

Dataset Development for Quran Memorizers: A Step Towards Data-Driven Personalized Learning Path

Miftah Farid Adiwisastra^{1,2}, Irfan Darmawan³, Dade Nurjanah⁴

¹Department of Information Systems, Universitas Bina Sarana Informatika, Tasikmalaya, Indonesia

²School of Computing, Telkom University, Bandung, Indonesia

³School of Computing, Telkom University, Bandung, Indonesia

⁴School of Computing, Telkom University, Bandung, Indonesia

Article Info

Article history:

Received Oct 6, 2025

Revised Mar, 26, 2026

Accepted Mar 31, 2026

Keywords:

Quran Memorization

Dataset Development

Personalized Learning

Machine Learning

Artificial Intelligence

ABSTRACT

Advances in artificial intelligence and machine learning have created new opportunities for improving Quran memorization methods. A key requirement for such innovation is the availability of structured and representative datasets specifically designed for Quran memorizers. This study presents the design and development of a dataset that captures demographic characteristics, daily learning behavior, and temporal memorization patterns from 350 students across three Islamic boarding schools in Indonesia. Through preprocessing stages, including normalization, discretization, and feature engineering, the dataset was prepared for Hidden Markov Model (HMM)-based analysis. Experimental results show that the model achieved an accuracy of 30.57%, precision of 71.46%, recall of 30.57%, and F1-score of 37.70% in predicting memorization states. These findings indicate that the proposed dataset provides a useful foundation for modeling memorization progress and supporting adaptive learning path recommendations. However, the study is limited by the relatively small dataset and the modest predictive performance of the initial HMM model. Overall, this work provides an important first step toward building an intelligent and personalized data-driven Quran memorization system.

Copyright © 2026 Institute of Advanced Engineering and Science.

All rights reserved.

Corresponding Author:

Miftah Farid Adiwisastra

Department of Information Systems, Universitas Bina Sarana Informatika

Tasikmalaya, Indonesia

Email: miftah.mow@bsi.ac.id

1. INTRODUCTION

Memorizing the Qur'an is a spiritual and intellectual process that requires perseverance and effective methods. In today's digital era, technology can be a supporting tool to increase the efficiency of the memorization process, especially with the existence of a learning system that automatically adapts itself based on individual needs and abilities. However, the development of such systems is highly dependent on adequate and quality datasets. Unfortunately, until now the availability of special datasets for those who memorize the Al-Quran is still very limited. The concept of personalized learning paths has become a major focus in modern educational innovation [1,2,3,4]. In the context of memorizing the Qur'an, a personalized learning path means that the system can adjust the sequence, intensity, and method of memorization based on the unique profile of each memorizer. This includes adjustments to learning speed, responses to certain verses, and the most effective murojaah strategies. A comprehensive and rich in information allows the learning system to construct more relevant and realistic learning paths for each individual [5,6,7].

In addition, the traditional approach to memorizing the Qur'an still relies on manual memorization-based methods which are not always effective for every individual [8]. Some memorizers have difficulty mem-

orizing certain verses, while others find it easier to remember certain verse patterns or structures. With a comprehensive dataset, learning systems can be developed to provide data-driven recommendations, adapt the curriculum to individual learning styles, and provide automated feedback on memorization progress. Furthermore, collecting data on memorization habits, the level of difficulty of each verse, and the emotional and cognitive responses of the memorizer can be a strong foundation for developing an adaptive learning system. Technologies such as machine learning, Natural Language Processing (NLP), and recommendation systems can be utilized optimally if relevant and high-quality data is available [9,10,11,12]. Therefore, it is important to design and develop datasets that are not only rich in quantity, but also have quality in terms of content and contextual relevance to the process of memorizing the Qur'an. This article proposes the development of a comprehensive dataset containing information on memorization behavior, personal characteristics, and other temporal and contextual data of Qur'an memorizers. This dataset aims to be a foundation for the development of a personalized and adaptive learning path recommendation system. By utilizing an approach such as the Hidden Markov Model (HMM), it is hoped that it will be able to automatically identify the memorization status of students and recommend appropriate learning strategies.

A number of studies have been conducted related to the use of technology in learning the Qur'an. Research [13] on the use of technology in Al-Quran education by developing a mobile application called "EzHifz" to memorize the Al-Quran based on Visual, Auditory, Reading/Writing, and Kinesthetic (VARK) learning styles. Research [14] proposes a conceptual solution to help students memorize the Quran more easily and independently. Previous research has proposed a system that can help memorizers of the Qur'an to test their memorization independently without a partner using the Group Decision Support System (GDSS) approach that combines the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) and Weighted Geometric Mean to sort the surahs based on the user data provided [15]. Development of a mobile-based Al-Quran memorization learning model, so that students can memorize independently, but teachers can still listen to reviews of reading examples, evaluate, and provide guidance [16]. A study has been conducted in proposing an interactive Al-Quran mobile application, namely Hafaz, to facilitate autistic children in reading and memorizing the Al-Quran [17]. Research utilizing Automatic Speech Recognition approach to prepare, design, develop and evaluate a continuous and speaker-independent speech recognizer with large vocabulary for the Qur'an based on the narration of Hafz from A'asim [18]. An empirical study was conducted on [19] research to identify the optimal combination of feature extraction techniques and machine learning algorithms, including deep learning, for automatic mispronunciation detection during Quran recitation. Utilization of Machine Learning techniques in [20] research to classify spoken Arabic phonemes into separate classes by utilizing Mel Frequency Cepstral Coefficients (MFCCs), audio characteristics, to build optimized classifiers.

Several previous studies have developed personalization approaches in the context of data-driven education [5,21,22]. Research focused on the creation and implementation of personalized learning paths driven by artificial intelligence [23]. presents an ontology-based and artificial intelligence-based framework for personalized learning paths in education [24]. The deep integration of artificial intelligence and education, educational change and innovation are facing unprecedented historical opportunities, and in the form of class organization, gradually moving towards a large-scale individual teaching model [25]. In the wave of data-driven educational innovation, personalized teaching has become a core issue in the field of education [26]. Research on intelligent learning systems and adaptive education has largely focused on academic subjects such as mathematics and languages [27,28,29,30]. However, of all the studies that have been conducted, not many have focused on developing a specific dataset to support personalized Quran memorization. This study aims to fill this gap by developing a dataset that can be used to track memorization progress and customize learning paths based on the individual needs of the memorizer.

Although previous studies have explored mobile applications, decision support systems, speech recognition, and machine learning techniques for Quran learning and recitation, limited attention has been given to the development of a structured dataset specifically designed to support personalized Quran memorization pathways. Existing studies mainly focus on application development or recitation assistance, rather than on dataset construction for adaptive memorization modeling. Therefore, the main contribution of this study is threefold. First, it develops a structured dataset for Quran memorizers by integrating demographic, behavioral, and temporal features. Second, it prepares the dataset for sequential probabilistic modeling through normalization, discretization, and feature engineering. Third, it demonstrates the feasibility of using the dataset as a foundation for Hidden Markov Model-based personalized learning path recommendation. To the best of our knowledge, this study is the first to develop a structured dataset specifically designed to support probabilistic

modeling for Quran memorization pathways.

To better position this study within the existing literature, Table 1 presents a comparison between previous works and the proposed approach, highlighting differences in objectives, methodologies, and contributions.

Table 1. Comparison of Previous Studies and the Proposed Work.

Study	Main Objective	Method	Dataset	Personalization	Limitation
Mustafa et al.[13]	Learning app	VARK-based learning model	No	Limited	Not adaptive
Rajagede et al.[15]	Decision system	GDSS with TOPSIS	No	Partial	Ranking only
Tantawi et al.[18]	Recitation	Deep learning	No	No	Speech only
This study	Dataset + PLP	Hidden Markov Model (HMM)	Yes	Yes	Initial performance

As shown in Table 1, most previous studies focus on application development or recitation assessment, whereas this study emphasizes dataset construction to support probabilistic and adaptive personalized learning path modeling.

2. METHOD

To build a personalized learning system for Quran memorizers, a structured methodological approach is crucial. This methodological process encompasses how data is collected, processed, and prepared to support AI-based decision-making. We collected data on the activities of Quran memorizers in Indonesia from three different Islamic boarding schools (pesantren). Data collection was conducted using a quantitative approach using surveys or memorization activity recording instruments from February 2025 to June 2025 with a total of 350 students. Data descriptions can be seen in the table 2.

Table 2. Dataset Description.

Variable	Description
student_id	Unique identifier for each student
gender	Student gender
Age	Student age in years
level	Student's educational background
total_memorized_juz	Number of Juz memorized
memorization_order	The sequence of juz that has been memorized
new_memorization_rate_day_page	rate of new memorization per day in pages
new_memorization_duration_min	new memorization duration per session in minutes
review_rate_day_juz	rate of reviews per day in juz
review_duration_min	review duration in minutes
review_time_morning	memorize memorization in the morning
review_time_noon	memorize memorization during the day
review_time_evening	memorize memorization in the afternoon
review_time_night	memorize memorization at night
years_memorizing_quran	long time to memorize in years
target_years_30juz	target time to memorize 30 juz

Table 2 presents a detailed description of the variables used in this study. These variables are categorized into three main groups: demographic data, memorization ability, and study habits. The demographic data includes basic information such as student_id, gender, age, and level, which are used to represent individual characteristics. Variables related to memorization ability include the total number of memorized juz (total_memorized_juz), memorization order (memorization_order), and indicators of memorization performance, such as the rate and duration of new memorization (new_memorization_rate_day_page and new_memorization_duration_min). The study habits category captures muroja'ah (review) activities, including frequency (review_rate_day_juz), duration (review_duration_min), and the distribution of review sessions across different times of the day, namely morning, noon, evening, and night (review_time_morning, review_time_noon, review_time_evening, and review_time_night). In addition, the dataset includes variables representing the overall memorization duration (years_memorizing_quran) and the target completion time for memorizing 30 juz (target_years_30juz). Overall, this structured dataset provides a comprehensive representation of student characteristics, memorization performance, and learning behaviors. These features form a strong foundation for subsequent modeling and analysis to support adaptive learning systems and personalized Quran memorization

pathways. Furthermore, the integration of behavioral and temporal features enables the modeling of dynamic learning patterns over time. This allows the system to better capture variations in individual memorization progress and review strategies. As a result, the dataset not only supports descriptive analysis but also facilitates predictive modeling using probabilistic approaches such as Hidden Markov Models. This capability is essential for developing adaptive and context-aware learning recommendations.

2.1. Data Collection Framework

The data collection framework was designed to ensure that any information obtained was truly relevant and authentically reflected the process of memorizing the Qur'an. Figure 1 illustrates the data collection framework in the context of Quran memorization by students. This framework consists of four main components that are interconnected and serve to integrate the data comprehensively: Student Information is the starting point component of the framework, which includes basic data about students. This data is used as an individual context to personalize the learning process and track progress. This information is the basis for understanding the background and capacity of each student. This data flows into the Memorization Progress component to be integrated with memorization performance data. Memorization Progress is a component that focuses on data that reflects the process and results of students' memorization. This data is used to track the progress of students in memorizing the Qur'an periodically. This information provides insight into how fast and how much students can memorize. Data from this component flows into Quality Evaluation for quality assessment and also to Progress Tracking and Reporting for the creation of progress reports. Quality Evaluation is a component that involves assessing the quality of the memorization process based on review data and student habits. Quality evaluation is used to measure the effectiveness of the memorization method and identify whether students need adjustments in the schedule or review strategy. This ensures that the memorization is of high quality and long-lasting. The results of this evaluation flow into Progress Tracking and Reporting to provide detailed feedback based on process quality. Progress Tracking and Reporting is the final component that integrates all the data from the previous stages to produce progress and tracking reports. This includes visualizations or summaries such as progress charts. Data from Student Information, Memorization Progress, and Quality Evaluation are synthesized here to produce actionable outputs, such as daily schedule recommendations or weekly progress reports.

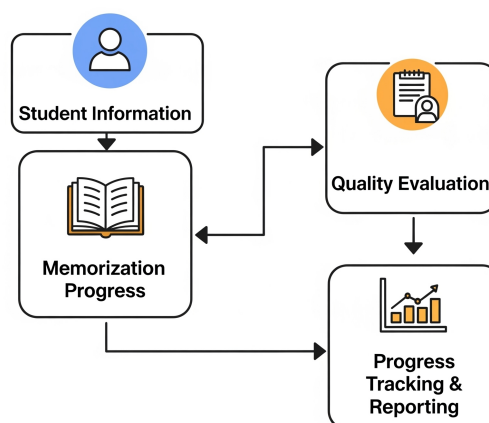


Figure 1. Data Collection Framework for Memorization.

In order to build a personalized learning system based on data, the dataset used consists of various categories of data that reflect the identity of students, educational background, as well as details of activities and performance in memorizing and reviewing the Al-Quran. Data description can be seen in the table 3.

Table 3. Dataset Variable Categories

Data Category	Variables
Demographic Data	student_id
	gender
	Age
	level
Memorization Ability	total_memorized_juz
	memorization_order
	years_memorizing_quran
	target_years_30juz
Study Habits	new_memorization_rate_day_page
	new_memorization_duration_min
	review_rate_day_juz
	review_duration_min
	review_time_morning
	review_time_noon
	review_time_evening
	review_time_night

Table 3 groups the research data into three main categories: Demographic Data, Memorization Ability, and Study Habits. Demographic data includes basic student identities such as student_id, gender, age, and level, which provide personal context. Memorization ability includes the number of juz (student chapters) memorized, the memorization sequence, the length of time devoted to memorization, the target completion of 30 juz (student chapters), the speed of new memorization per day, and the duration of new memorization. Study habits record muroja'ah patterns, including the frequency and duration of memorization repetitions per day, as well as the distribution of muroja'ah times in the morning, afternoon, evening, and night. These three categories complement each other to provide a comprehensive picture of student conditions, thus supporting the development of an adaptive learning system capable of recommending personalized memorization paths.

The raw data are stored in Excel format and include variables with varying data types, ranging from strings for identification to integers and decimals for quantitative metrics. The following table 4 describes each variable and its data type, based on the original structure of the data set before any transformations such as normalization, discretization, or feature engineering. This information is the basis for understanding the characteristics of the students and preparing the data for further analysis.

Table 4. Data Types of Dataset Variables

Variables	Data Type
student_id	String
gender	Integer (0/1)
Age	Integer
level	Integer (0/1)
total_memorized_juz	Integer
memorization_order	String
new_memorization_rate_day_page	Float
new_memorization_duration_min	Integer
review_rate_day_juz	Float
review_duration_min	Integer
review_time_morning	Integer (0/1)
review_time_noon	Integer (0/1)
review_time_evening	Integer (0/1)
review_time_night	Integer (0/1)
years_memorizing_quran	Float
target_years_30juz	Integer

2.2. Data Preprocessing

The next step in the data transformation process from the dataset is normalization, discretization, and feature engineering. These steps are designed to prepare the raw data to suit the analysis needs.

1. Normalization; Transform numeric variables to a uniform scale (e.g., [0,1]) to avoid bias due to differences in value ranges, especially when used in machine learning models. Normalization Method Using Min-Max Normalization.

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \quad (1)$$

where the following terms are used:

- X is the value of the years_memorizing_quran variable in the dataset
- $X_{\min}$ is the minimum value of the years_memorizing_quran variable in the dataset (0.125)
- $X_{\max}$ is the maximum value of the years_memorizing_quran variable in the dataset (7.0)

2. Discretization;

Transforming continuous variables into discrete categories to meet the assumptions of models such as HMMs that work with discrete states. Discretized Variables:

- total_memorized_juz :Divided into 3 categories (0-10, 11-20, 21-30 juz). Changed to category (0, 1, 2) to represent the level of student progress, which is suitable for the state in HMM.
 - new_memorization_rate_day_page:Divided into 3 categories (0-1, 1-2, ≥2 pages/day).Discretized to group memorization rates, for example students with 2 pages/day enter category 1.
3. Feature engineering. Create new features to improve model insights and performance by combining or transforming existing variables. New features added:
- efficiency_ratio:Calculated as total_memorized_juz divided by years_memorizing_quran to measure memorization speed per year.
 - freq_reviews:Total number of daily review sessions based on review_time_morning,review_time_noon, review_time_evening, and review_time_night

The transformation process includes normalization to equalize the scale of numeric variables to the range [0,1], discretization to transform continuous variables into discrete categories, and feature engineering to generate new features that provide insight into memorization performance. In addition, some variables undergo a format transformation. To assess the quality of the transformed data and ensure that the data are ready for modeling. This evaluation includes checking the integrity, consistency and quality of the performed transformations (normalization, discretization, and feature engineering).

- Data Integrity Check; The data integrity check, particularly for missing values, indicates that the dataset is in excellent condition in terms of data completeness. There are no missing values, so no imputation or row deletion is required. This is ideal for advanced statistical analysis and modeling, as it ensures that the analysis results will not be affected by data inconsistencies.
- Data Consistency; Verify that transformations (normalization, discretization) produce values that are reasonable and consistent with the expected range.Checks the range of values (min, max) for normalized variables (should be [0,1]) and the categorical distribution for discretized variables. Table 5 shows that all features (except one) have minimum value = 0.0 and maximum = 1.0 , indicating that Min-Max normalization was successfully applied correctly. One exception years_memorizing_quran: min = 0.001, max = 1.000 This is still reasonable, because the minimum value is very close to 0, and there may be no students who are really new (0 years), so the original data does not have a value of 0 and the normalization result is still valid.

Table 5. Transformation Results

Variables	Min	Max
age	0	1.0
new_memorization_duration_min	0.0	1.0
review_rate_day_juz	0.0	1.0
review_duration_min	0.0	1.0
year_memorizing_quran	0.001	1.000
target_years_30juz	0	1.0

- **Feature Engineering Quality:** Based on the descriptive statistical analysis of the `efficiency_ratio` and `freq_reviews` variables, it can be observed that the dataset exhibits consistent and reasonably well-distributed characteristics. The `efficiency_ratio`, which represents the level of memorization efficiency, has a mean value of 40.70, with values ranging from 1.77 to 80.91. This indicates noticeable variability among participants, reflecting differences in memorization speed and learning duration. The standard deviation further confirms moderate dispersion, suggesting that the variability of memorization efficiency remains within a controlled range. Furthermore, after data cleaning and preprocessing, the distribution becomes more stable and free from extreme outliers. Table 6 presents the descriptive statistical results of these newly engineered features.

Table 6. Descriptive Statistics for New Features

	efficiency_ratio	freq_reviews
count	350.00	350.00
mean	40.70	2.665714
std	32.18	0.63
min	1.77	1.00
25%	7.17	2.00
50%	18.33	3.00
75%	36.67	3.00
max	80.91	4.00

- **Distribution Validation;** The figure 2 shows four graphs that analyze distribution and efficiency in a specific context, likely related to learning or memory (based on labels such as "Memorized Juz Category" and "Memorizing Time"). Here's an explanation:

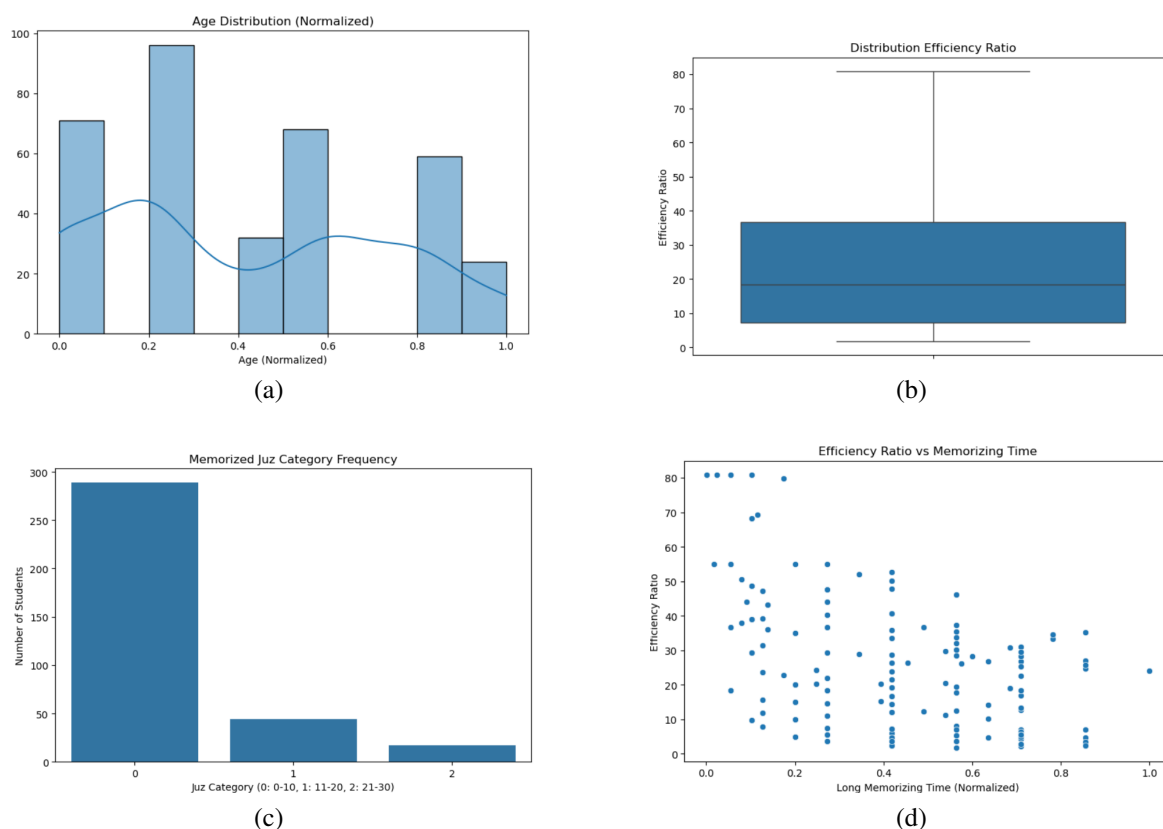


Figure 2. (a) Analysis of Age Distribution. (b) Distribution Efficiency Ratio. (c) Frequency of Memorized Juz Categories. (d) Relationship of Efficiency with Memorizing Time.

4. Feature selection

Selecting relevant features to model the sequence of students' memorization progress using Hidden Markov Model (HMM), which will be used to build a personal learning path (PLP). Features should reflect memorization efficiency, speed, and patterns. Figure 3 Heatmap that provides an initial overview of the relationship between features

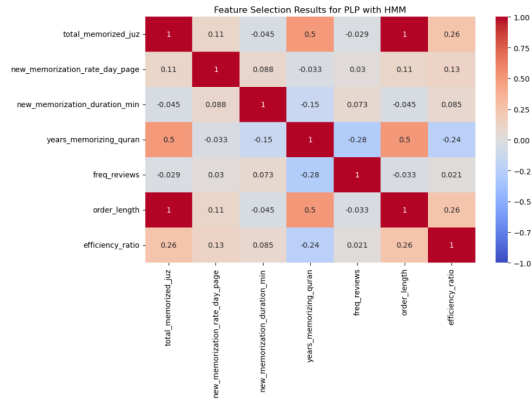


Figure 3. Feature Selection Results for PLP with HMM.

This process ensures clean data by handling extreme values that may arise due to calculations efficiency_ratio. Converting memorization_order to order_length simplifies the initial analysis. Heatmaps provide an initial overview of the relationships between features. For example, total_memorized_juz is likely to be positively correlated with efficiency_ratio, while years_memorizing_quran is likely to be negatively correlated. Table 7 shows Feature Selection Results for PLP with HMM Feature Scores and p-values

Table 7. Feature Selection Results for PLP with HMM

Feature	Score	p-value
total_memorized_juz	24.568554	0.000001
order_length	24.350081	0.000001
years_memorizing_quran	20.515740	0.000008
new_memorization_rate_day_page	5.934510	0.015348
new_memorization_duration_min	2.560047	0.110503
freq_reviews	0.149476	0.699273

Although the feature freq_reviews shows a low statistical significance compared with the other selected variables, it was retained in this study for conceptual and educational reasons. Review frequency reflects the consistency of muroja'ah practice, which is theoretically important in Quran memorization even if its direct contribution to the current model is limited. Its low significance in this dataset may be influenced by the relatively coarse measurement format and the limited variability across students. Therefore, the feature was preserved as a supporting behavioral variable and may become more informative when combined with richer temporal or quality-related review indicators in future studies.

3. RESULT AND DISCUSSION

The Hidden Markov Model (HMM) approach to model the memorization ability status of students that cannot be observed directly (hidden states), by relying on observations of daily memorization behavior. Figure 4 illustrates the architecture of an HMM-based adaptive recommendation system. This figure 4 explains how the memorization status of students (as states in HMM) can be determined based on observations of daily behavioral features. Transitions between states are made based on the probability of change learned from the data. The system then provides memorization path recommendations based on the current status and updated through feedback loops from teachers. Before a recommendation system can be developed comprehensively, it is important to understand the relationship between the memorization status of students and the most appropriate learning strategies. Based on initial data analysis and input from tahfizh educators, a mapping of

memorization status against suggested learning paths has been prepared as a basis for adaptive decision making. This mapping considers the frequency and duration of memorization, time consistency, and murojaah results as indicators of student development.

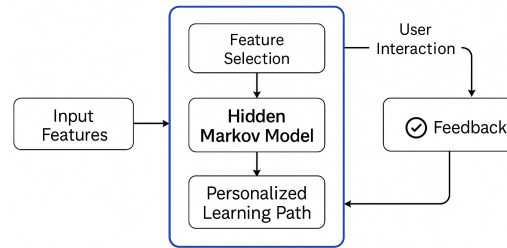


Figure 4. HMM-Based Adaptive Recommender System Architecture.

In this study, the teacher feedback loop is conceptualized as a mechanism for validating and refining system recommendations. After the model predicts a student's memorization state and suggests an adaptive learning path, teachers can review whether the recommendation matches the student's actual memorization condition. Their feedback can then be used to adjust the recommended memorization load, review schedule, or learning emphasis. Although this feedback process has not yet been implemented as an automatic retraining pipeline, it provides an important human-in-the-loop component to ensure that recommendations remain pedagogically appropriate and context-sensitive.

3.1. Initial State Distribution

Initial Distribution is to determine 3 states, namely Beginner, Intermediate, and Advanced, in the context of Hidden Markov Model (HMM) for Personalized Learning Path (PLP) based on the dataset. The initial distribution in HMM determines the initial probability of each hidden state at time $t=1$. The formula is defined as follows:

$$\pi_i = P(q_1 = s_i), \quad i = 1, 2, \dots, N \quad (2)$$

where the following terms are used:

- π_i : The probability that the initial state is s_i (in this case, $s_1 = \text{Beginner}$, $s_2 = \text{Intermediate}$, $s_3 = \text{Advanced}$)
- N : Number of hidden states (here $N = 3$).
- q_1 : Hidden state at time $t = 1$.
- Condition: $\sum_{i=1}^N \pi_i = 1$, meaning the total probability must be 100.

Since HMM is trained on observational data, the initial distribution can be estimated from the proportion of students belonging to each state category based on features such as total_memorized_juz. The empirical formula is:

$$\pi_i = \frac{n_i}{N_{\text{total}}} \quad (3)$$

The following figure shows the Initial State Distribution with 3 states: Beginner, Intermediate, and Advanced, based on the dataset.

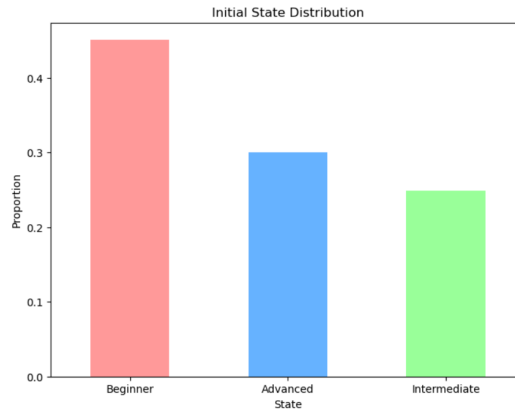


Figure 5. Initial State Distribution.

The figure 5 above shows the state distribution of participants in the learning model, categorized into three ability levels: Beginner, Intermediate, and Advanced. The results of this visualization show that the majority of participants are at the Beginner level with a proportion of around 45, followed by Advanced with a proportion of around 30, and Intermediate around 25. Table 8 shows the distribution results indicating that most participants start at a basic level of ability, so personalized learning strategies should focus on transitioning and supporting from the beginning level to higher levels gradually and adaptively.

Table 8. Initial Distribution

State	Probability (II)
Beginner	0.45
Advanced	0.30
Intermediate	0.25

3.2. State Transition Probability

The probability matrix of transitions from one state to another at time t to $t + 1$. Each element a_{ij} indicates the probability of transition from state s_i to state s_j . The transition matrix is calculated based on the sequence of states predicted by the HMM after training, using relative transition frequencies.

$$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i), \quad i, j = 1, 2, \dots, N \quad (4)$$

Figure 6 above is a visualization of the transition matrix in the form of a heatmap, which illustrates the probability of participants moving from one level of ability to another in a learning system. The vertical axis shows the state of origin, while the horizontal axis shows the destination state. Lighter colors indicate higher probabilities, with the color scale on the right side as a reference. It can be seen that the transition with the highest color intensity occurs on the main diagonal (Beginner $\rightarrow$ Beginner, Intermediate $\rightarrow$ Intermediate, Advanced $\rightarrow$ Advanced), which indicates the level of stability of participants in maintaining their respective levels.

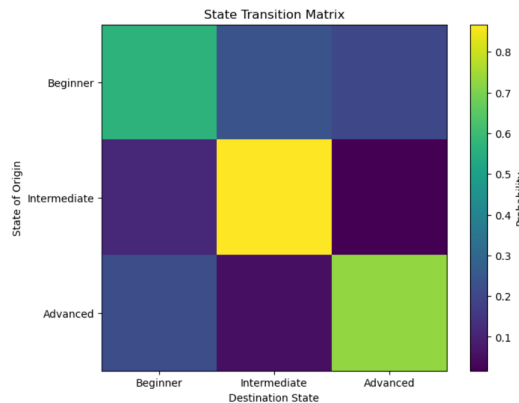


Figure 6. State Transition Matrix.

Table 9 shows the transition probabilities between participant ability levels: Beginner, Intermediate, and Advanced. Participants at the Beginner level have the greatest probability of remaining at their level at 56.7%, while the probability of transitioning to Intermediate is 23.3%, and to Advanced is 20.0%. For Intermediate participants, the probability of remaining at that level is very high at 86.7%, while only 11.6% go down to Beginner and 1.7% go up to Advanced. Meanwhile, participants at the Advanced level tend to be stable with a probability of remaining at that level of 73.1%, but there is a 21.8% chance of returning to Beginner and 5.1% to Intermediate. This pattern indicates that the transition to a higher level is not too easy and shows stability in the Intermediate and Advanced states, which can be a reference in designing a gradual learning strategy based on ability. This behavior suggests that the model prioritizes precision over recall, indicating that while predictions are reliable, many true transitions remain undetected.

Table 9. Transition Matrix (From Row to Column)

From / To	Beginner	Intermediate	Advanced
Beginner	0.567	0.233	0.200
Intermediate	0.116	0.867	0.017
Advanced	0.218	0.051	0.731

The transition matrix suggests that memorization development is relatively stable within each state, particularly for the Intermediate and Advanced groups. This indicates that once students reach a certain memorization level, they tend to remain within that level unless substantial behavioral changes occur. The relatively low upward transition probabilities also imply that progress in Quran memorization is gradual rather than abrupt, which supports the need for adaptive and stage-sensitive learning interventions.

3.3. Emission Probability

Emission probability reflects how each level of ability (hidden state) correlates with the type or quantity of observations (output) produced. High probability values in the diagonal column indicate that students with advanced abilities tend to produce larger or better quality output, while students at the beginner level produce simpler observations. This matrix is very important in the inference process, because it allows the model to map observations to the possible states that cause them. $b_j(o_t)$: Probability of observing o_t (e.g., a feature value such as total_memorized_juz) at state s_j at time t .

$$b_j(o_t) = P(o_t | q_t = s_j), \quad j = 1, 2, \dots, N \quad (5)$$

For GaussianHMM (as used), the emission distribution is assumed to be Gaussian:

$$b_j(o_t) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(o_t - \mu_j)^2}{2\sigma_j^2}\right) \quad (6)$$

Figure 7 shows a consistent trend that each feature has an average value that increases along with the state level, namely Beginner (blue) has the lowest value on all features such as total memorized juz, memorization time, sequence length, murajaah frequency, and efficiency ratio. Intermediate (orange) is in the middle, reflecting students who have developed enough but are not yet fully proficient. Advanced (green) shows the highest average value on all features, indicating optimal achievement in memorization and efficiency aspects.

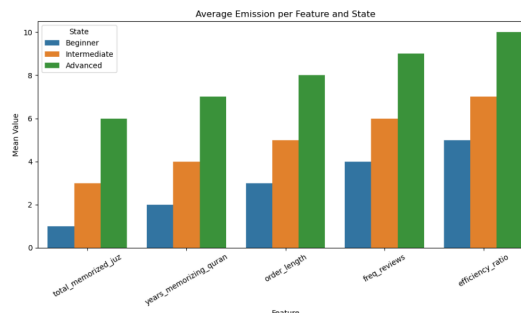


Figure 7. Average Emission.

3.4. Joint Probability Approach

Joint Probability ($P(q, o)$): In HMM, the joint probability combines the probabilities of the hidden state sequence (q) and the observation sequence (o) based on the model parameters. It is the basis for evaluation and decoding.

$$P(q, o) = \pi_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} \prod_{t=1}^T b_{q_t}(o_t) \quad (7)$$

Based on the HMM inference results in Figure 8, students experience a transition from state 0 (Beginner) to state 2 (Advanced) in just one time step, then maintain that position until the end of the observation. This transition spike is mainly driven by a significant increase in the total_memorized_juz and order_length features, each increasing from a standardized value of 0.10 to 0.65 in the second step. Meanwhile, the years_memorizing_quran and freq_reviews features remain stable in the range of 0.71 and 0.67, indicating that memorization progress quantitatively affects state shifts more than time experience or frequency of muraajaah. This finding indicates that the model is sensitive to the sharp increase in memorization achievement as an indicator of student level progress.

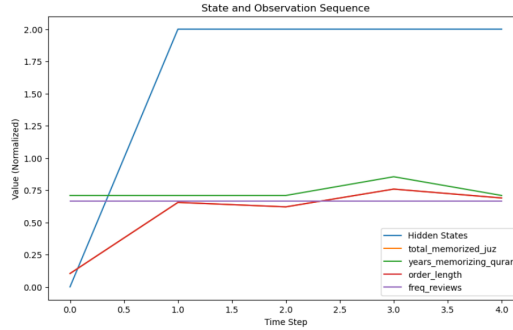


Figure 8. State and Observation Sequence.

3.5. Probability of Observation

Observation Probability ($P(o)$): This is the marginal probability of an observation sequence o without considering any hidden states explicitly. In HMM, $P(o)$ is efficiently computed using the Forward Algorithm, which sums the joint probabilities $P(q, o)$ over all possible state sequences q . Formula (Forward Algorithm):

- Partial probability up to time t with state s_i ;

$$\alpha_t(i) = P(o_1, o_2, \dots, o_t, q_t = s_i) \quad (8)$$

- Initialization;

$$\alpha_1(i) = \pi_i b_i(o_1) \quad (9)$$

- Iteration

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}) \quad (10)$$

- Termination

$$P(o) = \sum_{i=1}^N \alpha_T(i) \quad (11)$$

$$P(q, o) = \pi_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} \prod_{t=1}^T b_{q_t}(o_t) \quad (12)$$

Figure 9 depicts the sequence of observations showing a rapid and steady increase in order_length reaching a normalized peak of 2.0 from step 1 to 4, while years_memorizing_quran remains relatively constant

at 1.0, with a slight spike in step 3. The freq_reviews feature remains consistently 0.0, and total_memorized_juz is identical to the order_length pattern.

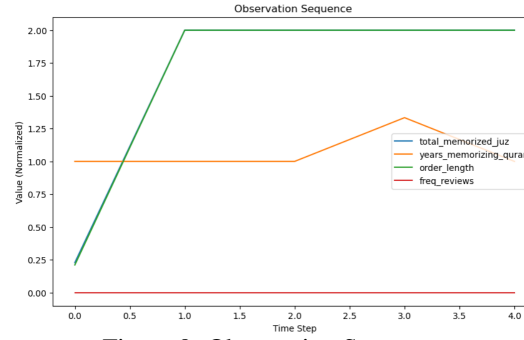


Figure 9. Observation Sequence.

3.6. Viterbi Algorithm

Implementation of the Viterbi Algorithm to find the most likely state sequence in the context of Hidden Markov Model (HMM) based on a dataset. The goal of the Viterbi algorithm is to find a sequence of states $q^* = (q_1^*, q_2^*, \dots, q_T^*)$ that maximizes the joint probability $P(q, o)$ for a sequence of observations $o = (o_1, o_2, \dots, o_T)$. Viterbi Algorithm Formula and Steps:

- Variable Definition:

$$V_t(i) = \max_{q_1, \dots, q_{t-1}} P(q_1, q_2, \dots, q_{t-1}, q_t = s_i, o_1, o_2, \dots, o_t). \quad (13)$$

- Initialization (t = 1):

$$V_1(i) = \pi_i \cdot b_i(o_1) \quad (14)$$

- Iteration (t = 2 to T): For each state s_j at time t :

$$V_t(j) = \max_i [V_{t-1}(i) \cdot a_{ij}] \cdot b_j(o_t) b_j(o_{t+1}) \quad (15)$$

Save backpointer:

$$\text{ptr}_t(j) = \arg \max_i [V_{t-1}(i) \cdot a_{ij}] \quad (16)$$

- Termination (t = T):

find the final state that gives the maximum probability:

$$q_T^* = \arg \max_i V_T(i) \quad (17)$$

Best path probability:

$$P^* = \max_i V_T(i) \quad (18)$$

- Backtracking:

$$q_t^* = \text{ptr}_{t+1}(q_{t+1}^*) \quad (19)$$

Figure 10 depicts the state transition from 0 to 1 in the first step along with a drastic increase in the observed feature values especially total_memorized_juz and order_length. This indicates that the HMM model recognized significant changes in the learning process early on and mapped these transitions to hidden state changes. After the transition, the features remained stable with little variation, and the state remained at 1. This illustrates the “consolidation” phase, where the student maintained his memorization progress consistently. The lack of repetitions (freq_reviews = 0) during all five time steps could be a concern for further coaching, as the lack of reviews could impact long-term retention.

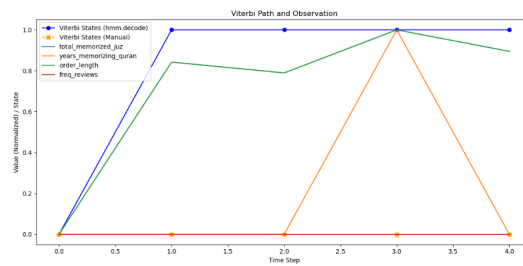


Figure 10. Viterbi Path and Observation.

3.7. Evaluation

The Hidden Markov model developed in table 10 shows quite varied performance based on the evaluation of the prediction of the status of the Quran memorizers. The accuracy level of the model is at 30.57% which indicates that only about one-third of the total status predictions match the actual conditions. However, the precision of the model is recorded as high at 71.46%, indicating that the status predictions given by the model are generally correct when deciding a state. On the other hand, recall is still low at 30.57%, which means that the model fails to recognize most of the correct actual statuses. As a result of this imbalance, the F1-score value only reaches 37.70%, reflecting a less than ideal trade-off between accuracy and coverage. These results imply that although the model is quite confident when predicting, improvements are still needed to increase its ability to recognize more variations in the actual statuses that occur. The evaluation results indicate that the model is more reliable when it assigns a memorization state than when it attempts to identify all actual state changes. This is reflected in the relatively high precision but low recall. In practice, this means the current model is conservative: it avoids frequent incorrect predictions, but it also misses many true transitions. Such a pattern is understandable in an initial HMM framework trained on limited observational variables, yet it highlights the need for richer sequential evidence and improved state calibration.

Table 10. Model Performance Evaluation

Metric	Score
Accuracy	0.305714
Precision	0.714620
Recall	0.305714
F1-score	0.376965

Figure 11 shows a comparison between the Viterbi prediction results and the target state, indicating that the model predicts all time steps to remain in the initial state (state 0), while the actual target undergoes a transition from state 0 to state 1 and finally reaches state 2. The model's inability to recognize this change indicates a weakness in capturing the dynamics of the Quran memorizer's state transition temporally. This is in line with the resulting evaluation metrics, where the accuracy only reaches 30.57% and the recall is also low at the same figure, reflecting that the model fails to identify most of the correct states. This condition requires improvements to the model structure, especially in the transition parameters between states and the use of more representative features to detect progress more accurately.

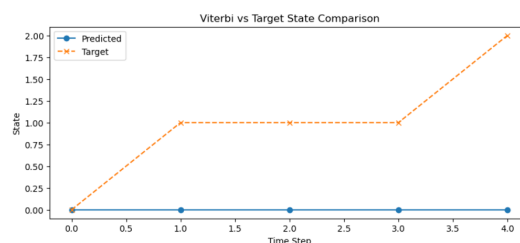


Figure 11. Viterbi Path and Observation.

This study emphasizes the importance of a data-driven approach in transforming conventional Quran memorization methods into a more personalized and adaptive learning system. By implementing the Hidden Markov model (HMM), we successfully model the progress of memorization ability through hidden statuses (Beginner, Intermediate, Advanced) derived from the daily behavioral features of students. The developed

adaptive recommendation system allows real-time status classification and memorization path suggestions based on the dynamics of individual development. Limited trials at tahfiz institutions showed that the system was able to increase the efficiency of muroja'ah time and memorization retention. The interactive dashboard used by teachers also received positive responses because it provided visualization of student development and supported data-based decision-making. The feedback loop mechanism allows the system to continuously adjust recommendations based on teacher intervention, making the system increasingly relevant in real learning contexts. Beyond predictive metrics, the practical value of the proposed system lies in its potential to support more personalized Quran memorization guidance. By identifying students' memorization states and mapping them to adaptive learning paths, the system can help teachers adjust review intensity, memorization targets, and learning schedules according to individual progress. In practical settings, this may improve the efficiency of muroja'ah planning, reduce mismatches between student ability and memorization load, and support more data-informed supervision in tahfiz education.

However, there are some limitations that need to be considered. Although the model precision is relatively high, the accuracy and recall levels are still low, indicating the need for improvements in capturing memorized state transitions. Adding temporal features, increasing data resolution, and integrating technologies such as speech recognition or affective analysis could be future development directions. The relatively low accuracy and recall indicate that the current HMM configuration is still limited in capturing the full dynamics of students' memorization transitions. One possible reason is that the observation sequence is constructed from a limited number of features, which may not fully represent changes in memorization quality, consistency, and retention over time. In addition, the transition structure may oversimplify the complexity of real memorization development, where progress is often non-linear. Therefore, future improvements should include richer temporal and behavioral features, more refined state definitions, and hybrid sequential models to better represent memorization trajectories. This study is also limited by the dataset scope, which includes 350 students from only three Islamic boarding schools in Indonesia. While the dataset is sufficient for an initial proof-of-concept study, it may not fully represent the diversity of memorization practices across broader educational, cultural, and institutional contexts. As a result, the generalizability of the findings remains limited. Future studies should validate the proposed framework using larger and more diverse datasets collected from different regions and memorization environments.

4. CONCLUSIONS

This study presents the development of a structured dataset for Quran memorizers as a foundation for data-driven personalized learning path modeling. By integrating demographic, memorization, and study habit variables and preparing them through normalization, discretization, and feature engineering, the dataset was made suitable for Hidden Markov Model (HMM)-based analysis. Experimental results indicate that the current HMM framework achieves relatively high precision but limited overall predictive performance, highlighting both its potential and current limitations. The findings demonstrate that the proposed dataset can serve as an initial resource for modeling memorization progress and supporting adaptive learning recommendations. However, the study is constrained by the limited dataset size and the simplicity of the modeling approach. Overall, this work provides an important step toward intelligent, adaptive, and context-aware Quran memorization systems, while future research should focus on richer features, hybrid modeling strategies, and broader validation scenarios.

Future research can be extended in several directions. First, additional behavioral and temporal variables, such as memorization error rate, review quality, retention score, and speech-based recitation feedback, should be incorporated to enrich the observation sequence. Second, hybrid modeling approaches, including HMM integrated with Bayesian Networks or recurrent neural networks, can be explored to better capture both sequential dependencies and contextual relationships. Third, validation on larger and more diverse student populations is necessary to improve robustness and generalizability. Finally, future implementations should incorporate an operational teacher feedback mechanism to dynamically refine recommendations over time.

REFERENCES

- [1] K. Pan and R. Sen, "Navigating the evolving landscape of personalized learning in aided: Recent trends and innovations," *Explainable AI for Education: Recent Trends and Challenges*, pp. 169–192, 2024.

- [2] M. Bendahmane, B. El Falaki, and M. Benattou, "Toward a personalized learning path through a services- oriented approach," *International Journal of Emerging Technologies in Learning (Online)*, vol. 14, no. 15, p. 52, 2019.
- [3] A. Bhutoria, "Personalized education and artificial intelligence in the united states, china, and india: A systematic review using a human-in-the-loop model," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100068, 2022.
- [4] K.-J. Laak and J. Aru, "Ai and personalized learning: bridging the gap with modern educational goals," *arXiv preprint arXiv:2404.02798*, 2024.
- [5] C. Song, S.-Y. Shin, and K.-S. Shin, "Implementing the dynamic feedback-driven learning optimization framework: a machine learning approach to personalize educational pathways," *Applied Sciences*, vol. 14, no. 2, p. 916, 2024.
- [6] B. Curum and K. K. Khedo, "Cognitive load management in mobile learning systems: principles and theories," *Journal of Computers in Education*, vol. 8, no. 1, pp. 109–136, 2021.
- [7] D. Kapadiya, C. Shekhawat, and P. Sharma, "A study on largescale applications of big data in modern era," in *Proceedings of the 5th International Conference on Information Management and Machine Intelligence*, 2023, pp. 1–6.
- [8] F. Alvindo, N. A. Setiawan, and H. A. Nugroho, "Al-quran recitation or sound processing analysis: A systematic literature review on methods," in *2024 International Conference on Electrical Engineering and Computer Science (ICECOS)*. IEEE, 2024, pp. 54–58.
- [9] A. Da'u and N. Salim, "Recommendation system based on deep learning methods: a systematic review and new directions," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2709–2748, 2020.
- [10] R. J. K. Almahmood and A. Tekerek, "Issues and solutions in deep learning-enabled recommendation systems within the e-commerce field," *Applied Sciences*, vol. 12, no. 21, p. 11256, 2022.
- [11] B. Ramzan, I. S. Bajwa, N. Jamil, R. U. Amin, S. Ramzan, F. Mirza, and N. Sarwar, "An intelligent data analysis for recommendation systems using machine learning," *Scientific Programming*, vol. 2019, no. 1, p. 5941096, 2019.
- [12] E. Hasan, M. Rahman, C. Ding, J. Huang, and S. Raza, "Based recommender systems: a survey of approaches, challenges and future perspectives," *ACM Computing Surveys*, 2024.
- [13] N. M. Mustafa, Z. Mohd Zaki, K. A. Mohamad, M. Basri, and S. Ariffin, "Development and alpha testing of ezhifz application: Al-quran memorization tool," *Advances in Human-Computer Interaction*, vol. 2021, no. 1, p. 5567001, 2021.
- [14] M. G. B. Musa, M. H. N. B. Yusop, M. M. B. M. Sopee, and N. A. M. Ali, "I-tasmik mobile platform– enabling tahfiz student to memorize al-quran independently," in *2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M)*. IEEE, 2018, pp. 24–29.
- [15] R. A. Rajagede, Y. H. Hisyam, M. I. A. Yulianto, F. A. Ridwanto, A. T. Putranto, and M. R. F. P. Danar, "Personalized al-quran memorization testing system using group decision support system," *Aceh International Journal of Science and Technology*, vol. 10, no. 3, pp. 171–181, 2021.
- [16] D. Purbohadi, B. R. Rahmawati, and H. Setiawan, "Development of qur'an memorization learning model based on mobile learning," in *Journal of Physics: Conference Series*, vol. 1381, no. 1. IOP Publishing, 2019, p. 012029.
- [17] N. Senan, W. A. W. Ab Aziz, M. F. Othman, and S. Suparjoh, "Embedding repetition (takrir) technique in developing al-quran memorizing mobile application for autism children," in *MATEC Web of Conferences*, vol. 135. EDP Sciences, 2017, p. 00076.
- [18] I. K. Tantawi, M. A. Abushariah, and B. H. Hammo, "A deep learning approach for automatic speech recognition of the holy qur'an recitations," *International Journal of Speech Technology*, vol. 24, no. 4, pp. 1017–1032, 2021.
- [19] Y. S. Alsaahafi and M. Asad, "Empirical study on mispronunciation detection for tajweed rules during quran recitation," in *2024 6th International Conference on Computing and Informatics (ICCI)*. IEEE, 2024, pp. 39–45.
- [20] S. M. Hussain, M. N. Jilani, M. U. Haq, and M. S. Shaikh, "A machine learning approach to arabic phoneme classification through ensemble techniques," in *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*. IEEE, 2024, pp. 1–7.
- [21] M. K. K. Alsou and R. A. I. Alsaraireh, "Data-driven decision-making in education: Leveraging ai for school improvement," in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, vol. 1. IEEE, 2024, pp. 1–6.
- [22] L. Tetzlaff, F. Schmiedek, and G. Brod, "Developing personalized education: A dynamic framework," *Educational Psychology Review*, vol. 33, pp. 863–882, 2021.
- [23] R. K. Yekollu, T. Bhimraj Ghuge, S. Sunil Biradar, S. V. Haldikar, and O. Farook Mohideen Abdul Kader, "Ai-driven personalized learning paths: Enhancing education through adaptive systems," in *International Conference on Smart data intelligence*. Springer, 2024, pp. 507–517.
- [24] I. Kabashkin, B. Mironovs, and O. Zervina, "Ai-driven and ontology-based framework for personalized learning pathways in education," in *International Conference on Reliability and Statistics in Transportation and Communication*. Springer, 2024, pp. 494–506.
- [25] Y. Luo, "Personalized learning paths based on big data analysis in digital english teaching," in *2024 3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*. IEEE, 2024, pp. 734–738.
- [26] Y. Guo and Y. Li, "Exploration of personalized teaching paths in english education based on big data analysis," *Journal of Computational Methods in Sciences and Engineering*, p. 14727978251322283, 2025.
- [27] C. G. Demartini, L. Sciascia, A. Bosso, and F. Manuri, "Artificial intelligence bringing improvements to adaptive learning in education: A case study," *Sustainability*, vol. 16, no. 3, p. 1347, 2024.
- [28] T. Kabudi, I. Pappas, and D. H. Olsen, "Ai-enabled adaptive learning systems: A systematic mapping of the literature," *Computers and education: Artificial intelligence*, vol. 2, p. 100017, 2021.
- [29] X. Wang, R. uang, M. Sommer, B. Pei, P. Shidfar, M. S. Rehman, A. D. Ritzhaupt, and F. Martin, "The efficacy of artificial intelligence-enabled adaptive learning systems from 2010 to 2022 on learner outcomes: a meta-analysis," *Journal of Educational Computing Research*, vol. 62, no. 6, pp. 1568–1603, 2024.
- [30] L. Lim, S. H. Lim, and W. Y. R. Lim, "A rasch analysis of students' academic motivation toward mathematics in an adaptive learning system," *Behavioral Sciences*, vol. 12, no. 7, p. 244, 2022.

BIOGRAPHY OF AUTHORS

Miftah Farid Adiwisatra    Earned a Bachelor of Engineering degree from BSI University, Bandung, Indonesia, and a Master of Computer Science degree in Computer Science from Nusa Mandiri University, Jakarta, Indonesia. Currently, he is pursuing a Doctoral degree in Informatics at Telkom University, Bandung, Indonesia. Research interests are in the fields of data science, adaptive learning, Information Systems. He is currently teaching at Bina Sarana Informatika University.



Irfan Darmawan    received the Ph.D. degree in informatics from the School of Electrical and Informatics Engineering, Bandung Institute of Technology. He is currently a Professor of open data with the Information Systems Department, Faculty of Industrial Engineering, Telkom University. His research interests include data mining, information systems, and the IoT.



Dade Nurjanah    He earned a Bachelor of Engineering and Master of Engineering degree from the Bandung Institute of Technology, Indonesia. He earned a Ph.D. in Computer Science from the University of Southampton, England. His research interests include Adaptive Learning, User Modeling, Computer-Supported Cooperative Work, and Social Computing. He is currently teaching at Telkom University.