JEEMECS (Journal of Electrical Engineering, Mechatronic and Computer Science ISSN 2614-4859
Vol. 8, No. 2, August 2025, pp. 91-97

Dataset Ratio Experiment for Tajweed Law Recognition Using MFCC and CNN Features

Mas Muhammad Aqil Salim [a,1,*], Ani Dijah Rahajoe [b,2], Anggraini Puspita Sari [a,3]

[a] Informatics Engineering, UPN Veteran Jawa Timur, Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, East Java (60294), Indonesia
[b] Department of Computer Science, UPN Veteran Jawa Timur, Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, East Java (60294), Indonesia
[1] aqilsalimm123@gmail.com [*]; [2] anidijah.if@upnjatim.ac.id; [3] anggraini.puspita.if@upnjatim.ac.id
[*] corresponding author

ABSTRACT

Keywords
Audio Classification, Convolutional Neural Network, Mel-Frequency Cepstral Coefficient, Nun Sukun, Tajwid Detection

This research aims to develop a tajwid classification system focusing on the detection of nun sukun and tanwin rules in Quranic recitation using the Convolutional Neural Network (CNN) and Mel-Frequency Cepstral Coefficient (MFCC) methods. The dataset used includes 1,344 audio samples collected from both direct recordings and YouTube observations. Audio preprocessing involved silence removal and noise filtering, followed by feature extraction using MFCC with 40 coefficients. These features were then classified into six categories: Idghom Bighunnah, Idghom Bilaghunnah, Idzhar Halqi, Ikhfa Haqiqi, Iqlab, and No Class. The CNN architecture implemented includes three convolutional layers with Batch Normalization and Leaky ReLU activation, optimized with a softmax classifier. Three different dataset split scenarios (80:10:10, 70:15:15, and 60:20:20) were evaluated to determine the best performance. The highest accuracy of 89% was achieved using the 80:10:10 data split, with macro-average F1-score reaching 0.87. Results show that CNN combined with MFCC provides reliable classification of tajwid rules, particularly in identifying distinctive acoustic patterns. The study confirms that data partitioning significantly influences model performance and highlights the importance of optimal preprocessing and architecture selection in deep learning-based speech recognition tasks.

1. Introduction

The rapid advancement of information technology has influenced various aspects of human life, including education, healthcare, politics, economics, and religion. For Muslims, this technological development opens new opportunities to access Islamic knowledge through digital tools. One key area of learning is the correct recitation of the Qur'an, which requires mastery of tajwid, the rules that govern pronunciation and articulation. Although most Muslims can read the Qur'an, a significant portion still lacks proper knowledge of tajwid. According to a 2023 survey by the Directorate General of Islamic Community Guidance, only about 44.57% of Indonesian Muslims are able to apply tajwid rules accurately when reading the Qur'an [1].

A fundamental rule often taught at the beginning of tajwid learning is hukum nun sukun and tanwin, which occurs frequently in the Qur'an. Its high frequency makes it a crucial element in understanding tajwid comprehensively. As Lailatul Husna (2023) and Rudi Dian Arifin (2023) explain, nun sukun and tanwin appear in many word structures throughout the Qur'an, thus requiring special attention in tajwid instruction [2][3]. Unfortunately, tajwid learning is still mostly manual, relying on limited face-to-face instruction, and therefore not always sufficient—especially for beginners.

Given these challenges, there is an increasing need for technology-based solutions that can help individuals learn tajwid more independently. One such approach involves the use of Artificial Intelligence (AI). According to Kaplan et al. (2019), AI enables computers to perform tasks that mimic or surpass human intelligence [4]. A subfield of AI, Computer Vision, is widely used in

recognizing patterns in image and audio data, including tasks like object detection, image classification, and speech recognition.

In this study, Convolutional Neural Networks (CNNs) are applied to recognize acoustic patterns in Qur'anic recitation. Although CNNs are traditionally known for image classification, they are increasingly used for audio data through visual representations such as spectrograms, mel-spectrograms, and cepstral coefficients. Combined with Mel-Frequency Cepstral Coefficients (MFCCs) a method that captures human auditory characteristics—CNNs can be trained to detect variations in tajwid pronunciation. MFCCs extract acoustic features that reflect differences in articulation, tone, and duration, which are essential in identifying tajwid patterns [5].

Previous studies have implemented CNN and MFCC for tasks such as emotion classification in Qur'anic recitation and general voice recognition. These studies demonstrate that MFCCs serve as strong input features for CNN models in audio-based classification, even in limited datasets [6]. This supports the development of tajwid classification systems based on these methods.

What differentiates this research is its focus on evaluating the impact of different dataset split ratios on the classification performance of the CNN-MFCC system. Deep learning models are highly influenced by how data is distributed between training, validation, and test sets. Therefore, this paper specifically investigates three different ratios: 80:10:10, 70:15:15, and 60:20:20, to determine which partition provides the most optimal accuracy for detecting hukum nun sukun and tanwin. The results aim to support future improvements in automated tajwid learning systems by ensuring the data configuration is well-optimized for model training and evaluation.

2. The Proposed Method

2.1. Flowchart Research Method

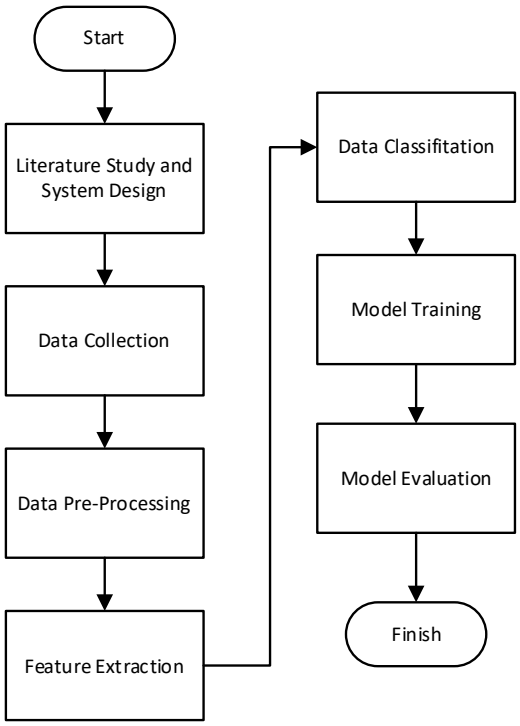

Fig 1. Flowchart research method

This research is structured into several key stages, starting with the collection of Qur'anic recitation audio data categorized into six classes: Idghom Bighunnah, Idghom Bilaghunnah, Idzhar Halqi, Ikhfa Haqiqi, Iqlab, and a comparison class labeled No Class. The audio data then undergo a pre-processing phase that includes trimming or padding to 2 seconds, removal of silence, and background noise reduction. These steps ensure a cleaner and more uniform input for feature extraction.

The processed audio is then converted into Mel-Frequency Cepstral Coefficients (MFCC), forming two-dimensional representations suitable for deep learning. These MFCC features are passed to a Convolutional Neural Network (CNN) model designed to classify the tajwid patterns based on the learned spectral features [7]. The model uses a softmax activation in the output layer to generate multi-class predictions. The final stage involves system evaluation using unseen data to determine model accuracy and classification performance.

3. Method

3.1. Mel-Frequency Cepstral Coefficient

In this study, the Mel-Frequency Cepstral Coefficients (MFCC) method was employed to extract relevant acoustic features from Qur'anic recitation audio. The process begins by segmenting the audio into overlapping frames (framing), followed by the application of a Hamming window to minimize spectral leakage. Each frame is then transformed into the frequency domain using the Fast Fourier Transform (FFT), and its power spectrum is calculated. This power spectrum is passed through a set of Mel-scale filter banks to mimic human auditory perception, emphasizing frequencies that are more perceptually relevant [8].

The output from the filter banks is converted into MFCCs using the Discrete Cosine Transform (DCT), which reduces data dimensionality while preserving the most salient frequency components [9]. A fixed number of coefficients (n_mfcc) are selected, and padding is applied if needed to standardize input dimensions for batch processing. The resulting MFCC representations are stored as .npz files and serve as input features for the CNN classifier. This approach enables the model to capture essential characteristics of tajwid pronunciation in a structured and compact format suitable for deep learning.

3.2. Split Dataset Scenario

In this study, the dataset comprises a total of 1,344 audio samples of Qur'anic recitation, which were collected through direct recordings and supplementary observations from publicly available YouTube channels. After merging both sources, the data underwent pre-processing to produce a standardized, noise-free dataset ready for feature extraction and classification. Each sample was then converted into MFCC representations for input to the CNN model [10].

To evaluate the effect of dataset distribution on model performance, three data-splitting scenarios were tested using a randomized shuffle method: $80:10:10$, $70:15:15$, and $60:20:20$, representing training, validation, and test sets respectively. All other training parameters were held constant to ensure consistency: 40 MFCC coefficients, a maximum input length of 400 frames (approx. 2 seconds), a batch size of $8,100$ pre-training epochs, and 50 fine-tuning epochs. The goal of this experiment was to determine the most effective data ratio for maximizing classification accuracy. According to recent studies, optimal data splitting plays a crucial role in improving a model's generalization capability [11].

3.3. Convolutional Neural Network

The CNN architecture in this study consists of three convolutional layers followed by fully connected layers, incorporating techniques such as Leaky ReLU and Batch Normalization to address challenges like vanishing gradients and overfitting. Leaky ReLU is used to prevent dead neurons often encountered with traditional ReLU activations [12] [13], while Batch Normalization is employed to accelerate training convergence and stabilize the internal covariate shift across mini-batches [14].

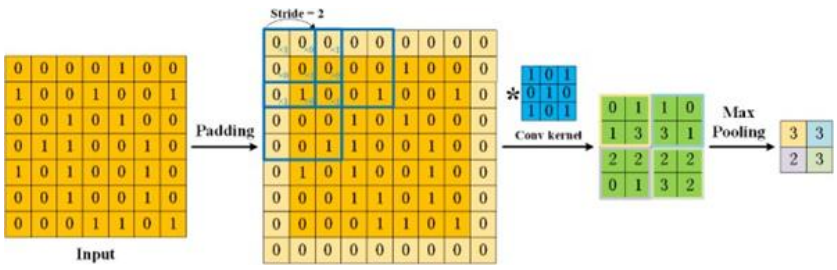

Fig 2. Convolutional neural network architecture

In Layer 1, the input MFCC features are processed through a Conv2D layer with 32 filters of size 4×4 and 'same' padding, followed by Leaky ReLU activation and Batch Normalization. Dimensionality is then reduced using MaxPooling2D with a pool size of 2×2. In Layer 2, 48 filters of size 4×10 are applied via Conv2D to extract more specific spectral patterns, followed by the same activation and normalization scheme, and reduced via AveragePooling2D. Layer 3 uses 64 filters of size 4×10, continuing the feature extraction process for complex patterns, and downscales via MaxPooling2D [15].

The output is passed through a GlobalAveragePooling layer to flatten the feature maps into a single dimension [16], followed by another Batch Normalization layer. It is then processed by a Dense layer with 64 neurons using Leaky ReLU activation and a dropout rate of 40% to prevent overfitting. Finally, the output layer consists of num_classes neurons activated by Softmax, allowing multi-class classification [12].

4. Results and Discussion

This section presents the evaluation of model performance based on three dataset splitting scenarios to assess how different training-validation-testing ratios affect CNN-based classification of tajwid audio. The dataset consisted of 1,344 MFCC-extracted audio samples representing six tajwid classes: Idghom Bighunnah, Idghom Bilaghunnah, Idzhar Halqi, Ikhfa Haqiqi, Iqlab, and a neutral class (No Class). The scenarios evaluated were: Scenario A (80:10:10) with 144 samples, Scenario B (70:15:15) with 208 samples, and Scenario C (60:20:20) with 280 samples.

Classification Report:

	precision	recall	f1-score	support
Idghom Bighunnah	0.78	0.75	0.77	24
Idghom Bilaghunnah	0.88	0.88	0.88	8
Idzhar Halqi	0.68	0.81	0.74	16
Ikhfa Haqiqi	0.92	0.88	0.90	40
Iqlab	0.94	1.00	0.97	16
No Class	1.00	0.97	0.99	40
accuracy			0.89	144
macro avg	0.87	0.88	0.87	144
weighted avg	0.89	0.89	0.89	144

Fig 3. Classification report on scenario A

Among the three, Scenario A (80: 10: 10) demonstrated the best overall classification result, achieving an accuracy of 89%, as presented in the classification report on Fig 3. Notably, No Class and Iqlab achieved F1-scores of 0.99 and 0.97, respectively, highlighting the model's strength in recognizing these acoustically distinct classes. Furthermore, the model maintained stable performance across other categories, such as Ikhfa Haqiqi (F1-score: 0.90) and Idghom Bilaghunnah (F1-score: 0.88). The macro and weighted averages for precision, recall, and F1-score all hovered around 0.87–0.89, indicating high reliability in class prediction consistency.

Classification Report:

	precision	recall	f1-score	support
Idghom Bighunnah	0.88	0.62	0.73	24
Idghom Bilaghunnah	0.79	0.82	0.80	40
Idzhar Halqi	0.67	0.75	0.71	16
Ikhfa Haqiqi	0.88	0.88	0.88	32
Iqlab	0.88	0.94	0.91	48
No Class	1.00	1.00	1.00	48
accuracy			0.87	208
macro avg	0.85	0.84	0.84	208
weighted avg	0.87	0.87	0.87	208

Fig 4. Classification report on scenario B

In Scenario B (70: 15: 15), as shown in Fig 4, the model achieved slightly lower accuracy at 87%. The performance drop was most noticeable in Idghom Bighunnah, which registered a recall of

0.62, suggesting a lower sensitivity to this class despite a relatively high precision. Nevertheless, the model maintained solid performance in recognizing Iqlab and No Class, both achieving F1-scores above 0.90, and showed robust average metrics with macro and weighted averages at 0.84– 0.87. This indicates moderate generalization, although further refinement may be required for more ambiguous phonetic patterns.

Classification Report:

	precision	recall	f1-score	support
Idghom Bighunnah	0.79	0.68	0.73	40
Idghom Bilaghunnah	0.83	0.75	0.79	72
Idzhar Halqi	0.71	0.71	0.71	48
Ikhfa Haqiqi	0.59	0.83	0.69	24
Iqlab	0.86	0.91	0.89	56
No Class	0.97	0.97	0.97	40
accuracy			0.80	280
macro avg	0.79	0.81	0.80	280
weighted avg	0.81	0.80	0.80	280

Fig 5. Classification report on scenario C

For Scenario C (60:20:20), depicted in Fig 5, the model recorded the lowest accuracy of 80%. While No Class and Iqlab again showed reliable classification results with F1-scores of 0.97 and 0.89, performance declined in categories like Ikhfa Haqiqi (F1: 0.69) and Idghom Bighunnah (F1: 0.73). This suggests that a greater dataset size although potentially more diverse introduced higher complexity that the model struggled to generalize from effectively. The weighted average F1-score of 0.80 reflects this dip in stability and precision across the board.

Table 1. Comparison result

No.	*Dataset Split*	*Accuracy (%)*	*Macro AVG F1-Score*	*Weighted AVG F1-Score*	*Loss*	*Total Samples*
1.	80 : 10 : 10	89	0,87	0,89	0,5061	144
2.	70 : 15 : 15	87	0,84	0,87	0,5103	208
3.	60 : 20 : 20	80	0,80	0,80	0,5168	280

As shown in Table 1, the best performance was achieved using the 80: 10: 10 data split, where the model reached an accuracy of 89% with a macro average F1-score of 0.87 and the lowest loss value of 0.5061. This scenario also had the smallest number of total samples, indicating that a balanced yet moderate training set may enhance the model's ability to learn distinguishing features more effectively. In contrast, the 60: 20: 20 scenario, despite using more data, resulted in the lowest accuracy and slightly degraded performance, suggesting a potential challenge in model generalization as the dataset becomes more complex.

In summary, the results confirm that Scenario A (80: 10: 10) provides the most optimal configuration for model training and evaluation, balancing training size with test set reliability. This experiment underlines that increasing dataset size does not always yield better performance and that thoughtful splitting strategy is crucial for achieving high and consistent classification accuracy in deep learning-based tajwid detection systems.

5. Conclusion

This study has demonstrated that the combination of Mel-Frequency Cepstral Coefficients (MFCC) and Convolutional Neural Network (CNN) is effective in classifying tajweed rules, particularly nun sukun and tanwin, from Qur'anic recitations. By evaluating three dataset split scenarios (80: 10: 10, 70: 15: 15, *and* 60: 20: 20), it was found that the 80: 10: 10 ratio yielded the highest performance with an accuracy of 89%, a macro F1-score of 0.87, and a low loss value of 0.5061. The model consistently performed well on classes with distinct phonetic features, such as No Class and Iqlab.

These results confirm that a well-balanced and moderate dataset split can significantly impact the CNN model's classification capability. Therefore, the 80: 10: 10 configuration is considered the optimal setup for training tajweed classification models using MFCC and CNN. Future work can

explore more advanced architectures or larger and more diverse datasets to further improve model generalization and robustness in real-world applications.

References

[1] N. Rihladhatul and A. Sayoga, "AL-AFKAR : Journal for Islamic Studies The Living Qur ' an : Ethnographic Study and Understanding of Al- Isra ': 9 in Urban Communities," vol. 7, no. 4, pp. 334–346, 2024, doi: 10.31943/afkarjournal.v7i4.1648.Abstract.

[2] H. Fajrussalam, W. Widaningsih, N. Nurdiansyah, L. N. Istighna, C. Nurjanah, and C. R. A. Widjaya, "Implementasi Program Maghrib Mengaji sebagai Sarana Pembelajaran Al-Quran pada Anak Usia Dini: Studi Deskriptif Kualitatif di TPQ Al Ittihad Kab. Cirebon," *At-Ta'lim Media Inf. Pendidik. Islam*, vol. 22, no. 2, p. 259, 2023, doi: 10.29300/attalim.v22i2.2546.

[3] R. D. Arifin, "No Title," Dianisa. [Online]. Available: https://dianisa.com/nun-sukun-dan-tanwin/

[4] M. H. Ashadiqi, A. Erlansari, and F. Farady, "Aplikasi Pembelajaran Tajwid Berbasis Android," *J. Rekursif*, vol. 8, no. 1, pp. 59–70, 2020.

[5] A. Wirdiani, S. N. Machetho, I. K. G. D. Putra, M. Sudarma, R. S. Hartati, and H. A. Ferdian, "Improvement Model for Speaker Recognition using MFCC-CNN and Online Triplet Mining," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, no. 2, pp. 420–427, 2024, doi: 10.18517/ijaseit.14.2.19396.

[6] L. Syafa'ah, R. Prasetyono, and H. Hariyady, "Enhancing Qur'anic Recitation Experience with CNN and MFCC Features for Emotion Identification," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 4, no. 2, 2024, doi: 10.22219/kinetik.v9i2.2007.

[7] J. J. M. Montemayor, G. P. G. Escuadra, M. A. G. Nambatac, and D. T. Tenoria, "Detecting Rice Weevils in Stored Grains using MFCC and CNN," *Procedia Comput. Sci.*, vol. 234, no. 2023, pp. 1681–1688, 2024, doi: 10.1016/j.procs.2024.03.173.

[8] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," *IEEE Access*, vol. 10, no. October, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[9] F. Wang and X. Shen, "Research on Speech Emotion Recognition Based on Teager Energy Operator Coefficients and Inverted MFCC Feature Fusion," *Electron.*, vol. 12, no. 17, 2023, doi: 10.3390/electronics12173599.

[10] S. Kulkarni, "Impact of Various Data Splitting Ratios on the Performance of Machine Learning Models in the Classification of Lung Cancer," in *Proceedings of the Second International Conference on Emerging Trends in Engineering (ICETE 2023)*, Springer Nature, 2023, p. 96.

[11] Y. Dokuz and Z. Tufekci, "Mini-batch sample selection strategies for deep learning based speech recognition," *Appl. Acoust.*, vol. 171, p. 107573, 2021, doi: 10.1016/j.apacoust.2020.107573.

[12] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal-frequency attention based convolutional neural network," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, 2021, doi: 10.1038/s41598-021-01045-4.

[13] A. P. Sari, H. Suzuki, T. Kitajima, T. Yasuno, D. A. Prasetya, and N. Nachrowie, "Prediction model of wind speed and direction using convolutional neural network - Long short term memory," *PECon 2020 - 2020 IEEE Int. Conf. Power Energy*, no. 1, pp. 356–361, 2020, doi: 10.1109/PECon48942.2020.9314474.

[14] Y. Furusho and K. Ikeda, "Theoretical analysis of skip connections and batch normalization from generalization and optimization perspectives," *APSIPA Trans. Signal Inf. Process.*, vol.

9, 2020, doi: 10.1017/ATSIP.2020.7.

[15] Y. Liu, X. Chen, M. Cheng, C. Hsieh, Y. You, and L. Angeles, "C ONCURRENT A DVERSARIAL L EARNING FOR L ARGE -," pp. 1–17, 2021.

[16] S. H. Golzar, H. Bagherpour, and J. A. Parian, "Research Article A New Method to Optimize Deep CNN Model for Classification of Regular Cucumber Based on Global Average Pooling," vol. 2024, no. Dl, 2024, doi: 10.1155/2024/5818803.