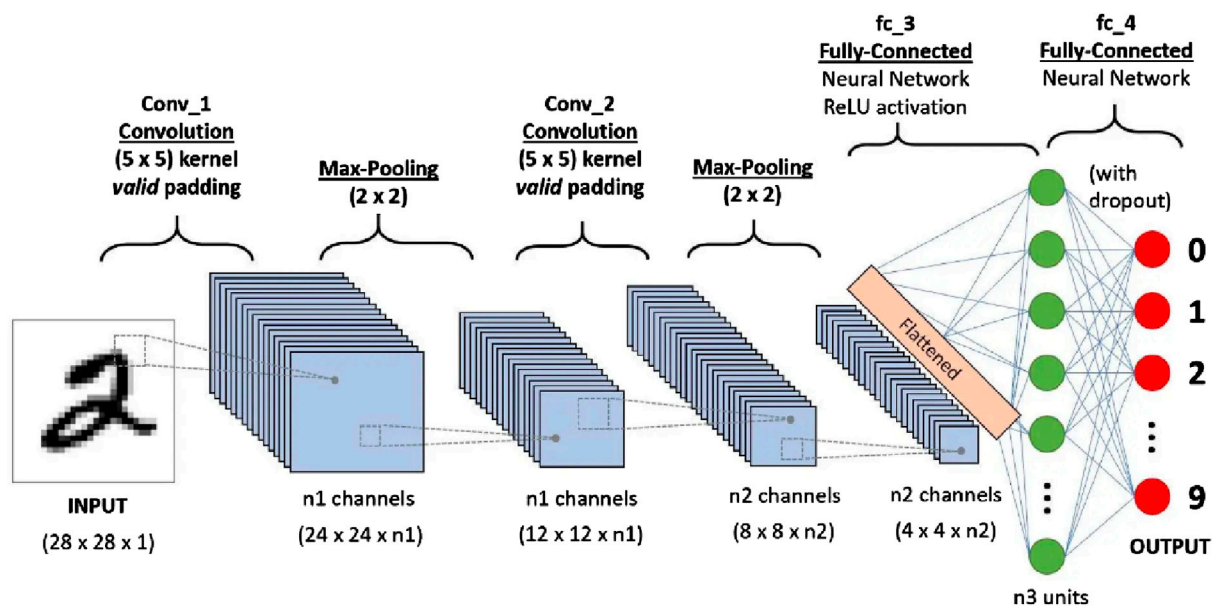


FIGURE 3
Convolutional Neural Network Framework. Picture taken from Sumit Saha (2023) <https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/>.

intervals. A memory cell is managed by three critical gating structures to handle the movement of information: This Forget Gate controls what to remember from the previous cell state. Input Gate: The Input Gate decides what new information is to be stored in the current cell state. Finally, the Output Gate decides what information will be passed from the current cell to impact future output. The addition of these gated components allows LSTMs to retain or disregard information, which proves to be particularly beneficial for learning long-range dependencies present in sequential datasets. A key reason why LSTM networks have become one of the most robust and widely used architectures for real-world sequence modeling tasks is their capability to address some severe limitations of traditional RNNs.

5.1.2.2 Gated Recurrent Unit (GRU)

The other popular variant of Recurrent Neural Network (RNN) is a Gated Recurrent Unit (GRU), introduced by Cho et al. (2014), that use gating mechanisms to control the flow of information over time in sequential data. The GRU is another variant of the Long Short-Term Memory (LSTM) architecture, overcoming the limitations of vanilla RNN but with a simpler design. In contrast to the LSTM, which has three gates (input, forget, outputs), the GRU has only two gates (reset gate and update gate). So, by having an output gate, GRU takes one less gate which reduces the number of parameters making it computationally less expensive. The reset gate decides how much information of the past should be forgotten, while the update gate decides how much of the present information should be passed on to the next state. The GRU reduces the dimensionality and introduces elements that control the accumulation of past information, which has made it an efficient architecture to exploit long-term dependencies in sequential data. The GRU has shown similar performance on smaller or less complicated datasets (Gruber and Jockisch,

2020), while requiring a simpler design than its predecessor LSTM (Gruber and Jockisch, 2020).

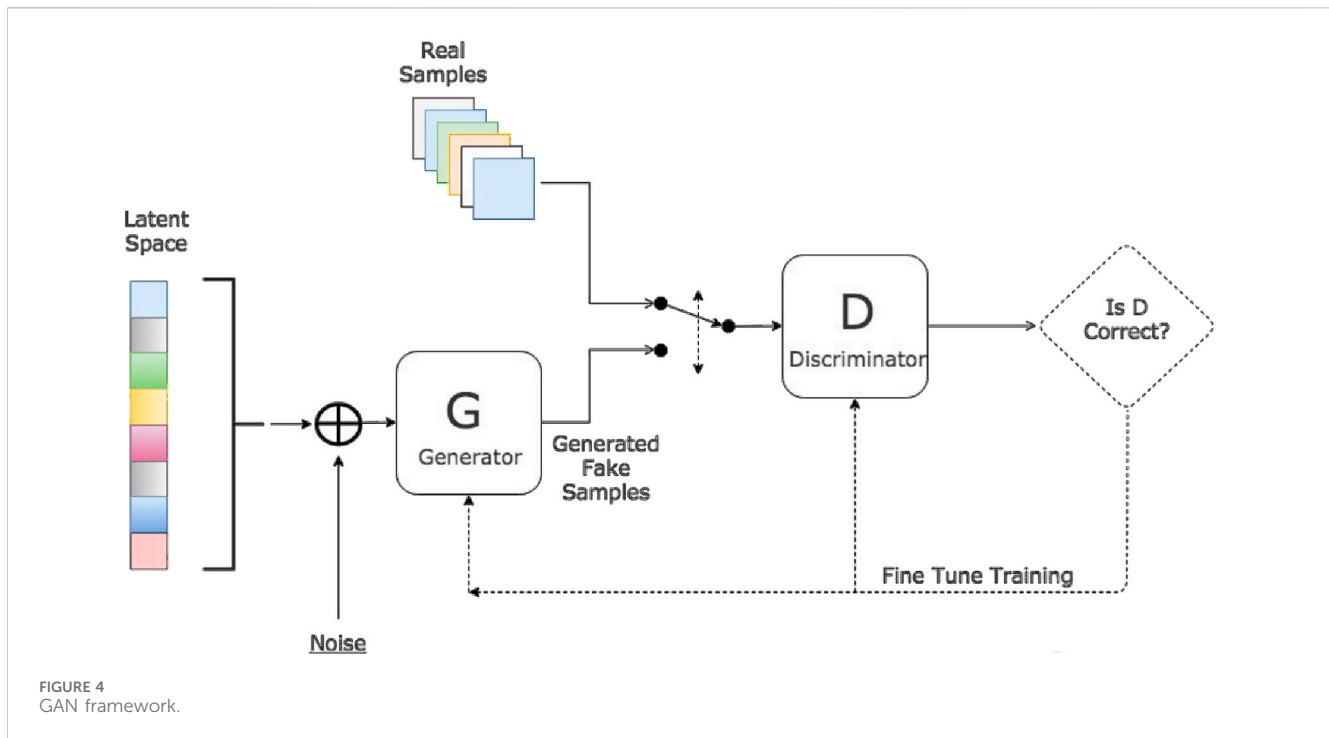
5.2 Unsupervised or learning

This class of deep learning (DL) techniques, known as unsupervised models, is primarily employed to capture high-order correlations and structural features within data for purposes such as pattern analysis, synthesis, and modeling the joint statistical distributions between observed inputs and their corresponding latent representations or categories (Da'u and Salim, 2020). A fundamental principle of unsupervised deep architectures is their independence from explicit supervisory signals—such as class labels—during the learning process. Consequently, these techniques are predominantly utilized in unsupervised learning, where the focus lies in feature extraction, data generation, and representation learning rather than direct classification or regression tasks (Da'u and Salim, 2020).

Unsupervised models can also serve as a preprocessing step for supervised learning, wherein the learned feature representations enhance the performance and accuracy of downstream discriminative models. Among the most widely adopted approaches in this domain are the Generative Adversarial Network (GAN) and the Autoencoder (AE), both of which have demonstrated strong capabilities in learning meaningful representations from unlabeled data and generating new samples that resemble the original data distribution.

5.2.1 Generative Adversarial Network (GAN)

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), represent a groundbreaking neural network architecture for generative modeling, enabling the



synthesis of realistic and novel data samples. GANs are designed to automatically learn the underlying distribution and patterns of the input data in an unsupervised manner, allowing the generation of synthetic instances that closely resemble those in the original dataset. As illustrated in Figure 4, the GAN framework comprises two competing neural networks: a generator (G) and a discriminator (D). The generator is tasked with producing synthetic data that mimic the real data distribution, while the discriminator evaluates whether a given sample is authentic (i.e., drawn from the true data distribution) or synthetic (produced by the generator). During training, these networks engage in a minimax game, where the generator continuously refines its output to “fool” the discriminator, and the discriminator simultaneously improves its ability to distinguish real from generated data. This adversarial process ultimately leads to the production of highly realistic synthetic data.

While GANs are primarily employed in unsupervised learning contexts, they have demonstrated remarkable versatility in semi-supervised and reinforcement learning applications as well (Aggarwal et al., 2021). Moreover, in transfer learning, GANs play a significant role by enforcing alignment in the latent feature space, thus enhancing domain adaptation capabilities. Extended architectures, such as the Bidirectional GAN (BiGAN) (Donahue et al., 2016), further enhance the utility of GANs by learning a mapping from the data space back to the latent space, offering bidirectional modeling capabilities that complement traditional GANs. GANs have been successfully applied across a wide array of domains, including but not limited to healthcare, medical imaging, data augmentation, video and speech synthesis, cybersecurity, epidemic modeling, traffic prediction, and digital art generation. Their adaptability and ability to generate high-fidelity synthetic data have made GANs a cornerstone in modern AI research, particularly for tasks that require data expansion,

creative content generation, or domain transfer. Collectively, GANs continue to define a robust paradigm in generative modeling and are central to the future of intelligent, data-driven systems.

5.2.2 Auto-encoder (AE)

An Autoencoder (AE) is an unsupervised learning approach that is commonly used, which utilizes neural networks to encode input data into lower-dimensional space without the need of labeled outputs (Goodfellow et al., 2016). Autoencoders are especially effective for high-dimensional datasets because they can achieve dimensionality reduction by determining which aspects of the input is salient and represent the input in a compact manner. An autoencoder generally consists of three parts: an encoder, a latent code (bottleneck layer), and a decoder. The encoder passes the input data (the thread, for example,) and converts it into a lower-dimensional representation, or code, thus compressing the input. Then the decoder tries to reconstruct the original input as closely as possible from this compressed code to the original data. The latter process allows the network to learn abstract and low-dimensional representations of the data distribution.

Initially, autoencoders were always used for unsupervised dimensionality reduction and feature extraction, but with recent works, they were extended and proposed as a generative model, in that they approximate the data distribution to perform data sampling [69]. Their ability to capture complex data distributions has resulted in broad applicability to a variety of unsupervised learning tasks, including data compression, denoising, anomaly (or outlier) detection, and efficient coding (Goodfellow et al., 2016). Autoencoders thus serve as a fundamental building block in representation learning, enabling the development of more complex generative and discriminative models by providing robust, compressed abstractions of raw input data.

5.3 Hybrid deep neural networks

Unsupervised models are adaptive and can learn from both labeled and non-labeled data. Supervised models, on the other hand, cannot learn from unlabeled data, but outperform their unsupervised counterparts at supervised tasks. A framework for training both deep unsupervised and supervised models simultaneously can benefit from the benefits of both, which is the rationale for hybrid networks. Hybrid deep learning models typically consist of several (two or more) deep basic models, where the underlying deep learning model is a supervised or unsupervised deep learning model, as discussed above. Based on the integration of different basic unsupervised and supervised models, the following hybrid deep learning model categories may be useful in solving real world problems. These are as follows:

CNN + LSTM: An integration of different supervised models to extract more meaningful and robust features.

GAN + CNN: An integration of an unsupervised model followed by a supervised model.

5.4 Data and their problems

Machine learning applications in all fields of technology, used for real-life problems, are diversifying and growing rapidly. The performance of machine learning models depends on the amount, quality and diversity of data. In order to increase the reliability of the algorithm, it is important to select target data that are selected from the original data set. Data can be obtained in the form of symbolic and numerical attributes which can be retrieved from human beings to sensors at different degrees of complexity and reliability. Data can be provided in various formats: structured tables, unstructured tables, images, audio files, and video files. The machine cannot read raw text, video or images directly; it has to convert data into 1 s and 0 s. Therefore, you cannot directly feed raw data into a machine learning model and expect it to learn. Data pre-processing is the first step in machine learning, where data is transformed or encoded to bring it to a point where the machine can process or analyse the data quickly. In other words, it can also be interpreted as meaning that the model algorithm is able to quickly analyze the properties of the data.

Data collected from different sources often contain incomplete, noise and inconsistencies, which may significantly hinder an efficient analysis of the data. It is therefore necessary to address these issues in advance. These data quality problems can be broadly classified into three main types: over-data, insufficient data, and fragmented or disjointed data.

Over data: In medicine, telecommunications, and space, the volume and speed of data is too great. It plays a crucial role as a constraint for performing analysis with a real-time data set. Additionally, corrupted data may impair the predictive power of the model. Pre-processing of data for correct interpretation is a function of the form, which pre-processes the input data to make subsequent extraction easier and more accurate (Davis, 2022). Reducing the data set dimension may help improve the performance of the model. It is also important to pay due

attention to data sets that contain numerical or other symbolic parameters, as this may increase the model complexity.

Insufficient data: If the available data does not include a sufficient amount of data of all kinds, then the reliability of the information may not be able to be inferred from it. In the case of missing attributes, this will decrease the model's precision. For example, we can deal with decision tree induction, where the backlog may be unequal in length due to missing attributes. They also divide the dataset into training and test collections. This may cause some characteristics to be unevenly distributed. If the data used has more than 20% of Data missing, it must be eliminated (Davis, 2022).

Fractured data: Data incompatibilities become a major problem when collected from multiple groups or different platforms. The purpose, depth and standard of data storage and management may vary according to the need. The level of detail at which data is stored in a database may also vary, which may lead to problems during modelling.

6 Practical implementations for Qur'anic education through technology

In recent years, the incorporation of modern technology into educational tools has significantly improved the way in which the Qur'an is taught and learned, particularly by non-native Arabic speakers and students in remote and self-directed learning settings. These innovations address language complexity, support accurate recitation and provide real-time, personalised feedback, bridging the gap between traditional instruction and modern accessibility. Two particularly impactful areas of implementation are automated translation tools and speech recognition systems, both of which have practical applications in promoting Quranic teaching and learning.

6.1 Automated translation tools

In educational settings, automated translation tools have become a valuable resource to increase the availability of the Qur'an to non-Arabic-speaking learners. These tools have been developed to deal effectively with the complex morphological and lexical problems inherent in the Arabic language. Integrating advanced techniques such as Earley's algorithm and text mining, these systems offer more accurate and relevant translations. This not only extends the scope of Qur'anic learning, but also allows students to understand the meaning of the verses in their own language, thus promoting a deeper and more inclusive understanding of the sacred text.

6.2 Speech recognition technology

Speech recognition systems are increasingly being integrated into Qur'anic learning platforms to improve the accuracy and authenticity of the recitation. Using advanced technologies such as hidden Markov models (HMM), Mel-Frequency Cepstral

Coefficients (MFCs) and deep learning algorithms, these systems can recognize, evaluate and verify recitation patterns with high accuracy. They provide real-time feedback on pronunciation, rhythm and tajweed, helping the learner correct mistakes and develop sound pronunciation. This technique not only promotes individual learning and self-education, but also helps to preserve the oral tradition and integrity of the Qur'an.

7 Potential ethical and cultural sensitivities in applying NLP to religious texts

The application of natural language processing (NLP) to religious texts presents a unique set of ethical and cultural sensitivities, which researchers and developers need to navigate carefully. Religious texts such as the Qur'an are not only revered by billions of believers as holy, but also have a profound theological, spiritual, and historical significance. Any distortion, misinterpretation or change of meaning caused by computer processing can cause offence, disinformation and even cultural and interfaith tensions (Church and Hovy, 2021). For example, translation errors, incorrect emotional analysis or biased semantic interpretations may distort the intended message of a religion, potentially compromising religious beliefs or practices (Samarati et al., 2020). Moreover, the development and use of artificial intelligence tools involving these texts must take account of the diversity of religious interpretations, sectarian differences and the contextual and metaphorical nature of the Scriptural language (Ahmed and Shuja, 2022). Ethically, transparency, inclusiveness and accountability in the way data is processed and NLP models trained are essential. Involving religious scholars, linguists, and community members in the design and assessment of these systems is crucial for respecting, authenticating and preserving cultural integrity (Ali et al., 2021). Ethical frameworks should also address issues of data privacy, open access and the possible misuse of automated religious interpretation tools in ways that could lead to misinformation or radicalisation. Ultimately, the interface between NLP and religious texts requires a multi-disciplinary and ethical approach to preserve the sanctity of the content while harnessing the benefits of technology.

8 Future research directions in Qur'anic NLP: insights and challenges

Rapid progress in natural language processing (NLP) has inspired applications in a wide range of areas, including religious and classical texts. One such area is Qur'anic NLP, where researchers are working to develop artificial intelligence tools to help us understand, analyze, translate, and recite the Qur'an. This paper, highlights major challenges and, most importantly, proposes key directions for future research. These recommendations are essential to guide researchers who want to make a meaningful contribution to the growing intersection of artificial intelligence and Qur'anic studies. Below are the potential future research directions based on our study.

8.1 Benchmarking and evaluation

A key direction emphasized in the paper is the urgent need for standardized, linguistically rich, and annotated Qur'anic data sets. Current research efforts are hindered by fragmented and limited corporates, which limit the training, evaluation, and reproducibility of models. The development of shared, open access annotation resources such as word marks, syntactic structure, semantic relationships and thematic categorization would significantly improve the quality of Quranic NLP models. Benchmarking and evaluation tools such as AYTEC are also proposed as key elements for standardising performance comparisons and promoting collaborative progress.

8.2 Deep learning architectures

As discussed earlier throughout the paper the adoption of state-of-the-art deep learning architectures, especially transformer-based models like BERT and AraBERT. These models can dramatically improve performance in semantic understanding, classification, and translation tasks. The combination of architectures—such as CNNs, LSTMs, and GRUs—is also recommended, especially for complex applications like recitation analysis and Tajweed error detection. Furthermore, hybrid deep learning models are viewed as a promising avenue to harness strengths from different approaches in multimodal settings, where both textual and audio data are processed.

8.3 Ontology integration

Integration and development of ontologies is another promising direction. Ontology helps to structure the semantic relationships and hierarchies contained in the Qur'an, allowing for more efficient semantic search, understanding of concepts, and automated reasoning. Future research should consider the integration of the existing fragmented ontologies into a single, coherent framework. Such integration would provide a solid basis for applications such as concept-based search systems and contextually aware verse interpretation.

8.4 Cross-lingual and context-aware translation

The paper also recommends extending research into multilingual and context-sensitive translation systems. Traditional machine translation techniques often fail to capture the deep, layered meanings contained in Quranic verses. By incorporating semantic context, historical chronology, and theological interpretations, future systems could provide more accurate and accurate translations. In addition, multimodal systems combining text-based and speech-based analysis could play a key role in developing adaptive learning tools for the learning of the Quran, the practice of recitation and Tajweed.

In conclusion, this paper serves as a valuable model for future research into the topic of Qur'anic NLP. It outlines concrete

technical, linguistic and ethical challenges, while offering concrete paths forward for the divestment process. By embracing interdisciplinary cooperation, focusing on data-driven development, integrating modern deep learning methods, and respecting religious contexts, the research community can build powerful, respectful, and meaningful artificial intelligence tools that can contribute to both a technological and spiritual understanding of the Qur'an.

9 Conclusion

In conclusion, the Qur'an presents unique challenges and opportunities for the NLP research domain, including its unique orthography, the need for technical and linguistic support, as well as the need for advanced NLP techniques to address the challenges listed above. Given that the Qur'an is a sacred book for nearly two billion Muslims across this globe, Qur'anic NLP is a significant field of study. The development of AI-based NLP technologies for the Arabic language has garnered significant attention lately. In contrast to Arabic NLP, which is considered a low-resource language with fewer tools and data available as compared to studies focusing on the English language, research work revolving around Qur'anic NLP appears to be less developed.

Author contributions

AAI: Writing – original draft. FN: Writing – review and editing. IM: Writing – review and editing. MY: Writing – review and editing. IE: Writing – review and editing. EA: Writing – review and editing.

References

- Abouelnour, M. M., Sarhan, O., and Alsharef, A. K. A. (2023). Linguistic structure and its impact on diversifying positive discourse in the Holy Qur'an (A semantic Grammatical Study). *Migr. Lett.* 20 (S1), 1098–1112. doi:10.59670/ml.v20iS1.3878
- Abro, B., Naqvi, A. B., and Hussain, A. (2012). "Qur'an recognition for memorization using Speech Recognition technique," in *2012 15th international multitopic conference (INMIC)* (IEEE), 30–34.
- Afzal, H., and Mukhtar, T. (2019). Semantically enhanced concept search of the holy quran: qur'anic English WordNet. *Arabian J. Sci. Eng.* 44, 3953–3966. doi:10.1007/s13369-018-03709-2
- Aggarwal, A., Mittal, M., and Battineni, G. (2021). Generative adversarial net-work: an overview of theory and applications. *Int. J. Inf. Manag. Data Insights* 1, 100004. doi:10.1016/j.jjime.2020.100004
- Ahmed, A. H., and Abdo, S. M. (2017). Verification system for Quran recitation recordings. *Int. J. Comput. Appl.* 163 (4), 6–11. doi:10.5120/ijca2017913493
- Ahmed, R., and Atwell, E. S. (2016). Developing an ontology of concepts in the Qur'an. *Int. J. Islamic Appl. Comput. Sci. Technol.* 4 (4), 1–8. Available online at: <https://eprints.whiterose.ac.uk/id/eprint/112967/1/ahmed16ijasat.pdf>.
- Ahmed, Z., and Shuja, J. (2022). Ethical AI: challenges and implications in religious contexts. *J. AI Ethics* 3 (1), 34–47.
- Ahsiah, I., Noor, N. M., and Idris, M. Y. I. (2013). "Tajweed checking system to support recitation," Proceedings of the 2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS). *IEEE*, 189–193.
- Al Anazi, M. M., and Shahin, O. R. (2022). A machine learning model for the identification of the holy Quran reciter utilizing k-nearest neighbor and artificial neural networks. *Inf. Sci. Lett.* 11 (4), 1093–1102. Available online at: <https://digitalcommons.aaru.edu.jo/isl/vol11/iss4/10>.
- Al Bakri, T., and Mallah, M. (2016). Musical performance of the holy Quran with the assistance of the Arabic Maqams. *Turkish j. educational tech.* Turkey: Sakarya University, 121–132.
- Al-Bakeri, A. A., and Basuhail, A. A. (2017). ASR for Tajweed rules: integrated with self-learning environments. *Int. J. Inf. Eng. Electron. Bus.* 9 (6), 1–9. doi:10.5815/ijeeb.2017.06.01
- Al-Jarrah, M. A., Al-Jarrah, A., Jarrah, A., AlShurbaji, M., Magableh, S. K., Al-Tamimi, A. K., et al. (2022). Accurate reader identification for the Arabic holy quran recitations based on an enhanced VQ algorithm. *Rev. d'Intelligence Artif.* 36 (6), 815–823. doi:10.18280/ria.360601
- Al-Kabi, M. N., Kanaan, G., Shalabi, R. A., Nahar, K. M., and Ismail, B. M. B. (2005). Statistical classifier of the holy Quran verses (Fatiha and Yaseen chapters). *J. Appl. Sci.* 5 (3), 580–583. doi:10.3923/jas.2005.580.583
- Al-Khalifa, H. S., Al-Yahya, M. M., Bahanshal, A., and Al-Odah, I. (2009). "SemQ: a proposed framework for representing semantic opposition in the Holy Quran using Semantic Web technologies," in *2009 international conference on the current trends in information technology (CTIT)* (IEEE), 1–4.
- Al-Yahya, M., Al-Khalifa, H., Bahanshal, A., Al-Odah, I., and Al-Odah, N. (2010). An ontological model for representing semantic lexicons: an application on time nouns in the holy Quran. *Arabian J. Sci. Eng.* 35 (2), 21. Available online at: https://www.researchgate.net/publication/228955782_An_Ontological_Model_for_Representing_Semantic_Lexicons_An_Application_on_Time_Nouns_in_the_Holy_Quran.
- Alargrmi, A. M., and Eljazzar, M. M. (2020). "Imam: word embedding model for islamic Arabic NLP," in *2020 2nd novel intelligent and leading emerging sciences conference (NILES)* (IEEE), 520–524.
- Ali, M., Farooq, U., and Khan, N. (2021). Context-aware NLP for sacred texts: ethical and technical perspectives. *Int. J. Humanit. AI* 2 (2), 15–28.
- Aljohani, N. F., and Jaha, E. S. (2023). Visual lip-reading for quranic Arabic Alphabets and words using deep learning. *Comput. Syst. Sci. and Eng.* 46 (3), 3037–3058. doi:10.32604/csse.2023.037113
- Alkhateeb, J. H. (2020). A machine learning approach for recognizing the Holy Quran reciter. *Int. J. Adv. Comput. Sci. Appl.* 11 (7). doi:10.14569/ijacsa.2020.0110735

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Alqahtani, M., and Atwell, E. (2016). "Arabic quranic search tool based on ontology," in *Natural Language processing and information systems: 21st international conference on applications of Natural Language to information systems, NLDB 2016, salford, UK, June 22-24, 2016, proceedings 21* (Springer), 478–485.
- Alqahtani, M., and Atwell, E. (2017). Evaluation criteria for computational Quran search. *Int. J. Islamic Appl. Comput. Sci. Technol.* 5 (1), 12–22. Available online at: <https://eprints.whiterose.ac.uk/id/eprint/111427/>.
- Alshammeri, M., Atwell, E., and Alsalka, M. A. (2021a). "Quranic topic modeling using paragraph vectors," in *Intelligent systems and applications: Proceedings of the 2020 intelligent systems conference (IntelliSys) volume 2* (Springer), 218–230.
- Alshammeri, M., Atwell, E., and Ammar Alsalka, M. (2021b). Detecting semantic-based similarity between verses of the Quran with Doc2vec. *Procedia Comput. Sci.* 351–358. doi:10.1016/j.procs.2021.05.104
- Alsubhi, K., Jamal, A., and Alhothali, A. (2021). *Pre-trained transformer-based approach for Arabic question answering: a comparative study*. arXiv preprint arXiv:2111.05671.
- Altalmas, T., Ahmad, S., Nik Hashim, N. N. W., Shahbudin Hassan, S., and Sediono, W. (2022). Characteristics with opposite of quranic letters mispronunciation detection: a classifier-based approach. *Bull. Electr. Eng. Inf.* 11 (5), 2817–2827. doi:10.11591/eei.v11i5.3715
- Altammami, S., and Atwell, E. (2022). "Challenging the transformer-based models with a classical Arabic dataset: quran and Hadith," in *Proceedings of the Thirteenth language resources and evaluation conference*, 1462–1471.
- Atwell, E. (2015). A review of semantic search methods to retrieve information from the qur'an corpus. *Technology* 6, 4492–4498. Available online at: <https://eprints.whiterose.ac.uk/id/eprint/88193/>.
- Balula, N. O., Rashwan, M., and Abdou, S. (2021). Automatic speech recognition (ASR) systems for learning Arabic language and Al-Quran recitation: a review. *Int. J. Comput. Sci. Mob. Comput.* 10 (7), 91–100. doi:10.47760/ijcsmc.2021.v10i07.013
- Beirade, F., Azzoune, H., and Zegour, D. E. (2021). Semantic query for Quranic ontology. *J. King Saud University-Computer Inf. Sci.* 33 (6), 753–760. doi:10.1016/j.jksuci.2019.04.005
- Bentrcia, R., Zidat, S., and Marir, F. (2018). An analytical study on the holy Quran based on the order of words in Arabic AND conjunction. *Malays. J. Comput. Sci.* 31 (1), 1–16. doi:10.22452/mjcs.vol31no1.1
- Brierley, C., Sawalha, M., and Atwell, E. (2014). "Tools for Arabic Natural Language Processing: a case study in qalqalah prosody," in 9th international Conference on language Resources and evaluation. *European language resources association*, 283–287.
- Chakroborty, S., Roy, A., and Saha, G. (2008). Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. *Int. J. Electron. Commun. Eng.* 2 (11), 2554–2561. Available online at: https://www.researchgate.net/publication/306364973_Improved_closed_set_text-independent_speaker_identification_by_combining_MFCC_with_evidence_from_flipped_filter_bank.
- Cho, K., Van, M. B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). *Learning phrase representations using rnn encoder-decoder for statistical machine translation*. arXiv preprint arXiv:1406.1078.
- Chollet, F. (2017). "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 1251–1258.
- Church, K., and Hovy, E. (2021). "Ethical issues in NLP and religious corpora," in *Proceedings of the ACL workshop on ethics in NLP*, 45–54.
- Da'u, A., and Salim, N. (2020). Recommendation system based on deep learn-ing methods: a systematic review and new directions. *Artif. Intel. Rev.* 53 (4), 2709–2748. doi:10.1007/s10462-019-09744-1
- Davis, J. F. (2022). Process data analysis and interpretation. *Adv. Chem. Eng.* 25. doi:10.1016/S0065-2377(08)60108-8
- Dimitrakakis, C., and Bengio, S. (2011). Phoneme and sentence-level ensembles for speech recognition. *EURASIP J. Audio, Speech, Music Process.* 2011, 1–17. doi:10.1155/2011/426792
- Donahue, J., Krähenbühl, P., and Darrell, T. (2016). *Adversarial feature learn-ing*. arXiv preprint arXiv:1605.09782.
- Dukes, K., and Buckwalter, T. (2010). "A dependency treebank of the Quran using traditional Arabic grammar," in *2010 the 7th international conference on Informatics and systems (INFOS)* (IEEE), 1–7.
- Dukes, K., and Habash, N. (2010). "Morphological annotation of quranic Arabic," in *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)* (Valletta, Malta: European Language Resources Association), 2530–2536.
- Dukes, K., Atwell, E., and Sharaf, A.-B. M. (2010). "Syntactic annotation guidelines for the quranic Arabic dependency treebank," in *Proceedings of the Seventh international conference on language resources and evaluation (LREC'10)* (Valletta, Malta: European Language Resources Association).
- Dukes, K., Atwell, E., and Habash, N. (2013). Supervised collaboration for syntactic annotation of Quranic Arabic. *Lang. Resour. Eval.* 47, 33–62. doi:10.1007/s10579-011-9167-7
- Dupond, S. (2019). A thorough review on the current advance of neural network structures. *Annu. Rev. Control* 14, 200–230.
- El Amrani, M. Y., Rahman, M. H., Wahiddin, M. R., and Shah, A. (2016a). Building CMU Sphinx language model for the Holy Quran using simplified Arabic phonemes. *Egypt. Inf. J.* 17 (3), 305–314. doi:10.1016/j.eij.2016.04.002
- El Amrani, M. Y., El Amrani, M. Y., El Amrani, M. Y., Rahman, M. M., Wahiddin, M. R., and Shah, A. (2016b). Towards using CMU sphinx tools for the holy Quran recitation verification. *Int. J. Islam. Appl. Comput. Sci. Technol.* 4 (2), 10–15. Available online at: https://www.semanticscholar.org/paper/Towards-Using-CMU-Sphinx-Tools-for-the-Holy-Quran-Amrani-Rahman/ca7d434ba0cb67a4cd1805a7660b2dc34b7acc79?utm_source=direct_link.
- Elaffendi, M. A., Abuhaimeid, I., and AlRajhi, K. (2021). A simple Galois Power-of-Two real-time embedding scheme for performing Arabic morphology deep learning tasks. *Egypt. Inf. J.* 22 (1), 35–43. doi:10.1016/j.eij.2020.03.002
- Elbarougy, R., Behery, G., and El Khatib, A. (2020). A proposed natural language processing preprocessing procedures for enhancing Arabic text summarization. *Recent Adv. NLP Case Arabic Lang.*, 39–57. doi:10.1007/978-3-030-34614-0_3
- Fadlun, S. (2016). *The analysis of conjunctions as cohesive device in the Quranic translation of Surah Al-Baqarah by Abdullah Yusuf Ali: a corpus based study*. IAIN Syekh Nurjati Cirebon.
- Fakhrezi, M. F., Bijaksana, M. A., and Huda, A. F. (2021). Implementation of automatic text summarization with TextRank method in the development of Al-qur'an vocabulary encyclopedia. *Procedia Comput. Sci.* 179, 391–398. doi:10.1016/j.procs.2021.01.021
- Gales, M., and Young, S. (2008). The application of hidden Markov models in speech recognition. *Found. Trends® Signal Process.* 1 (3), 195–304. doi:10.1561/20000000004
- Géron, A. (2019). "Hands-on machine learning with Scikit-Learn, Keras and TensorFlow", in Concepts, Tools, And Techniques to Build Intelligent Systems. O'Reilly Media.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in neural information processing systems*, 2672–2680.
- Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, vol. 1. Cambridge: MIT Press.
- Gruber, N., and Jockisch, A. (2020). Are gru cells more specific and lstm cells more sensitive in motive classification of text? *Front. Artif. Intell.* 3, 40. doi:10.3389/frai.2020.00040
- Hadiyansah, R., and Andamira, R. (2023). Convolutional neural network (CNN) for detecting Al-Qur'an reciting and memorizing. *Khazanah J. Relig. Technol.* 1 (2), 44–48. doi:10.15575/kjrt.v1i2.235
- Hadwan, M., Alsayadi, H. A., and Al-Haggee, S. (2023). An end-to-end transformer-based automatic speech recognition for qur'an reciters. *CMC-COMPUTERS Mater. and CONTINUA* 74 (2), 3471–3487. doi:10.32604/cmc.2023.033457
- Hammo, B., Sleit, A., and El-Haj, M. (2007). "Effectiveness of query expansion in searching the Holy Quran," in *Proceeding of the second international conference on Arabic Language Processing*.
- Hanum, H. M., Bakar, Z. A., and Ismail, M. (2013). "Evaluation of Malay grammar on translation of Al-Quran sentences using Earley algorithm," in *2013 5th international conference on information and communication technology for the muslim world (ICT4M)* (IEEE), 1–4.
- Harere, A. Al, and Jallad, K. Al (2023). *Quran recitation recognition using end-to-end deep learning*. arXiv preprint arXiv:2305.07034.
- Hassan, G. S., Mohammad, S. K., and Alwan, F. M. (2015). 'Categorization of 'Holy Quran-Tafseer using K-nearest neighbor algorithm. *Int. J. Comput. Appl.* 129 (12), 1–6. doi:10.5120/ijca2015906909
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pat- Tern. Anal. Mach. Intell.* 37 (9), 1904–1916. doi:10.1109/tpami.2015.2389824
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 770–778.
- Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735
- Iqbal, R., Mustapha, A., and Mohd. Yusoff, Z. (2013). An experience of developing Quran ontology with contextual information support. *Multicultural Educ. and Technol. J.* 7 (4), 333–343. doi:10.1108/metj-03-2013-0009
- Jabar, F. H., Mohammad, J. I., Zain, A. F., Hasan, A. B., Zakaria, Z., Yusoff, A. M., et al. (2018). "Preliminary results of Quranic maqamat profiling using W-DFT technique," in *Proc. 25th Int. Congr. Sound Vib.*, 4572–4579.
- Jamaliah Ibrahim, N., Yamani Idna Idris, M., Razak, Z., and Naemah Abdul Rahman, N. (2013). Automated tajweed checking rules engine for Quranic learning. *Multicultural Educ. and Technol. J.* 7 (4), 275–287. doi:10.1108/metj-03-2013-0012
- Khadhim, S. A., and Chiad, H. W. (2022). Parsing Facets of letters inflicted by the three Diacritics in the holy quran. *Resmilitaris* 12 (2), 4092–4101. Available online at: <https://resmilitaris.net/uploads/paper/5f847ca48f0ae00976c30cf6251f0443.pdf>.

- Khan, H. U., Muhammad Saqlain, S., Shoaib, M., and Sher, M. (2013). Ontology based semantic search in Holy Quran. *Int. J. Future Comput. Commun.* 2 (6), 570–575. doi:10.7763/ijfcc.2013.v2.229
- Khazani, M. M. M., Mohamed, H., Tengku Sembok, T. M., Mohd Yusop, N. M., Wani, S., Gulzar, Y., et al. (2021). Semantic graph knowledge representation for Al-Quran verses based on word dependencies. *Malays. J. Comput. Sci.*, 132–153. doi:10.22452/mjcs.sp2021no2.9
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 1097–1105.
- Kuppusamy, K., and Eswaran, C. (2020). *Speaker recognition system based on age-related features using convolutional and deep neural networks*.
- Larabi-Marie-Sainte, S., Alnamlah, B. S., Alkassim, N. F., and Alshathry, S. Y. (2022). A new framework for Arabic recitation using speech recognition and the Jaro Winkler algorithm. *Kuwait J. Sci.* 49. doi:10.48129/kjs.v49i1.11231
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (11), 2278–2324. doi:10.1109/5.726791
- Mahmudin, H. M., and Akbar, H. (2023). Qur'an recitation correction system using Deepspeech. *Indonesian J. Multidiscip. Sci.* 2 (11), 4010–4022. doi:10.55324/ijoms.v2i11.638
- Malhas, R., and Elsayed, T. (2020). Ayatec: building a reusable verse-based test collection for Arabic question answering on the holy qur'an, *ACM Trans. Asian Low-Resource Lang. Inf. Process. (TALLIP)* 19 (6), 1–21. doi:10.1145/3400396
- Mannaa, Z. M., Azmi, A. M., and Aboalsamh, H. A. (2022). Computer-assisted i'raab of Arabic sentences for teaching grammar to students. *J. King Saud University-Computer Inf. Sci.* 34 (10), 8909–8926. doi:10.1016/j.jksuci.2022.08.020
- Mohamed, R., Ragab, M., Abdelnasser, H., El-Makky, N. M., and Torki, M. (2015). “Al-Bayan: a knowledge-based system for Arabic answer selection,” in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 226–230.
- Mohammed, A., Sunar, M. S., and Salam, M. S. H. (2015). Quranic verses verification using speech recognition techniques. *Jurnal Teknologi Sci. and Eng.* 73 (2). doi:10.11113/jt.v73.4200
- Mohammed, A., Sunar, M. S. B., and Salam, M. S. H. (2018). “Recognition of holy Quran recitation rules using phoneme duration,” in *Recent trends in information and communication technology: proceedings of the 2nd international conference of reliable information and communication technology (IRICT 2017)* (Springer), 343–352.
- Muhammad, A., Qayyum, Z., Mirza, M. W., Tanveer, S., Martinez-Enriquez, A. M., and Syed, A. Z. (2012). E-hafiz: intelligent system to help muslims in recitation and memorization of Quran. *Life Sci. J.* 9 (1), 534–541. Available online at: https://www.researchgate.net/publication/275966170_EHafiz_Intelligent_System_to_Help_Muslims_in_Recitation_and_Memorization_of_Quran.
- Nahar, K. M., Al-Khatib, R. M., Al-Shannaq, M. A., and Barhoush, M. M. (2020). An efficient holy Quran recitation recognizer based on SVM learning model. *Jordanian J. Comput. Inf. Technol. (JJCIT)* 6 (04), 394–414. doi:10.5455/jjcit.71-1593380662
- Nassourou, M. (2012). Towards a knowledge-based learning system for the quranic text. *opus.Bibl. uni-wuerzburg*.
- Putra, B., Atmaja, B. T., and Prananto, D. (2012). Developing speech recognition system for Quranic verse recitation learning software. *IJID Int. J. Inf. Dev.* 1 (2), 14–18. doi:10.14421/ijid.2012.01203
- Putra, S. J., Mantoro, T., and Gunawan, M. N. (2017). “Text mining for Indonesian translation of the Quran: a systematic review,” in *2017 international conference on computing, Engineering, and design (ICCED)* (IEEE), 1–5.
- Qayyum, A., Latif, S., and Qadir, J. (2018). “Quran reciter identification: a deep learning approach,” in *2018 7th international conference on computer and communication engineering (ICCCE)* (IEEE), 492–497.
- Razak, Z., Ibrahim, N. J., Idris, M. Y. I., Tamil, E. M., Yusoff, M., Abdul Rahman, N. N., et al. (2008). Quranic verse recitation recognition module for support in j-QAF learning: a review. *Int. J. Comput. Sci. Netw. Secur. (IJCSNS)* 8 (8), 207–216. Available online at: http://paper.ijcsns.org/07_book/200808/20080831.pdf.
- Rostam, N. A. P., and Malim, N. H. A. H. (2021). Text categorisation in Quran and Hadith: overcoming the interrelation challenges using machine learning and term weighting. *J. King Saud University-Computer Inf. Sci.* 33 (6), 658–667. doi:10.1016/j.jksuci.2019.03.007
- Saad, S., Noah, S. A. M., Salim, N., and Zainal, H. (2013). “Rules and natural language pattern in extracting Quranic knowledge,” in *2013 taibah university international conference on advances in information technology for the holy quran and its sciences (IEEE)*, 381–386.
- Saleh, M. F. (2021). Appropriateness of attention and its connection with the maqam in verses from the Holy Quran. *AL-ADAB J.* 1 (139), 53–70. doi:10.31973/aj.v1i139.1156
- Samara, G., Al-Daoud, E., Swerki, N., and Alzu'bi, D. (2023). The recognition of holy qur'an reciters using the MFCCs' technique and deep learning. *Adv. Multimedia* 2023, 1–14. doi:10.1155/2023/2642558
- Samarati, P., Zannone, N., and Mancini, L. (2020). Responsible AI in textual analysis of sacred scriptures. *AI Soc.* 35 (4), 897–910.
- Sarker, I. H. (2021). Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. *SN Computer. Science.* 2 (3), 1–16. Available online at: <https://www.preprints.org/manuscript/202102.0340/v1>.
- Satori, H., Harti, M., and Chenfour, N. (2007). *Introduction to Arabic speech recognition using CMUSphinx system*. arXiv preprint arXiv:0704.2083.
- Sawalha, M., Brierley, C., Atwell, E., and Dickens, J. (2017). Text analytics and transcription technology for quranic Arabic. *Int. J. Islamic Appl. Comput. Sci. Technol. (IJASAT)* 5 (2), 45–51. Available online at: https://www.researchgate.net/publication/318876054_Text_Analytics_and_Transcription_Technology_for_Quranic_Arabic.
- Shahriar, S., and Tariq, U. (2021). Classifying maqams of Qur'anic recitations using deep learning. *IEEE Access* 9, 117271–117281. doi:10.1109/access.2021.3098415
- Sherif, M. A., and Ngonga Ngomo, A. C. (2015). Semantic quran. *Semantic Web* 6 (4), 339–345. doi:10.3233/SW-140137
- Shoaib, M., Nadeem Yasin, M., Hikmat, U. K., Imran Saeed, M., and Khiyal, M. S. H. (2009). “Relational WordNet model for semantic search in holy quran,” in *2009 international conference on emerging technologies (IEEE)*, 29–34.
- Siddiqui, M. A., Faraz, S. M., and Sattar, S. A. (2013). “Discovering the thematic structure of the Quran using probabilistic topic model,” in *2013 taibah university international conference on advances in information technology for the holy quran and its sciences (IEEE)*, 234–239.
- Stevens, R., Rector, A., and Hull, D. (2010). *What is an ontology? Ontogenesis* [Preprint].
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 1–9.
- Tabbaa, H. M. A., and Soudan, B. (2015). Computer-aided training for Quranic recitation. *Procedia-Social Behav. Sci.* 192, 778–787. doi:10.1016/j.sbspro.2015.06.092
- Tabbal, H., El Falou, W., and Monla, B. (2006). “Analysis and implementation of a “Quranic” verses delimitation system in audio files using speech recognition techniques,” in *2006 2nd international conference on information and communication technologies (IEEE)*, 2979–2984.
- Tabrizi, A. A., Mahmud, R., Idris, N., and Tohidi, H. (2016). A rule-based approach for pronoun extraction and pronoun mapping in pronominal anaphora resolution of Quran English translations. *Malays. J. Comput. Sci.* 29 (3), 207–224. doi:10.22452/mjcs.vol29no3.4
- Ta'a, A., Abidin, S. Z., Abdullah, M. S., Ali, A. B. M., and Ahmad, M. (2013). “Al-Quran themes classification using ontology,” in *4th international conference on computing and informatics*, 383–389.
- Ta'a, A., Abed, Q. A., and Ahmad, M. (2017). Al-Quran ontology based on knowledge themes. *J. Fundam. Appl. Sci.* 9 (5S), 800–817. doi:10.4314/jfas.v9i5s.57
- Touati-Hamad, Z., Laouar, M. R., and Bendib, I. (2020). “Quran content representation in NLP,” in *Proceedings of the 10th international conference on information systems and technologies*.
- Yauri, A. R., Kadir, R. A., Azman, A., and Murad, M. A. (2012). “Quranic-based concepts: verse relations extraction using Manchester OWL syntax,” in *2012 international conference on information retrieval and knowledge management (IEEE)*, 317–321.
- Yekache, Y., Mekelleche, Y., and Kouninef, B. (2012). *Towards Quranic reader controlled by speech*. arXiv preprint arXiv:1204.1566.
- Yunus, M. A., Zainuddin, R., and Abdullah, N. (2010). “Visualizing Quran documents results by stemming semantic speech query,” in *2010 international conference on user science and engineering (i-USEr)* (IEEE), 209–213.