Analysis and Prediction of the Occurrence of an Earthquake Using ARIMA and Statistical Tests

Rabbani Nur Kumoro[*1]**, Audrey Shafira Fattima**[2]**, William Hilmy Susatyo**[3]**,**
Dzikri Rahadian Fudholi[4]
[1,2]Program Studi S1 Ilmu Komputer FMIPA UGM, Yogyakarta, Indonesia
[2]Departemen Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta, Indonesia
e-mail: [*1]rabbani.nur.kumoro@mail.ugm.ac.id, [2]audrey.shafira1503@mail.ugm.ac.id,
[3]william.hilmy.susatyo@mail.ugm.ac.id, [4]dzikri.r.f@ugm.ac.id

Abstrak

Gempa bumi menghadirkan risiko yang signifikan terhadap keselamatan manusia dan infrastruktur, sehingga menekankan perlunya model prediksi yang tepat untuk meminimalisir dampak buruknya. Penelitian ini bertujuan untuk mengatasi tantangan dalam memprediksi waktu terjadinya gempa bumi secara akurat dengan memanfaatkan dataset LANL Earthquake, yang terdiri dari sinyal seismik model laboratorium yang meniru patahan tektonik. Dalam penelitian ini, kami menggunakan model ARIMA dan membandingkannya dengan Linear Regression untuk memprediksi kejadian gempa bumi. Temuan kami menunjukkan bahwa model ARIMA (1,1,1) mengungguli model-model lainnya, dengan MAE terendah sebesar 0,110628. Validitas model telah dikonfirmasi melalui uji Ljung-Box dan Jarque-Bera, yang memverifikasi tidak adanya autokorelasi dan distribusi normal dari residual.

Katakunci *— Earth Sciences, Forecasting, Machine Learning, Seismic Signals, Time Series*

Abstract

Earthquakes present significant risks to both human safety and infrastructure, emphasizing the need for precise prediction models to minimize their adverse effects. This study seeks to tackle the challenge of accurately forecasting the occurrence time of earthquakes by utilizing the LANL Earthquake dataset, which comprises seismic signals from a laboratory model emulating tectonic faults. In this study, we employed the ARIMA model and compared it with Linear Regression to predict earthquake occurrences. Our findings demonstrate that the ARIMA (1,1,1) model surpasses other models, achieving the lowest MAE of 0.110628. The validity of the model's assumptions is confirmed through the Ljung-Box and Jarque-Bera tests, which verify the absence of autocorrelation and the normal distribution of residuals, respectively.

Keywords *— Earth Sciences, Forecasting, Machine Learning, Seismic Signals, Time Series.*

1. INTRODUCTION

Earth sciences research primarily focuses on forecasting the occurrence and severity of earthquakes. Earthquakes are natural events resulting from movement or vibration within the earth's layers, and they can occur globally, often caused by tectonic activities such as plate movements or volcanic events. Anticipating the timing of earthquakes is crucial due to the potentially devastating impact of these disasters. For instance, the earthquake that struck Turkey and Syria in February 2023, resulted in a minimum of 41.232 casualties and estimated material losses of up to US$1 billion [1].

In predicting the timing of earthquakes, one of the factors used is information derived from seismic signals through time series analysis. Seismic signals are vibrations or waves that travel through the Earth's interior due to geological activities like earthquakes, explosions, or magma movements within volcanoes. These signals are crucial in seismology for studying the earth's nature and structure, as well as for detecting and monitoring seismic activity that may endanger human safety and the environment. Various types of seismic signals, such as primary waves (P-waves), secondary waves (S-waves), and surface waves, are commonly utilized in seismological research [2].

Our study comprises several key stages. We begin by working with the Los Alamos National Laboratory (LANL) Earthquake dataset, which stems from an experiment conducted on rocks in a laboratory earthquake model mirroring the tectonic faults in the earth's layers. This dataset exhibits periodic and realistic behavior, signifying irregular earthquake occurrences. Data pre-processing involves identifying missing values, addressing data skewness, and organizing the raw data for analysis. Exploratory Data Analysis (EDA) is then conducted to visualize and summarize the dataset's main characteristics, reveal correlations and outliers, and assess data stationarity, which informs the modeling process. Next, feature engineering is undertaken to create new features that enhance the predictive power of the models, encompassing segmentation, feature manipulation, and feature scaling.

Moving on to the modeling stage, we employ the AutoRegressive Integrated Moving Average (ARIMA) stochastic modeling approach and Linear Regression to the pre-processed and engineered dataset to predict the timing of earthquakes. The primary metric used to assess our models' performance is the Mean Absolute Error (MAE), which quantifies the average magnitude of errors between the predicted and actual values without considering their direction.

2. METHODS

The research methodology for predicting earthquakes involves several stages illustrated in Figure 1. It commences with the acquisition of raw data from the LANL Earthquake dataset, which features seismic signals from a laboratory model simulating tectonic faults. This raw data is initially pre-processed to remove noise, normalize, and partition it into training and testing sets. Cleaning involves removing any noise or irrelevant information, normalization ensures the data is scaled appropriately, and the data is then split into 80% for training and 20% for testing. The training data is then subjected to the Augmented Dickey-Fuller (ADF) test to check for stationarity.

Following the stationarity check, different ARIMA models are developed to capture the underlying patterns in the seismic data. To validate the model's assumptions, the Ljung-Box test is applied to verify the absence of autocorrelation in the residuals, and the Jarque-Bera test is used to confirm the normality of the residuals. These statistical tests are essential for ensuring the reliability of the ARIMA models. The best-performing ARIMA model is chosen as the final model, and its performance is assessed on the testing set using the MAE to ensure accurate earthquake predictions.

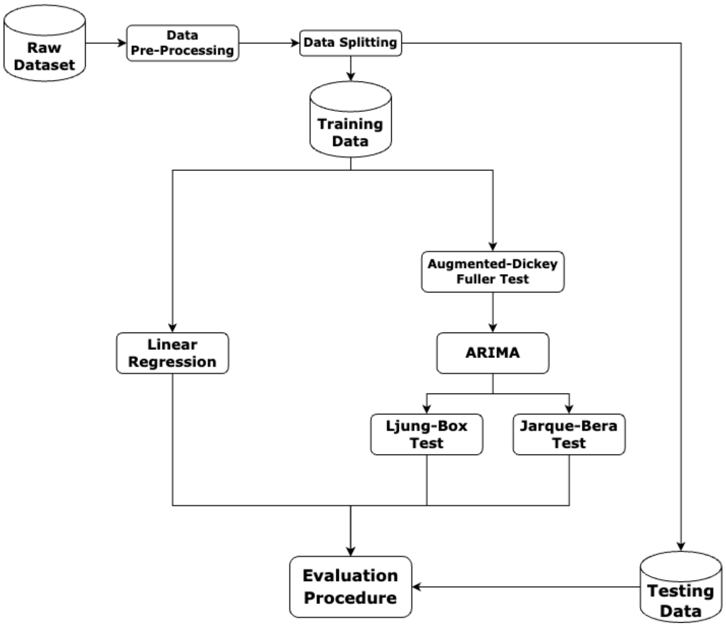

Figure 1. Flowchart Diagram of the Research Methodology

2.1 Dataset

The dataset utilized in this research was sourced from the Los Alamos National Laboratory (LANL) via the Kaggle platform [3]. This extensive dataset comprises a total of 577.060.473 rows. It was generated synthetically during an experimental study on rocks conducted by LANL using the "classic lab earthquake model," a device designed to replicate the loading and failure cycles of a tectonic fault, thereby simulating the processes involved in natural earthquakes [4].

The laboratory model mimics tectonic faults within the earth's layers and while it represents a simplified version of an actual earthquake, it is purported to encompass most of the physical characteristics of a real earthquake. The data exhibits periodic, realistic behavior, encompassing irregularly occurring earthquakes. The dataset encompasses two primary features: *acoustic_data* and *time_to_failure*, as depicted in Table 1. For this study, the focus is on predicting the *time_to_failure* feature.

Table 1. Features in the Dataset

No.	**Feature**	**Data Type**	**Explanation**
1.	*acoustic_data*	Integer	Seismic signal
2.	*time_to_failure*	Float	Time until the occurrence of an earthquake

2.3 *Data Pre-Processing*

To ensure accurate results from the machine learning model, it is essential to conduct data preprocessing as a preliminary step. This involves checking for missing values and skewness. Upon conducting the missing value check, it was found that there are no missing data in the dataset. Additionally, the skewness check revealed a slight right-skewed distribution with a skewness value of 0.82 for the *acoustic_data* feature. However, this skewness was considered insignificant, suggesting a generally balanced data distribution.

2.3 *Exploratory Data Analysis*

In this stage, an exploration of the utilized dataset is conducted. The first step involves visualizing the correlation between the two features. It was observed through Figure 2 that there is no apparent correlation between the two features, indicating that they possess distinct information and are suitable for use within the same machine learning model.

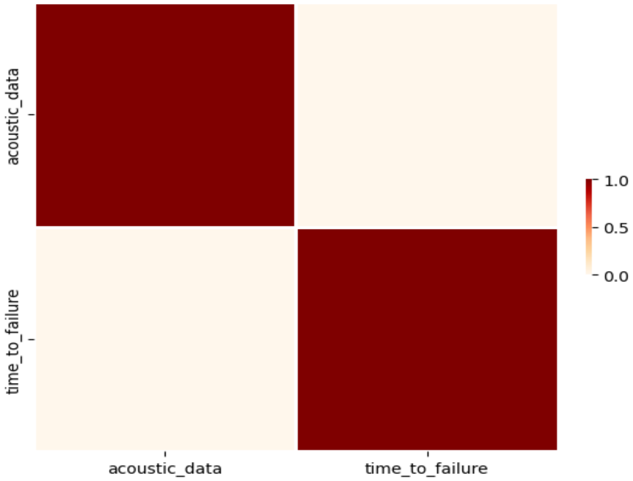

Figure 2. Correlation Matrix Visualization

Next, a scatter plot visualization is performed to identify any outliers in the distribution of the *acoustic_data* feature with respect to the *time_to_failure* feature. Figure 3 illustrates the presence of outliers in the *acoustic_data* feature, particularly as *time_to_failure* approaches 0, indicating increased variability in the distribution.

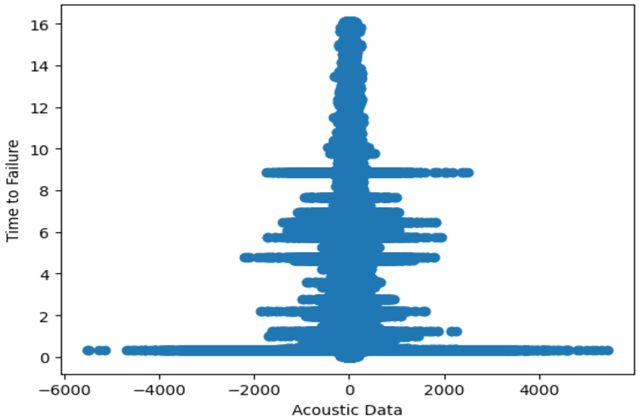

Figure 3. Scatter Plot Visualization

Following this, the Augmented Dickey-Fuller (ADF) Test was performed to assess the stationarity of the time series data. The ADF Test utilizes Linear Regression to determine if the data is stationary, with a constant mean and variance over time [5, 6]. The Test yielded a low p-value of (8.7221 x 10^{-7}), leading to the rejection of the null hypothesis and indicating that the data is stationary. Consequently, no differencing is required for prediction.

2.4 *Feature Engineering*

During this stage, the process involves dividing the dataset into segments, manipulating features within the dataset, and applying feature scaling using the Robust Scaler.

2.4.1 Segmentation

The EDA process revealed that the dataset contains an extremely large number of rows, leading to a highly complex computational process. To address this, the dataset, which contains 577.060.473 rows, is divided into 3.847 segments, with each segment consisting of 15.000 rows.

2.4.2 Feature Manipulation

It became evident in the previous stages that only one feature, namely *acoustic_data*, can be used as a predictor. This may lead to suboptimal learning processes for the model, affecting its ability to accurately predict the target variable. Therefore, new features are added in this stage, including the average, standard deviation, maximum value, and minimum value of the *acoustic_value* feature for each row of the dataset within a specific segment. Additionally, the *rseismic* feature is also added, obtained from the difference between the *acoustic_mean* feature in two adjacent rows.

2.4.3 Feature Scaling

The scatter plot shown in Figure 3 indicates an outlier in the train acoustic feature. To address this, the Robust Scaler is implemented to eliminate the outliers in this feature. The implementation of the Robust Scaler involves calculating the median and quartiles of the feature with outliers, subtracting the median from each feature value, and dividing the difference by the interquartile range (IQR). The description of features in the dataset after the feature engineering process can be seen in Table 2.

Table 2. Features in the Dataset after Feature Engineering

No.	**Feature**	**Data Type**	**Explanation**
1.	*acoustic_mean*	Float	The mean seismic signal on a specific segment.
2.	*acoustic_std*	Float	The standard deviation of seismic signal on a specific segment.
3.	*acoustic_max*	Float	The maximum value of the seismic signal on a specific segment.
4.	*acoustic_min*	Float	The minimum value of the seismic signal on a specific segment.

5.	*rseismic*	Float	The difference in the acoustic mean between two consecutive records.
6.	*time_to_failure*	Float	The time until the occurrence of an earthquake.

2.5 Modeling

2.5.1 AutoRegressive Integrated Moving Average

AutoRegressive Integrated Moving Average (ARIMA) is a statistical method used for forecasting time series data by examining the relationship between the current value and previous values [7]. This model combines regression and moving averages to enhance its predictive accuracy [8, 9]. The model's key parameters include the differencing level, autoregression, and moving average. Generally, ARIMA predictions with orders (1, 1, 1), (4, 1, 1), or (2, 1, 2) closely approximate actual data, as depicted in Figure 4.

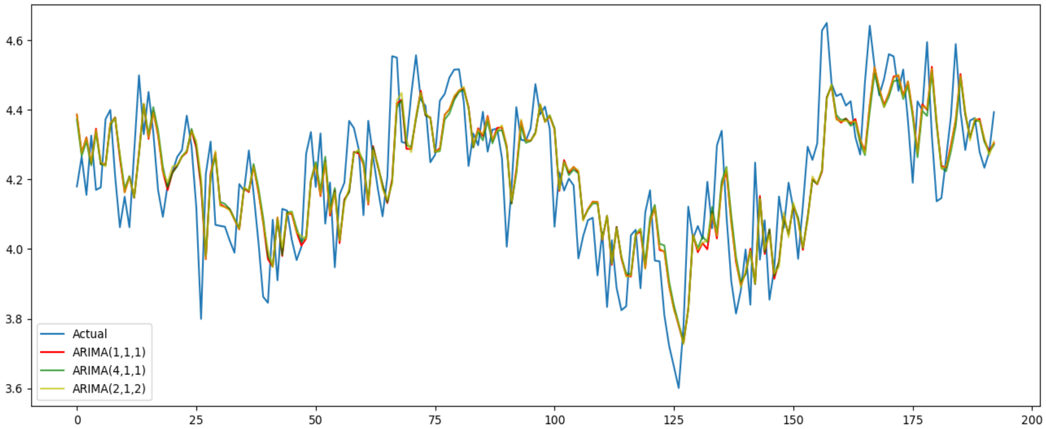

Figure 4. Comparison of 3 ARIMA Models Prediction Results

2.5.2 Linear Regression

Linear Regression (LR) is a method for predicting the relationship between a dependent variable and one or more independent variables by assuming a linear relationship and creating the best-fitting line to represent that relationship. The result is an equation of the line that can be used to predict the value of the dependent variable based on the independent variables [10, 11]. From the visualization in Figure 5, it is evident that the LR model's predictions tend to deviate from the actual data. When the actual data values decrease, the LR model's predictions do not decrease proportionally.

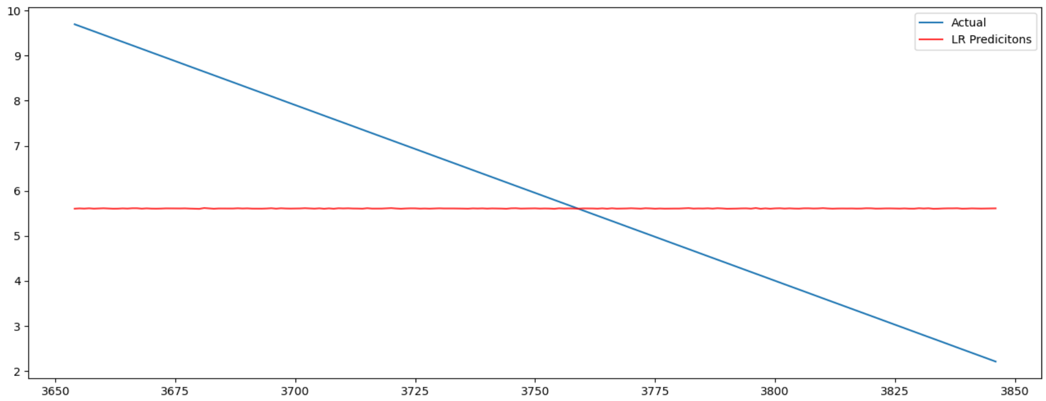

Figure 5. Linear Regression Model Prediction Result

3. RESULTS AND DISCUSSIONS

3.1 Validation

The custom train test split method is a crucial step in the preprocessing and feature engineering stages of the dataset. This method involves dividing the dataset into training and validation data based on a specified percentage, offering enhanced flexibility compared to regular train test splits. The custom train test split is implemented as a function, allowing for easy modification of the separation percentage and addressing issues such as class imbalance and overfitting in the dataset [12].

3.2 Evaluation

In this study, the Mean Absolute Error (MAE) is employed as the evaluation metric for prediction results. MAE measures the average absolute differences between prediction results and actual values, representing an effective indicator of model performance [13, 14]. A lower MAE value signifies better model performance. The formula used to calculate MAE is as follows:

$$\boldsymbol{MAE = \frac{1}{n}\Sigma\,|y - \hat{y}|} \tag{1}$$

with n representing the quantity of the dataset, y denoting the actual values in the dataset, and $\hat{y}$ signifying the predicted outcome. The MAE scores for each model used in the research are compared in Table 3. The results demonstrate that the ARIMA (1, 1, 1) model yields the lowest MAE score, indicating its superior predictive performance compared to both Linear Regression and ARIMA models with different parameters.

Table 3. Comparison of the Models' Performance

Model	**Mean Absolute Error (MAE)**
ARIMA (1,1,1)	**0.110628**
ARIMA (2,1,2)	0.110654
ARIMA (4,1,1)	0.111187
Linear Regression	1.895926

The ARIMA models effectively capture temporal dependencies and patterns in seismic data, outperforming the linear assumptions of Linear Regression, which overlook the inherent time-dependent structure of the data. These findings emphasize the potential of ARIMA models in time series analysis for earthquake forecasting.

Moreover, the presence of autocorrelation in the prediction results using the best model is determined through the Ljung-Box Test, while the Jarque-Bera Test is conducted to assess whether the prediction results follow a normal distribution or not [15, 16]. These tests add significant depth to the analysis and contribute to the robustness of the findings. Further explanations regarding each of these tests are as follows:

I. Ljung-Box Test
The Ljung-Box Test is a statistical method used to assess the presence of autocorrelation in a dataset by comparing its variance to the expected variance if the data points were independent. A positive result from the Ljung-Box Test indicates the presence of autocorrelation. In this study, the Ljung-Box Test was conducted on the *time_to_failure* predictions, yielding a p-value of 0.44. This suggests that the null hypothesis is accepted, indicating that the predictions are independent and lack autocorrelation.

II. Jarque-Bera Test
The Jarque-Bera (JB) Test is employed to test if a sample data distribution follows a normal distribution by analyzing skewness, kurtosis, and then comparing them to the standard normal distribution. A significant test result implies a departure from normality. The JB Test performed on the *time_to_failure* feature estimates yielded a p-value close to 0.24, suggesting that the predictions follow a normal distribution.

4. CONCLUSION

The research findings suggest that the ARIMA (1,1,1) model is adept at predicting the occurrence time of earthquakes, boasting an outstanding MAE score of 0.110628 in comparison to other models, including ARIMA with different parameters and Linear Regression. This superiority underscores the ARIMA (1,1,1) model's ability to accurately capture the temporal dependencies and patterns within seismic data, which Linear Regression is unable to do. The Ljung-Box Test results revealed no autocorrelation in the prediction outcomes, ensuring that the residuals are white noise. Furthermore, the JB Test results indicated that the distribution of the prediction outcomes follows a normal distribution, validating the model assumptions.

This study demonstrates the effectiveness of the ARIMA model in predicting earthquake occurrences using the LANL Earthquake dataset, which replicates realistic seismic activity characterized by irregular occurrences. The ARIMA model's capability to handle the periodic and irregular nature of the dataset underscores its robustness and applicability to real-world seismic data. These findings provide valuable insights for institutions aiming to leverage the ARIMA model for earthquake prediction and to strategize preventive measures to mitigate the adverse effects of earthquakes on communities and the environment.

5. FUTURE WORKS

The study proposes several methods to enhance research outcomes. Firstly, it suggests creating new features from seismic signal information by utilizing the real and imaginary values of Fast Fourier Transform on the *acoustic_value* feature in each segment. Additionally, the study recommends implementing the Deep Neural Network (DNN) architecture as a predictive model due to its effective capability to learn complex and large-sized feature representations. Finally, the study encourages the utilization of Vector Autoregressive (VAR) as a predictive model that can simultaneously utilize multiple features as predictors. These methods are expected to contribute to improving the overall research outcomes in the field.

REFERENCES

[1] N. Christiastuti, "Korban Jiwa GEMPA Turki-Suriah Bertambah jadi 41.000 orang," detiknews, https://news.detik.com/internasional/d-6569601/korban-jiwa-gempa-turki-suriah-bertambah-jadi-41-000-orang (accessed Feb. 15, 2023).

[2] K. Hirose, S. Labrosse, and J. Hernlund, "Composition and state of the Core," Annual Review of Earth and Planetary Sciences, vol. 41, no. 1, pp. 657–691, 2013. doi:10.1146/annurev-earth-050212-124007.

[3] A. Howard, B. Rouet-Leduc, and L. J. Pyrak-Nolte, "Lanl earthquake prediction," Kaggle, https://kaggle.com/competitions/LANL-Earthquake-Prediction (accessed Feb. 7, 2023).

[4] P. A. Johnson et al., "Laboratory earthquake forecasting: A machine learning competition," Proceedings of the National Academy of Sciences, vol. 118, no. 5, 2021. doi:10.1073/pnas.2011362118.

[5] E. Paparoditis and D. N. Politis, "The asymptotic size and power of the augmented dickey–fuller test for a unit root," Econometric Reviews, vol. 37, no. 9, pp. 955–973, 2016. doi:10.1080/00927872.2016.1178887.

[6] A. K.P, A. S. Oluwaseun, and V. G. Jemilohun, "Test for Stationarity on Inflation Rates in Nigeria using Augmented Dickey Fuller Test and Phillips-Persons Test," IOSR Journal of Mathematics, vol. 16, no. 3, pp. 11–14, 2020. doi:10.9790/5728-1603031114.

[7] X. Wang, Y. Kang, R. J. Hyndman, and F. Li, "Distributed Arima models for ultra-long Time Series," International Journal of Forecasting, vol. 39, no. 3, pp. 1163–1184, 2023. doi:10.1016/j.ijforecast.2022.05.001.

[8] B. Dey, B. Roy, S. Datta, and T. S. Ustun, "Forecasting ethanol demand in India to meet future blending targets: A comparison of Arima and various regression models," Energy Reports, vol. 9, pp. 411–418, 2023. doi:10.1016/j.egyr.2022.11.038.

[9] I. Unggara, A. Musdholifah, and A. K. Sari, "Optimization of Arima forecasting model using Firefly algorithm," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 13, no. 2, p. 127, 2019. doi:10.22146/ijccs.37666.

[10] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear regression," Springer Texts in Statistics, pp. 69–134, 2023. doi:10.1007/978-3-031-38747-0_3.

[11] H. K. Prakosa and N. Rokhman, "Anomaly detection in hospital claims using K-means and linear regression," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 15, no. 4, p. 391, 2021. doi:10.22146/ijccs.68160.

[12] L. Wynants et al., "Prediction models for diagnosis and prognosis of covid-19: Systematic Review and Critical Appraisal," BMJ, p. m1328, 2020. doi:10.1136/bmj.m1328.

[13] S. M. Robeson and C. J. Willmott, "Decomposition of the mean absolute error (mae) into systematic and unsystematic components," PLOS ONE, vol. 18, no. 2, 2023. doi:10.1371/journal.pone.0279774.

[14] M. D. Fauzi, A. E. Putra, and W. Wahyono, “Estimation of average car speed using the haar-like feature and Correlation Tracker method,” IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 14, no. 4, p. 353, 2020. doi:10.22146/ijccs.57262.

[15] H. Hassani and M. R. Yeganegi, “Selecting optimal lag order in ljung–box test,” Physica A: Statistical Mechanics and its Applications, vol. 541, p. 123700, 2020. doi:10.1016/j.physa.2019.123700.

[16] D. Abdellatif, K. El Moutaouakil, and K. Satori, “Clustering and Jarque-bera normality test to face recognition,” Procedia Computer Science, vol. 127, pp. 246–255, 2018. doi:10.1016/j.procs.2018.01.120.