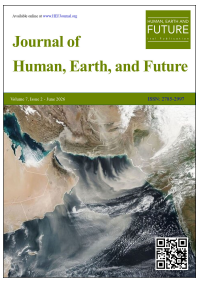
Available online at www.HEFJournal.org

ISSN: 2785-2997

Journal of
Human, Earth, and Future

Vol. 7, No. 2, June, 2026



Modeling and Forecasting LQ45 Stock Index Dynamics with a Hidden Markov Model

Arip Ramadan [1], Fadya Amalia Zahra [2], Dwi Rantini [2, 3*], Fazidah Othman [4], Mochammad Fahd Ali Hillaby [2]

[1] *Information System Study Program, School of Industrial and System Engineering, Telkom University, Surabaya Campus, Jl. Ketintang No.156, Surabaya 60231, East Java, Indonesia.*

[2] *Data Science Technology Study Program, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya 60115, Indonesia.*

[3] *Research Group of Data-Driven Decision Support System, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya 60115, Indonesia.*

[4] *Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia.*

Received 11 June 2025; Revised 02 March 2026; Accepted 16 March 2026; Published 01 June 2026

Abstract

Investing has become increasingly popular with the advancement of digital technology. A thriving stock market, reflecting strong investor confidence and capital formation, often serves as a leading indicator of robust economic growth. One of the main indicators used in evaluating the performance of the Indonesian stock market is the LQ45 index on the Indonesia Stock Exchange (IDX). The objective of this research is to determine the predictive capability of the LQ45 stock index in identifying buying and selling opportunities. This research employs the hidden Markov model method, which has the ability to capture unobservable observations and predict the uncertainty that occurs. The research, which utilized training data in 2023, made predictions for the first quarter of 2024. The results of the model were evaluated using a paired t-test with a significance value of 5%, obtaining a p-value of 0.51747 for the Open index, 0.28551 for Close, and not significant for High and Low. These values indicate that there is no difference between actual data and predicted data on Open and Close data. Applying a Hidden Markov Model (HMM) to LQ45 stock analysis offers a significant improvement over traditional Markov Chain methods by accounting for the unobservable factors that influence stock prices.

Keywords: Hidden Markov Model (HMM); Prediction; Stock; LQ45 Index; Economic Growth.

1. Introduction

A stock index is a statistical measure that reflects the overall price movement of a group of stocks selected based on specific standards and methods and analyzed periodically [1]. The index serves as an indicator of stock market performance, describing changes in the stock prices of a particular group of companies as a representation of the market as a whole [2]. In investment, stock indices help monitor financial markets and serve as a basis for decision-making [3].

One of the leading indicators used to evaluate the performance of the Indonesian stock market is the LQ45 index on the Indonesia Stock Exchange (IDX). The stocks included in this index are highly demanded and are a preferred

* Corresponding author: dwi.rantini@ftmm.unair.ac.id

https://doi.org/10.28991/HEF-2026-07-02-013

choice for investors. This is because LQ45 stocks have large market capitalization and high trading frequency, indicating strong growth prospects and sound financial conditions [4, 5]. The LQ45 index is calculated based on a selection of 45 stocks [6].

Stock movements are represented as time series, which describe data over sequential time intervals [7]. Changes in the stock index are fluctuating, random, and unpredictable, and can therefore be considered stochastic processes [8]. This allows analysts to estimate the probability of future stock prices and assess risks in investment decision-making. Information about the stock market can be obtained through statistical analysis, particularly by examining the behavior and trends of stock prices [9]. However, stock prices tend to fluctuate randomly, and predictions cannot be reliably made based solely on historical data, making forecasting increasingly complex [10]. Predicting stock price behavior remains an important and interesting topic due to its relevance for generating profit opportunities for companies and investors [11, 12]. Various methods have been developed to analyze and predict stock price movements, primarily through the application of statistical analysis [13].

The research by Yang et al. introduces a novel analytical pipeline integrating autoencoders, arithmetic optimization clustering, and principal component analysis to address the limitations of traditional real estate forecasting methods [14]. The pipeline identifies five critical clusters of industrial market stock indices related to real estate in Taiwan between 2009 and 2020, providing actionable insights for policymakers. In the research by Adlakha et al., a stock forecasting system is developed using a combination of stacked LSTM, linear regression, random forest, and K-nearest-neighbor algorithms to predict stock trends based on price history [15]. The proposed framework analyzes each company's stock using mathematical technical metrics, aiming to reduce the risk of failure in stock exchanges while increasing potential benefits. Ayala et al.'s research proposes a hybrid approach for generating trading signals by combining technical indicators with machine learning techniques such as Linear Models, Artificial Neural Networks, Random Forests, and Support Vector Regression [16].

The resulting technique, tested on daily trading data from three major indices, demonstrates that integrating machine learning with technical analysis improves trading signals and the competitiveness of trading rules. For LQ45 data, several researchers have analyzed the data. The research by Syukur & Istiawan proposed a comparison framework to benchmark the performance of various classification models in predicting the LQ45 index using transaction-level and capitalization-size data from the Indonesian Stock Exchange [17]. Results indicate that the Random Forest algorithm performs best for predicting the LQ45 index, while traditional statistical-based learners underperform. Simahatie & Inuzula's research used an associative approach to examine the effect of independent variables on dependent variables within the Indonesia Stock Exchange, drawing data from LQ-45 index issuers between 2017 and 2020 [6]. Purposive sampling was employed based on specific criteria aligned with the study's objectives.

Based on the existing research, mentioned above, several possibilities can be explored. Arithmetic optimization clustering is a clustering algorithm. Clustering algorithms, in general, treat data points as independent and identically distributed [18, 19]. They do not account for the temporal order of data points. Principal component analysis is a dimensionality reduction technique that identifies the principal components of the data [20]. It does not account for the temporal order of data points. LSTMs are designed for sequential data and model temporal dependencies [21]. Linear regression is typically used for static prediction or classification tasks [22]. Random Forest is an ensemble learning method that aggregates the predictions of multiple decision trees [23]. Each tree is trained on a subset of the data and a subset of the features. While Random Forest can capture complex relationships between features, it treats each data point as independent. It does not explicitly model the temporal order or dependencies between data points in a sequence. While various machine learning techniques offer different strengths, Hidden Markov Models (HMMs) provide a more robust and interpretable solution for explicitly modeling temporal dependencies and underlying states in sequential data, addressing the limitations of methods that treat data points as independent or lack a probabilistic framework for understanding dynamic relationships [24].

Based on the problems explained in the background, this research will examine the prediction of the LQ45 stock index by applying the Hidden Markov Model (HMM). The reason for choosing the LQ45 stock index as the subject of analysis is that it has a large market capitalization [25]. HMMs might be chosen over ARIMA, LSTM, or GARCH models primarily when the underlying time series is believed to be driven by unobserved states, such as market regimes or customer journey stages, offering interpretability by examining state transitions and emission probabilities [26]. While ARIMA focuses on autocorrelation, LSTM requires extensive data and can be a black box. GARCH specializes in volatility modeling; HMMs excel at capturing regime-switching behavior. They can be more suitable with limited data or a need for model transparency, though the optimal choice depends on the specific problem, data characteristics, and goals of the analysis [27].

This paper is structured as follows: Section 2 presents a review of the relevant literature about this research, including an introduction to the LQ45 index, explanations of Markov Chains, the Hidden Markov Model (HMM) process, the Baum-Welch algorithm, the Paired-Samples t-test, and a description of the Akaike Information Criterion

(AIC) and Bayesian Information Criterion (BIC) as measures of model fit. Section 3 details the methods employed, providing information regarding data preparation prior to analysis, model initialization, determination of the optimal number of states, HMM training, prediction, evaluation, and a visual representation of the research process in the form of a flowchart. Section 4 presents the results of the analysis and subsequent discussion, encompassing the characteristics of the LQ45 index, model initialization, and so forth; the findings presented in this section directly address the steps outlined in Section 3. Finally, Section 5 provides a comprehensive conclusion based on the analyses conducted throughout this study.

2. Literature Review

2.1. LQ45 Index

In February 1997, the LQ45 stock index was first introduced to the public. At that time, the LQ45 stock index was released with an index value of 100, precisely on July 13, 1994. The LQ45 index was created by the Indonesia Stock Exchange (IDX), which aims to complement the IDX composite, namely by providing a means for investors, financial analysts, and investment managers to observe the movement of stocks with high transactions [28]. LQ45 stocks are a combined calculation of 45 stocks, which will be assessed and selected through several selection criteria from the stock market [17, 29]. These criteria are based on liquidity, market capitalization, and transaction activity indicated by the volume and number of transactions on the market. In addition, the shares must have been listed on the IDX for at least 3 months. The shares must also be included in the top 60 companies with the highest market capitalization and transaction value in the last 12 months.

Every 3 months, the LQ45 index is evaluated to examine the stocks included in the LQ45. In this evaluation, the stocks included in the LQ45 index will be assessed for their suitability, considering whether they are still relevant or no longer in accordance with the established criteria. Every February and August, there is a change of stocks exactly once every 6 months. If certain stocks in the LQ45 no longer meet the LQ45 index selection criteria, then the stocks are removed from the index calculation and replaced with other stocks that meet the criteria. Evaluation of the list of stocks that meet the criteria is carried out periodically to ensure that the LQ45 index represents active and liquid stocks in the market. The 45 companies whose shares are listed on the IDX for the constituent period from August 2023 to January 2024. The list of stocks included in LQ45 can be seen on: *https://www.idxchannel.com/*.

In the research by Pradana, the LQ45 Indonesian stock index is analyzed and forecasted using an ARIMA (4, 3, 6) model, demonstrating accurate short- to medium-term predictions [30]. The findings offer practical implications for investors and contribute a robust approach to stock index forecasting, encouraging further research incorporating advanced models and external economic factors. In the research by Hidayat & Suhendri, nine machine learning algorithms are comparatively evaluated for predicting LQ45 stock movements, finding that Random Forest and XGBoost perform best with smoothed technical indicators for continuous data, while Naive Bayes excels in binary classification tasks [31]. The research highlights the importance of data representation and preprocessing techniques, particularly smoothing, for enhancing the accuracy and robustness of stock prediction models. Reyzan & Abdurrohim's research investigated the relationship between CR, DER, ROA, and PER in nine companies, revealing inconsistencies between conventional theory and empirical data [32].

The results indicate that CR and DER do not affect ROA. However, all three independently influence PER, with ROA mediating the effect of DER on PER, highlighting the need for companies to balance CR and DER for increased PER and investor confidence. Arief & Hidayat's research used multiple linear regression to analyze the effects of Bitcoin and macroeconomic factors on the LQ45 index from 2018 to 2022 [33]. The results indicate that Bitcoin, exchange rate, BI Rate, and money supply negatively impact the LQ45 index, while inflation and the IDX have positive effects. In the research by Astuti & Anggraini, they measured and compared the Value at Risk (VaR) of single stocks (BBRI, BBCA, ASII) and a portfolio within the LQ45 index using the Variance-Covariance method [34]. The results show that the combined portfolio has a lower VaR than the individual stocks, indicating that diversification effectively reduces investment risk.

2.2. Markov Chain

A Markov Chain is a stochastic process that can be mathematically defined as a sequence of possible events, with the probability of each event occurring only based on the current state and not affected by past events [35]. The Markov chain model was discovered by a Russian mathematician named Andrei Andreyevich Markov in 1906. Markov chains are used to see the possibility of a state occurring in the future. The Markov process has a "memoryless" nature, namely, the probability of a state occurring is only influenced by today's events without looking at past events [36].

The main components used in the development of the Markov Chain model are the state transition matrix and probability [37]. These components summarize all the important parameters that make it the basis for developing a

mathematical representation for the dynamics of change in a system that involves elements of randomness. The transition matrix on an element describes the changes that occur from one state to another. Probability explains the chance of an event occurring. So, the transition probability shows the change from one state to another, which is a random process expressed in probability.

Stock price movements are described as time series, a representation of data in sequential time intervals [38]. Changes in stock prices are random and unpredictable, so they can be said to be stochastic processes [39]. In this context, the condition of stock prices is a random process, namely, all information about the future is contained in the current state. Therefore, the process of predicting future events depends on current price conditions without being influenced by previous events.

In a Markov chain, several basic assumptions must be known [40]. The assumptions include the following:

- Initial State;
- Transition State;
- Steady State.

The stochastic process $X_t = X_0, X_1, X_2, \ldots$ represents a state that transitions over time. The Markov process (X_t) is a stochastic process with the property that if given the value of X_t, the value of X_s for $s > t$ is not affected by the values of X_u for $u < t$. With s referring to the state of time in the future, t referring to the state at time t, and u referring to the state of time in the past. This shows that when the current condition is known, knowledge of past behavior does not change the probability of a particular event in the future.

A Markov chain refers to a process whose state space consists of countable numbers, and the time index $T = (0, 1, 2, \ldots)$, and $X_n \geq 0$ states an event that occurs in a certain time period. Markov properties are notated as in Equation 1.

$$P(X_{t+1} = j | X_t = i, X_{t-1}, X_{t-2}, \ldots, X_1 = i, X_0 = i_0) = P(X_{t+1} = j | X_t = i) \quad (1)$$

where P is the probability of event, i is the index for initial state, j is the index for the initial state that may follow after the initial state, t is the current time, $t + 1$ is the next time after t, X_t is the state at time t, X_{t+1} is the state at time $t + 1$, X_{t-1} is the state at time $t - 1$, and X_{t-2} is the the state at time $t - 2$, and X_0 is the state at time 0 and i_0 is the index for state 0.

The probability of X_{t+1} being in state j with X_t being in state i is called the one-step transition probability and is denoted by $P_{ij}^{t,t+1}$, or described in Equation 2.

$$P_{ij}^{t,t+1} = P(X_{t+1} = j | X_t = i) \quad (2)$$

where, P_{ij} is the probability of transition from state i to state j. The mathematical notation emphasizes that the transition probability function is a function that depends not only on the initial and final states but also on the transition time. Transition time refers to the change that occurs between one state and another or the change that occurs in time between the initial and final states. Most Markov chains have stationary transition probabilities, which occur when the one-step transition probability is independent of the time variable t, that is, when the probability of an event does not change over time. In this research, the stationary probability is used to see the time picture when the stock index no longer changes. In $P_{ij}^{t,t+1} = P_{ij}$ is independent or does not depend on t. Meanwhile, P_{ij} explains the conditional probability of the transition state from i to j in one trial, so it must meet the conditions in Equation 3.

$$P_{ij} \geq 0 \text{ for } i, j = 0, 1, 2, \ldots \quad (3)$$

and Equation 4:

$$\sum_{j=0}^{\infty} P_{ij} = 1 \text{ for } i = 0, 1, 2, \ldots \quad (4)$$

The transition matrix A is expressed in non-negative numbers because it is a probability. The transition matrix contains the information needed to predict what will happen, based on previous knowledge. The transition matrix is written using the mathematical formula in Equation 5.

$$A = \begin{array}{c} \\ p_0 \\ p_1 \\ \vdots \\ p_n \end{array} \begin{array}{c} \begin{array}{cccc} p_0 & p_1 & \cdots & p_n \end{array} \\ \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0n} \\ p_{10} & p_{11} & \cdots & p_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n0} & p_{n1} & \cdots & p_{nn} \end{bmatrix} \end{array} \quad (5)$$

The transition matrix A is of $N \times N$ with each element a_{ij} representing the probability of transitioning from state i to state j. In the transition matrix A, the rows represent the initial conditions of state i. The columns represent the conditions after the transition or state j, for example, p_{01} shows the transition probability from state 0 to state 1. Then, p_{ij} is the probability of moving from initial state i to state j.

2.3. Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a method developed from the Markov Chain method discovered by A.A Markov in 1906. In this model, the system is considered as a Markov process by considering unknown observations (hidden) [41]. This method is based on unobserved states and transitions between states [24, 42]. Although the transitions between states are not directly observed, they have an impact on the observed results [43]. In HMM, the model integrates several Markov chains, one of which is a chain consisting of observable states, while the others form unobservable states (hidden) [44]. In the framework of the conventional Markov model, each state can be clearly observed by the observer. In this case, the probability of transition between states becomes a single observable parameter. However, there are situations where there is a sequence of states that can be understood but cannot be observed directly [45]. The advantage of the HMM is its ability to model hidden states. In this model, the state cannot be observed directly, but the output produced depends on the state that can be observed. The sequence of steps generated by HMM provides information about the underlying state sequence. It is important to note that the term "hidden" refers to the states that the model passes through, not to the model parameters themselves. Even when the model parameters are known, the hidden nature of the model remains. An HMM model $\lambda = (\pi, A, B)$ has the following component parameter characteristics.

1. N, the number of hidden state elements S_t contained in the model.
2. M, the number of observable state components.
3. $A = \{a_{ij}\}$, the transition probability matrix A. a_{ij} represents the probability of moving from state i to state j.
4. $B = \{b_i(k)\}$, the emission matrix. Describes the probability distribution of observations at state i and at time t. Represented as $b_i(k) = P(O_t = k|X_t = i)$, for $1 \leq i \leq N$ and $1 \leq k \leq M$. Then, $b_i(k) = N(\mu_i, \Sigma_i)$, for μ_i and Σ_i are means and covariances based on state S_i.
5. $\pi = \{\pi_i\}$, the set of initial state probabilities.

2.4. Forward and Backward Algorithm

The forward algorithm is an algorithm commonly used in the HMM. This algorithm is used to calculate the probability of observation by summing the probabilities of all possible hidden state paths that can produce a sequence of observations [46]. In calculating the probability of observation, this algorithm considers the possibilities that occur from all combinations of hidden states. This algorithm calculates the likelihood $P(O|\lambda)$ with HMM parameters $\lambda = (\pi, A, \pi, \Sigma)$ and with observations o_t. The stages in the forward algorithm are divided into 3 (three), namely initialization, recursion, and termination. In Equation 6, $\alpha_t(j)$ represents the probability of being in state j after the first t observations. The notation $q_t = j$ explains that at time t, the state in a system is j.

$$\alpha_t(j) = P(o_1, o_2, \dots, o_t, q_t = j|\lambda) \tag{6}$$

1) Initialization:

$$\alpha_t(j) = P(\pi_j b_j(o_1)), \qquad 1 \leq j \leq N \tag{7}$$

2) Recursion:

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t), \qquad 1 \leq j \leq N, 1 \leq t \leq T \tag{8}$$

3) Termination:

$$P(O|\lambda) = \sum_{j=1}^{N} \alpha_T(j) \tag{9}$$

where, $\alpha_t(j)$ is the probability of observing a sequence of observations and being in state j at time t and given model λ, λ is the HMM parameters, o_t is the observation value at time t, q_t is the hidden state at time t, π_j is the initial probability at state j, $b_j(o_t)$ is the probability of emission resulting in observation o_t at state j, a_{ij} is the probability of transition from state i to state j, N is the total number of states, and T is the order of observations.

The backward algorithm is an algorithm commonly used in the HMM. This algorithm is used to calculate the probability of an observation, like the Forward Algorithm. There is a difference between the two algorithms in the

initialization stage; for the backward algorithm, the initialization is based on all observation data [46]. Backward probability (β) is the probability of seeing observations from time $t + 1$ to the end, with the current position at state i and at time t. There are 3 (three) stages in the backward algorithm, namely initialization, recursion, and termination as in Equations 10 until 13.

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \dots, o_T | q_t = i|\lambda) \tag{10}$$

1) Initialization:

$$\beta_t(i) = 1, \qquad 1 \le j \le N \tag{11}$$

2) Recursion:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1})\, \beta_{t+1}(j), \qquad 1 \le i \le N, 1 \le t \le T \tag{12}$$

3) Termination:

$$P(O|\lambda) = \sum_{j=1}^{N} \pi_j b_j(o_1) \beta_1(j) \tag{13}$$

where, $\beta_t(i)$ is the probability of observing the remaining sequence of observations from time $t + 1$ to T.

2.5. Baum-Welch Algorithm

The Baum-Welch algorithm is a special case of the Expectation-Maximization (EM) algorithm [47]. The Baum-Welch algorithm iteratively updates model parameters to maximize the likelihood of observing data. The initial parameter estimates are refined through iterations to obtain an optimal model for further analysis. The EM algorithm aims to find the maximum likelihood estimate of the model parameters. In the first stage (E-step), the EM algorithm estimates the expected value of the unobserved variables, and in the second stage (M-step), it finds the maximum likelihood parameter estimates [48]. To estimate the HMM, the Baum-Welch algorithm uses forward and backward algorithms to complete the data. The forward algorithm is used to calculate the probability of observing a sequence up to a specific time step and being in a particular hidden state at that step. Meanwhile, the backward algorithm is used to calculate the probability of observing the rest of the sequence based on the current hidden state [49].

This algorithm begins by initializing the HMM parameters and iterating until it reaches the convergence stage. This indicates that the changes in the model parameters are no longer significant, and the algorithm has found a local maximum of the likelihood function. The stages of the Baum-Welch algorithm process are in Equations 14 and 15.

In the Expectation Step (E-Step) stage, the observed data and current model parameters are used to calculate the expected value of the hidden state variables. Forward probability and backward probability are calculated for each state at each time step. Forward probability determines the probability of observing the output sequence up to a certain point in time. In contrast, backward probability determines the probability of observing the remaining output sequence given by the state at a particular time.

$$\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\alpha_T(q_F)} \forall\ t \text{ and } j \tag{14}$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\alpha_T(q_F)} \quad \forall \quad t, i \text{ and } j \tag{15}$$

where, $\gamma_t(j)$ is the probability that the hidden state at time t is j, $\alpha_t(j)$ is the probability of observing a sequence of observations and being in state j at time t, $\beta_t(j)$ is the probability of observing the remaining sequence, $\alpha_T(q_F)$ is the probability of observing all observations, $\xi_t(i,j)$ is the probability that the hidden state at time t is i and the hidden state and state at time $t + 1$ are j, $\alpha_t(i)$ is the probability of observing an observation at state i and time t, a_{ij} is the probability of transition from state i to state j, $b_j(o_t)$ is the probability of emission producing observation o_t at state j.

In the Maximization Step (M-Step) stage, the model parameters are re-estimated using the expected values that have been calculated in the E-Step. This stage aims to maximize the likelihood of the observed data provided by the current model. The notation of the M-Step stage process is shown in Equations 16 and 17.

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^{N} \xi_t(i,j)} \tag{16}$$

$$\hat{b}_j(v_k) = \frac{\sum_{t=1\ s.t.O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{17}$$

where, $\hat{a}_{ij}$ is the estimated transition probability from state i to state j, $\xi_t(i,j)$ is the probability of hidden state at time t is i and state and hidden state at time $t+1$ is j, N is the total number of states, T is the sequence of observations, $\hat{b}_j(v_k)$ is the estimated probability of emission resulting in observation v_k of state j, $\gamma_t(j)$ is the probability that the hidden state at time t is j, v_k specific value of the observation being considered, and o_t is the value of the observation at time t. In Equation 17, $s.t.\, O_t = v_k$ explains that this addition operation is performed on the condition that the observation O_t must be equal to v_k. The addition is performed on the expected values $\gamma_t(j)$ for each time t, where the observed observation O_t is v_k.

2.6. Paired-t Test

The t-test is a test used to see if there is a difference between 2 groups [50]. Paired t-test is a test carried out on two paired groups where each group has a different treatment [51]. The paired t-test formula for comparing the analysis results is denoted in Equation 18.

$$t_{test} = \frac{D_{\hat{y}-y}}{\frac{S_{\hat{y}-y}}{\sqrt{n}}} \tag{18}$$

where, $D_{\hat{y}-y}$ is the average difference between $\hat{y}$ and y, $S_{\hat{y}-y}$ is the standard deviation of the difference between $\hat{y}$ and y, and n is the number of samples.

Hypothesis testing is carried out on two paired groups to determine the difference between the averages of the two groups. The hypothesis used in the decision-making process is as follows.

H_0: There is no significant difference between the stock price index prediction and the actual stock price.

H_1: There is a significant difference between the stock price index prediction and the actual stock price.

Decision making from the results of the paired t-test can be seen by comparing the t_{test} with the t-table or with the p-value compared to the level of significance α. Decisions can be made by referring to the following decision-making principles: reject H_0, if the p-value $< \alpha$ or fail to reject H_0, if the p-value $> \alpha$.

2.7. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Akaike Information Criterion (AIC) is a criterion for determining the best model from the number of parameters in the model [52]. The AIC criterion was developed by Hirotugu Akaike in 1973. To measure the suitability between the estimated model and the actual model, AIC uses Kullback-Leibler Information (KLI) [53]. KLI measures how well the estimated model can imitate the actual data distribution. Maximum likelihood estimation is used as a calculation in AIC. AIC can be defined as -2 times the maximum log-likelihood added to 2 times the number of parameters, and can be formulated as in Equation 19.

$$AIC = -2\log L + 2k \tag{19}$$

where, L is the likelihood and k is the number of parameters in the candidate model.

Bayesian Information Criterion (BIC) is a criterion for determining the closest model. In considering the number of parameters in the model, BIC uses a larger penalty term than AIC [54]. In AIC, the penalty for the number of parameters is $2k$, while in BIC it is $k \log n$ with n being the number of observations. This aims to prevent overfitting by giving a larger penalty to more complex models. The formula for calculating BIC is denoted in Equation 20.

$$BIC = -2\log L + k\log n \tag{20}$$

3. Methods

3.1. Data Preparation

Data preparation is the process of collecting, cleaning, and changing the format of data to continue further analysis. The method used to collect secondary data in this research was carried out through online data sources obtained from the Yahoo Finance website, which presents stock portfolios. The data to be used in this research includes the LQ45 stock portfolio from January 2023 to March 2024. The data obtained from the Yahoo website contains the variables Date, Close, Open, High, and Low. Data cleaning is carried out to avoid data inconsistencies and to adjust the data format.

3.1.1. Data Cleaning

In the data cleaning stage, the identification and handling of missing values are carried out. In addition, if there is data with a value, data deletion will be carried out if necessary. This process is carried out to ensure that the data is ready for the following analysis process, so that the model obtained is more accurate. In this research, the data used

were 293 observations. These observations were obtained from active days in the market from January 2023 to March 2024. The training data used were observations in 2023, and the testing data used were 58, or the same as the number of observations from January to March 2024.

3.1.2. Defining Variables

The variables to be used in this study include the Close, Open, High, and Low indices. At this stage, the number of predictions made is also determined by selecting 58 final data points to be used as testing data to see the performance of the model. The maximum number of iterations used for HMM training is 1,000. This indicates how many times the algorithm will try to refine the model parameters. Furthermore, the data will be divided into parts consisting of 29 observations. In HMM terms, the data is divided into 29 windows. This size is used to divide the data into smaller parts in the training or prediction process. This allows the model to learn patterns in subsets and make predictions based on those patterns.

3.2. Model Initialization

Model initialization at this stage is done by initializing the base model to determine the optimal number of states. The Baum-Welch algorithm is used for the model initialization process. Model initialization uses the "hmmlearn" package in Python. This model iterates with an undetermined number of states to find the optimal state value. In this initialization, the state space value interval is defined to determine the best model. The state space shows the number of possible hidden states that the system can take.

3.3. Determining Optimal Number of States

At this stage, the optimal number of hidden states will be determined. The number of hidden states is a parameter that can affect the performance and ability of the model to understand hidden patterns in the observation data. To determine the optimal number of states, the metrics used are AIC and BIC. These metrics can measure how well the estimated model can imitate the actual data distribution. From the selection of the optimal number of states, a good and accurate model will be obtained in explaining the data. In evaluating model quality, AIC and BIC consider the suitability and complexity of the model. The main goal is to find a model that achieves the optimal balance for both aspects. The evaluation process is carried out by testing various numbers of hidden states in the model and then training it on the data to obtain the metric value. The AIC and BIC values are used to compare models and determine the best number of hidden states. The model with the lowest AIC or BIC value indicates the optimal model.

3.4. HMM Training

HMM training involves estimating model parameters. This can be achieved using the Baum-Welch algorithm, a variant of the Expectation-Maximization (EM) algorithm explicitly designed for HMM. The Baum-Welch algorithm iteratively updates model parameters to find the optimal model. HMM training using the Baum-Welch Algorithm consists of three stages, namely initialization, E-Step, and M-Step. In the initialization stage, the initial values of the model parameters are automatically determined, such as transition probabilities, means, and covariance. Initialization is done automatically using "hmm.GaussianHMM" which is available in the HMM packages in Python. The initial values of the parameters are determined randomly or based on historical data. In the E-step stage, the expected probability value of the hidden state is calculated using the forward algorithm and backward algorithm. The model parameter estimates are updated based on the calculated expected probabilities. In the M-Step stage, the model parameters are re-estimated using the expected values that have been calculated in the E-Step. Iterations are carried out on the three stages until they converge. The iteration process is carried out to study the patterns and structures of the data.

3.4.1. E-Step

In the E-step stage, the expected probability value of the hidden state is calculated by involving the forward algorithm and backward algorithm. The model parameter estimate is updated based on the calculated expected probability. In HMM, the forward algorithm is an algorithm used to calculate the probability of an observation given by the model. The approach using this algorithm begins with the defined alpha function, which describes the probability of a particular state at time t. The recursion process is carried out on the calculation between the alpha probability (at time t) by accumulating the probability of each previous state (at time $t-1$), the transition probability, and the emission probability. This process is carried out by iteration until time T. The recursion process produces the probability of a series of observation sequences in the HMM model.

The backward algorithm is an algorithm used to calculate the probability of an observation based on the current state. In this algorithm, the beta function is defined: this function describes the probability of an observation from time $t+1$ to the end, with the position being in state i and time t. The recursion process is carried out on the calculation

between the beta probability (at time t) by accumulating the probability of each previous state (at time $t+1$), the transition probability, and the emission probability. The process is done by iterating from time $T-1$ to 1. This algorithm is called a backward algorithm because the iteration process is moving backward. The process uses the beta value from the next step to calculate the beta value at the current step.

3.4.2. M-Step

In the M-Step, the model parameters are re-estimated using the expected values calculated in the E-Step. Iterations are performed on the three stages until they converge. The iteration process is carried out to study the patterns and structures of the data.

3.5. Prediction

The optimization process that has been carried out aims to study the pattern and structure of the data. After the process runs optimally, the log-likelihood value of the model is calculated for the last K observations. The calculated log-likelihood value provides an overview of how well the optimized model can explain the observation data. From the calculated log-likelihood value, one day in the past will be searched for which has a log-likelihood value for the previous K observations that is similar to the log-likelihood value of the latest K observations. The log-likelihood value of the latest K observations is compared with all sub-sequences of the same size in the past. This process is carried out by gradually shifting the observation window one day back and calculating the log-likelihood for each sub-sequence. The day with the closest log-likelihood to the latest K observations is identified as the reference day for prediction. The process of finding similar log-likelihood values aims to find historical patterns that are similar to current market conditions as a reference for stock predictions. Predictions are made by calculating changes in stock prices from days in the past that have similar log-likelihood values to the next day. The difference in the stock price is added to the current stock price to produce a prediction of the stock price for the next day.

Prediction is done by calculating the log-likelihood value of the model for the last K observations. K represents the number of observations or the length of the sub-sequence or group of data used for the log-likelihood calculation. The last K observations mean that there are K observations of data before time t that will be predicted. The number of observations used in this study is 29, meaning that there are 29 last observations before time t or today used as training data. In 1 (one) window, there is one group of 29 observations.

In calculating the log-likelihood value for day $t+1$, a window with 29 previous observations is used, namely from $t-28$ to t. For day $t+1$, observations from $t-27$ to $t+1$ and so on are used. The log-likelihood value explains the value used for the testing data. The log-likelihood value on the training data will then be called the past likelihood, and for the testing data, it is called the current likelihood. At the current likelihood or time t, the log-likelihood value is sought that has similarities to all past likelihood values. This process is done by gradually shifting the observation window one day back. The process of finding similar log-likelihood values aims to find historical patterns that are similar to current market conditions.

The difference between the current likelihood at time t and the past likelihood of each group is calculated as Equation 21. The group with the most minor difference will be selected to be used as a prediction.

$$j = argmin_i(|P(O_t, O_{t-1}, O_{t-2}, \dots, O_{t-K}|\lambda) - P(O_{t-i}, O_{t-i-1}, O_{t-i-2}, \dots, O_{t-i-K}|\lambda)|) \quad (21)$$

where, i: $1, 2, \dots, T/K$, j is the index of the past likelihood sub-sequence that is similar to the current likelihood, O_t is the observation at time t, and λ is the HMM parameter. Prediction is done by adding the difference between the stock price and the current stock price to produce a prediction of the stock price on the next day. This explanation can be expressed as Equation 22.

$$O_{t+1} = O_t + (O_{t-j+1} - O_{t-j}) \quad (22)$$

where, O_{t+1} is the observation at the next time, after O_t, O_t is the observation at time t, O_{t-j+1} is the observation at time $t-j+1$, and O_{t-j} is the observation at time $t-j$.

3.6. Evaluation

To validate the performance of the prediction model in prediction, the accuracy value is used as an evaluation. When the accuracy of the model is measured, the model explains important information about the quality of the prediction by measuring how close the prediction is to the facts. The accuracy metric can be defined as the percentage of accurate predictions compared to the total number of predictions. The higher the accuracy value obtained, the better the model is at predicting future conditions.

In the predictions made, the accuracy value is obtained by the paired t-test value. The paired t-test is carried out to determine whether or not there is a difference in the actual data and the predicted data generated by the HMM. From

both evaluations, each model will be produced on open, close, high and, low data. In interpreting, from the analysis that has been carried out, an explanation will be given regarding the patterns that occur, accuracy, and important findings that refer to ideas or concepts related to the research. The results of the study are linked to previous research, and further explanations are given regarding the existing findings. Then the study ends with the creation of conclusions and suggestions for further research. Briefly, the conclusion will describe the findings in the study and how this study can provide an understanding of the problems raised.

3.7. Research Flow Chart

The analysis steps carried out in this research can be seen briefly in the flow diagram in Figure 1.

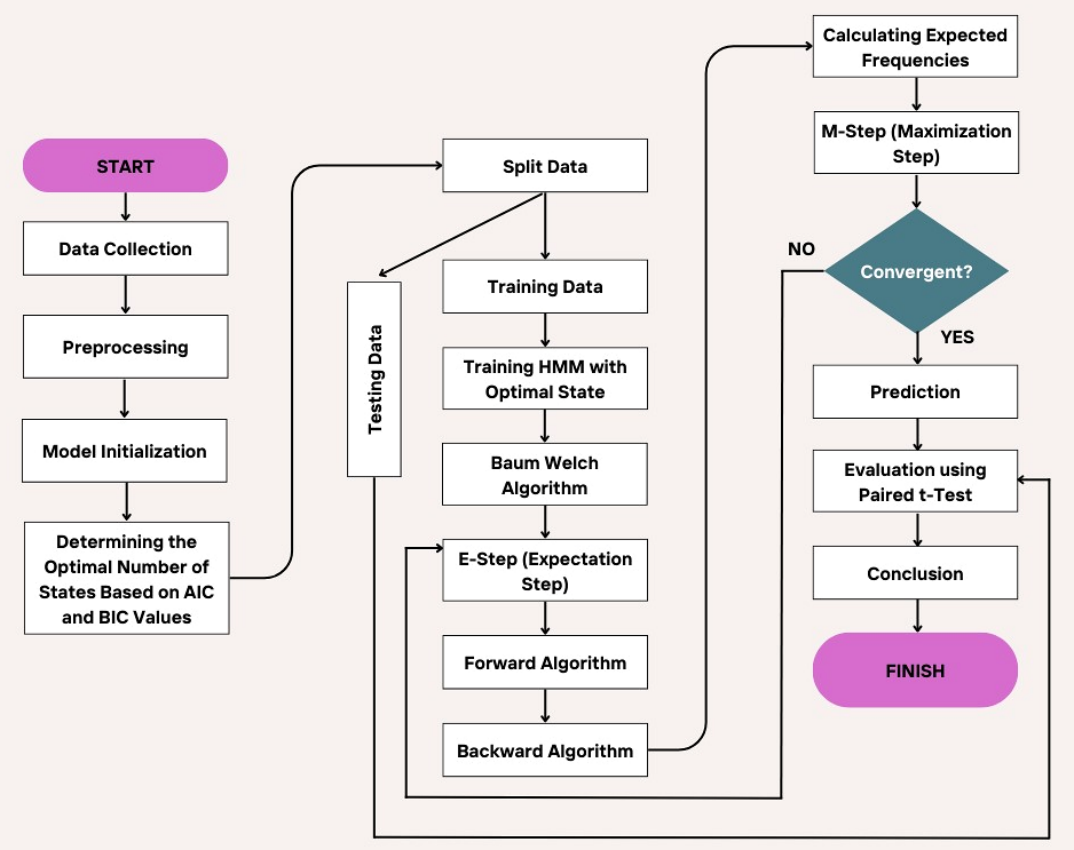


Figure 1. Research Flow Chart

4. Results and Discussion

4.1. Characteristics of the LQ45 Index

Figure 2 shows the trend of the LQ45 stock index from January 2023 to March 2024. The trend shows a fluctuating condition with a downward trend in January, March, May 2023, and February 2024. There was a drastic downward trend in November 2023. In December 2023, there was a significant increase compared to the stock index from January to November 2023. In early November 2023, there was a drastic decline in the LQ45 stock index. The LQ45 and IDX composite indices are closely related, where the movement of the IDX composite often affects the movement of the LQ45. This is because the LQ45 consists of stocks that are also part of the IDX composite. The decline in the IDX composite on November 1, 2023, reflects a decline in the LQ45, which explains the negative impact experienced by the stocks listed in the index. The decline in early November 2023 was caused by several factors that also had an impact on the LQ45. One of them is external sentiment; there is uncertainty in the global market regarding the monetary policy decision of the Federal Reserve (The Fed) in the United States, which makes investors tend to wait for the results of the Fed meeting, which is expected to maintain high interest rates. Then, there was a decline in the US stock exchange which also affected the IDX, where investors responded negatively to the minutes of the Fed meeting, which indicated the possibility of interest rates remaining high in the near future. In addition, the release of inflation data, which showed an increase in domestic goods prices affected stock price fluctuations. Overall, the weakening of the IDX in early November 2023 was caused by global uncertainty, such as international conflicts and changes in trade policies that had an impact on the domestic stock market. Internal factors, such as political uncertainty and domestic economic conditions, also affect investor confidence, which can cause fluctuations in the stock market.

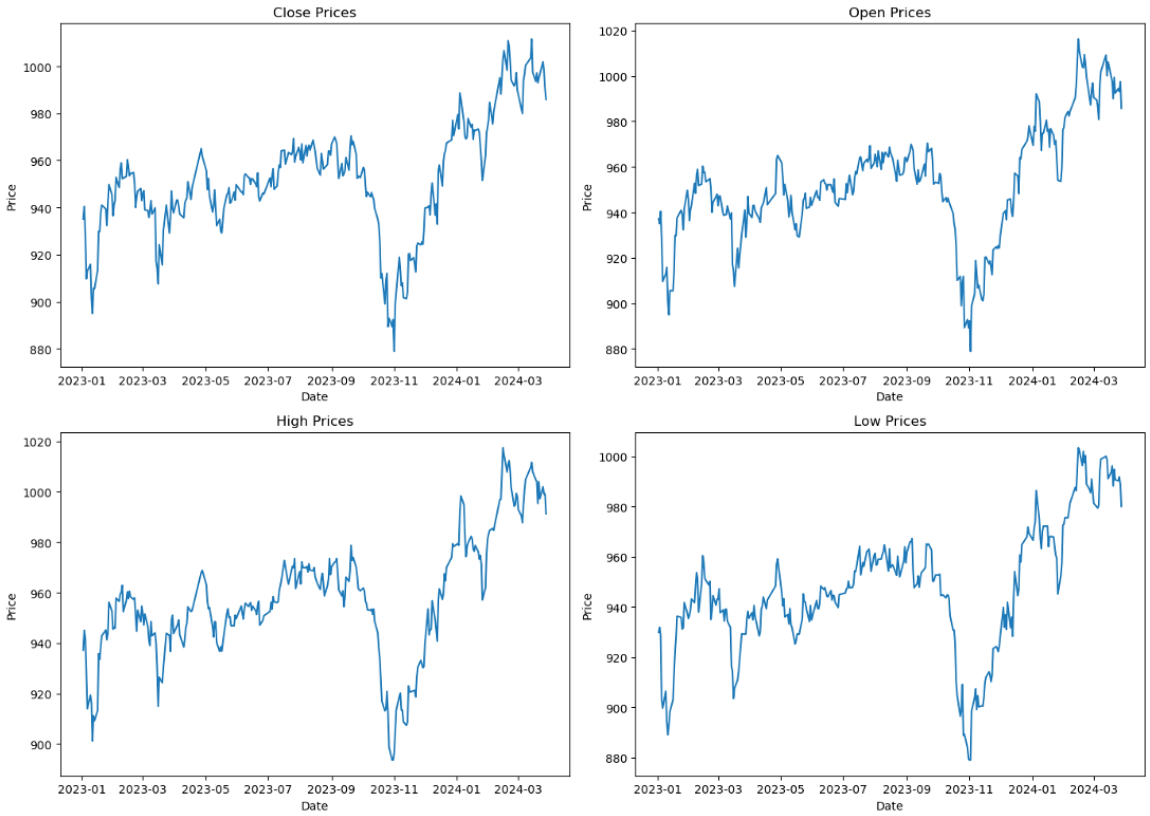


Figure 2. LQ45 Index Plot

4.2. HMM Initialization

Model initialization at this stage involves initializing the base model to determine the optimal number of states. The Baum-Welch algorithm is used for the model initialization process. Model initialization uses the "hmmlearn" package in Python. In this initialization, the state space value interval is defined to determine the best model. State space in HMM refers to the number of possible hidden states that the system can take. State space refers to a series of possible conditions that a system can have. This process involves determining the optimal number of hidden states in the system. In this stage, the state space is set between the range of 2 and 14. The HMM will explore the determination of the range of the number of hidden states. The range states that the model will be tested with various possible numbers of hidden states, ranging from 2 to 14. This aims to determine the optimal number of hidden states in capturing data patterns. The model will be evaluated using criteria such as AIC and BIC to assess how well a model with a certain number of hidden states can explain the data.

4.3. Determining the Optimal Number of States

The state shows a condition at a particular time. The use of state aims to see a picture of the transition from one state to another that can help to understand the movement of the stock index. State transition describes the changes that occur in the stock index on the stock exchange. The formation of the state is determined by looking at the difference between the stock index and the index on the previous day. Initialization of the HMM base model using the Baum-Welch algorithm is carried out to find the optimal number of states. At this stage, the optimal number of hidden states will be determined from the state space range that has been explained in the previous section. The number of hidden states is a parameter that can affect the performance and ability of the model to understand hidden patterns found in observation data. To determine the optimal number of states, the metrics used are AIC and BIC. In this research, the optimal number of states was obtained from the AIC and BIC scores, as shown in Table 1.

Table 1. Optimal Number of States

Metric	Optimal State	Score
AIC	10	8,221.9288
BIC	5	8,432.4978

Table 1 shows the number of optimal states with scores obtained from metric calculations using AIC and BIC. The results of these calculations show that the lowest score was obtained from the state space range of 2 to 14. The BIC metric can be used if the priority taken is to avoid overfitting and the model obtained is simple, while for AIC, the

resulting model is more complex. In this research, the state to be used is 5 based on the results of the BIC calculation with, consideration that the smaller number of states can facilitate the interpretation. In addition, a simpler and more general BIC allows the model to be better generalized when applied to new data.

4.4. HMM Training

HMM training involves estimating model parameters. This can be achieved using the Baum-Welch algorithm, which is part of the Expectation-Maximization (EM) algorithm. The Baum-Welch algorithm iteratively updates model parameters to find the optimal model. HMM training using the Baum-Welch Algorithm consists of three stages, namely initialization, E-Step, and M-Step. These stages produce optimal model parameters. The HMM parameters used in this modeling are denoted as $\lambda = (\pi, A, \mu, \Sigma)$, which include the initial probability π, transition matrix A, means μ, and covariance Σ.

4.4.1. Hidden State

The hidden state in HMM describes an internal state in the system that cannot be observed directly. The use of hidden state aims to see the picture of the transition from one internal state to another internal state that can help in understanding the movement of the stock index. Although the hidden state cannot be observed directly, the hidden state affects the output that can be seen. A hidden state allows the system to understand and model the dynamics of relationship patterns based on existing observations. A hidden state describes a pattern of changes that occur in the stock index on the stock exchange.

The hidden state is generated from patterns identified by the HMM model. The formation of a hidden state is based on changes in the stock index on a particular day compared to the index on another day. The model analyzes changes in sequential data to identify hidden states that reflect various market conditions. The hidden state regime describes a state of hidden patterns in the data identified by HMM. Regime reflects different situations or dynamics in a system. This refers to a state that cannot be observed directly but can be understood through analysis of changes in stock prices. Figure 3 shows the hidden state pattern obtained internally from the HMM modeling process. The pattern in state 1 shows low fluctuation with index changes ranging from 925 to 950. The pattern in state 2 shows a period of index decline ranging from 878 to 926. The pattern in state 3 shows an increase in the index, indicating a recovery after fluctuation ranging from 951 to 986. The pattern in state 4 shows a fairly high upward trend with an index value range of 973 to 1016. The pattern in state 5 shows a tendency for the index to fluctuate very low or can be said to be stable, with a movement range of 937 to 959.

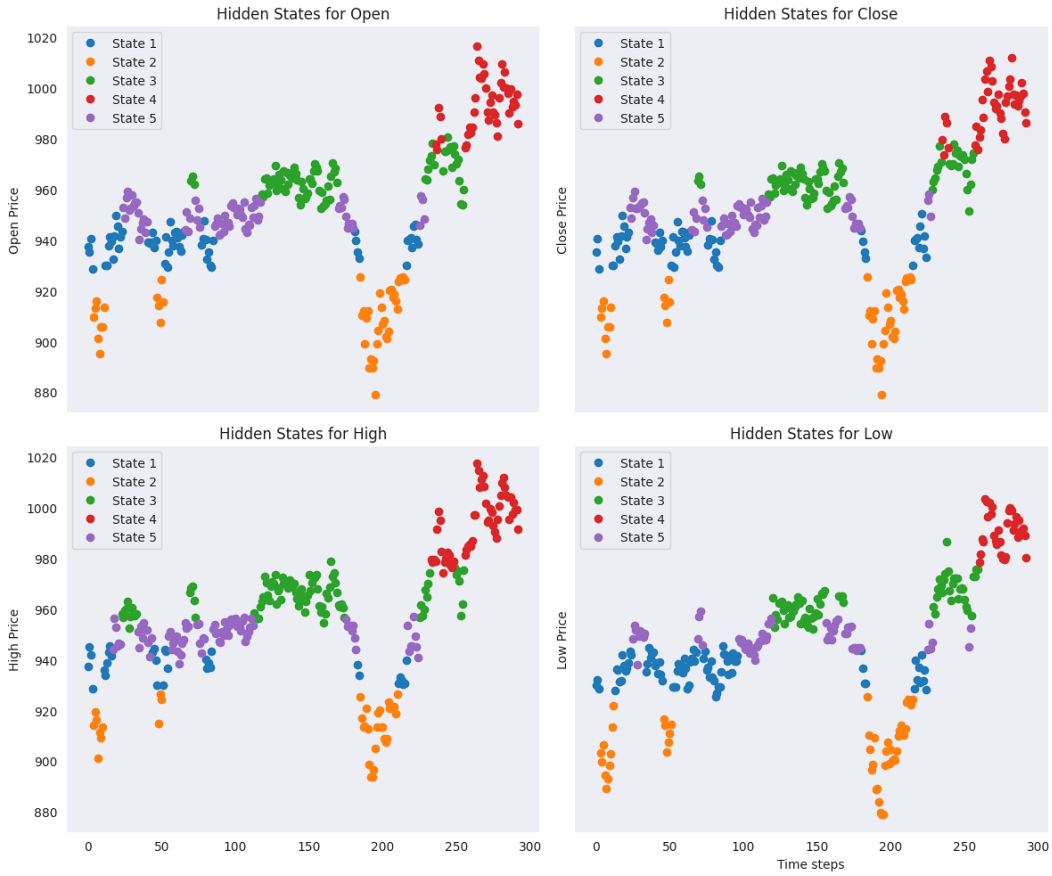


Figure 3. Hidden State Plot

4.4.2. Initial Probability

The optimal number of states obtained will then be used to determine the probability of the initial state that will be used as a model parameter. The probability of the initial state is obtained from the Baum-Welch algorithm for each variable. This probability is obtained from the average of the initial probabilities at time $t = 1$ from each observation sequence. The initial probability for Open data shows that the probability of starting a state that occurs most often is in state 3.

Initial probability for Open data:

$$\pi_{Open} = [0 \quad 0 \quad 1 \quad 0 \quad 0]$$

Initial probability for Close data:

$$\pi_{Close} = [0 \quad 0 \quad 1 \quad 0 \quad 0]$$

Initial probability for High data:

$$\pi_{High} = [0 \quad 0 \quad 1 \quad 0 \quad 0]$$

Initial probability for Low data:

$$\pi_{Low} = [0 \quad 0 \quad 1 \quad 0 \quad 0]$$

4.4.3. Transition Probability Matrix

The optimal number of states is used as a basis for viewing the transition probability. Based on the identification of the optimal number of states, the daily stock index movement pattern is categorized into five states. The movement of the stock index can be seen from the pattern of changes that occur in the daily stock index. The transition probability matrix is used to find out information about the state of stock prices in the past. Each row and column of the transition matrix shows the probability of transition from one state to another. The transition probability matrix shows the probability of a change in state in the LQ45 stock index. The transition probability matrix between states is shown in the following equation in the $N \times N$ matrix.

Transition probability matrix for Open data:

$$A_{Open} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.6862 & 0 & 0.12574 & 0 & 0.18806 \\ 0 & 0.08278 & 0.91722 & 0 & 0 \\ 0 & 0 & 0 & 0.93667 & 0.06333 \\ 0.02522 & 0.02472 & 0 & 0.03658 & 0.91348 \end{bmatrix}$$

Transition probability matrix for Open data Close:

$$A_{Close} = \begin{bmatrix} 0.81254 & 0.0809 & 0 & 0 & 0.10655 \\ 0.09805 & 0.90195 & 0 & 0 & 0 \\ 0 & 0.3359 & 0.6641 & 0 & 0 \\ 0 & 0 & 0 & 0.9371 & 0.0629 \\ 0.05066 & 0 & 0 & 0.03643 & 0.91291 \end{bmatrix}$$

Transition probability matrix for data High:

$$A_{High} = \begin{bmatrix} 0 & 0.03562 & 0 & 0.05145 & 0.91292 \\ 0.04641 & 0.95359 & 0 & 0 & 0 \\ 0 & 0.50313 & 0.49687 & 0 & 0 \\ 0 & 0 & 0 & 0.93929 & 0.06071 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Transition probability matrix for data Low:

$$A_{Low} = \begin{bmatrix} 0.70689 & 0 & 0 & 0.21154 & 0.08158 \\ 0 & 0.86859 & 0.05295 & 0 & 0.07846 \\ 0 & 0.07378 & 0.92622 & 0 & 0 \\ 0 & 0 & 0 & 0.92816 & 0.07184 \\ 0.06534 & 0.052 & 0 & 0 & 0.88266 \end{bmatrix}$$

In the transition probability matrix, the row represents the current state, and the column represents the next state. The transition matrix at the Open index for Row 1 shows: from state 1, there is a probability of 1 (or 100%) of moving to state 2. There is no probability of staying in state 1 or moving to states 3, 4, or 5. Row 2 shows: from state 2, there is a probability of 0.6862 (68.62%) of moving to state 1, 0.12574 (12.574%) of moving to state 3, and 0.18806 (18.806%) of moving to state 5, there is no probability of staying in state 2 or moving to state 4. Row 3 shows: from state 3, there is a probability of 0.08278 (8.278%) to move to state 2 and 0.91722 (91.722%) to stay

in state 3, there is no probability of moving to states 1, 4, or 5. Row 4 shows: from state 4, there is a probability of 0.93667 (93.667%) to stay in state 4 and 0.06333 (6.333%) to move to state 5, there is no probability of moving to states 1, 2, or 3. Row 5 shows: from state 5, there is a probability of 0.02522 (2.522%) to move to state 1, 0.02472 (2.472%) to move to state 2, 0.03658 (3.658%) to move to state 4, and 0.91348 (91.348%) to remain in state 5, there is no probability of moving to state 3. The transition matrices for Close index, High index, and Low index also have the same interpretation.

4.4.4. Means

The average parameter or mean describes the center of data distribution. The average value shows the center of data distribution for each state. In this research, 5 states were used. For each variable (Open, Close, High, and Low), the average parameter provides information about the center of data distribution in each state. For each state in each variable, the average obtained between states describes the difference in data distribution for each state. This fairly varied difference indicates that the center of data distribution is good. In the means, each row represents each existing state. The mean value for each state in the variable is shown as follows.

Means for Open Data:

$$\mu_{Open} = \begin{bmatrix} 953.23985 \\ 952.92526 \\ 963.92852 \\ 911.59925 \\ 941.39226 \end{bmatrix}$$

Means for Close data:

$$\mu_{Close} = \begin{bmatrix} 953.13910 \\ 963.39843 \\ 971.98931 \\ 911.46044 \\ 941.28141 \end{bmatrix}$$

Means for High data:

$$\mu_{High} = \begin{bmatrix} 950.11034 \\ 966.52756 \\ 978.83148 \\ 918.67670 \\ 949.41037 \end{bmatrix}$$

Means for Low data:

$$\mu_{Low} = \begin{bmatrix} 926.30493 \\ 946.95143 \\ 959.70034 \\ 902.26508 \\ 936.52212 \end{bmatrix}$$

4.4.5. Covariance Matrix

The Covariance Matrix describes the spread that occurs in observations around the mean in each state. Each element in the Covariance Matrix represents the variance of the distribution for each particular state. A small variance value indicates that the data is centered around the mean and has more minor fluctuations. At the same time, a considerable variance value indicates that the data has a wider spread around the mean with larger fluctuations. In the matrix, each row represents the covariance matrix value for each state. The covariance matrix for each state and for each variable is denoted as follows.

Covariance matrix for Open data:

$$\Sigma_{Open} = \begin{bmatrix} 13.54301 \\ 8.39134 \\ 20.53987 \\ 145.42967 \\ 25.04813 \end{bmatrix}$$

Covariance matrix for Close data:

$$\Sigma_{Close} = \begin{bmatrix} 10.15253 \\ 13.94872 \\ 14.17294 \\ 143.9587 \\ 27.6355 \end{bmatrix}$$

Covariance matrix for High data:

$$\Sigma_{High} = \begin{bmatrix} 40.01064 \\ 24.67967 \\ 0.32425 \\ 128.56984 \\ 36.72947 \end{bmatrix}$$

Covariance matrix for High data Low:

$$\Sigma_{Low} = \begin{bmatrix} 8.11048 \\ 15.14652 \\ 22.66807 \\ 110.68889 \\ 20.1324 \end{bmatrix}$$

4.5. Prediction

The prediction of the LQ45 stock index using the HMM method involves calculating the likelihood value. The log-likelihood of the model is used to predict the stock index in the future. The value is determined by comparing the log-likelihood of the previous observation with the log-likelihood of the current observation. This process involves comparing the log-likelihood of the previous observation with the log-likelihood of the new observation. The log-likelihood of the new observation is then identified as a reference for prediction. This process is used to estimate the historical probability of a future event, which is then used as a reference for predicting future prices.

Figure 4 shows a comparison between the results of the stock index prediction and the actual stock data. The blue line on the graph represents the movement trend of the prediction results using the HMM method, while the red dotted line shows the movement trend for the actual stock index. Figure 4 shows a comparison graph for the open, close, high, and low variables. The lines produced by the prediction model show a tendency to follow the same pattern as the actual data line. This indicates that the model prediction has a pretty good match with the actual data. However, certain periods show prediction results that do not match the actual data observed. This may indicate that there are several factors that are not well captured by the prediction model. The goodness of the model can be measured by the metrics used in this research.

Figure 4. Comparison of Actual Data with Predictions for Open, Close, High, and Low Price

4.6. Evaluation

The paired t-test value obtains the accuracy value. The paired t-test is conducted to determine whether there is a difference between the actual data and the predicted data generated by the HMM. From both evaluations, each model will be generated on Open, Close, High, and Low data. The t-test is used to see the accuracy obtained by the model. The test calculates the difference between the predicted value and the actual value. The metric is used to calculate the extent to which the predicted value differs from the actual value observed.

The paired t-test performed on the data produces a p-value as shown in Table 2. The paired t-test shows that the p-value for Open and Close is greater than the specified alpha of 5%. The p-value is 0.51747 for the Open index, 0.28551 for Close, 0.00 for High, and 0.00 for Low. From the results of the test, it can be concluded that there is no significant difference in the predicted data when compared to the actual data for Open and Close. There is a difference between the actual data and the predicted data for the High and Low variables. This shows that the prediction results for the Open and Close variables using the HMM method show good results.

Table 2. Paired t-test calculation results

	p-value	Decision	Conclusion
Open	0.51747	Failed to Reject H0	There is no difference between actual and predicted data
Close	0.28551	Failed to Reject H0	There is no difference between actual and predicted data
High	0.00	Reject H0	There is a difference between actual and predicted data
Low	0.00	Reject H0	There is a difference between actual and predicted data

Table 3 shows the Root Mean Squared Error (RMSE) for different stock prices. Notably, the Open and Close prices exhibit the lowest RMSE values at 4.7503 and 7.8992, respectively. This suggests that the model demonstrates a comparatively higher accuracy in predicting the Open and Close prices when compared to the High and Low prices. The lower RMSE values for Open and Close may be attributed to factors such as increased market activity and liquidity during these periods, which can lead to more predictable price movements. Further investigation into the underlying patterns and features associated with Open and Close prices could lead to even more refined predictive models.

Table 3. RMSE for Open, Close, High, and Low-Price Stock

Price	RMSE
Open	4.7503
Close	7.8992
High	11.7197
Low	16.7362

In the research by Hansun & Young, presented RMSE for BBCA, BBNI, BBRI, BBTN, BMRI, BTPS [55]. All stock codes are considerably high, exceeding 200 in each case, suggesting a notable degree of prediction error across the board. The research conducted by Wibisono et al. used LSTM and CNN to predict the price of BBCA, BMRI, BBRI, BBNI, BBTN, and BRIS stocks [56]. The results showed that all stock indexes had an RMSE value of more than 45. Research conducted by Adiatmaja & Indraswari showed that the RMSE values for BBCA and ASII are 91.4 and 80.1, respectively, using the GRU method [57]. Previous studies have shown that the RMSE obtained is greater than the RMSE in this study. Therefore, it can be concluded that the Open and Close prices have a significant influence and provide a better model than the Low and High prices.

5. Conclusion

The confluence of stock market dynamics and predictive modeling techniques necessitates a nuanced approach to forecasting the LQ45 index on the Indonesia Stock Exchange (IDX). As demonstrated by the studies reviewed, the LQ45 index, comprising 45 stocks with significant market capitalization and liquidity, serves as a key indicator of Indonesia's economic health. However, the inherent complexity of stock price movements, influenced by a myriad of factors ranging from macroeconomic indicators to investor sentiment, poses a significant challenge to accurate prediction. Traditional methods often fall short due to their inability to capture the temporal dependencies and hidden patterns within stock market data. This research, employing a Hidden Markov Model (HMM), marks a significant step forward by explicitly addressing these limitations. HMMs, with its capacity to model unobservable states and transitions, offer a more robust framework for understanding the stochastic processes governing stock prices. The

blend of the HMM framework, as discussed in the uploaded document, with findings from other studies, such as those using ARIMA, Random Forest, and regression models, underscores the importance of integrating multiple perspectives to create a more comprehensive understanding of LQ45 index behavior. Further research should explore hybrid models that leverage the strengths of various techniques to improve predictive accuracy and inform investment strategies.

The practical implications of accurate LQ45 index prediction are far-reaching, extending from individual investors seeking data-driven portfolio management strategies to policymakers aiming to foster a stable and predictable economic environment. The ability to anticipate stock price movements allows investors to make informed decisions, allocate capital efficiently, and mitigate risk. Moreover, a stable stock market, underpinned by reliable forecasting models, can attract both domestic and foreign investment, fueling business expansion, job creation, and innovation. The success of the HMM in capturing the probabilistic patterns within the LQ45 index, as evidenced by the paired t-test results in this research, highlights the potential of this method for practical application. However, it is essential to acknowledge the limitations of any predictive model. External factors, such as geopolitical events and sudden shifts in investor sentiment, can introduce unforeseen volatility into the stock market. Therefore, ongoing research is crucial to refining existing models, exploring new techniques, and developing comprehensive risk management strategies. Future research should focus on integrating external economic indicators, refining the determination of optimal hidden states, and exploring the application of HMMs in combination with other advanced analytical tools to enhance predictive accuracy and provide valuable insights for stakeholders in the Indonesian stock market.

6. Declarations

6.1. Author Contributions

Conceptualization, A.R., F.A.Z., D.R., and F.O.; methodology, A.R., F.A.Z., D.R., F.O., and M.F.A.H.; writing—original draft preparation, A.R., F.A.Z., D.R., and M.F.A.H.; writing—review and editing, A.R., F.A.Z., D.R., F.O., and M.F.A.H. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

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

6.3. Funding

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

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that there are no conflicts of interest concerning the publication of this manuscript. Furthermore, all ethical considerations, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7. References

[1] Verma, P., Dumka, A., Bhardwaj, A., Ashok, A., Kestwal, M. C., & Kumar, P. (2021). A Statistical Analysis of Impact of COVID19 on the Global Economy and Stock Index Returns. SN Computer Science, 2(1), 1-13. doi:10.1007/s42979-020-00410-w.

[2] Gavrilakis, N., & Floros, C. (2023). ESG performance, herding behavior and stock market returns: evidence from Europe. Operational Research, 23(1), 3. doi:10.1007/s12351-023-00745-1.

[3] Goldstein, I. (2023). Information in Financial Markets and Its Real Effects. Review of Finance, 27(1), 1–32. doi:10.1093/rof/rfac052.

[4] Rantini, D., Fakhruzzaman, M. N., Ningrum, R. A., Othman, F., Choir, A. S., Ramadan, A., Alya, N. A., Putri, E. R., & Pratama, M. A. (2024). Modeling the Percentage of NEET in Indonesia with Spatial Cauchy Regression through the Bayesian Analysis Approach. IAENG International Journal of Applied Mathematics, 54(7), 1288–1295.

[5] Alya, N. A., Almaulidiyah, Q., Farouk, B. R., Rantini, D., Ramadan, A., & Othman, F. (2024). Comparison of Geographically Weighted Regression (GWR) and Mixed Geographically Weighted Regression (MGWR) Models on the Poverty Levels in Central Java in 2023. IAENG International Journal of Applied Mathematics, 54(12), 2746–2757.

[6] Simahatie, M., & Inuzula, L. (2022). Effect of Fundamental Factors on Stock Return (On LQ-45 Index Companies in Indonesia Stock Exchange 2017-2020). Journal of Accounting Research, Utility Finance and Digital Assets, 1(1), 11–18. doi:10.54443/jaruda.v1i1.2.

[7] Wen, X., & Li, W. (2023). Time Series Prediction Based on LSTM-Attention-LSTM Model. IEEE Access, 11, 48322–48331. doi:10.1109/ACCESS.2023.3276628.

[8] Jallow, M. A., Weke, P., Nafiu, L. A., & Ogutu, C. (2021). Application of a Discrete-Time Semi-Markov Model to the Stochastic Forecasting of Capital Assets as Stock. Far East Journal of Theoretical Statistics, 63(1), 1–18. doi:10.17654/TS063010001.

[9] Sarsour, W. M., & Sabri, S. R. M. (2020). Forecasting the long-run behavior of the stock price of some selected companies in the Malaysian construction sector: A Markov chain approach. International Journal of Mathematical, Engineering and Management Sciences, 5(2), 296–308. doi:10.33889/IJMEMS.2020.5.2.024.

[10] Ayo, A. S., & Uwabor, E. S. (2021). Markovian Approach to Stock Price Modelling in the Nigerian Oil and Gas Sector. Central Bank of Nigeria Journal of Applied Statistics, 12(1), 23–43. doi:10.33429/cjas.12121.2/6.

[11] Lakshmi, G., & Jyothi, M. (2020). Application of Markov process for prediction of stock market performance. International Journal of Recent Technology and Engineering (IJRTE), 8(6), 1516–1519. doi:10.35940/ijrte.f7784.038620.

[12] Zakaria, N. N., Othman, M., Sokkalingam, R., Daud, H., Abdullah, L., & Kadir, E. A. (2019). Markov chain model development for forecasting air pollution index of miri, Sarawak. Sustainability (Switzerland), 11(19), 5190. doi:10.3390/su11195190.

[13] Rao Padi, T., Farooq Dar, G., & Rekha, S. (2022). Stock Market Trend Analysisand Prediction using Markov Chain Approach in the Context of Indian Stock Market. IOSR Journal of Mathematics, 18(4), 40–48. www.iosrjournals.org

[14] Yang, C. H., Lee, B., Lee, Y. I., Chung, Y. F., & Lin, Y. Da. (2025). An autoencoder-based arithmetic optimization clustering algorithm to enhance principal component analysis to study the relations between industrial market stock indices in real estate. Expert Systems with Applications, 266, 126165. doi:10.1016/j.eswa.2024.126165.

[15] Adlakha, N., Ridhima, & Katal, A. (2021). Real Time Stock Market Analysis. 2021 International Conference on System, Computation, Automation and Networking, ICSCAN 2021, 1–5. doi:10.1109/ICSCAN53069.2021.9526506.

[16] Ayala, J., García-Torres, M., Noguera, J. L. V., Gómez-Vela, F., & Divina, F. (2021). Technical analysis strategy optimization using a machine learning approach in stock market indices [Formula presented]. Knowledge-Based Systems, 225, 107119. doi:10.1016/j.knosys.2021.107119.

[17] Syukur, A., & Istiawan, D. (2020). Prediction of LQ45 Index in Indonesia Stock Exchange: A Comparative Study of Machine Learning Techniques. International Journal of Intelligent Engineering and Systems, 14(1), 453–463. doi:10.22266/IJIES2021.0228.42.

[18] Zhang, J., Wang, J., & Kongruang, C. (2025). Empirical Analysis of Executive Capital, Innovation, and Risk-Taking in A-Share Tech Firms. Emerging Science Journal, 9(3), 1610–1631. doi:10.28991/ESJ-2025-09-03-024.

[19] Anand, S. K., & Kumar, S. (2023). Experimental Comparisons of Clustering Approaches for Data Representation. ACM Computing Surveys, 55(3), 1–33. doi:10.1145/3490384.

[20] Hasan, B. M. S., & Abdulazeez, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining, 2(1), 20–30. doi:10.30880/jscdm.2021.02.01.003.

[21] Ahmed, D. M., Hassan, M. M., & Mstafa, R. J. (2022). A Review on Deep Sequential Models for Forecasting Time Series Data. Applied Computational Intelligence and Soft Computing, 6596397. doi:10.1155/2022/6596397.

[22] Bemporad, A. (2023). A Piecewise Linear Regression and Classification Algorithm with Application to Learning and Model Predictive Control of Hybrid Systems. IEEE Transactions on Automatic Control, 68(6), 3194–3209. doi:10.1109/TAC.2022.3183036.

[23] Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. Babylonian Journal of Machine Learning, 2024, 69–79. doi:10.58496/BJML/2024/007.

[24] Glennie, R., Adam, T., Leos-Barajas, V., Michelot, T., Photopoulou, T., & McClintock, B. T. (2023). Hidden Markov models: Pitfalls and opportunities in ecology. Methods in Ecology and Evolution, 14(1), 43–56. doi:10.1111/2041-210X.13801.

[25] Indrayana, K., Hamzah, D., & Aswan, A. (2020). The Effect of Net Income, Equity, Cash Dividend, Average Price and Volume to Corporate Market Capitalization Stocks in LQ45 Index of Indonesia Stock Exchange Period 2008 – 2018. Hasanuddin Journal of Applied Business and Entrepreneurship, 3(4), 117–135. doi:10.26487/hjabe.v3i4.382.

[26] Sivakumar, G. (2025). HMM-LSTM fusion model for economic forecasting. arXiv Preprint, arXiv:2501.02002. doi:10.48550/arXiv.2501.02002.

[27] Khazaeiathar, M., & Schmalz, B. (2025). Addressing Volatility and Nonlinearity in Discharge Modeling: ARIMA-iGARCH for Short-Term Hydrological Time Series Simulation. Hydrology, 12(8), 197. doi:10.3390/hydrology12080197.

[28] Solihin, I., Sugiarto, S., Ugut, G. S. S., & Hulu, E. (2022). LQ45 Stock Index Abnormal Return Reaction to the Covid-19 Pandemic: the Even Study Methodology. Indonesian Interdisciplinary Journal of Sharia Economics, 5(1), 342–355. doi:10.31538/iijse.v5i1.2051.

[29] Qin, X. (2025). Application of Deep Learning for Stock Prediction within the Framework of Portfolio Optimization in Quantitative Trading. HighTech and Innovation Journal, 6(2), 598–614. doi:10.28991/HIJ-2025-06-02-016.

[30] Pradana, B. L. (2025). Time Series Forecasting of LQ45 Stock Index Using ARIMA: Insights and Implications. Review of Management, Accounting and Tourism Studies, 1(1), 27-40.

[31] Hidayat, A., & Suhendri, A. P. P. (2025). Comparative Analysis of Machine Learning Algorithms for Predicting LQ45 Stock Index Prices. bit-Tech, 8(1), 1099-1108. doi:10.32877/bt.v8i1.2853.

[32] Reyzan S.A, A., & Abdurrohim, A. (2025). The Effect of Current Ratio and Debt to Equity Ratio on Company Value Mediated by Return on Assets in the Mining Sector Listed on the LQ45 Index for the Period 2019-2023. Formosa Journal of Applied Sciences, 4(8), 2723–2742. doi:10.55927/fjas.v4i8.307.

[33] Arief, F., & Hidayat, R. A. (2025). Analysis of the Influence of Bitcoin and Macroeconomic Fundamentals on the LQ45 Index for the 2018-2022 Period. International Journal of Business and Applied Economics, 4(3), 1059–1076. doi:10.55927/ijbae.v4i3.105.

[34] Astuti, T. H., & Pramuditha Dwi Anggraini. (2025). Risk Analysis of Single Stocks and Portfolios in Lq45 Index. Journal of Economic and Economic Policy, 2(4), 422–430. doi:10.61796/ijecep.v2i4.82.

[35] El-Awady, A., & Ponnambalam, K. (2021). Integration of simulation and Markov Chains to support Bayesian Networks for probabilistic failure analysis of complex systems. Reliability Engineering and System Safety, 211, 107511. doi:10.1016/j.ress.2021.107511.

[36] Liao, T. F., Zhao, W., Hackney, J., Tang, H., Xu, H., & He, J. Sequence analysis: Its past, present, and future. Social Science Research, 107, 102772.

[37] Stewart, W. J. (2021). Introduction to the Numerical Solution of Markov Chains. Princeton University Press, Princeton, New Jersey, United States. doi:10.2307/j.ctv182jsw5.

[38] Zhao, C., Hu, P., Liu, X., Lan, X., & Zhang, H. (2023). Stock Market Analysis Using Time Series Relational Models for Stock Price Prediction. Mathematics, 11(5), 1130. doi:10.3390/math11051130.

[39] Musaev, A., Makshanov, A., & Grigoriev, D. (2023). The Genesis of Uncertainty: Structural Analysis of Stochastic Chaos in Finance Markets. Complexity, 1302220. doi:10.1155/2023/1302220.

[40] Seabrook, E., & Wiskott, L. (2023). A Tutorial on the Spectral Theory of Markov Chains. Neural Computation, 35(11), 1713–1796. doi:10.1162/neco_a_01611.

[41] Palupi, I., Wahyudi, B. A., & Putra, A. P. (2021). Implementation of Hidden Markov Model (HMM) to Predict Financial Market Regime. 2021 9th International Conference on Information and Communication Technology, ICoICT 2021, 639–644. doi:10.1109/ICoICT52021.2021.9527459.

[42] Eddy, S. R. (1996). Hidden Markov models. Current opinion in structural biology, 6(3), 361-365. doi:10.1016/S0959-440X(96)80056-X.

[43] Boyko, J. D., & Beaulieu, J. M. (2021). Generalized hidden Markov models for phylogenetic comparative datasets. Methods in Ecology and Evolution, 12(3), 468–478. doi:10.1111/2041-210X.13534.

[44] Trichilli, Y., Boujelbène Abbes, M., & Masmoudi, A. (2020). Predicting the effect of Googling investor sentiment on Islamic stock market returns: A five-state hidden Markov model. International Journal of Islamic and Middle Eastern Finance and Management, 13(2), 165–193. doi:10.1108/IMEFM-07-2018-0218.

[45] Bhar, R., & Hamori, S. (2004). Linking Inflation and Inflation Uncertainty: Hidden Markov models: applications to financial economics. Springer, Boston, United States. doi:10.1007/1-4020-7940-0_5.

[46] Boutazart, Y., Ezzine, A., & Satori, H. (2024). A Rich and Balanced MSA Corpus for HMM-Based Consonant-Vowel Segmentation. 2024 3rd International Conference on Embedded Systems and Artificial Intelligence, ESAI 2024, 1–14. doi:10.1109/ESAI62891.2024.10913508.

[47] Pereira, D., Nunes, C., & Rodrigues, R. (2024). A new algorithm for inference in HMM's with lower span complexity. Computational Statistics and Data Analysis, 195, 107955. doi:10.1016/j.csda.2024.107955.

[48] Zhang, Y. M., Wang, H., Wan, H. P., Mao, J. X., & Xu, Y. C. (2021). Anomaly detection of structural health monitoring data using the maximum likelihood estimation-based Bayesian dynamic linear model. Structural Health Monitoring, 20(6), 2936–2952. doi:10.1177/1475921720977020.

[49] Singh, R., Zhang, Q., & Chen, Y. (2022). Learning hidden Markov models from aggregate observations. Automatica, 137, 110100. doi:10.1016/j.automatica.2021.110100.

[50] Lee, S. W. (2022). Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee. Life Cycle, 2. doi:10.54724/lc.2022.e1.

[51] Afifah, S., Mudzakir, A., & Nandiyanto, A. B. D. (2022). How to Calculate Paired Sample t-Test using SPSS Software: From Step-by-Step Processing for Users to the Practical Examples in the Analysis of the Effect of Application Anti-Fire Bamboo Teaching Materials on Student Learning Outcomes. Indonesian Journal of Teaching in Science, 2(1), 81–92. doi:10.17509/ijotis.v2i1.45895.

[52] Sutherland, C., Hare, D., Johnson, P. J., Linden, D. W., Montgomery, R. A., & Droge, E. (2023). Practical advice on variable selection and reporting using Akaike information criterion. Proceedings of the Royal Society B: Biological Sciences, 290(2007), 20231261. doi:10.1098/rspb.2023.1261.

[53] Diaz Rubio, G. A. (2022). Model selection and the vectorial misspecification-resistant information criterion in multivariate time series. Dissertation thesis, Alma Mater Studiorum Università di Bologna, Bologna, Italy. doi:10.48676/unibo/amsdottorato/10488.

[54] Zhang, J., Yang, Y., & Ding, J. (2023). Information criteria for model selection. Wiley Interdisciplinary Reviews: Computational Statistics, 15(5), 1607. doi:10.1002/wics.1607.

[55] Hansun, S., & Young, J. C. (2021). Predicting LQ45 financial sector indices using RNN-LSTM. Journal of Big Data, 8(1), 104. doi:10.1186/s40537-021-00495-x.

[56] Wibisono, S. S., Leviana, K., Muliadiredja, T. S., Margaretha, H., Ferdinand, F. V., & Utomo, J. (2024). Stock Price Prediction on LQ45's Banking Sector Using Long-Short-Term Memory and Convolutional Neural Network. Proceedings - 2024 2nd International Conference on Technology Innovation and Its Applications, ICTIIA 2024, 1–5. doi:10.1109/ICTIIA61827.2024.10761865.

[57] Adiatmaja, G., & Indraswari, R. (2024). Comparative Analysis of LSTM and GRU Models for Indonesian Stock Price Prediction with Integrated Technical and Fundamental Indicators. 2024 International Conference on Information Technology Systems and Innovation, ICITSI 2024 - Proceedings, 206–211. doi:10.1109/ICITSI65188.2024.10929370.