CNN-LSTM with Multi-Acoustic Features for Automatic Tajweed Mad Rule Classification

Nenny Anggraini[1,*], Yusuf Rahman[2], Achmad Nizar Hidayanto[3], Husni Teja Sukmana[4]

[1,4]*Faculty of Science and Technology, State Islamic University Syarif Hidayatullah, Jakarta, Indonesia*

[2]*Graduate School, State Islamic University Syarif Hidayatullah, Jakarta, Indonesia*

[3]*Faculty of Computer Science, University of Indonesia, Depok, Indonesia*

(Received: July 1, 2025; Revised: August 20, 2025; Accepted: December 5, 2025; Available online: January 14, 2026)

Abstract

The rules of mad recitation in the Qur'an are a crucial aspect of tajwīd, governing the lengthening of vowel sounds that affect both meaning and recitational accuracy. Despite its importance, there is currently no reliable automatic system capable of classifying mad rules based on voice input. This study proposes a deep learning-based approach using a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) model to automatically classify mad rules from Qur'anic recitations. The research follows the CRISP-DM methodology, covering data understanding, preparation, modeling, and evaluation stages. Acoustic features were extracted from 3,816 annotated audio segments of Surah Al-Fātiḥah, combining Mel-Frequency Cepstral Coefficients (MFCC), Chroma, Spectral Contrast, and Root Mean Square (RMS) to represent phonetic and prosodic attributes. The CNN layers captured spatial characteristics of the spectrum, while LSTM layers modeled temporal dependencies of the audio. Experimental results show that the combination of all four features achieved an accuracy of 97.21%, precision of 95.28%, recall of 95.22%, and F1-score of 95.25%. These findings indicate that multi-feature integration enhances model robustness and interpretability. The proposed CNN-LSTM framework demonstrates potential for practical deployment in voice-based tajwīd learning tools and contributes to the broader field of Qur'anic speech recognition by offering a systematic, ethically grounded, and data-driven approach to mad classification.

Keywords: Mad Classification, Tajweed, CNN-LSTM, MFCC, Spectral, Chroma, RMS

1. Introduction

Speech recognition technology plays a crucial role in advancing human–computer interaction, as it enables the conversion of spoken language into machine-readable text for various applications from virtual assistants to smart home systems. In assistive contexts such as in [1], speech recognition has even been applied to support communication for individuals with speech impairments. A key stage in any speech recognition system is featuring extraction, which transforms complex audio signals into simplified representations that retain essential phonetic and acoustic information including frequency, intensity, and temporal dynamics [2]. Building on these insights, this study employs a combination of four complementary acoustic features MFCC, Chroma, Spectral Contrast, and RMS. Together, these features form a richer acoustic representation capable of capturing the complex variations inherent in Qur'anic recitations.

From an ethical perspective, the use of Qur'anic recitations in artificial intelligence must comply with Islamic ethical principles emphasizing respect, privacy, and responsibility. AI should promote maslahah (benefit) and avoid darar (harm) or misuse of sacred content [3]. Guided by the Qur'an and Hadith, this study upholds amanah (trustworthiness) and adl (justice) as ethical foundations in developing an AI model for Qur'anic recitation classification. Focusing on the mad rule that governs vowel elongation in Surah Al-Fatihah, the research employs a CNN-LSTM model trained on four acoustic features (MFCC, Chroma, Spectral Contrast, and RMS) to capture spectral and temporal dynamics. As mad influences both the beauty and meaning of recitation [4], the study uses expert-labeled data, stratified cross-validation, and systematic feature integration to ensure methodological rigor and interpretability. The proposed

***Corresponding author: Nenny Anggraini (nenny.anggraini@uinjkt.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i1.1062

framework establishes a validated approach to automatic mad classification, advancing the field of Qur'anic speech processing.

2. Literature Review

The accuracy of speech classification models depends heavily on selecting suitable acoustic features. Prior studies have shown that combining multiple features enhances model performance. For example, [5] reported that merging MFCC and Mel Spectrogram features improved accuracy from 86.3% to 91.7%, while [6] found that combining GFCC and short-term energy increased accuracy from 83.6% to 89.3% in environmental sound classification. MFCC captures phonetic details by mapping signals to the Mel frequency scale, imitating human auditory perception [7]. Chroma encodes pitch class energy, making it useful for tonal pattern identification [8]. Spectral Contrast reflects differences in energy across frequency bands, distinguishing "bright" and "dark" sounds [9], while RMS measures the overall signal energy [10].

In addition to feature selection, the choice of classification algorithm plays an important role in mapping acoustic features to their corresponding labels [11]. One widely used approach in deep learning-based speech processing is the hybrid CNN-LSTM model, which combines the Convolutional Neural Network's (CNN) ability to extract spatial patterns from spectral representations with the Long Short-Term Memory (LSTM) network's ability to capture temporal patterns in sequential data [12], [13]. The CNN-LSTM combination has been proven effective in various audio analysis tasks, including emotion recognition [14] and Qur'anic verse classification [15].

Several previous studies have demonstrated the effectiveness of MFCC in tajwid rule classification using deep learning models. Study [16] developed an MFCC and LSTM-based approach with an accuracy of 90%, although the model faced challenges in handling class imbalance. MFCC has also been applied in related Qur'anic recitation domains such as Hijaiyah letter articulation recognition [4], tajwid classification [17], and Qur'anic qari recognition [18], achieving accuracy rates of up to 99.66%. Spectral Contrast has also been reported to complement MFCC features in tajwid and recitation analysis by enhancing robustness to timbre and dynamic variations, which are important in differentiating types of mad elongations [19]. Existing studies using CNN or LSTM with single-feature inputs such as MFCC have not adequately captured the phonetic and prosodic complexity of mad recitation. This study fills that gap by proposing a CNN-LSTM model integrating four complementary acoustic features (MFCC, Chroma, Spectral Contrast, and RMS) establishing a novel multi-feature deep learning framework specifically designed for tajwīd analysis with expert-annotated data and stratified cross-validation.

3. Methodology

The research workflow is presented in figure 1.

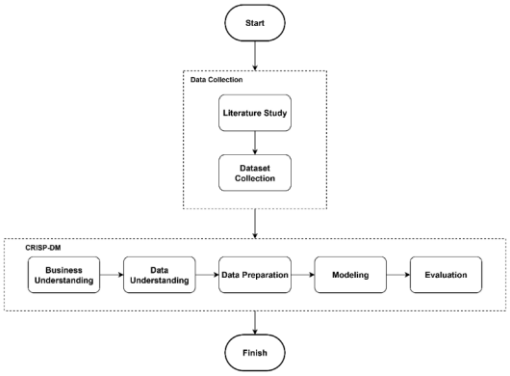

Figure 1. Research Flow

 ISSN 2723-6471

3.1. Data Collection Method

The data in this study were obtained through two main approaches, literature review and the collection of audio datasets. The literature review was conducted to strengthen the theoretical foundation and to explore previous studies relevant to the topic of tajwid classification and speech signal processing based on machine learning. Literature sources included scientific journals, books, conference proceedings, and credible online resources. The findings from this review served as the basis for developing the conceptual framework, selecting features, and determining the model architecture.

3.2. Development Method

This study adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) development approach, which consists of five main stages: research understanding, data understanding, data preparation, modeling, and evaluation.

3.2.1. Research Understanding

In this study, the Business Understanding phase from the CRISP-DM framework is conceptually adapted into a Research Understanding stage to align with an academic research context rather than an industrial application. The main problem addressed in this study is how to automatically classify the types of mad rules (mad thabi'i, mad 'aridh lissukun, mad lazim mutsaqqal kilmi) from recordings of Qur'anic recitation. The challenge in this classification arises from the dynamic nature of audio signals, both in terms of frequency and temporal context. To address this, a combination of four acoustic features (MFCC, Chroma, Spectral Contrast, and RMS) is employed and processed using a CNN-LSTM–based deep learning model. The main objective of this stage is to develop an accurate classification system with evaluation metrics including accuracy, precision, recall, and F1-score.

3.2.2. Data Understanding

The dataset consists of 212 recordings of Surah Al-Fatihah obtained from the Quran Central website. From these recordings, a total of 3,816 audio segments were produced, each containing a single type of mad rule that was manually annotated by the author. Although the recordings were obtained from multiple qari, the dialectal variations among recitations available on the Quran Central platform are relatively limited. Therefore, dialect diversity was not considered a primary focus of this study. Moreover, the recitations generally exhibit uniform tempo characteristics, and as such, tempo variation was likewise excluded from the scope of analysis. Surah Al-Fatihah was chosen as the focus of this study because it is the opening chapter of the Qur'an and the most frequently recited surah in daily prayers [20], making it the most representative and pedagogically relevant for initial model development. The study specifically examined three mad types (mad ṭabī 'ī, mad 'āridh lissukūn, and mad lāzim mutsaqqal kilmī) that naturally occur within this surah. Nevertheless, it is acknowledged that using only Surah Al-Fatihah may limit the model's generalizability to other recitation contexts. Future research will therefore expand the dataset to include additional surahs and various qari/qari'ah to enhance robustness and adaptability across broader Qur'anic readings.

3.2.3. Data Preparation

The first step of data preparation is data labeling. At this stage, the author conducted a data annotation process to mark the significant segments of the audio signal that contain mad recitations. This process represents a crucial step in the construction of the dataset that will be utilized for training the classification model. The annotation was performed manually to ensure the accuracy of the timestamps for each occurrence of mad recitation within the audio data. The figure 2 below serves as a reference for the positions of the mad rules in Surah Al-Fatihah.

Figure 2. Mad Rule Positions in Surah Al-Fatihah

The annotation process for both audio and textual transcripts was validated by Dr. Ahmad Fudhaili, M.Ag., Head of the Qur'anic Studies and Exegesis Program and a national MTQ judge, ensuring strict adherence to tajwid principles. Although labeling was performed by a single annotator, all annotations were verified and cross-checked by the expert to maintain accuracy and consistency. Due to the specialized nature of mad classification, expert validation was prioritized over multi-annotator labeling; hence, inter-annotator agreement was not calculated, and reliability was established through expert consensus and repeated validation sessions.

Annotation was conducted using Label Studio, an open-source platform supporting time-based audio labeling. Annotators marked the start and end points of each mad segment in accordance with tajwid standards, and the resulting annotations were exported in CSV format to segment audio files into class-specific folders, each labeled with timestamp and mad type. Figure 3 illustrates the annotation interface in Label Studio, showing how each mad segment was visually marked and categorized during the labeling process.

Figure 3. Data Labeling using Label Studio

After the data labeling process, the subsequent step involves extracting acoustic features from the audio signals. In this study, four types of acoustic features are extracted from each audio segment using the Librosa library implemented in Python. These features include Mel Frequency Cepstral Coefficients which represent the spectral characteristics of audio signals based on the Mel scale, Chroma features which describe the distribution of energy across twelve pitch classes, Spectral Contrast which captures the difference between spectral peaks and valleys across frequency bands, and Root Mean Square energy which reflects the overall intensity of the audio signal. All extracted features are stored in JSON format and subsequently normalized to ensure compatibility and stability during the model training process. The features were stored in JSON format and normalized for use as model input. Figure 4 presents an example of the extracted feature dataset, showing how the MFCC, Chroma, Spectral Contrast, and RMS values are structured alongside their corresponding labels and file paths before being processed by the CNN-LSTM model.

Figure 4. Extracted Features Stored in JSON File

The extracted features will be used as an input data for the model training and testing. The data were split into 70% for training and 30% for testing using train_test_split with class stratification. In addition, stratified 8-Fold Cross-Validation was applied to ensure fair and consistent model evaluation.

3.2.4. Modeling

ISSN 2723-6471

Following the completion of data preparation, the next step was to develop and train a classification model capable of recognizing mad rules based on the extracted acoustic features. The objective was to construct a system that automatically classifies mad types by capturing both spatial and temporal patterns in the audio signals. To achieve this, a hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks was implemented. The proposed architecture is as shown in the figure 5.

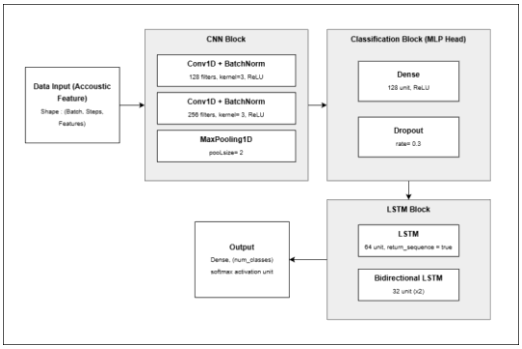

Figure 5. Proposed CNN-LSTM architecture

The proposed CNN-LSTM architecture was designed using the Keras library within TensorFlow. Two convolutional layers (128 and 256 filters, kernel size = 3, ReLU activation, L2 regularization) with batch normalization were employed to extract spatial representations, followed by a max-pooling layer to reduce temporal resolution. The output was then processed by a two-stage recurrent module consisting of an LSTM layer (64 units, return_sequences=True) and a bidirectional LSTM (32 units), enabling the model to capture temporal dependencies in both forward and backward directions. Finally, the network included a fully connected dense layer (128 units, ReLU) with dropout (0.3) and a softmax output layer corresponding to the three mad classes. This architecture was designed to balance complexity and performance while improving generalization through the use of regularization, normalization, and dropout.

The architecture design consisting of two convolutional layers followed by an LSTM and a Bidirectional LSTM (Bi-LSTM) was chosen to balance feature extraction depth and computational efficiency. Two convolutional layers were sufficient to capture the local spectral patterns of Qur'anic audio signals without overfitting, as deeper stacks tended to cause loss of temporal resolution in preliminary tests. The subsequent LSTM and Bi-LSTM layers were included to model the sequential and bidirectional temporal dependencies inherent in recitation patterns, such as gradual vowel elongation in mad sounds. This hybrid structure has been widely adopted in audio and speech-related studies for achieving a good trade-off between spatial and temporal representation power while maintaining training stability and interpretability. Hence, the selected configuration provided an effective yet computationally practical solution for mad rule classification.

The model received input tensors with dimensions of (batch_size, 100, 33), where 100 represents the number of temporal frames and 33 corresponds to the combined set of acoustic features. A GridSearchCV technique was employed to determine the optimal hyperparameters for model training [21], ensuring that the following configuration represents the best-performing setup obtained. This approach ensures that the selected parameters correspond to the best performing configuration identified during the optimization process.

Based on the results of the hyperparameter tuning, the model is trained using the Adam optimizer with a learning rate of 0.001. The training process is conducted for 20 epochs with a batch size of 16. In addition, an EarlyStopping callback is implemented to prevent overfitting by terminating the training process when the model performance on the validation data no longer shows improvement

To evaluate the contribution of each acoustic feature, fifteen training and testing scenarios were designed using different combinations of MFCC, Chroma, Spectral Contrast, and Root Mean Square (RMS). MFCC captures

perceptual spectral details, Chroma represents tonal information, Spectral Contrast reflects energy distribution, and RMS indicates signal intensity. All scenarios, covering single to multi-feature configurations, were trained under identical hyperparameters to ensure fair comparison. This design allowed a systematic analysis of individual and joint feature effects on mad classification performance, as summarized in table 1.

Table 1. Experimental Scenario

Scenario	MFCC	Spectral	Chroma	RMS	Description
1	✓				Single feature: MFCC
2		✓			Single feature: Spectral
3			✓		Single feature: Chroma
4				✓	Single feature: RMS
5	✓	✓			Two-feature combination
6	✓		✓		Two-feature combination
7	✓			✓	Two-feature combination
8		✓	✓		Two-feature combination
9		✓		✓	Two-feature combination
10			✓	✓	Two-feature combination
11	✓	✓	✓		Three-feature combination
12	✓	✓		✓	Three-feature combination
13	✓		✓	✓	Three-feature combination
14		✓	✓	✓	Three-feature combination
15	✓	✓	✓	✓	All features combined

The design of the 15 experimental scenarios was intentionally structured to provide a systematic and interpretable assessment of the contribution of each acoustic feature and their possible interactions. Instead of relying on automated feature selection techniques such as Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA), this study employed a comprehensive factorial combination strategy. Each of the four acoustic features captures distinct and complementary aspects of Qur'anic recitation, including phonetic, tonal, spectral, and energy-based information. Therefore, constructing all possible single-, dual-, triple-, and full-feature combinations (4C1 + 4C2 + 4C3 + 4C4 = 15) allowed for a complete evaluation of how each feature, individually or in combination, influences model performance. This design ensures that the effect of each feature inclusion or exclusion is empirically observed under identical model and training conditions, thereby maintaining methodological rigor while preserving domain interpretability in the context of Qur'anic speech analysis.

4. Results and Discussion

4.1. Model Evaluation

At this stage, the performance of the CNN-LSTM model was evaluated based on 15 scenarios of different acoustic feature combinations. Each scenario involved either a single feature, a combination of two features, or the complete combination of the four features used in this study: MFCC, Chroma, Spectral Contrast, and Root Mean Square (RMS). The evaluation was conducted using the test dataset, measuring the average accuracy, precision, recall, and F1-score metrics. The following table 2 presents the complete results for all scenarios.

Table 2. Testing Result

Scenario	Feature Combination	Accuracy	Precision (Avg)	Recall (Avg)	F1-Score (Avg)
1	MFCC	96.33%	91.78%	95.30%	93.35%
2	Spectral	92.93%	92.56%	82.25%	86.03%
3	Chroma	93.19%	89.00%	89.42%	89.19%

4	RMS	91.62%	84.43%	85.14%	84.78%
5	MFCC + Spectral	96.77%	96.10%	94.57%	95.30%
6	MFCC + Chroma	96.51%	94.29%	94.25%	94.27%
7	MFCC + RMS	96.24%	95.03%	93.95%	94.48%
8	Spectral + Chroma	93.19%	88.64%	91.37%	89.92%
9	Spectral + RMS	92.75%	86.46%	90.40%	88.17%
10	Chroma + RMS	91.88%	84.81%	85.75%	85.25%
11	MFCC + Spectral + Chroma	96.59%	95.37%	94.14%	94.73%
12	MFCC + Spectral + RMS	97.03%	95.09%	95.13%	95.11%
13	MFCC + Chroma + RMS	96.16%	94.73%	94.20%	94.44%
14	Spectral + Chroma + RMS	93.19%	88.20%	92.56%	90.11%
15	MFCC + Spectral + Chroma + RMS	97.21%	95.28%	95.22%	95.25%

The results demonstrate that MFCC consistently delivered the most stable and accurate performance, both individually and in combination with other features. Nearly all MFCC-based configurations achieved accuracy above 96% and F1-scores above 94%, confirming its effectiveness in representing the phonetic characteristics of mad recitations. Feature combinations generally outperformed single features, indicating the complementary nature of acoustic attributes. Among all configurations, MFCC combined with Spectral Contrast and RMS (Scenario 12) and the full-feature model (Scenario 15) achieved the best results, with accuracies exceeding 97% and F1-scores of 95.11% and 95.25%.

The superior performance of MFCC stems from its mel-scale representation, which mimics human auditory perception and effectively captures key spectral cues [7] such as vowel resonances and formant transitions crucial to distinguishing subtle phonetic variations in Qur'anic recitations. In contrast, Chroma, Spectral Contrast, and RMS emphasize tonal, spectral, or amplitude characteristics that only partially represent the complex articulatory dynamics of mad recitations. Consequently, MFCC provides a more discriminative and perceptually relevant representation, resulting in superior classification accuracy.

On the other hand, RMS and Spectral Contrast performed weakest when used independently, with F1-scores of only 84.78% and 86.03%. Although both features capture relevant aspects of signal intensity and spectral variation, they were insufficient on their own for robust classification. Furthermore, while combining multiple features generally improved performance, the improvements were not always substantial. For instance, MFCC combined with Spectral Contrast (Scenario 5) produced an F1-score of 95.30%, which was only slightly lower than the full-feature model (Scenario 15) at 95.25%, suggesting diminishing returns from adding more features beyond certain combinations. Another notable finding is that the classification performance for Mad Lazim Mutsaqqal Kilmi was consistently lower than for the other two classes, with precision, recall, and F1-scores frequently below 91%. This weakness was most evident in single-feature experiments, such as Scenario 2 (F1-score: 71.03%) and Scenario 4 (F1-score: 69.63%). The performance gap is likely attributable to class imbalance, as the dataset contained significantly fewer samples for Mad Lazim. Addressing this limitation through data augmentation or oversampling is suggested for future studies.

Although the lower recognition rate of *Mad Lāzim Mutsaqqal Kilmī* was mainly due to class imbalance, no oversampling or synthetic data generation was performed to maintain the natural phonetic characteristics of Qur'anic recitations. This ensured that the model learned from authentic vocal variations rather than artificial samples. Future studies, however, could apply data augmentation techniques such as pitch shifting, time-stretching, or background noise addition to enhance minority class representation and improve model generalization, especially for rare *tajwīd* patterns like *Mad Lāzim Mutsaqqal Kilmī*.

Overall, the findings confirm that combining multiple acoustic features leads to more informative and discriminative audio representations. Among all tested scenarios, the CNN-LSTM model trained with the complete set of features demonstrated the most stable and superior performance across evaluation metrics, proving its effectiveness in Qur'anic recitation classification.

ISSN 2723-6471

4.2. Best Scenario

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations. Scenario 15, which used the complete combination of all features, MFCC, Chroma, Spectral Contrast, and RMS, achieved the highest performance across all evaluation metrics. Figure 6 illustrates the training and validation performance trends of the 15th scenario.

Figure 6. 15th scenario training and validation loss and accuracy graph

The model demonstrated a stable training trend. The training accuracy increased consistently, while the validation accuracy remained high. Early stopping was triggered at around the 14th epoch. Figure 7 shows the confusion matrix of the best-performing scenario, table 3 summarizes the corresponding evaluation metrics.

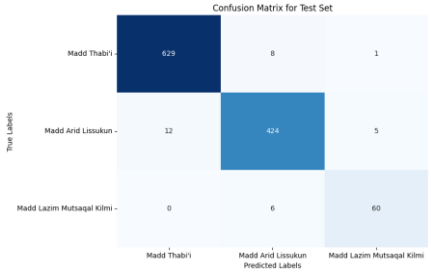

Figure 7. Confusion Matrix from the Best Scenario

Table 3. Evaluation Metric Result from the 15th Scenario

Metric	Mad Thabi'i	Mad Aridh	Mad Lazim	Average
Precision	0.98	0.97	0.91	0.95
Recall	0.99	0.96	0.91	0.95
F1-score	0.98	0.96	0.91	0.95
Accuracy				97.21%

A detailed examination of the confusion matrix reveals that most misclassifications occurred between *mad 'āridh lissukūn* and *mad lāzim mutsaqqal kilmī*, indicating that the model occasionally struggles to differentiate between these two categories. This confusion can be attributed to their acoustic resemblance, as both exhibit prolonged vowel durations and similar spectral envelopes, especially when recited with minimal pauses or at moderate tempo. Moreover, *mad lāzim* segments were underrepresented in the dataset, which likely contributed to reduced discriminative power for that class. These findings highlight that the model's main limitation lies in distinguishing acoustically overlapping mad types rather than in recognizing *mad ṭabī 'ī*, which consistently achieved near-perfect classification. Consequently, future work should emphasize dataset balancing and the inclusion of more diverse recitation samples to mitigate this bias and further enhance model robustness across all mad categories.

ISSN 2723-6471

4.3. Result Analysis

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

The evaluation results demonstrate that MFCC consistently yields the best performance in mad rule classification, both individually and when combined with other features. Its strong phonetic representation makes it the core feature in the classification process. Integrating MFCC with Spectral Contrast and RMS further enhances performance, with the best configurations MFCC + Spectral Contrast + RMS and the full-feature model achieving F1-scores of 95.11% and 95.25% respectively. These results highlight the benefit of combining spectral, tonal, and energy-based features to improve data representation and model generalization. However, performance gains diminish beyond certain combinations, suggesting that once the model receives sufficient representative features, additional complexity offers only marginal improvement.

Although this study utilized 3,816 annotated audio segments, the dataset size remains relatively limited for deep learning and may constrain model generalization. To address this, data were collected from multiple *qari* and *qari'ah* with diverse recitation speeds, pronunciation styles, and recording qualities, while stratified 8-fold cross-validation ensured consistent evaluation. These measures improved generalization despite the dataset's modest scale. Future work aims to expand the corpus with additional *surahs* and reciters to strengthen robustness and adaptability. The current evaluation primarily focused on classification accuracy across different acoustic feature combinations, with attention also given to model robustness under varying reciters and recording conditions. Since all recordings originated from the Quran Central platform, dialectal and environmental diversity remained limited. Subsequent studies will incorporate broader regional and acoustic variations to enable a more comprehensive assessment of the CNN-LSTM model's robustness and real-world applicability across diverse Qur'anic recitation contexts. Figure 8, figure 9, and figure 10 illustrate the precision, recall, and F1-score comparisons across the 15 experimental scenarios.

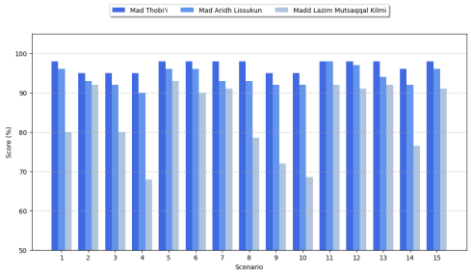

Figure 8. Precision Comparison of Each Mad class Across the 15 Experimental Scenarios

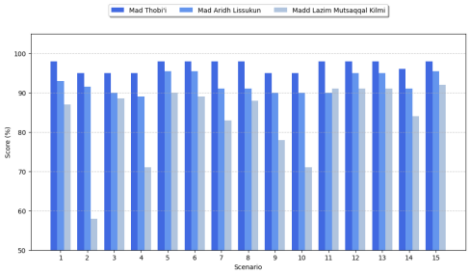

Figure 9. Recall Comparison of Each Mad class Across the 15 Experimental Scenarios

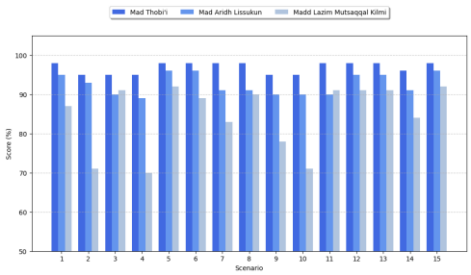

Figure 10. F1-Score Comparison of Each Mad class Across the 15 Experimental Scenarios

ISSN 2723-6471

Several strategies were applied to prevent overfitting, given the model's hybrid CNN-LSTM architecture and moderately sized dataset. The network depth was limited to two convolutional and two recurrent layers to maintain a balance between representation capacity and generalization. L2 regularization, batch normalization, and a 0.3 dropout rate were used to control weight magnitudes and prevent neuron over-reliance, while early stopping halted training when validation loss plateaued. Additionally, stratified 8-fold cross-validation ensured robust evaluation across independent data partitions, minimizing bias and inflated accuracy.

A per-class evaluation further revealed consistent performance trends across mad ṭabī 'ī, mad 'āridh lissukūn, and mad lāzim mutsaqqal kilmī. The mad ṭabī 'ī class achieved the highest precision and recall, while mad lāzim showed lower scores due to class imbalance and acoustic similarity to mad 'āridh. However, scenarios 12 and 15 significantly improved mad lāzim recognition, with F1-scores exceeding 90%, confirming that integrating complementary acoustic features enhances inter-class balance and model robustness. The inclusion of integrating multiple acoustic features improved model accuracy but increased computational cost, as training time rose from 18 minutes per epoch with MFCC alone to 31 minutes with all features, while GPU memory usage grew from 2.1 GB to 4.2 GB; given the marginal performance gain of only +0.18% accuracy, the MFCC + Spectral Contrast + RMS configuration offers the most efficient balance between accuracy and resource usage.

4.4. Comparison with Previous Study

This research demonstrates a significant advancement in the classification accuracy of mad rules. The proposed CNN-LSTM model leveraging four acoustic features (MFCC, Chroma, Spectral Contrast, and RMS) achieved a testing accuracy of 97.21%. This score surpasses the findings presented in [16] where the LSTM algorithm and the MFCC method were utilized to detect mad rules in Surah Al-Fatihah, obtaining an accuracy of 90.00%. Furthermore, it also exceeds another study [22], which employed the Dynamic Time Warping (DTW) method to measure the duration of Harakaat within a single syllable, showing an overall testing accuracy of 80.47%.

However, these comparisons should be interpreted with caution, as variations in datasets [23], verse coverage, and labeling strategies [24] across studies affect performance outcomes. Previous works often relied on small, single-surah datasets and limited evaluation methods, while this study employed stratified 8-fold cross-validation, multiple feature combinations, and refined preprocessing, including normalization, expert annotation, and balanced scaling. These methodological improvements enhanced both fairness and robustness in evaluation.

Overall, the superior accuracy obtained in this research reflects not only the architectural strength of the CNN-LSTM hybrid in combining spatial and temporal feature extraction [25] but also the comprehensive experimental design and feature integration that improved the model's generalization and reliability in automatic mad rule classification.

5. Conclusion

Based on the results of this study, it can be concluded that a mad rule classification model for Qur'anic recitation was successfully developed using the CNN-LSTM architecture with the combined input of four acoustic features: MFCC, Chroma, Spectral Contrast, and RMS. Among the 15 tested feature combination scenarios, the best performance was achieved when all four features were used together, yielding an accuracy of 97.21% and an F1-score of 95.25%. These results indicate that feature combination makes a significant contribution to accurately recognizing the phonetic patterns of mad recitation. MFCC proved to be the most dominant feature, while other features such as RMS and Spectral Contrast provided a meaningful impact only when combined. The addition of acoustic features also improved model performance, although not always significantly, indicating that the relevance of information from each feature is key to building an effective and efficient classification model.

Despite its promising results, this study acknowledges several limitations. The dataset was restricted to recitations of Surah Al-Fatihah, which may constrain the model's ability to generalize across other Qur'anic chapters. Future research is therefore recommended to employ a larger and more diverse dataset encompassing multiple Surahs. Additionally, exploring recitations with varying dialects and tempos could further improve model robustness and adaptability to different recitation styles. Beyond research settings, the proposed model holds potential for integration into mobile or web-based platforms, providing accessible tools for learners, educators, and researchers in Qur'anic recitation studies. Such deployment would enable real-time classification and feedback, enhancing the model's practical applicability.

6. Declarations

6.1. Author Contributions

Conceptualization: F.A., and S.A.; Methodology: F.A., S.A., and T.A.; Software: T.A., and L.A.; Validation: S.A., T.A., and L.A.; Formal Analysis: F.A., and S.A.; Investigation: F.A.; Resources: S.A., and L.A.; Data Curation: F.A.; Writing Original Draft Preparation: F.A., S.A., and T.A.; Writing Review and Editing: F.A.; Visualization: F.A.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Commented [JADS1]: Please confirm or you can modify it directly.

References

[1] N. Anggraini, A. Kurniawan, L. K. Wardhani, and N. Hakiem, "Speech Recognition Application for the Speech Impaired using the Android-based Google Cloud Speech API," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 16, no. 6, pp. 2733–2739, Dec. 2018, doi: 10.12928/telkomnika.v16i6.9638.

[2] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 20, no. 1, pp. 61–79, Oct. 1998, doi: 10.1023/A:1008066223044.

[3] H. Ilhami, A. Khafizah, and H. Nor, "Islamic ethical perspectives on AI development and use," *in Towards Resilient Societies: The Synergy of Religion, Education, Health, Science, and Technology, CRC Press*, vol. 2025, no. 1, pp. 562–567, 2025.

[4] N. Aljohani and E. Jaha, "Visual Lip-Reading for Quranic Arabic Alphabets and Words Using Deep Learning," *CSSE*, vol. 46, no. 3, pp. 3037–3058, 2023, doi: 10.32604/csse.2023.037113.

[5] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An Ensemble of Convolutional Neural Networks for Audio Classification," *Applied Sciences*, vol. 11, no. 13, pp. 1-16, June 2021, doi: 10.3390/app11135796.

[6] R. Li, B. Yin, Y. Cui, Z. Du, and K. Li, "Research on Environmental Sound Classification Algorithm Based on Multi-feature Fusion," *in 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 2020, no. Dec., pp. 522–526, 2020. doi: 10.1109/ITAIC49862.2020.9338926.

[7] Md. A. Hossan, S. Memon, and M. A. Gregory, "A Novel Approach for MFCC Feature Extraction," *in 2010 4th International Conference on Signal Processing and Communication Systems*, vol. 2010, no. Dec., pp. 1–5, 2010. doi: 10.1109/ICSPCS.2010.5709752.

[8] C. Weiss and M. Müller, "Tonal Complexity Features for Style Classification of Classical Music," *in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2015, no. Apr., pp. 688–692, 2015. doi: 10.1109/ICASSP.2015.7178057.

[9] M. L. Massar, M. Fickus, E. Bryan, D. T. Petkie, and A. J. Terzuoli, "Fast Computation of Spectral Centroids," *Adv Comput Math*, vol. 35, no. 1, pp. 83–97, July 2011, doi: 10.1007/s10444-010-9167-y.

[10] R. Nekoutabar, F. S. Ghaheri, and H. Jalilvand, "The Effect of Root-Mean-Square and Loudness-Based Calibration Approach on the Acceptable Noise Level," *Auditory and Vestibular Research*, vol. 33, no. 4, pp. 1-12, Oct. 2024, doi: 10.18502/avr.v33i4.16654.

[11] O. Theobald, "The Machine Learning Toolbox," *in Machine Learning For Absolute Beginners*, 3rd ed., London: Scatterplot Press, 2021, pp. 25–33.

[12] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artif Intell Rev*, vol. 53, no. 8, pp. 5929–5955, Dec. 2020, doi: 10.1007/s10462-020-09838-1.

[13] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[14] J. Zhao, X. Mao, and L. Chen, "Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks," *Biomedical Signal Processing and Control*, vol. 47, no. Jan., pp. 312–323, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035.

[15] S. Al-Hagree, A. Alawdi, H. A. Alsayadi, M. Alsurori, and M. A. Al Sabri, "Machine Learning Techniques for Identifying Tajweed Rules: A Case Study on Noon Saakin and Tanween," *in 2024 1st International Conference on Emerging Technologies for Dependable Internet of Things (ICETI)*, vol. 2024, no. Nov., pp. 1–7, 2024. doi: 10.1109/ICETI63946.2024.10777164.

[16] N. Anggraini, Zulkifli, Y. Rahman, A. N. Hidayanto, and H. T. Sukmana, "Modeling Madd Reading Classification in Surah Al-Fatihah with MFCC Feature Extraction and LSTM Algorithm," *in 2024 12th International Conference on Cyber and IT Service Management (CITSM)*, vol. 2024, no. Oct., pp. 1–7, 2024. doi: 10.1109/CITSM64103.2024.10775370.

[17] F. Ahmad, S. Z. Yahya, Z. Saad, and A. R. Ahmad, "Tajweed Classification Using Artificial Neural Network," *in 2018 International Conference on Smart Communications and Networking (SmartNets)*, Aachen, Germany, vol. 2018, no. Nov., pp. 1–4, 2018. doi: https://doi.org/10.1109/SMARTNETS.2018.8707394.

[18] G. Samara, E. Al-Daoud, N. Swerki, and D. Alzu'bi, "The Recognition of Holy Qur'an Reciters Using the MFCCs' Technique and Deep Learning," *Advances in Multimedia*, vol. 2023, no. Mar., pp. 1–14, Mar. 2023, doi: 10.1155/2023/2642558.

[19] G. Tzanetakis, "Audio Feature Extraction," in Music Data Mining, Boca Raton: CRC Press, 2011, pp. 41–73.

[20] N. Suri and M. Tanjung, "Metaphor and Symbolism in the Language of the Quran: A Linguistic Study on the Concept of Tauhid (Analysis of Surah al-Fatihah).," *Pharos Journal of Theology*, vol. 106, no. 1, pp. 1-12, Jan. 2025, doi: 10.46222/pharosjot.106.3.

[21] G. S. K. Ranjan, A. Kumar Verma, and S. Radhika, "K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries," *in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, vol. 2019, no. Mar., pp. 1–5, 2019. doi: 10.1109/I2CT45611.2019.9033691.

[22] N. Shafie, A. Azizan, M. Z. Adam, H. Abas, Y. M. Yusof, and N. A. Ahmad, "Dynamic Time Warping Features Extraction Design for Quranic Syllable-Based Harakaat Assessment," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, pp. 1-12, 2022, doi: https://dx.doi.org/10.14569/IJACSA.2022.0131207.

[23] A. Bailly et al., "Effects of Dataset Size and Interactions on the Prediction Performance of Logistic Regression and Deep Learning Models," *Computer Methods and Programs in Biomedicine*, vol. 213, no. Jan., pp. 1-14, Jan. 2022, doi: 10.1016/j.cmpb.2021.106504.

[24] Y. Wang, A. E. Mendez Mendez, M. Cartwright, and J. P. Bello, "Active Learning for Efficient Audio Annotation and Classification with a Large Amount of Unlabeled Data," *in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2019, no. may, pp. 880–884. doi: 10.1109/ICASSP.2019.8683063.

[25] V. Passricha and R. K. Aggarwal, "A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition," *Journal of Intelligent Systems*, vol. 29, no. 1, pp. 1261–1274, Dec. 2019, doi: 10.1515/jisys-2018-0372.