MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN : 0854-5138 || E-ISSN 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

Between Machine Feedback and *Talaqqī*: Noise Robustness of WavLM in Tajweed Learning and Its Implications for Religious Knowledge Transmission

Dewi Khairani[1*], Gina Sagita[2], Siti Ummi Masruroh[3], Ibrahim Shehu Usman[4], and Ajang Sopandi[5]

[1,2,3]Informatics Engineering Study Program, Faculty of Science and Technology, Syarif Hidayatullah State Islamic University Jakarta, Indonesia
[4]Jigawa State College of Education, Gumel, Nigeria
[5]School of Computing, University Malaysia of Computer Science & Engineering, Selangor, 46200 Malaysia

Open Access

Article Type: **Research Paper**

Article Info
Article history:
Received: April 12, 2026
Revised: July 17, 2026
Accepted: September 24, 2026
Published: October 30, 2026

Keywords:
Tajweed; WavLM; Noise robustness; Digital religion; Religious authority; Transmission of religious knowledge

Correspondence Address:
dewi.khairani@uinjkt.ac.id

© Khairani et al. (2026)

This is an open-access article under the CC BY-SA license

ABSTRACT

Artificial intelligence is increasingly used to support Qur'an recitation learning, yet its pedagogical value depends on reliable sound processing and on distinguishing machine output from the authority of tajweed teachers. Prior tajweed classification studies have mainly reported performance on clean recordings. This study evaluates the noise robustness of WavLM-Base-Plus in classifying three tajweed rules (Mad Jaiz Munfasil, Ghunnah, and Ikhfa) and discusses its implications for religious knowledge transmission. The method used QDAT recordings, MUSAN noise, and a speaker-independent split; the model was tested on clean audio and at 20, 10, 5, and 0 dB SNR. The results show that F1-Macro remained between 0.855 and 0.880, only marginally above a majority baseline (0.842), and on clean audio specificity was only 18.2% for Mad and 19.2% for Ikhfa, meaning many incorrect readings were accepted as correct. Read through recitation traditions and digital religion, the model is best positioned as preliminary feedback within *talaqqī*, not as a substitute for teachers.

Please cite this article in APA Style as:
Khairani, D., Sagita, G., Masruroh, S. U., Usman, I. S., & Sopandi, A. (2026). Between machine feedback and *talaqqī*: Noise robustness of WavLM in tajweed learning and its implications for religious knowledge transmission. *Mimbar Agama dan Budaya*, 43(2), 364-388. https://doi.org/10.15408/mimbar.v43i2.54472

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

1. INTRODUCTION

The recitation of the Qur'an is a religious practice centered on text as well as sounds. As a text, the Qur'an is read, studied, and interpreted, as a chanted reading, it is maintained through listening, repetition, articulation correction, and habituation (Ayyad, 2022; Idriz, 2022; Minan & Mustaqim, 2026). In tajweed teaching, mastery of the rules does not stop at understanding definitions. Learners need to produce sounds, listen to examples, receive corrections, and then repeat the pronunciation (Minan & Mustaqim, 2026). At the level of Indonesian tajweed pedagogy, Khodijah (2023) emphasizes the practice of makhraj and the nature of letters as well as the importance of pentashihan through deliberation, while Genç and Yılmaz (2025) show that Qur'anic education with the help of digital media still faces the obstacle of tajweed correction without direct interaction. Therefore, the use of artificial intelligence for tajweed is not solely a matter of voice recognition, but also touches on the process of knowledge transmission and the responsibility of teachers.

The development of digital media is changing the way Qur'an reading and knowledge are accessed and taught (Campbell & Tsuria, 2021; Harnadi et al., 2021). Ghozali et al. (2022), through TikTok content analysis, found that short videos are a new medium for showing examples of reading, teaching tajweed, and circulating interpretation of the Qur'an. In a device-mediated learning environment, users can playback, slow down, or compare recordings independently. However, the ability to obtain sample readings does not by itself guarantee the availability of adequate explanations and corrections. Research by Sari and Moore (2024) on Indonesian Muslim families abroad shows that digital spaces can expand the network of learning the Qur'an language while relying on the relationship between family and teachers. Machine learning-based audio processing adds a new stage, recordings are not only stored or listened to, but can be grouped automatically according to the acoustic patterns learned by the model.

Technically, automatic pronunciation assessment builds on mispronunciation detection research in computer-assisted language learning (Çalık et al., 2024), and artificial intelligence research on recitation is progressing through sound signal processing, acoustic feature extraction, and the use of deep learning. Self-supervised speech representations such as HuBERT and WavLM (Hsu et al., 2021; Chen et al., 2022) have become common backbones for Qur'anic pronunciation assessment, as reflected in the systems benchmarked in the Iqra'Eval shared task (El Kheir et al., 2025), while Alsahafi and Asad (2024) examined mispronunciation detection specifically for tajweed rules. Al Harere and Al Jallad (2023) and Shaiakhmetov et al. (2025) reported high model performance in the classification of several categories of tajweed using QDAT recordings (Osman et al., 2021). The study is relevant to show that certain sound cues can be recognized computationally. However, performance measures in controlled data do not automatically describe the reliability of usage when recordings are made in homes, lecture halls, or other spaces with background sound. In everyday learning practice, the audio

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

background can come from fans, electronic devices, vehicles, or ambient sounds, and these interferences may alter the model's ability to maintain its classification decisions.

The issue of reliability is important because the wrong decision can result in two different consequences. First, the model can accept as true a sample that is annotated as false in the dataset, so that the user has the potential to gain a false sense of security. Second, the model can reject a sample that is annotated in the dataset correctly, so that the learner can receive unnecessary corrections. These two errors are not equal from an educational point of view: the first can perpetuate errors, whereas the second can reduce trust in the tool or interfere with practice. Therefore, the discussion must go beyond one overall accuracy score and look at the changes in false positives and false negatives by tajweed rule and noise level.

In the context of religious and cultural studies, the interdisciplinary contribution of this research lies in the placement of computational performance in the larger questions regarding the sustainability of Qur'anic reading skills. The review of Kusnadi et al. (2025) shows that Qur'anic learning research in Indonesia includes pedagogical practices and technological developments, while studies of talaqqī affirm that correct pronunciation is acquired through corrective interaction between teacher and learner (Minan & Mustaqim, 2026; Widayati, 2026). The study of digital religious authorities also reminds that mediums and platforms can shape the way users perceive expertise and legitimacy (Maemonah et al., 2022; Ats Tsaqofi et al., 2025; Hidayatullah, 2025). Therefore, the evaluation of the robustness of AI models is placed as an examination of technical prerequisites, rather than direct evidence that a learning tradition has changed or become better.

This study draws on two complementary frames. The first is the tradition of Qur'anic knowledge transmission. Qur'anic recitation has historically been passed on orally through *talaqqī* (receiving the reading directly from a teacher) and *mushāfahah* (face-to-face oral exchange), in which the learner listens to the teacher's model, reproduces it, and is corrected until the pronunciation conforms to the accepted reading (Idriz, 2022; Minan & Mustaqim, 2026). This process is anchored in *sanad*, an unbroken chain of teachers linking a reader's recitation back to the Prophet, and is formalised through *ijāzah*, the teacher's certification that the student may recite and teach (Idriz, 2022). Contemporary sanad institutions in Indonesia continue to validate a reader's competence by examining ijāzah chains, the continuity (ittiṣāl) of transmission, and correction (taṣḥīḥ) by certified muqriʾ, even when instruction takes place online (Minan & Mustaqim, 2026). In Indonesia, this pattern persists in *pesantren* and Qur'anic learning circles, where correction is embedded in a relationship of trust between teacher and student (Harnadi et al., 2021; Minan & Mustaqim, 2026). Within this frame, the correctness of a recitation is not merely an acoustic property; it is a judgement that derives its authority from a recognised teacher and a verifiable chain of transmission. Viewed through the cultural elements that frame religious and cultural studies, tajweed learning

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

thus sits at the intersection of a knowledge system, a technology of transmission, and religious practice, which is why the arrival of automated assessment is a cultural as well as a technical question.

The second frame is digital religious authority. Studies of digital religion show that online media do not simply transmit religious content but also reconfigure who is recognised as knowledgeable and how legitimacy is established (Campbell & Tsuria, 2021). Harnadi et al. (2021) show that, even as digital media open alternative channels of religious guidance, pesantren continue to reproduce the traditional authority of the kiai. Widayati (2026) likewise describes Zoom-based qirā'at communities as a form of mediated talaqqī that maintains sanad continuity while reconfiguring its spatial form. Malik (2024) extends this question to artificial intelligence, asking whether AI systems can take part in Islamic knowledge production and in the chains of transmission on which religious authority rests, while Niam (2024) cautions that reliance on AI-generated religious guidance may blur the boundaries of scholarly authority. Combining the two frames, this study treats an automatic tajweed classifier as a potential new node in the transmission process: its feedback may be perceived by learners as a form of correction, yet it carries no *sanad* or *ijāzah*. The model's error patterns, particularly the acceptance of incorrect readings, therefore matter not only as technical limitations but also as questions about where corrective authority should reside.

Research on automatic tajweed classification has shown the ability of models to recognize error patterns in annotated recordings (Al Harere & Al Jallad, 2023; Shaiakhmetov et al., 2025), and studies of Arabic and Qur'anic mispronunciation detection increasingly rely on self-supervised encoders such as wav2vec 2.0, HuBERT, and WavLM (Çalık et al., 2024; El Kheir et al., 2025). WavLM itself was pre-trained with a masked speech denoising objective intended to improve robustness to overlapping speech and background noise (Chen et al., 2022). Recent studies on digital Qur'an learning have highlighted the access, teaching practices, and integration of distance learning with hands-on guidance (Ghozali et al., 2022; Genç & Yılmaz, 2025). Meanwhile, research on religious authority has addressed changes in the legitimacy and behavior of users in the digital space (Chair et al., 2024; Hidayatullah, 2025). These three bodies of work play different roles: experimental studies show how reliably a system performs, whereas religious and educational studies help define what its outputs mean and how they may be used. The classification experiment does not by itself prove a change in the culture of reciting or an improvement in learners' abilities. However, to the best of our knowledge, tajweed classification studies on QDAT have reported performance on clean recordings only; the noise robustness of a self-supervised speech model for tajweed rule classification, and the asymmetry of its errors in relation to the transmission of recitation knowledge, have not yet been examined.

This study evaluated WavLM-Base-Plus for classifying the correct application of three tajweed rules, namely Mad Jaiz Munfasil, Ghunnah, and Ikhfa, on recordings that

Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

experienced a variety of acoustic disturbances. The model was trained using clean audio, while the test was conducted on original audio and audio that was given stationary noise with several levels of Signal-to-Noise Ratio (SNR). This clean-to-noisy design allowed the identification of performance changes as well as types of misclassification in different noise conditions. In addition to technical testing, the study examined the significance of the findings for the possible use of sound technology in tajweed learning and the transmission of religious knowledge.

On that basis, this article aims to: (1) explain the performance of the WavLM-Base-Plus model for annotations of three tajweed rules on clean and noisy audio data; (2) examine the implications of the pattern of misclassification for the possible use of learning aids; and (3) establish the principle boundaries of the relationship between the digital system and the teacher's authority in the transmission of tajweed knowledge. The argument on the first objective is based directly on the results of the experiment; the second and third objectives are conceptual readings that rely on these findings and on the scientific literature. This basic distinction of evidence is important to maintain the interdisciplinary validity of the writing. The novelty of this article is threefold. First, it tests the noise robustness of a self-supervised speech model for tajweed rule classification across controlled SNR conditions, rather than reporting performance on clean recordings alone. Second, it analyses the asymmetry of false positives and false negatives through sensitivity and specificity, showing that an apparently stable F1 score can conceal a model that accepts many incorrect readings. Third, it offers an interpretive framework that relates these error patterns to talaqqī, sanad, and teacher authority, positioning machine feedback as a supplement to, rather than a substitute for, the transmission of Qur'anic recitation knowledge.

2. METHODS

Study design and limits

This study uses a quantitative experimental design based on audio processing and interdisciplinary reading through literature studies. The experimental component is directed to evaluate the robustness of the tajweed classification model on test data that experiences different acoustic disturbances. The empirical analysis unit is an audio recording along with the three binary tajweed rule labels, while the conceptual discussion unit is the relationship between the limitations of computational aids and the learning principles of Qur'an recitation. This design combines technical measurements with scientific interpretations of technological mediation in religious practice, but does not include ethnography, user surveys, or direct measurement of learning effectiveness.

The literature component followed a structured search procedure. Sources were retrieved from Scopus, Google Scholar, the ACL Anthology, and Indonesian journal indexes (SINTA and Garuda) using combinations of the keywords "tajweed" OR "Qur'anic recitation", "mispronunciation detection", "self-supervised speech" OR

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

"WavLM", "talaqqī" OR "sanad", "digital religion", "religious authority" AND "artificial intelligence", and "AI in education". The core search was limited to 2021–2026; earlier works were retained only as dataset or method sources (e.g., MUSAN). Inclusion criteria were peer-reviewed journal articles, conference papers, or scholarly books in English or Indonesian that addressed (a) automatic tajweed or Qur'anic pronunciation assessment, (b) the transmission of Qur'anic knowledge and religious authority, or (c) the ethics and governance of AI in education. Opinion pieces, sources without accessible full text, and studies unrelated to Qur'anic recitation or religious learning were excluded; preprints were used only as supporting evidence. After screening, 31 sources were retained for interpretation: 7 on automatic tajweed and Qur'anic pronunciation assessment, 19 on Qur'anic knowledge transmission, digital religion, and religious authority, and 5 on the ethics and governance of AI in education; a further 7 sources were cited as dataset, model, or evaluation-method references. The selected literature was coded thematically into these three clusters and used in the discussion to interpret the experimental error patterns, particularly the acceptance of incorrect readings, in relation to talaqqī, sanad, and teacher authority, rather than as additional empirical data.

The data processing stages follow the CRISP-DM framework to the evaluation stage, namely problem understanding, data understanding, data preparation, modeling, and evaluation. This framework is still used in the development of cutting-edge AI projects and emphasizes the relationship between objectives, data quality, modeling, and outcome review (Bokrantz et al., 2024) and has been widely adapted beyond its industrial origin to data science and machine learning projects (Martínez-Plumed et al., 2021; Shimaoka et al., 2024). The operational focus of the research is the measurement of the model's resistance to additive stationary noise. The evaluation was conducted in a clean-to-noisy design: the model was studied using clean recordings and tested on recordings where the level of interference was varied. The design was intended to compare prediction stability, not to test the educational impact of an application. Figure 1 summarises the research flow.

Table 1. Research flow of the study (stages proceed from left to right)

1. Problem & data understanding	2. Data preparation	3. Modelling	4. Noise injection	5. Cross-condition evaluation	6. Interdisciplinary interpretation
QDAT corpus; three tajweed rules (Mad, Ghunnah, Ikhfa)	16 kHz WAV; multi-label annotation; speaker-independent split	Fine-tuning WavLM-Base-Plus on clean audio (weighted BCE)	MUSAN stationary noise at SNR ∞, 20, 10, 5, 0 dB	F1/F1-Macro, confusion matrix, sensitivity, specificity	Error patterns read through talaqqī, sanad, and digital religious authority

Source: Adapted from the CRISP-DM process (Martínez-Plumed et al., 2021) and the research design of this study.

Data sources and label preparation

Qur'anic recitation data was obtained from the Quranic Recitation Dataset (QDAT) (Osman et al., 2021; available at https://www.kaggle.com/annealdahi/quran-recitation),

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

an annotated corpus that was also used in the testing of the classification of three tajweed rules by Al Harere and Al Jallad (2023) and Shaiakhmetov et al. (2025). The corpus was accessed for this experiment in August 2025. According to its developers, QDAT contains more than 1,500 recordings of a single verse (Surah al-Mā'idah 5:109) collected from over 150 different readers (350 recordings by male and 1,159 by female readers), with age and gender metadata for each speaker. The recordings were submitted online via WhatsApp and stored as WAV files at an 11 kHz sampling rate, so device type and recording conditions were not controlled (Osman et al., 2021). Each recording was manually annotated by an expert for the correctness of the three tajweed rules; however, the dataset documentation does not report the number of annotators, their tajweed qualifications (e.g., ijāzah), or inter-annotator agreement. The labels are therefore treated here as computational targets defined by the corpus, not as independently validated religious judgements, and this is acknowledged as a limitation. The Kaggle repository does not specify a licence (listed as "Unknown"); the corpus was therefore used only for non-commercial research, and no audio is redistributed with this article. The three outputs evaluated were Mad Jaiz Munfasil, Ghunnah, and Ikhfa. Each output was treated as a binary classification task based on the corpus annotations; a positive value indicates a correct pronunciation annotation and a negative value indicates an incorrect annotation. Therefore, the model's capabilities can only be interpreted within the framework of the quality and scope of the annotations used.

For the noise simulation, the study used files from the noise category in the MUSAN audio corpus (Snyder et al., 2015). Three files with relatively stationary characters were selected according to the energy stability criteria at the data preparation stage. All audio was standardized to the WAV format with a sampling rate of 16 kHz. Because the original QDAT files were recorded at 11 kHz (Osman et al., 2021), they were upsampled to 16 kHz, the input rate expected by WavLM (Chen et al., 2022). Upsampling changes the file format but adds no acoustic information: the resulting signals contain no speech content above approximately 5.5 kHz, the Nyquist limit of the original recordings, so the model operates on band-limited speech and any pronunciation cues carried in higher frequencies are unavailable to it. In addition, because the MUSAN noise is sampled at 16 kHz, part of the injected noise energy may fall above 5.5 kHz, where the speech carries no signal; the effective in-band SNR may therefore differ slightly from the nominal level. Both points are treated as limitations of the experimental setup. The study compared the conditions without added noise (SNR ∞ dB) with the audio from the noise injection at SNR levels of 20, 10, 5, and 0 dB, following the common practice of controlled noise augmentation for robust speech recognition (Chen et al., 2022). Noise was added to each recording additively as $y = s + \alpha n$, where s is the speech signal, n the noise segment, and α a scaling factor chosen so that $10 \cdot \log10(P_s / (\alpha^2 \cdot P_n))$ equals the target SNR, with Ps and Pn denoting the mean power of the speech and noise signals. The speaker identity and reading label remained the same in each scenario, so the change in

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

metrics could be attributed to differences in acoustic conditions in this test design. The MUSAN corpus is referred to as the source of experimental materials, while the original WavLM study provides a reference for sound processing architecture and the use of representational learning that considers denoising (Chen et al., 2022).

The data division uses a speaker-independent split to reduce the leakage of speaker identity information between the training data and the test data. After data preparation, the final dataset comprised 1,497 recordings, partitioned into 1,197 for training, 144 for validation, and 156 for testing based on a speaker-independent split ratio of approximately 80:10:10. The actual post-processing ratio is close to 80:10:10, but is not exactly identical due to the division by group of speakers. The difference in the number of positive labels in each category indicates an unbalanced distribution; this is one of the reasons for the use of the average score between labels as the main measure.

Table 2. Distribution of annotated labels on experimental datasets

Dataset	Audio count	Mad Jaiz Munfasil (correct / incorrect)	Ghunnah (correct / incorrect)	Ikhfa (correct / incorrect)
Train	1,197	631 / 566	944 / 253	596 / 601
Validation	144	112 / 32	126 / 18	88 / 56
Test	156	101 / 55	138 / 18	104 / 52

Source: Results of research data processing.

Evaluation models and procedures

The model used is WavLM-Base-Plus, which is a speech representation model pre-trained via self-supervised learning that can be adapted through transfer learning (Chen et al., 2022). The pretrained model is adjusted to a multi-label binary classification so that each record has three independently predicted target indicators. The Weighted Binary Cross-Entropy loss function is applied to handle unbalanced label distribution. The training settings include batch size 8, learning rate 1×10^{-5}, weight decay 0.05, warmup ratio 0.1, and maximum limit of 15 epochs; the best configuration was selected based on the validation results on the fourth epoch. The implementation utilized Python, PyTorch, Transformers, and relevant audio processing libraries. A sigmoid output with a decision threshold of 0.5 was used for each label, and the positive-class weights in the Weighted BCE loss were set to 0.897 for Mad, 0.268 for Ghunnah, and 1.008 for Ikhfa. The checkpoint from the fourth epoch was selected because it achieved the best validation F1-Macro, and training was stopped early when this metric did not improve for 3 consecutive epochs. All experiments used a fixed random seed of 42 and were run on an NVIDIA Tesla T4 GPU (Google Colab) using Python 3.12, PyTorch, Transformers, Librosa, and Scikit-learn.

The Cross-Condition Evaluation scheme tests the model on fixed test data, but changes its acoustic conditions through noise mixing. By retaining the recording and reference labels, the comparison focuses on how the model's decisions change when the quality of the input decreases. The main score is F1-Macro, which is the arithmetic

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

average of the F1 score for all three tajweed rules. The F1 score combines precision and recall, making it more informative than just the percentage of the right decision in an unbalanced class distribution (Chicco & Jurman, 2020). However, the aggregate score also needs to be paired with a binary confusion matrix to determine whether the change in performance is due to a false positive or false negative (Chicco & Jurman, 2020). Because this positive-class F1 rewards the acceptance of correct readings, whereas the pedagogical purpose is to detect incorrect readings, three complementary measures were also derived from the confusion matrices: the F1 score of the error (negative) class, the Matthews correlation coefficient (MCC), which uses all four cells of the confusion matrix and is robust to class imbalance (Chicco & Jurman, 2020; Chicco et al., 2021), and a majority-class baseline that always predicts the correct label. AUROC could not be computed because only thresholded predictions were available.

The confusion matrix of the study is presented for each label separately. True positive indicates a positive prediction that matches a positive annotation; true negative indicates a negative prediction that matches a negative annotation. False positive occurs when the model gives a positive prediction to negatively labeled data, while false negative occurs when the model gives a negative prediction to the positively labeled data. Because the label represents an accuracy annotation in the QDAT, false positive errors in hypothetical usage scenarios can be understood as the risk of receiving readings that should be corrected, and false negative errors as the risk of providing unnecessary corrections. The pedagogical definition is a conceptual analogy, not the result of user interaction observations.

The analysis was carried out through three stages. First, the experimental results were summarized in a cross-condition SNR performance table and a confusion matrix for each tajweed rule. Second, score changes and false positive and false negative patterns were interpreted by paying attention to the acoustic character of the three categories of tajweed, without treating untested phonetic hypotheses as causal explanations. Third, quantitative results were discussed through knowledge transmission literature, digital mediation, and pedagogical authority. This interdisciplinary discussion was based on measurement results and literature studies; the research did not conduct teacher interviews, student observations, or direct application tests in the classroom.

3. RESULTS AND DISCUSSION

Model performance at different noise levels

Table 2 shows the main results of the SNR cross-condition test. In clean audio, the WavLM-Base-Plus model obtained an F1-Macro of 0.874. The same values were reported at 20 and 10 dB conditions. At 5 dB, the F1-Macro increased slightly to 0.880 before dropping to 0.855 at 0 dB. Overall, the difference between the clean and 0 dB conditions reached 0.019 F1-Macro points. However, these scores must be read against a majority-class baseline that always predicts the correct (positive) label. Because the test

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

labels are identical across SNR conditions, this baseline yields F1 values of 0.786 for Mad (101/156 positives), 0.939 for Ghunnah (138/156), and 0.800 for Ikhfa (104/156), or an F1-Macro of approximately 0.842 in every condition. The model therefore exceeds the trivial baseline by only 0.013–0.038 F1-Macro points. Combined with a sensitivity of 100% and low specificity on clean audio (Table 4), this indicates that much of the apparent stability under noise reflects the model's bias towards the positive class rather than a robust ability to discriminate correct from incorrect readings. The results support, at most, the narrow statement that the positive-class F1 changed little under the stationary noise tested, with degradation when the noise energy approached the speech energy.

Table 3. F1 score on SNR cross-condition testing

Conditions	F1 Mad	F1 Ghunnah	F1 Ikhfa	F1-Macro
Noiseless (∞ dB)	0.818	0.972	0.832	0.874
20 dB	0.818	0.972	0.832	0.874
10 dB	0.818	0.972	0.832	0.874
5 dB	0.821	0.975	0.844	0.880
0 dB	0.797	0.953	0.815	0.855
Majority baseline (all SNR)	0.786	0.939	0.800	0.842

Source: Model test results on research data. The majority baseline always predicts the correct (positive) label; its values are identical across SNR conditions because the test labels do not change.

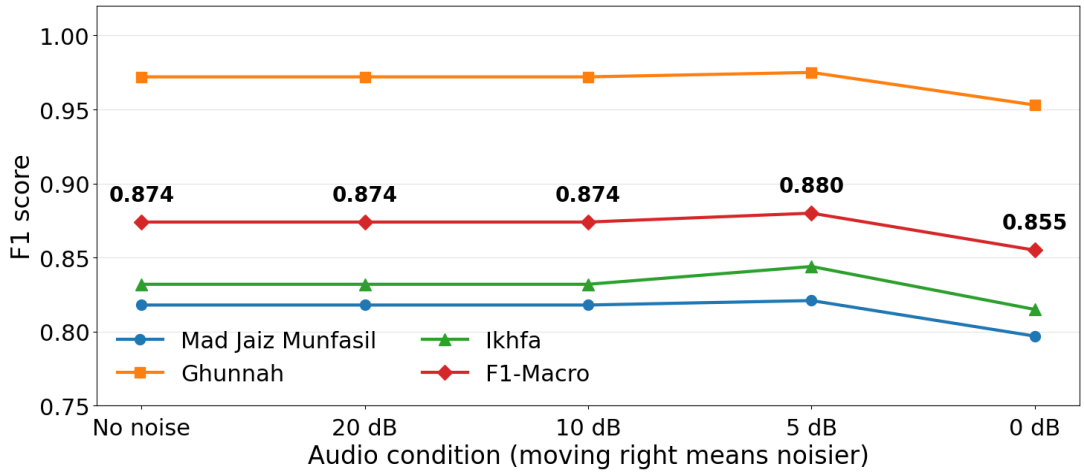


Figure 1. Changes in F1 scores for each tajweed rule and F1-Macro according to SNR

Source: Re-visualization of the results in Table 2.

The consistency of the score in the initial three conditions needs to be read precisely. The same value does not mean that all recordings have identical sound quality or acoustic characteristics; it indicates that the summary metric does not change at the rounding rate presented. Similarly, a slight increase in F1-Macro at 5 dB does not mean that the noise improves the quality of pronunciation or has an educational benefit. Based on changes in the confusion matrix, the reduction of a few false positives can increase the combination of precision and recall despite the addition of false negatives in Ikhfa.

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

Without the repetition of the experiment, confidence intervals, and randomization variation checks, the 0.006-point difference of F1-Macro cannot be treated as a statistically significant improvement. To give a first indication of sampling uncertainty, 95% bootstrap confidence intervals (10,000 resamples of the 156 test recordings, drawn from the confusion-matrix counts) were computed for the per-rule F1 scores. On clean audio the intervals were 0.762–0.867 for Mad, 0.951–0.990 for Ghunnah, and 0.778–0.879 for Ikhfa; at 0 dB they were 0.737–0.849, 0.926–0.976, and 0.759–0.865, respectively. The clean and 0 dB intervals overlap substantially for every rule, so even the largest drop observed cannot be distinguished from sampling variation. An exact McNemar test comparing clean and 0 dB decisions likewise did not reach significance for any rule (Mad $p = 0.22$, Ghunnah $p = 0.06$, Ikhfa $p = 0.45$). It should be noted that these p-values are conservative estimates derived from the confusion-matrix net differences, which serves as a limitation of this post-hoc analysis. Furthermore, because the model was trained with a single random seed, these metrics capture test-set sampling uncertainty only; estimating full training variability through multiple seeds remains an area for future research.

Across the three rules, Ghunnah had the highest F1 in all test conditions: 0.972 in clean audio and 0.953 in 0 dB. Mad Jaiz Munfasil started at 0.818 and decreased to 0.797, while Ikhfa moved from 0.832 to 0.815. This difference shows that the number of one model should not be interpreted as an equal ability for all categories of tajweed. Higher performance in Ghunnah may be related to the combination of acoustic characters, sample composition, and label distribution, but this study did not conduct control experiments to determine the causal weight of each factor. Therefore, the superiority of a particular class should be reported as a quantitative observation, rather than an inherent superiority that can be ascertained in each reading context.

Other studies have used a variety of tajweed classification approaches. In a preprint that has not yet undergone peer review, Al Harere and Al Jallad (2023) reported high accuracy through MFCC and LSTM, while in a peer-reviewed conference paper Shaiakhmetov et al. (2025) used mel-spectrogram and EfficientNet architecture on the related dataset. The figures of the two studies cannot be directly compared with the F1-Macro in this study because the metrics, data sharing, annotations, and test scenarios may differ. The contribution of this experiment is to extend the evaluation from the ability to recognize labels in clean conditions to predictive robustness when the quality of the record changes. The robustness dimension is important as a prerequisite for the development of tools that will later be used outside of a controlled testing environment. WavLM was chosen partly because its pre-training combines masked speech prediction with denoising of simulated noisy and overlapped speech, an objective intended to make its representations more robust to background interference than those of earlier self-supervised models (Chen et al., 2022). The present results suggest, however, that such representation-level robustness does not automatically translate into reliable rule-level

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

error detection on a small, imbalanced tajweed corpus. A structured comparison with earlier QDAT studies is given in Table 6 in the Discussion.

Changes in errors by tajweed rule

The confusion matrix analysis provides a more detailed picture than the aggregate score. In the clean condition, the test showed that there were no false negatives on the three labels in the test set, but false positives were still quite high for Mad and Ikhfa. The results were Mad: 101 true positive, 10 true negative, 45 false positives, and 0 false negatives; Ghunnah: 138, 10, 8, and 0; and Ikhfa: 104, 10, 42, and 0. All three rules therefore have exactly 10 true negatives in the clean condition; because each label was evaluated independently on different negative subsets (55, 18, and 52 incorrect readings), this identical True Negative value is an empirical result of the specific 80:10:10 random split configuration used in this study. The Ghunnah estimate is particularly fragile: with only 18 incorrect readings in the test set, a single recording changes its specificity by about 5.6 percentage points, so the reported 55.6% should be read as an unstable estimate. This pattern reminds us that high recall on positive labels does not guarantee the model's ability to adequately recognize negative annotations. If it is ever used for learning activities, the error of receiving readings that should be corrected needs to be specially evaluated.

At 20 and 10 dB conditions, the main confusion matrix showed no change compared to clean audio. When the SNR decreased to 5 dB, one false negative appeared in Ikhfa, while Mad and Ghunnah still retained the initial true positive count. At the same time, the false positives of several categories decreased: from 45 to 44 for Mad and from 8 to 7 for Ghunnah. In Ikhfa, false positives decreased from 42 to 37. The reduction of false positive errors can drive an increase in F1-Macro even if a single positive sample of Ikhfa is no longer detected. These findings prevent the erroneous conclusion that the addition of noise in itself improves the quality of the Qur'anic recitation.

At SNR 0 dB, five positive samples per label shifted to false negatives. Mad true positive dropped from 101 to 96; Ghunnah from 138 to 133; and Ikhfa from 104 to 99. Under the same conditions, Mad false positives totaled 44, Ghunnah 8, and Ikhfa 40. Thus, exposure to heavy noise decreased the ability to recognize positive labels in all three categories at once, rather than just one category. The shift was in line with the common notion that strong interference could mask the acoustic information required for classification, but this experiment did not provide an explainability analysis of the model (e.g., saliency or attention maps) to establish the specific internal features that were missing.

Table 4. Summary of the confusion matrix per label on three conditions

SNR	Law	TP	TN	FP	FN
∞ dB	Mad	101	10	45	0
∞ dB	Ghunnah	138	10	8	0
∞ dB	Ikhfa	104	10	42	0
5 dB	Mad	101	11	44	0

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

SNR	Law	TP	TN	FP	FN
5 dB	Ghunnah	138	11	7	0
5 dB	Ikhfa	103	15	37	1
0 dB	Mad	96	11	44	5
0 dB	Ghunnah	133	10	8	5
0 dB	Ikhfa	99	12	40	5

Source: Model test results on research data.

From a pronunciation point of view, the Ikhfa characters involving the process of camouflage and nasal transitions may make some of the sound patterns sensitive to interference. However, this test did not include detailed phonetic analysis or testing of formant, duration, and resonance contributions to each prediction. Statements about the relationship between articulated camouflage and model vulnerability should remain submitted as hypotheses for further research. What can be ascertained from the table is the sequence of occurrence of predictive changes: at 5 dB, a new false negative is first found on the Ikhfa label, while at 0 dB the same phenomenon has occurred on the entire label.

These error patterns reveal a risk asymmetry that needs to be thought about from the design stage. If a hypothetical application only displays "true" or "false" indicators, the user does not know whether the decision was made on a clear signal or a recording with strong interference. More importantly, prediction errors do not have the same pedagogical costs. False positive predictions can cause users to skip parts that actually need to be trained; false negative predictions can cause users to repeat parts that have been annotated. Here, pedagogical accountability demands the evaluation of class-wise (per-label) errors and audio condition handling mechanisms, not just the delivery of one seemingly high model score.

Sensitivity, specificity, and consequences of correction

In addition to reading F1, this study needs to distinguish the model's ability to accept correctly annotated examples and reject incorrect annotated examples. Based on the QDAT label definition, sensitivity or true positive rate is calculated as TP/(TP+FN), while specificity or true negative rate is calculated as TN/(TN+FP). Balanced accuracy is the average of both measures. All three are recalculated from the confusion matrix presented in Table 3; thus, this section does not present new experiments or change the main results of the study.

Table 5. Sensitivity and specificity derived from the confusion matrix

Conditions	Tajweed rule	Sensitivity	Specificity	Balanced accuracy
No noise	Mad	100.0%	18.2%	59.1%
No noise	Ghunnah	100.0%	55.6%	77.8%
No noise	Ikhfa	100.0%	19.2%	59.6%
0 dB	Mad	95.0%	20.0%	57.5%
0 dB	Ghunnah	96.4%	55.6%	76.0%
0 dB	Ikhfa	95.2%	23.1%	59.1%

Source: Author's calculation based on TP, TN, FP, and FN in Table 3. Numbers are rounded to one decimal.

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

Under noise-free conditions, the sensitivity of all labels reached 100%, but the specificity for Mad was only 18.2% and Ikhfa was 19.2%; Ghunnah's specificity was 55.6%. This difference means that many incorrectly annotated recordings are labeled positive by the model, especially on Mad and Ikhfa. Therefore, F1 that looks good for a positive class should not be read as evidence that the system is capable of detecting misreadings just as well (Chicco & Jurman, 2020; Chicco et al., 2021). The balanced accuracy of only 59.1% for Mad and 59.6% for Ikhfa in clean audio shows the limitations that are not immediately apparent from F1-Macro 0.874.

Error-class metrics make the same pattern explicit (Table 5). On clean audio, the F1 score of the error class was only 0.308 for Mad and 0.323 for Ikhfa, against 0.714 for Ghunnah, and the macro-averaged MCC was 0.483. At 5 dB the error-class scores rose slightly (macro MCC 0.523), whereas at 0 dB they fell (macro MCC 0.359), mainly because false negatives appeared in all three rules. These values show that the model's ability to detect incorrect readings is weak to moderate and degrades under strong noise, even though the positive-class F1 changes little.

Table 6. Error-class F1 and MCC derived from the confusion matrix

SNR	Mad F1-error	Mad MCC	Ghunnah F1-error	Ghunnah MCC	Ikhfa F1-error	Ikhfa MCC	Macro F1-error	Macro MCC
∞ dB	0.308	0.355	0.714	0.725	0.323	0.370	0.448	0.483
5 dB	0.333	0.373	0.759	0.763	0.441	0.433	0.511	0.523
0 dB	0.310	0.237	0.606	0.563	0.348	0.276	0.421	0.359

Source: Calculated from Table 3. F1-error = 2TN/(2TN + FP + FN); MCC = (TP·TN − FP·FN)/√[(TP+FP)(TP+FN)(TN+FP)(TN+FN)].

At the noisiest conditions (0 dB), sensitivity decreased to 95.0% for Mad, 96.4% for Ghunnah, and 95.2% for Ikhfa. The specificity was 20.0%, 55.6%, and 23.1%, respectively. Partial changes in specificity values should not be considered a meaningful improvement in general because the negative sample size per class was relatively small and the study did not perform a statistical uncertainty test. Instead, this measure shows that the robustness of positive predictions and the ability to reject inaccurate readings are two different evaluation questions.

In the context of tajweed learning, these differences have an impact on the feedback design. A model with very high sensitivity remains at risk of providing a false sense of security if many pronunciation errors are accepted as correct. Therefore, at the stage of learning tool development, tajweed expert assessment and specificity evaluation need to be used as a condition before positive feedback is displayed as a sign of reading mastery. The teacher-student relationship in the assessment process remains important, as the model only studies corpus labels in a limited scope. This conclusion is the interpretation of the potential use of existing data, not the results of testing an application on students.

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

Interpreting the robustness of technology in learning practice

Model stability in some noise conditions provides an initial technical prerequisite for device-assisted learning; the assessment of Qur'anic applications using AI does show the potential for training support, but it also emphasizes the issue of feedback accuracy (Nasir et al., 2025). A learner may practice using a mobile phone at home or a classroom with mild noise; systems that are too volatile predictions in such situations will be difficult to use as training material. However, stationary noise injected in a controlled manner does not represent the overall real environment. Classrooms can contain changing conversations, sound reflections, changes in microphone distance, music, and device interference. The results under MUSAN noise should therefore not be read as a guarantee of performance in all religious education settings.

There is a difference between the reliability of the instrument and the success of education. The review of Qur'an learning by Kusnadi et al. (2025) and the online classroom study by Genç and Yılmaz (2025) show that educational evaluation encompasses processes, contexts, and interactions that cannot be reduced to a single model metric. This study produces information about the limited reliability of the classifying instrument under signal perturbation; it does not measure the improvement of students' reading, the intensity of independent learning, user satisfaction, or teacher acceptance. To assess the real benefits in the transmission of religious knowledge, the next stage should be an evaluation with students and tajweed teachers. The test should measure the accuracy of corrections compared to expert assessments, changes in performance before–after use, the type of feedback understood, and the level of learner's dependence on the application. Without these steps, the word "effective" can only refer to the performance of the experiment's predictions, not to the achievement of the learning process.

From a cultural perspective, voice recognition technology has the potential to change the distribution of learning time: some repetition of basic exercises may be done outside of the meeting with the teacher, while face-to-face meetings can be focused on more complex corrections. Genç and Yılmaz (2025) discuss the model of integrating traditional and digital Qur'anic education; Azzahrowaini et al. (2025) highlight the opportunities for AI to support tahfiz learning as well as the importance of teachers; and Nasir et al. (2025) show that AI-assisted Qur'anic learning applications face issues of technical accuracy, bias, privacy, and pedagogical appropriateness. The tajweed tradition demands more than machine-recognized sound patterns: it also involves explaining the teacher's rules and responsibilities in providing contextual corrections. Therefore, the design of the application should provide an escalation mechanism to teachers when the prediction is doubtful.

A more responsible use approach is to make machine decisions temporary feedback, in line with the caution over accuracy and ethical aspects of AI applications for Qur'an memorization discussed by Nasir et al. (2025). For example, after a user records

Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

a reading, the system can display that an indication of pronunciation in one of the targets requires review, accompanied by an admonition to listen to examples and consult with the teacher. Such a design avoids claims that statistical labels are the final determinant of reading norms. It is also in line with caution in the study of digital religion: authority is not just an attribute that can be transferred from an individual to an algorithm, but is related to legitimacy, social relations, and the practice of knowledge (Chair et al., 2024; Harnadi et al., 2021; Hidayatullah, 2025). Ahmed (2024) similarly argues that AI may assist, but not replace, the muftī, whose authority rests on scholarly formation and accountability, while Abdelnour (2025) frames technology in Islamic theology as a means whose value depends on the meanings and ends it serves. In the terms of digital religion, media do not simply carry religious content but are negotiated by communities according to their existing values and structures of authority (Campbell & Tsuria, 2021). Debates on AI-powered religious authority likewise caution that algorithmic systems cannot enter the chain of transmission or carry the accountability expected of scholars (Malik, 2024; Niam, 2024). For tajweed, this implies that automated feedback should sit outside the sanad: it may support practice between sessions, but validation of a recitation remains the work of a teacher within talaqqī.

The issue of authority becomes sharper when the platform or interface displays percentages and labels convincingly. High-precision numbers can be understood by the user as certainty, even though evaluations show false positives and false negatives still appear. Therefore, communication of limitations needs to be part of the design. Applications that adopt similar models should state three things: that the system compares sounds with patterns from a particular dataset, that noise can change predictions, and that teachers or experts still determine corrections according to the rules taught. This explanation does not automatically eliminate errors, but builds usage expectations that are more in line with the nature of the instrument.

In terms of knowledge, the preservation of the tradition of reading the Qur'an cannot be equated with the digitization of sound archives. Archives record examples; classification models extract patterns; while knowledge transmission requires the process of forming competencies, understanding, and habits recognized by the learning community (Idriz, 2022; Minan & Mustaqim, 2026). Historical research on early Qur'an memorization shows that this embodied, teacher-mediated learning has long been central to Muslim cultures of transmission (Ayyad, 2022), and recent studies of virtual talaqqī suggest that digital media can extend, rather than replace, this relationship when sanad continuity and teacher correction are preserved (Widayati, 2026). These three functions can complement each other without having to replace each other. The framework provides a more appropriate position for artificial intelligence in cultural studies: technology is one of the means of mediation of practice, while the norms of pronunciation and the process of validating abilities are still determined through the rules of science and relevant pedagogical institutions.

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

From acoustic data to socio-cultural responsibility

The use of data for Qur'anic recitation models raises questions about representation; research on the application of AI for Qur'anic learning also identified the problem of algorithmic bias and the model's suitability to the user's needs (Nasir et al., 2025). QDAT provides records and annotations for model development, but cannot be assumed to cover the entire diversity of readers, variations in educational practices, or all tajweed rules. The evaluation of this study was limited to three labels from a specific corpus. If the application is extended to groups that are not adequately represented, the level of error can be unevenly distributed. Therefore, before wider implementation, testing is needed across speaker groups, device quality, noise conditions, and expert agreement on annotations. From a cultural perspective, dataset inequality is also related to the fairness of access to trusted educational feedback. Ethical frameworks for AI in education identify such uneven performance across groups as a core fairness concern (Holmes et al., 2022). A related question is data sovereignty: recordings of recitation are produced within Muslim learning communities, and decisions about who collects, owns, and reuses these voices, and for what purposes, should involve those communities and their teachers rather than be settled by developers alone. Scholarship on Indigenous data sovereignty offers a useful analogy: the CARE Principles (collective benefit, authority to control, responsibility, and ethics) hold that communities whose knowledge is captured in data should share in its benefits and retain authority over its use, including in machine learning (Carroll et al., 2020).

The design of the system also needs to distinguish the assessment of rules from the uniformity of tradition. The variety of Qur'an readings has a framework of knowledge and history that cannot be reduced to a single universal pronunciation format. In Indonesia, qirā'āt knowledge has been transmitted through dense networks of sanad and talaqqī–musyāfahah; Ḥafṣ ʿan ʿĀṣim became the national standard through print and curricular standardization, while other readings such as Warsh and Qālūn persist in specialized communities (Kharisma & Rozaq, 2026), and qirā'āt sabʿah have been sustained by scholarly sanad networks (Ulya & Alkaff, 2023; Minan & Mustaqim, 2026; Widayati, 2026). This study follows the categories and annotations of QDAT, not verifying all possible qiraat practices. Therefore, the use of labels must be accompanied by information about the scope of the rules established from the beginning. Without an explanation of the scope, the model risks being understood as a judge of every variation of the reading, even though it works for the purpose of narrow classification. The determination of the scope with tajweed experts is a methodological necessity as well as a religious responsibility. Framed in this way, dataset bias is not only a technical issue but also a question of respect for the legitimate diversity of recitation traditions. A system trained on a single riwāya or on a narrow annotation scheme could, if deployed uncritically, mark valid variant readings as errors and thereby privilege one tradition over others. Designing such tools to state their scope explicitly, and to defer to teachers when

Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

a reading falls outside it, is consistent with the commitment to tolerance and religious moderation that recognizes plural and authentic paths of transmission rather than enforcing a single standard (Ats Tsaqofi et al., 2025).

Privacy protection is also important if the system transitions from an experimental simulation to a service that records the learner's voice. Voice recordings can be personal data that contains the characteristics of the speaker along with a specific context. This study utilizes a secondary corpus and does not develop a user audio storage mechanism. Therefore, the test results cannot be considered as evidence of an application's privacy compliance. Subsequent development requires user consent, retention limits, access settings, governance of rights to audio sources, and data deletion mechanisms. UNESCO's guidelines on AI in education and teacher competency frameworks emphasize human-centered governance, educator AI literacy, and learner protection (Miao & Holmes, 2023; Miao & Cukurova, 2024). Peer-reviewed scholarship reaches similar conclusions: Holmes et al. (2022) call for a community-wide ethical framework for AI in education that addresses fairness, accountability, and pedagogy, while Kasneci et al. (2023) caution that AI-generated feedback requires teacher oversight and critical learner engagement to avoid overreliance.

At the level of use, the teacher's responsibility remains not erased by the presence of a sound detection system. Teachers need to know the types of errors that are common in the system in order to be able to explain when the results of the application and the direct assessment differ. On the other hand, developers must provide performance records per class, information on test conditions, and possible uncertain predictions. Dialogue between teachers, developers, and learners is more promising than one-way application introduction. This corresponds to a human-in-the-loop design, in which the model flags uncertain or likely incorrect readings and a teacher confirms or overrides the decision, an arrangement that the literature on AI in education recommends to limit overreliance on automated feedback (Kasneci et al., 2023). If one day the technology is tested in the tahsin classroom, the cooperation could be an important research variable: whether the teacher's explanation of the AI results makes learners more able to make corrections independently than using an unsupervised application.

Attention to limitations is not a rejection of innovation. The quality of religious learning innovation depends on the ability to link computational advantages with pedagogical interests. WavLM's results show that cross-condition evaluation and error matrix analysis can be the basis for designing when automated feedback is feasible to display and when the system needs to request re-recording. Thus, the results of technical measurements provide preliminary evidence for further testable designs, rather than automatic legitimacy for changing educational practices.

Contributions and limits of research inference

The first contribution of this article is to maintain clear empirical measures. The F1 value and the confusion matrix allow the reader to examine the model's stability claims

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

on a given scenario and understand the consequences of different types of mispredictions. Second, the article explains why acoustic findings are in limited but important relation to the study of religion and culture: sound assessment tools can be part of the new ecology of recitation learning, provided that the information it provides is not equated with authoritative decisions. Third, the paper formulates a cross-disciplinary evaluation agenda that combines technical robustness testing with expert validation and learning experience measurement.

Table 6 summarises how this study differs from earlier work on the same corpus. Previous QDAT studies reported accuracy on clean recordings; the present study adds controlled noise conditions, error-class metrics, and a majority baseline, which together show that high aggregate scores can coexist with weak detection of incorrect readings.

Table 7. Comparison with previous tajweed classification studies on QDAT

Study	Dataset	Features / architecture	Metric	Test condition	Reported result
Al Harere & Al Jallad (2023), preprint	QDAT	MFCC + LSTM	Accuracy per rule	Clean audio only	96% (Mad), 95% (Ghunnah), 96% (Ikhfa)
Shaiakhmetov et al. (2025)	QDAT (1,505 recordings)	Mel-spectrogram + EfficientNet	Average accuracy	Clean audio only	97.23%
This study	QDAT (1,497 recordings after preparation)	WavLM-Base-Plus (self-supervised), fine-tuned	F1-Macro, F1-error, MCC, sensitivity/specificity, majority baseline	Clean and MUSAN noise at 20, 10, 5, 0 dB; speaker-independent split	F1-Macro 0.874 (clean) to 0.855 (0 dB); baseline 0.842; specificity 18.2–55.6% (clean)

Source: Compiled from Al Harere and Al Jallad (2023), Shaiakhmetov et al. (2025), and the results of this study. Accuracy and F1 are not directly comparable.

Its limitations also need to be explicitly stated. Experiments were conducted on QDAT and selected stationary noise from MUSAN, using one architecture and three labels, without testing differences in dialect, age group, recording quality, or qirā'āt variation. The model was trained in a single run with one random seed, so training variability could not be estimated; the bootstrap intervals and McNemar tests reported above capture only test-set sampling uncertainty. There was also no systematic comparison with human raters (tajweed teachers). The study did not conduct phonetic analysis of the causes of class differences, independent assessment of religious validity, application trials, user behavior surveys, or observation of learning traditions. In addition, the test set contained few incorrect readings (55 for Mad, 18 for Ghunnah, and 52 for Ikhfa), so specificity estimates are unstable, particularly for Ghunnah, where a single recording shifts specificity by about 5.6 percentage points. Finally, the model's F1-Macro exceeds a majority-class baseline by only 0.013–0.038 points, which limits any claim of practical robustness. Accordingly, the pedagogical scenarios, interface mechanisms, and notions of cultural preservation, user acceptance, and teacher authority discussed above are literature-based interpretations and design hypotheses, not tested findings, and require validation through classroom trials and field research.

MIMBAR AGAMA DAN BUDAYA
Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

By limiting ourselves to the level of evidence that is actually available, the article can still make a contribution to interdisciplinary debates. Religious and cultural studies do not have to ignore technological metrics; rather, they can be used to clarify the material requirements of a digital mediation. A proportionate perspective recognizes the mutual relationship: computational research provides measurement, tradition studies provide an interpretive framework, and field research will later serve as a link to test whether technology actually improves tajweed learning.

CONCLUSION

This study showed that WavLM-Base-Plus maintained an F1-Macro of 0.874 at clean audio conditions, 20 dB, and 10 dB; the value increased slightly to 0.880 at 5 dB and decreased to 0.855 at 0 dB. These scores are only marginally above a majority-class baseline (F1-Macro ≈ 0.842), so the small variation across SNR levels should not be read as evidence of robust discrimination; it largely reflects the model's tendency to accept readings as correct. Analysis per category showed that the first additional false negatives appeared at Ikhfa at SNR 5 dB, while at 0 dB each category experienced five false negatives. Therefore, the assessment of model readiness needs to consider the distribution of errors and should not rely solely on a single aggregate score.

Derivative calculations of the confusion matrix confirm another limitation relevant to learning: in the clean audio condition, the specificity of Mad and Ikhfa was only 18.2% and 19.2%, respectively, although the sensitivity of positive labels reached 100%. This means that the aggregate score does not fully describe the model's ability to identify improper annotated readings. Therefore, the evaluation of correction quality should include sensitivity, specificity, and expert judgment, not just F1-Macro.

The interpretive contribution of this research lies in the placement of AI as a means of supporting the transmission of Qur'anic reading skills which are limited by data, acoustic context, and design of its use. The tajweed tradition contains a teacher-student relationship and scientific authority that is not adequately represented by binary classification indicators. Therefore, the model should not be treated as a substitute for the teacher or as a final normative decision-giver. Its more appropriate role to be tested is to provide a transparent initial indication of limitations and provide room for examination by humans.

Further research is suggested to combine the addition of noise variations and real-world evaluations with tajweed experts' assessments of model predictions. In the field of education, it is necessary to conduct controlled tests that compare exercises with teachers, exercises with digital feedback, and blended learning models. Assessments need to include the accuracy of corrections, the development of students' abilities, the comfort of use, the privacy of recordings, and the level of teacher acceptance. Only through these tests can the contribution of artificial intelligence to the sustainability of tajweed learning practices be assessed as empirical results, not solely conceptual possibilities.

Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

DECLARATIONS

Author Contributions: All authors contributed substantially to this study. The first author contributed to the conceptualization, methodology, and supervision of the study. The second author led the data collection and curation, including the acquisition and preparation of the recitation and noise datasets, and contributed to the experiments. The third author contributed to the methodology, formal analysis, and validation of the results. The fourth author contributed to the interpretation of the findings and the discussion of their implications for religious knowledge transmission. The fifth author contributed to the analysis and visualization of the results. All authors participated in writing, reviewing, and editing the manuscript and approved the final version.

Funding: The author(s) did not receive any financial support from any party during the preparation of the current study.

Conflict of Interest: The author(s) of this article declare no conflict of interest.

Generative Artificial Intelligence Use Statement: During the preparation of this manuscript, the authors used generative artificial intelligence tools to assist with translation and with the analysis of the research findings. The authors reviewed, verified, and edited all AI-assisted outputs and take full responsibility for the content of this article.

Ethics Approval and Informed Consent: This study did not involve the recruitment of human participants or the collection of new personal data. The analysis used only secondary, publicly available datasets, namely the Quranic Recitation Dataset (QDAT) (Osman et al., 2021) and the MUSAN audio corpus (Snyder et al., 2015). Therefore, ethical approval and informed consent were not required for this study.

Data Availability: The data analysed in this study are publicly available. The Quranic Recitation Dataset (QDAT) can be accessed at https://www.kaggle.com/annealdahi/quran-recitation, and the MUSAN audio corpus (Snyder et al., 2015) is available at https://www.openslr.org/17/. Processed data and analysis scripts supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Abdelnour, M. G. (2025). Artificial intelligence and the Islamic theology of technology: From “means” to “meanings” and from “minds” to “hearts”. Religions, 16(6), 796. https://doi.org/10.3390/rel16060796

Ahmed, S. (2024). Fatwā and AI: A literature-based assessment on how AI impacts the role of the muftī. Journal for Islamic Studies, 42(1–2). https://doi.org/10.25159/2957-9163/14400

Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

Al Harere, A., & Al Jallad, K. (2023). Mispronunciation detection of basic Quranic recitation rules using deep learning [Preprint]. arXiv. https://arxiv.org/abs/2305.06429

Alsahafi, Y. S., & Asad, M. (2024). Empirical study on mispronunciation detection for tajweed rules during Quran recitation. In 2024 6th International Conference on Computing and Informatics (ICCI) (pp. 39–45). IEEE.

Ats Tsaqofi, W., Satori, A., Mahmudah, S. H., Saehudin, A., Abshar, U., & Li, W. (2025). Digital religious discourse: Constructing legitimacy and counter-narratives in Islami.co. Mimbar Agama dan Budaya, 42(1), 225–238. https://doi.org/10.15408/mimbar.v42i1.48924

Ayyad, E. (2022). Re-evaluating early memorization of the Qurʾān in medieval Muslim cultures. Religions, 13(2), 179. https://doi.org/10.3390/rel13020179

Azzahrowaini, L., Shohib, M. W., & Maksum, M. N. R. (2025). Digitization of AI-based Qur'an tahfidz learning in challenge innovation and implementation. Educatio, 20(2), 253–265. https://doi.org/10.29408/edc.v20i2.30044

Bokrantz, J., Subramaniyan, M., & Skoogh, A. (2024). Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM. Production Planning & Control, 35(16), 2234–2254. https://doi.org/10.1080/09537287.2023.2234882

Çalık, Ş. S., Küçükmanisa, A., & Kilimci, Z. H. (2024). A novel framework for mispronunciation detection of Arabic phonemes using audio-oriented transformer models. Applied Acoustics, 215, 109711.

Campbell, H. A., & Tsuria, R. (Eds.). (2021). Digital religion: Understanding religious practice in digital media (2nd ed.). Routledge. https://doi.org/10.4324/9780429295683

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE principles for Indigenous data governance. Data Science Journal, 19(1), 43. https://doi.org/10.5334/dsj-2020-043

Chair, B. M., Wawaysadhya, & Oktafiani, T. U. (2024). Cyber-religion and the issue of religious authority: How Indonesian youth learn religion through social media? Al'Adalah, 27(1), 1–12. https://doi.org/10.35719/aladalah.v27i1.440

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., & Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6. https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining, 14, 13. https://doi.org/10.1186/s13040-021-00244-z

El Kheir, Y., Meghanani, A., Toyin, H. O., Almarwani, N., Ibrahim, O., Elshahawy, Y. A., Shahin, M., & Ali, A. (2025). Iqra'Eval: A shared task on Qur'anic pronunciation assessment. In Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks (pp. 443–452). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.61

Genç, H., & Yılmaz, A. (2025). The teaching of the Quran in the digital age: The role of technology and social media in Quran education. Mütefekkir, 12(23), 239–267. https://doi.org/10.30523/mutefekkir.1733231

Ghozali, M., Mursyid, A. Y., & Fitriana, N. (2022). Al-Qur'an (re)presentation in the short video app TikTok: Reading, teaching, and interpretive. Pertanika Journal of Social Sciences & Humanities, 30(3), 1263–1282. https://doi.org/10.47836/pjssh.30.3.18

Harnadi, D., Siahaan, H., & Hilmy, M. (2021). Pesantren and the preservation of traditional religious authority in the digital age. Masyarakat, Kebudayaan dan Politik, 34(3), 272–280. https://doi.org/10.20473/mkp.V34I32021.272-280

Hidayatullah, R. (2025). Digital religious authority: The formation of a new Islamic authority in the digital space. Ushuluna: Journal of Ushuluddin Science, 10(2), 1–16. https://doi.org/10.15408/ushuluna.v10i2.42831

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. International Journal of Artificial Intelligence in Education, 32(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

Idriz, M. (2022). Educational tradition of ijāzah in Islamic history with reference to Persian milieu. Journal of Islamic Thought and Civilization, 12(2), 179–195.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kharisma, S. N. L., & Rozaq, M. F. (2026). The transmission network of qirā'āt knowledge: The genealogy of the sanad of Arab-Indonesian scholars and the dominance of the Riwayat Hafs in national standardization. Al-Fahmu: Jurnal Ilmu Al-Qur'an dan Tafsir, 5(1), 460–472. https://doi.org/10.58363/alfahmu.v5i1.743

Khodijah, S. (2023). Tahsin Al-Qur'an: A guide to reciting the Qur'an with the rules of tajweed. Deepublish.

Kusnadi, Sulaeman, M., & Zulfikar, E. (2025). Systematic literature review (SLR) on Qur'anic learning in Indonesia (2014–2024). QURANICA: International Journal of Quranic Research, 17(2), 28–60. https://doi.org/10.22452/quranica.vol17no2.28

Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

Maemonah, M., Purnama, S., Rohinah, R., Aziz, H., Bastian, A. B. F. M., & Syafii, A. (2022). The shift in the authority of Islamic religious education: A qualitative content analysis on online religious teaching. The Qualitative Report, 27(9), 1830–1846. https://doi.org/10.46743/2160-3715/2022.5325

Malik, S. A. (2024). Artificial intelligence and Islamic thought: Two distinctive challenges. Journal of Islamic and Muslim Studies, 8(2), 108–115. https://doi.org/10.2979/JIMS.00020

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. https://www.unesco.org/en/articles/ai-competency-framework-teachers

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Minan, M. A., & Mustaqim, A. (2026). Preserving sanad in the digital age: A case study of virtual talaqqī for qirā'āt al-Qur'ān at Markaz Qiroat Indonesia. Jurnal Pendidikan Agama Islam, 23(1), 210–231. https://ejournal.uin-suka.ac.id/tarbiyah/jpai/article/view/v23i1.12798

Nasir, K., Zaman, R. K., Hilmi, A. B. A., Mahadzir, A. H., & Mustaffa, A. M. (2025). Artificial intelligence integration in mobile applications: Innovation and challenges in supporting Quran memorization and review. QURANICA: International Journal of Quranic Research, 17(2), 612–645. https://doi.org/10.22452/quranica.vol17no2.24

Niam, M. F. (2024). Does artificial intelligence go beyond the limits of religious authority? An ethical review on IslamGPT. Al'Adalah, 27(1), 71–84. https://doi.org/10.35719/aladalah.v27i1.477

Osman, H. M., Mustafa, B. S., & Faisal, Y. (2021). QDAT: A data set for reciting the Quran. International Journal on Islamic Applications in Computer Science and Technology, 9(1), 1–9. https://sign-ific-ance.co.uk/index.php/IJASAT/article/view/2370

Sari, A. P., & Moore, L. C. (2024). Learning Qur'anic Arabic in a virtual village: Family religious language policy in transnational Indonesian Muslim families. International Journal of Bilingualism. https://doi.org/10.1177/13670069241256194

Shaiakhmetov, D., Gimaletdinova, G., Momunov, K., & Cankurt, S. (2025). Evaluation of the pronunciation of tajweed rules based on DNN as a step towards interactive recitation learning. In 2025 International Conference on Computer Systems and Technologies (CompSysTech) (pp. 1–5). IEEE. https://doi.org/10.1109/CompSysTech65493.2025.11137272

Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for data science: Methods, processes and frameworks. SBC Computing Reviews, 4(1), 28–43. https://doi.org/10.5753/reviews.2024.3757

Vol. 43 No. 2 – October 2026 (364-388)
P-ISSN: 0854-5138 (Print) | E-ISSN: 2715-7059 (Online)
DOI: https://doi.org/10.15408/mimbar.v43i2.54472

Snyder, D., Chen, G., & Povey, D. (2015). MUSAN: A music, speech, and noise corpus [Preprint]. arXiv. https://arxiv.org/abs/1510.08484

Ulya, M. N., & Alkaff, S. A. R. (2023). An analysis of the sanad transmission by K.H. Muhammad Arwani (1905–1994) and his role in the dissemination of qiraat sab'ah knowledge in Indonesia. QOF: Jurnal Studi Al-Qur'an dan Tafsir, 7(2), 245–262. https://doi.org/10.30762/qof.v7i2.1400

Widayati, R. (2026). Digitalizing sanad: A virtual ethnography of Zoom-based qira'at learning communities in post-pandemic Indonesia. Jurnal Ilmiah Al-Mu'ashirah, 23(2). https://doi.org/10.22373/jim.v23i2.35181