Jurnal Statistika Dan Komputasi (STATKOM), Vol. 5 No. 1 (June 2026)
ISSN : 2963-038X (print), ISSN : 2963-0398
DOI : https://doi.org/10.32665/statkom.v5i1.6427

Time Series Anomaly Detection of International Tourist Arrivals Using Copula-Based Outlier Detection Method

Nabil Bintang Prayoga[1], Yenni Angraini[2], Akbar Rizki[3]

[1, 2, 3]Study Program in Statistics and Data Science, IPB University
E-mail: bintangprayoga@apps.ipb.ac.id

Submitted: April 8, 2026 ***Revised:*** June 14, 2026 ***Accepted:*** June 18, 2026

Abstract

Background: *An outlier or anomaly is an observation that deviates from normal historical patterns. An observation may appear normal individually, but can be identified as anomalous when evaluated through its dependence on other variables. Copula-based outlier detection (COPOD) accommodates multiple variables using empirical marginal distributions and tail dependence structures to identify anomalies.*
Objective: *This study aims to detect anomalies in the number of international tourists in Indonesia by considering the variables of inflation and the rupiah exchange rate, as well as to evaluate the handling of anomalies on the forecasting performance of long short-term memory (LSTM).*
Methods: *Monthly data from January 2000 to December 2025 obtained from CEIC Data, Statistics Indonesia, and Bank Indonesia were used in the study. The analysis includes data exploration and the development of feature engineering, anomaly detection using COPOD, followed by LSTM forecasting.*
Results: *Detection was carried out based on twelve variables resulting from feature engineering, and eleven periods were identified as anomalies. The forecasting results show better accuracy in the model after handling with a mean absolute percentage error, root mean square error, and Pearson correlation between actual and predicted data, which were 7.494%, 99233, and 0.864, respectively, on the test data.*
Conclusion: *That good accuracy result can't be separated from the precision in detecting anomalies. Further research is expected to add relevant variables and develop feature engineering.*

Keywords : *Anomaly, Copulas-Based Outlier Detection, International Tourists, LSTM.*

Abstrak

Latar Belakang: *Outlier* atau anomali adalah pengamatan yang menyimpang dari pola historis normal. Sebuah pengamatan mungkin terlihat normal secara individu tetapi dapat dikenali sebagai anomali ketika dievaluasi melalui ketergantungannya dengan peubah lain. *Copula-based outlier detection* (COPOD) mengakomodasi beberapa peubah menggunakan distribusi marginal empiris dan struktur ketergantungan ekor untuk mengidentifikasi anomali.
Tujuan: Penelitian ini bertujuan mendeteksi anomali data jumlah wisatawan mancanegara di Indonesia dengan mempertimbangkan peubah inflasi dan nilai tukar rupiah serta mengevaluasi penanganan anomali terhadap performa peramalan *long short-term memory* (LSTM).
Metode: Data bulanan dari Januari 2000 hingga Desember 2025 yang diperoleh dari CEIC Data, Badan Pusat Statistik, dan Bank Indonesia digunakan dalam penelitian. Analisis mencakup eksplorasi data dan pembentukan *feature engineering*, pendeteksian anomali menggunakan COPOD, dilanjutkan peramalan LSTM.
Hasil: Pendeteksian dilakukan berdasarkan dua belas peubah hasil *feature engineering* dan teridentifikasi sebelas periode anomali. Hasil peramalan menunjukkan akurasi lebih baik pada model sesudah penanganan dengan *mean absolute percentage error*, *root mean square error,* dan korelasi Pearson antara data aktual dan prediksi sebesar 7.494%, 99233, dan 0.864 pada data uji.
Kesimpulan: Hasil akurasi yang baik tersebut tidak dapat dipisahkan dari ketepatan pendeteksian anomali. Penelitian lanjutan diharapkan dapat menambah peubah relevan dan mengembangkan *feature engineering.*

Kata kunci: Anomali, *Copula-Based Outlier Detection,* LSTM, Wisatawan Mancanegara.

INTRODUCTION

Data quality is a fundamental factor that affects the accuracy of estimation and modeling results. One of the challenges in maintaining data quality is the presence of outliers, also commonly referred to as anomalies, which are data that deviate from the pattern of the majority of observations due to unusual behavior or rare events. The presence of anomalies can disrupt the fulfillment of model assumptions, increase the variance of parameter estimates, and cause bias in inference results if not handled properly (Arimie et al., 2020). Anomaly detection and management become an important stage, especially in time series data that have temporal dependencies.

Time series anomaly detection requires an approach that accounts for temporal dependence, seasonal patterns, and long-term trends. In this context, anomalies are not only observations that deviate from historical patterns and probability distributions (Zhao et al., 2019), but can also arise when observations that appear normal individually violate temporal structures or dependency relationships with other variables (Wang et al., 2024). Therefore, anomaly detection should consider both temporal dynamics and intervariable dependence to identify deviations more accurately and support stable forecasting model learning.

Anomaly detection is usually carried out using parametric approaches, such as residual testing of autoregressive integrated moving average (ARIMA) models, Z-score, and seasonal decomposition based on the interquartile range (IQR). These methods identify anomalies based on the patterns in the data, so they work well for univariate data assuming a normal distribution, but they're limited when it comes to data with many variables and nonlinearity (Schmidl et al., 2022). One method that can handle this limitation is copula-based outlier detection (COPOD), which uses copula theory to model dependencies between variables without assuming any distribution (Li et al., 2020). This method uses information and dependencies from other variables related to the main variable (Shahrestani et al., 2026). COPOD is nonparametric, does not require labels, and is capable of handling many variables with diverse scales and distributions, making it potentially applicable to time series data involving multiple variables (Schmidl et al., 2022).

Data in the field of tourism, such as the number of international tourists, is closely related to economic conditions. Travel costs and the purchasing power of international tourists are related to the macroeconomic conditions of a country, including Indonesia, which include inflation stability and the strength or weakness of the rupiah exchange rate (Wicaksono, 2022). Changes in their combination can be reflected as patterns that deviate from normal historical dynamics. Anomalies identified in the data represent the impact of external events that are not explicitly modeled. Unhandled anomalies can mess up historical patterns and hurt forecasting performance, including long short-term memory (LSTM) models that are sensitive to extreme values, making the information less reliable for supporting tourism planning and policy. Anomaly detection helps improve the accuracy and robustness of the model against prediction errors (Karnick et al., 2025).

Previous research generally focused on forecasting the number of international tourists using ARIMA or LSTM without anomaly detection (Fitri et al., 2019; Lisnawati et al., 2025). A multivariable approach has been carried out using the seasonal autoregressive integrated moving average with exogenous regressors (SARIMAX) method, with the rupiah exchange rate as one of the exogenous variables; however, the study has not yet applied anomaly detection (Sutisna et al., 2026).

Previous studies on international tourist forecasting in Indonesia have primarily focused on prediction performance and have rarely addressed anomaly detection as a dedicated stage. Moreover, no study was found that applies COPOD to international tourist arrivals or integrates anomaly detection results with LSTM to evaluate their impact on forecasting accuracy. Therefore, this study aims to identify anomalies in international tourist arrivals using COPOD and assess the effect of anomaly handling on LSTM forecasting performance.

METHODS

Research Design

This study uses a quantitative approach with an observational design using multivariate time series data. The study focuses on analyzing numerical data to detect anomaly periods and the impact of handling them on forecasting results. The data used consists of monthly data from January 2000 to December 2025. The research stages include data exploration, feature engineering, anomaly detection using COPOD based on the interdependence between variables, and evaluating the impact of handling on LSTM forecasting results.

Population and Sample

The study population consists of all monthly data on the number of international tourists in Indonesia, the inflation rate, and the rupiah exchange rate. The study sample includes data on all three variables from January 2000 to December 2025, totaling 312 periods. The study utilizes the entire dataset within the observation period for the analysis process without selecting a subset of the data.

Sampling Techniques

The sampling technique used was purposive time-based sampling, which is a method of selecting a sample based on a specific time period that aligns with the research objectives (Syahzaqi et al., 2025). This technique is appropriate for research data in the form of time series, where there are temporal relationships between periods in each observation.

Research Subjects

The research subjects in this study consist of multivariate time series data related to international tourists. This research uses secondary data obtained from the CEIC Data website, the Central Bureau of Statistics, and Bank Indonesia. Monthly data was used for 312 periods, from January 2000 to December 2025. The list of variables used is presented in Table 1.

Table 1. List of variables.

Variables	Unit	Reference
International tourists (y_{1_t})	People	(Sutisna et al., 2026)
Inflation rate (y_{2_t})	Percent	(Nabila & Munifatussa'idah, 2022)
Rupiah exchange rate (y_{3_t})	Rp/USD	(Sutisna et al., 2026)

Data Analysis Techniques

Data analysis was conducted using Python and R software. The stages of data analysis include exploration and feature engineering formation, anomaly detection using the COPOD method, and LSTM forecasting. The analysis procedure carried out is as follows.

1) Data exploration and preparation.
   a. Identifying missing data for each variable.
   b. Performing missing data imputation if there are missing values in the data.
   c. Conducting data exploration to identify patterns and characteristics of each variable using time series plots and descriptive statistics.
2) Time dependency analysis and determination of lag.
   a. Testing variance stationarity using a Box-Cox plot. Data that is stationary in

variance has a lambda (λ) value of 1 at a 95% confidence interval. The Box-Cox transformation needs to be performed if the data is not stationary in variance.

b. Conducting a mean stationarity test using an autocorrelation function (ACF) plot and the augmented Dickey-Fuller (ADF) test. Data that is not stationary in mean is indicated by an ACF plot that declines slowly (tails off slowly) and a p-value greater than the significance level in the ADF test (Maharani et al., 2023). Handling is done through the differencing process (Syahzaqi et al., 2025).
c. Identifying significant lags for each variable using ACF and partial autocorrelation function (PACF) plots (Khoiriyah et al., 2023).
d. Identifying dominant lags that are significant simultaneously using the VAR based on the smallest Hannan-Quinn (HQ) and Schwarz (SC) criteria.

3) Formation of feature engineering.
   a. Performing a logarithmic transformation of the number of international tourists to reduce skewness.
   b. Converting the rupiah exchange rate into the logarithm of monthly changes (log-return) to explain relative changes (Ahmar et al., 2024).
   c. Creating lag features according to the results in step 2.
   d. Creating month-on-month (MoM) change features for the number of international tourists to capture short-term dynamics.
   e. Creating seasonal pattern features represented by sine and cosine transformations of the month to maintain the continuity of the annual cycle (Taylor & Letham, 2018).
4) Anomaly detection using the COPOD method (Li et al., 2020).
   a. Conceptually, dividing the data into two periods, namely the reference period to build the COPOD empirical distribution and determine anomaly thresholds, and the temporal evaluation period to identify anomalies in observations against historical patterns without updating the distribution.
   b. Transforming each observation value y_{tj} into a uniform scale [0,1] using the empirical distribution function of the j-th variable ($F_j(.)$) in Equation 1. The transformation is carried out to eliminate the influence of differences in data scale and marginal distributions of the variables, with u_{tj} being the result of the transformation and n being the number of observations in the reference period. y_{kj} is the data to be transformed, while y_{tj} is the data in the reference period.

$$u_{tj} = F_j(y_{tj}) = \frac{1}{n+1}\sum_{k=1}^{n} 1(y_{kj} \le y_{tj}) \quad (1)$$

   c. Calculating the lower-tail (l_{tj}) and upper-tail (r_{tj}) dependence scores for each variable using Equations 2 and 3. A large l_{tj} indicates observations with low extremes, while a large r_{tj} indicates observations with high extremes.

$$l_{tj} = -\log(u_{tj}) \quad (2)$$

$$r_{tj} = -\log(1-u_{tj}) \quad (3)$$

   d. Selecting the anomaly score (s_{tj}) based on skewness (γ_j) of the j-th variable distribution over the reference period through Equation 4 to ensure that anomalies are detected on the relevant side of the distribution. This approach allows COPOD to adapt to the characteristics of the marginal distributions of variables.

$$s_{tj} = \begin{cases} r_{tj}, & \gamma_j \ge 0 \\ l_{tj}, & \gamma_j < 0 \end{cases} \quad (4)$$

   e. Calculating the total anomaly score (S_t) for the t-th observation as described in Equation 5 as the summation of contributions from all variables. A higher S_t value indicates that the observation at time t deviates more simultaneously from the historical pattern of all variables, with q denoting the number of variables.

$$S_t = \sum_{j=1}^{q} s_{tj} \quad (5)$$

f. Determining the anomaly threshold based on the 99th percentile of the COPOD score distribution in the reference period to avoid future information bias. Observations with S_t values exceeding this threshold are classified as anomalies.

g. Repeat steps 4b to 4f during the evaluation period with $F_j(.)$, γ_j, and the anomaly threshold based on calculations from the reference period.

5) Handling anomalies becomes y_t^* based on the amount of anomaly data (Iwueze et al., 2018).

a. If there is one anomalous data point, use the neighbor mean method according to Equation 6, with y_{t-1} and y_{t-2} being normal data among the anomalous data.

$$y_t^* = \frac{y_{t-1}+y_{t+1}}{2} \quad (6)$$

b. If there are two to four consecutive anomalous data points, use Equation 7 to handle them, with y_1 and t_1 representing the data and period before the anomaly also y_2 and t_2 representing the data and period after the anomaly.

$$y_t^* = y_1 + \frac{y_2 - y_1}{t_2 - t_1}(t - t_1) \quad (7)$$

6) Forecasting the number of international tourists using LSTM on data before and after anomaly handling.

a. Normalizing the number of international tourists data using min-max to maintain training stability (Pranolo et al., 2024).

b. Splitting the data into training data and test data based on data patterns.

c. Constructing a time series k-fold cross-validation scheme on the training data while maintaining the time order and $k = 5$ using an expanding window approach. The training data is divided into training and validation data with an 80:20 ratio in each fold without shuffling the order. The model is only trained on past data and validated on future data.

d. Determining the LSTM hyperparameter scenario on the training data according to Table 2. The parameters set as the same are the use of two hidden layers with 128 and 64 neurons, the adaptive moment estimation (Adam) optimizer, a dropout of 0.2, and the loss function as mean squared error (MSE) (Lisnawati et al., 2025).

Table 2. Hyperparameter scenario.

Scenario	Hyperparameter		
	Epoch	Learning Rate	Batch Size
1	100	0.01	16
2	100	0.01	32
3	100	0.001	16
4	100	0.001	32
5	150	0.01	16
6	150	0.01	32
7	150	0.001	16
8	150	0.001	32

e. Train the LSTM model on each fold with all combinations of hyperparameters.

f. Select the best hyperparameter combination based on the smallest root mean square error (RMSE).

g. Retrain the final LSTM model with all training data based on the best hyperparameter combination.

h. Validate the LSTM model with test data to obtain forecast values.

i. Perform denormalization to the original scale using the inverse transform method.

j. Evaluate the performance of the LSTM model on test data by measuring mean absolute percentage error (MAPE), RMSE, and correlation.

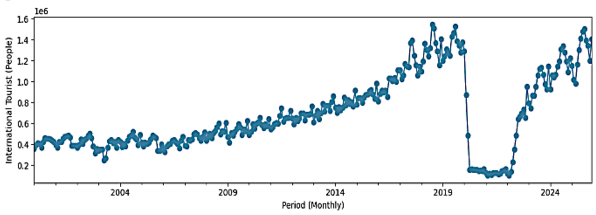
7) Compare the MAPE, RMSE, and correlation accuracy of the LSTM model before and after anomaly handling.

8) Interpret the results and draw conclusions.

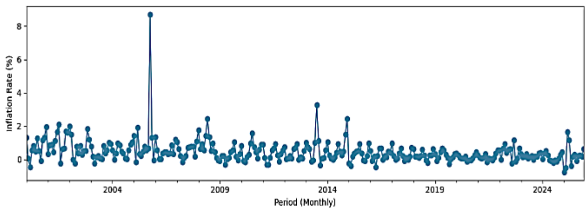
RESULTS AND DISCUSSION

Data Exploration

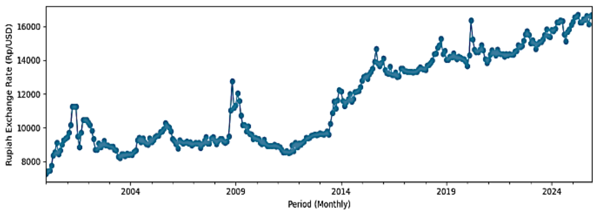
Data exploration was carried out using time series graphs and descriptive statistics to identify patterns, trends, and characteristics of each variable's movement. Figure 1 shows that the number of international tourists has an increasing trend from the early 2000s until before the pandemic, with a seasonal pattern every 12 months. A sharp decline occurred in April 2020 due to the COVID-19 pandemic, followed by a gradual recovery that became visible again in 2023. This increase serves as evidence that Indonesian tourism has a high level of adaptability despite being heavily affected by the COVID-19 pandemic (Rasyidi, 2024). Inflation value fluctuations can be seen around certain mid-range values, with several large spikes, such as in 2005 and periods of low pressure during the pandemic. Long-term depreciation occurred in the rupiah's exchange rate against the US dollar, accompanied by several periods of volatility, especially during global crises and the pandemic.



(a)



(b)



(c)

Figure 1. Time series graphs of (a) international tourists, (b) inflation rate, and (c) the rupiah exchange rate

Descriptive statistical results show that the number of international tourists averages around 686 thousand visits per period. The standard deviation is relatively large, reflecting high variability. The average monthly inflation is around 0.471 percent, with a fairly wide range of values, including periods of deflation and price surges. The rupiah exchange rate has an average of around 16 thousand rupiah per USD, with fluctuations indicating depreciation and appreciation dynamics throughout the observation period. The variability of these three variables indicates the presence of quite sharp changes at certain times.

Based on Figure 1, anomalies in the number of international tourists can only be recognized to a limited extent through visible pattern changes, like the sharp drop around April 2020 and the rise again around April 2022. However, not all anomalies always appear as clear spikes or drops in a single variable. A period might seem normal when looking only at tourist data, but it becomes unusual when linked with other related variables, such as inflation or structural changes in the exchange rate. Therefore, anomaly detection needs to take into account the relationships between variables so that deviations that are not visually obvious in a single time series can still be identified.

The three variables are macroeconomically interrelated through exchange rate and inflation mechanisms. Rupiah depreciation can increase inflationary pressure through the exchange rate pass-through mechanism, while high inflation can affect competitiveness and exchange rate movements (Syamad & Handoyo, 2023). Exchange rate changes also affect cost attractiveness for tourists. (Nugraha & Naylah, 2023). This interrelation indicates that the dynamics of the three variables do not stand alone and have the potential to cause extreme fluctuations during certain periods.

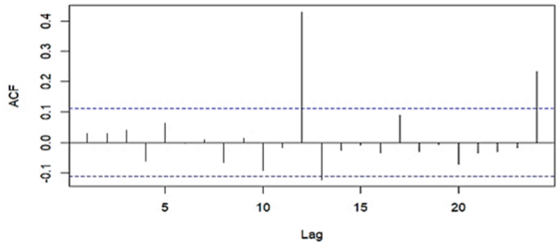
Data Stationarity

Stationarity checks were conducted on the variance and mean. Variance stationarity was examined through a Box-Cox plot, while mean stationarity was checked via an ACF plot and the ADF test (Nurhambali et al., 2024). Box-Cox requires positive data values, so the inflation variable needed a data shift by adding a constant equal to the minimum value plus 0.1 due to the presence of negative values in the form of deflation (Atkinson et al., 2021). The results showed that all variables were non-stationary in variance, with λ values for the number of international tourists, the inflation rate, and the rupiah exchange rate being 0.747, 0.343, and 0.505, respectively. Because $\lambda = 1$ was not within the 95% confidence interval, the Box-Cox transformation was applied according to each variable's λ value to make the data variance-stationary.

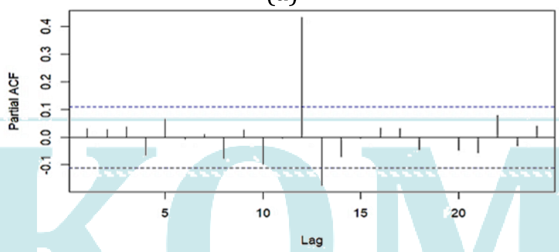
The mean stationarity test indicates that the number of international tourists and the rupiah exchange rate are not stationary, marked by ACF plots that tail off slowly. The p-values from the ADF test for both variables are 0.345 and 0.420, which are greater than 0.05, so a handling of one-time differencing was applied. Different results are shown for the inflation rate data, which did not require differencing because it was already mean-stationary with an ADF test p-value of 0.01, so the handling was only done with a Box-Cox transformation since variance stationarity had not yet been met. The final condition of all variables after handling shows that the stationarity assumption has been satisfied. The value $\lambda = 1$ is already within the 95% confidence interval, with a normally declining ACF plot pattern and an ADF test p-value of less than 0.05.

Lag Identification

Lag identification is performed to determine statistically relevant time delays before feature engineering is constructed. The analysis uses ACF and PACF plots for each variable to identify cut-off and tails-off patterns indicating short-term and seasonal dependencies. Simultaneous dynamic dependence among variables is also examined using the VAR approach to obtain dominant lags representing temporal interactions.



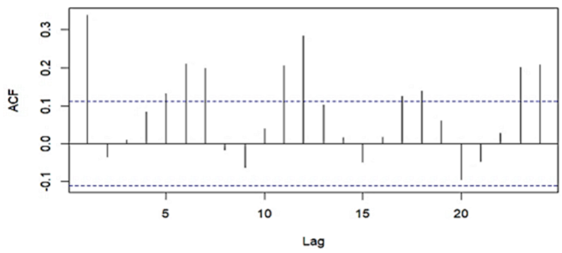
(a)



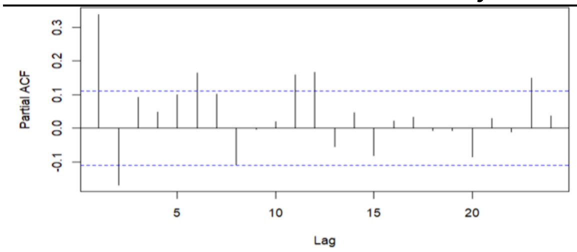
(b)

Figure 2. Plot (a) ACF and (b) PACF for the international tourists

The identification results for the number of international tourists are shown in Figure 2, which displays significant autocorrelation values at lag 12. This pattern indicates a strong annual seasonal component, with values in one month highly correlated with the same month in the previous year. No significant cut-off is found at short lags, indicating that short-term dependence is relatively weak compared to the influence of the annual seasonality.



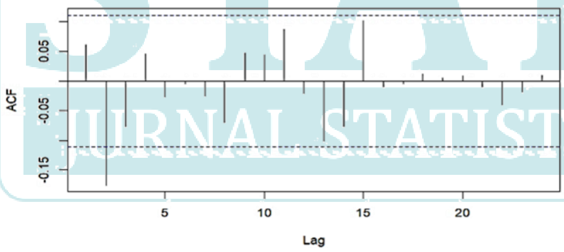
(a)



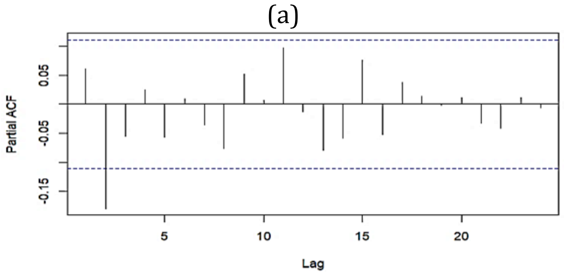
(b)

Figure 3. *Plot* (a) ACF and (b) PACF of the inflation rate

Figure 3 shows the autocorrelation pattern of the inflation rate variable, which tends to tail off without a clear cut-off. Seasonal exploration is further carried out through periodic ACF and shows significance at lags 6, 12, 18, and 24, indicating a semi-annual pattern. The ACF and PACF plots of the rupiah exchange rate variable in Figure 4 show a significant spike at lag 2. The pattern after lag 2 indicates that the autocorrelation values are within the confidence bounds, thus indicating lagged effects occurring in short-term dynamics.



(a)



(b)

Figure 4. *Plot* (a) ACF and (b) PACF of the rupiah exchange rate

Simultaneous identification using VAR with the smallest SC and HQ criteria shows lag 1 as the optimal lag. The main interactions among variables occur in the previous period. Adding higher lags does not provide an improvement proportional to complexity. These results strengthen the lag selection in the feature construction stage.

Feature Engineering Construction

Feature engineering is carried out because the COPOD method does not explicitly consider temporal dependencies, but rather operates in a cross-sectional framework. The extraction of relevant features plays an important role in improving the method's performance as well as understanding data patterns (Horvath et al., 2020). Anomalies may be wrongly detected without feature transformation because information on trends, seasonality, and inter-period lags risks being unrepresented. These features are created based on the results of lag identification and adjustments to variable shapes to better capture time series dependency patterns.

The engineered lag features may exhibit multicollinearity because temporal observations in time series data are naturally correlated. However, unlike regression-based approaches, COPOD does not estimate model coefficients and therefore is less sensitive to multicollinearity. While correlated features may introduce a degree of redundancy, they can also strengthen the representation of persistent temporal shocks, which is desirable in anomaly detection. The feature engineering used is presented in Table 3.

Table 3. Feature engineering

Variables	Feature Engineering
International tourists (y_{1_t})	Log of international tourists MoM log of international tourists Lag 1 log of international tourists Lag 12 log of international tourists
Inflation rate (y_{2_t})	Lag 1 inflation Lag 6 inflation
Rupiah exchange rate (y_{3_t})	Log-return of rupiah exchange rate Lag 1 log-return of rupiah exchange rate Lag 2 log-return of rupiah exchange rate
Addition (seasonal effect)	Seasonal sine Seasonal cosine

Logarithmic transformation was applied to the variable of the number of tourists to reduce the skewness of the distribution and control the influence of extreme values on the anomaly scores (West, 2022). The 1st lag feature $\left(y_{1_{t-1}}\right)$ and the 12th lag feature $\left(y_{1_{t-12}}\right)$ were formed based on previous lag identification, representing the annual seasonal pattern. MoM was also added because it allows the method to recognize short-term growth slowdowns or spikes that are not always visible in the actual values. The inflation rate variable still uses its original value due to its relatively narrow distribution and the presence of negative values during deflation. The 1st lag $\left(y_{2_{t-1}}\right)$ and the 6th lag $\left(y_{2_{t-6}}\right)$ were formed based on identification results to represent short-term dynamics and semi-annual seasonal patterns.

The rupiah exchange rate variable was transformed into log-returns to eliminate strong deterministic trends, making it more suitable for short-term dynamics analysis (Ahmar et al., 2024). The features used include log-return values as well as lag 1 and 2, allowing sudden changes that are truly abnormal to be recognized, compared to only long-term depreciation trends. Annual seasonal effects are also modeled using sine and cosine functions so that the representation of monthly cycles is continuous (Taylor & Letham, 2018). This approach preserves the mathematical closeness between December and January and helps distinguish phases within a single annual cycle. Creating features with a maximum lag of 12 months results in the loss of the first 12 periods, so the number of observations drops from 312 to 300 periods, and the analysis period used becomes January 2001 to December 2025.

Anomaly Detection with the COPOD Method

Anomaly detection using the copula-based outlier detection (COPOD) method was conducted on twelve variables, namely eleven feature engineering results and one inflation variable, without additional transformation. The data were divided into a reference period and an evaluation period. The reference period, from January 2001 to December 2019, with 228 periods, was used to build an empirical distribution representing the historical normal pattern before the structural disruption of the COVID-19 pandemic. The evaluation period, from January 2020 to December 2025, with 72 periods, was assessed relative to the reference distribution, which remained fixed and was not updated. This strategy ensures that structural changes during the pandemic do not shift the historical comparison baseline, so the anomalies detected genuinely reflect deviations from pre-pandemic normal patterns.

Each variable was transformed into a uniform [0,1] scale using the empirical cumulative distribution function. This transformation eliminates the influence of differences in scale and data units as well as the shape of the marginal distribution of variables. These characteristics make copulas very flexible and suitable for empirical data that is non-normal, asymmetric, or has very heavy-tailed distributions (Genest et al., 2024). The analysis focuses on the joint dependence structure (copula), rather than on absolute values. The rarity of combinations of relative positions between variables is evaluated in COPOD, rather than on the extremity of a single variable. Each observation is compared with the entire reference data to obtain its relative position proportion in the historical distribution. Division by $n+1$ is used so that the transformed values are never exactly 0, thus the logarithm in the next stage remains defined.

The dependence scores of the lower and upper tails are calculated for each variable. Negative logarithmic transformation increases sensitivity to values that are very close to 0 or 1, so positions that are very rare in the historical distribution result in large scores. The selection of the tail side used as the partial anomaly score is determined based on the direction of the distribution's tailing in the reference period. Variables with right-skewed distributions are evaluated on the upper tail, while the lower tail is used otherwise, making COPOD adaptive to the shape of the variable's distribution and preventing irrelevant detection on the side of the distribution with long spread (Shahrestani et al., 2026). The rupiah exchange rate and seasonal cosine have negative skewness values, so anomalies are identified in the lower tail and l_{tj} is chosen as the partial anomaly score. The partial anomaly score with r_{tj} values is applied to other variables.

The partial anomaly scores were summed to obtain the total score for each period. This aggregation detects anomalies from several variables simultaneously, indicating extreme positions rather than the contribution of a single variable. The threshold is set based on the 99th percentile of the reference period score distribution, which is 22.521, and is supported by IQR-based assessment (Li et al., 2020). Using the 99th percentile focuses the detection process on the top 1% of extreme observations, making sure that only rare events with significant deviations from the historical multivariate dependency structure are classified as anomalies. This approach reduces the chance of false alarms while still keeping accuracy against structural disturbances in the data. All stages are repeated in the evaluation period using the empirical distribution function, extreme values, and the same threshold from reference period data.

The periods identified as anomalies using the COPOD method were 11 periods spread across the early 2000s, 2008, and the COVID-19 pandemic period, namely from 2020 to 2022. The highest concentration and scores occurred during the post-pandemic recovery period, namely in March and April 2022. These values indicate the overall deviation of the most extreme variables throughout the observation period. Figure 5 shows the anomaly periods resulting from COPOD visualized on the international tourists variable.

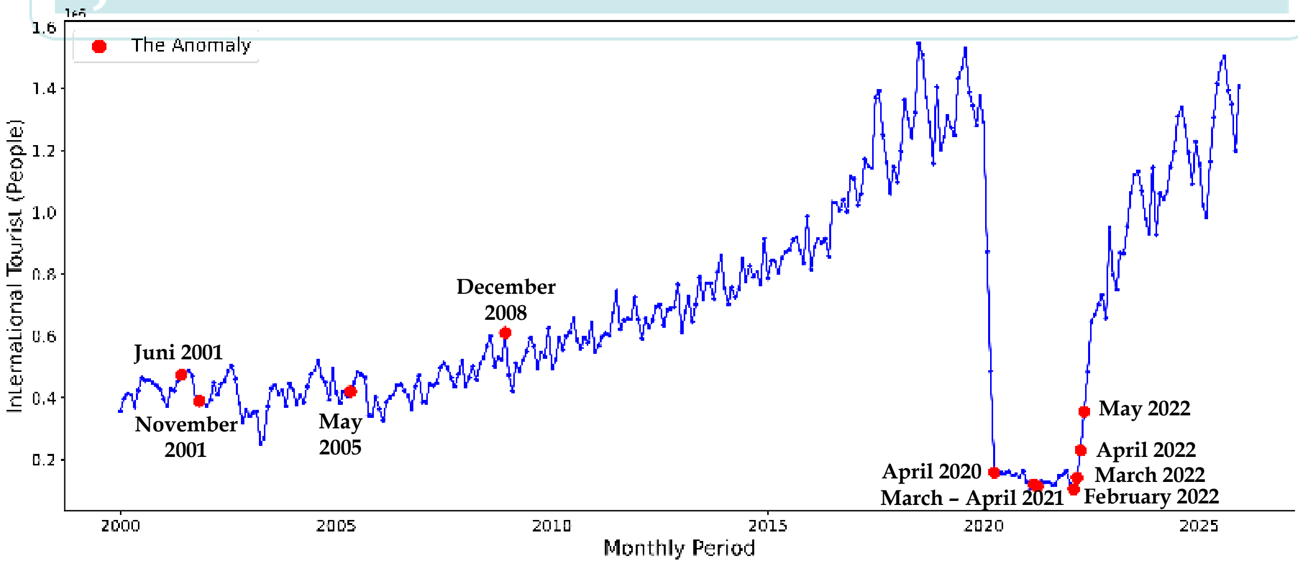


Figure 5. COPOD anomaly periods

Anomalous characteristics from March to April 2022 represent a phase of rapid post-pandemic recovery. Month-over-month tourist numbers and seasonal lags during this period show a very high surge due to the pandemic comparison effect (Rasyidi, 2024). The detected anomalies are not in the form of extreme declines but rather a sharp transition from low levels to a high rebound. These results are reinforced by changes in the inflation lag structure and the normalization of the rupiah exchange rate log-returns after pandemic volatility. This finding indicates the sensitivity of COPOD to structural dynamic changes in the form of sudden contractions and rapid expansions. The entire procedure in the COPOD method allows it to identify structural disturbances across the entire dataset collectively.

Table 4. COPOD scores and handling data

Period	COPOD Score	International Tourists	
		Before Handling	After Handling
June 2001	22.793	474527	466387
November 2001	22.786	388739	390172
May 2005	22.875	419747	430355
December 2008	25.711	610452	498664
April 2020	23.789	158066	323999
March 2021	23.754	119979	113670
April 2021	23.169	112756	121552
February 2022	24.796	105195	194359
March 2022	27.804	142007	266740
April 2022	27.507	230076	339121
May 2022	23.233	354920	411502

The identified anomalous observations were handled using a local smoothing approach according to Equations 6 and 7. The procedure was applied only to the international tourist arrivals variable because the forecasting stage focused on univariate LSTM modeling, whereas inflation and the rupiah exchange rate served solely as contextual variables in anomaly detection. Rather than removing observations or replacing them through global imputation, local smoothing adjusts anomalous values using information from neighboring periods, thereby preserving the temporal continuity, trend, and seasonal structure of the series. Consequently, the anomalous periods remain identifiable and retain their practical relevance for policy evaluation and crisis monitoring, while the reduced influence of extreme observations helps the forecasting model learn the underlying data pattern more effectively and produce more stable predictions.

Table 4 shows the COPOD scores for the anomaly periods along with the values of international tourists before and after handling. The neighbor mean method was used by replacing a single anomalous value at time t with the average of the two surrounding observations. Sequential anomalies of two to four periods were handled through time interpolation, distributing proportional changes based on the time distance between the two surrounding normal points. Extreme spikes or drops were smoothed without changing the overall trend direction, so data continuity remained intact and was more stable for LSTM forecasting.

Data Forecasting with LSTM

LSTM was applied to two different sets of international tourists data, namely before and after processing, with 276 training periods from January 2000 to December 2022. LSTM was used based on its ability to capture short and long-term dependencies in nonlinear time series. The architecture consisted of two hidden layers with 128 and 64 neurons, dropout of 0.2, a loss function of MSE, and Adam optimizer. The use of two hidden layers with different neurons aims to enhance the representation of nonlinear patterns in the time series data and reduce noise in the data (Aldi & Nathasia, 2025).

The selection of the hyperparameter combination in Table 2 was performed through 5-fold time series cross-validation using an expanding window approach. The same best scenario was obtained in both datasets based on the lowest average RMSE, ensuring fair model comparison. The combination consisted of 150 epochs, a learning rate of 0.01, and a batch size of 32 in scenario 6 with RMSE values of 8574 and

8411, respectively. The model was retrained with all training data and validated using test data to predict the period from January 2023 to December 2025.

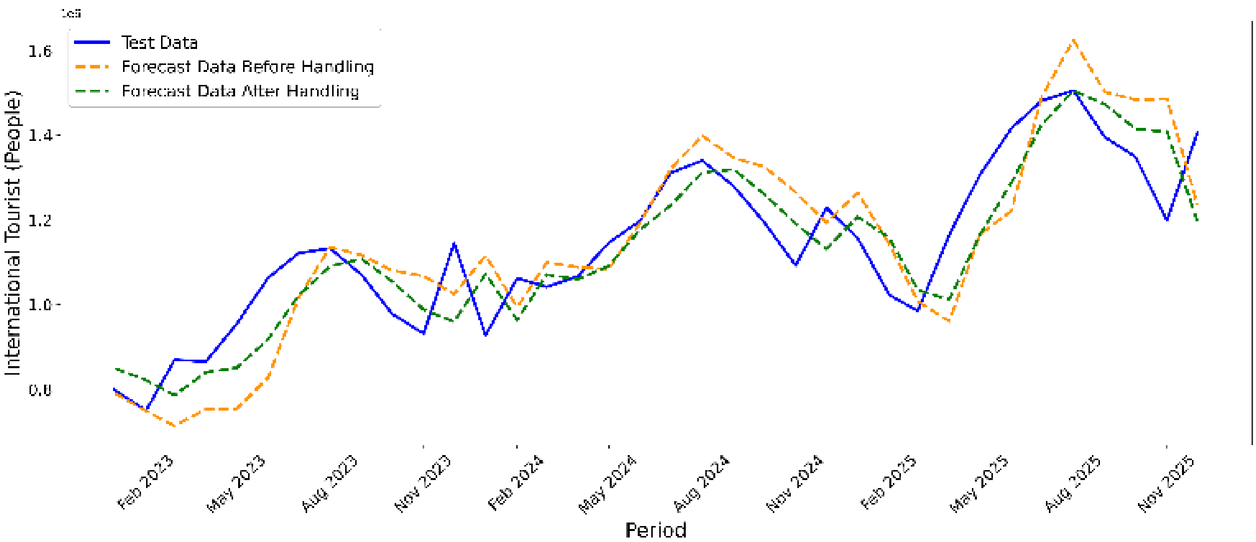


Figure 6. Forecasting graph of test data

Visual evaluation in Figure 6 shows that both models are able to follow the upward trend after the COVID-19 recovery. The model after anomaly handling produces a smoother and more stable prediction pattern, with more consistent around peaks and turning points, although occasionally still lower than the actual data. The model before handling tends to deviate more during sharp changes and shows higher predictions at some peaks. The presence of anomalies, especially in the form of extreme values, can cause large prediction errors in square-based loss functions, thus dominating the LSTM learning process. This condition causes the gradients in backpropagation to become unstable, as well as triggers fluctuations in validation loss, slowing down model convergence (Yidan et al., 2023). These results indicate that the presence of anomalies affects the model's sensitivity to certain variations.

Table 5. LSTM evaluation metrics

Accuracy	Before Handling		After Handling	
	Train Data	Test Data	Train Data	Test Data
MAPE (%)	11.921	9.225	9.483	7.494
RMSE	72878	126214	64409	99233
Correlation	0.984	0.838	0.980	0.864

The evaluation metrics in Table 5 show that anomaly handling provides better performance on both training and test data with different characteristics. MAPE and RMSE on the training data decreased by 2.483% and around 8.5 thousand, respectively, although the correlation slightly decreased by 0.004. These results indicate that the historical patterns are maintained while the predictions become numerically closer to the actual values. A more noticeable difference is seen on the test data, with MAPE decreasing from 9.225% to 7.494% and RMSE from 126214 to 99233, accompanied by an increase in the correlation between actual data and predictions on the test data from 0.838 to 0.864. The consistent decrease in error and increase in correlation on the test data indicate that the model trained after anomaly handling has better generalization. The presence of anomalies in the training data causes the model to learn incorrect representations, resulting in decreased prediction performance on new data (Choi et al., 2021).

The improvement in model performance is noticeable after anomaly handling, but its effectiveness cannot be separated from the accuracy of the detection stage using COPOD. Without accurate identification of anomalous periods, handling risks is applied to normal observations, which can obscure the data structure. COPOD plays a role in isolating periods that truly deviate from the historical distribution of many variables, so interventions are carried out selectively based on cross-variable information. Beyond forecasting, the detected anomalies provide practical insights for policymakers by identifying periods of exceptional disruption in tourism activity, thereby supporting policy evaluation, risk monitoring, and the formulation of more responsive tourism and economic policies.

CONCLUSION

Conclusion

This study identified eleven periods of anomalies in international tourist arrivals based on information from tourist arrivals, inflation, and the exchange rate of the rupiah using the COPOD method. The anomalies detected correspond to periods of major disruptions and structural changes in tourism activity. Handling these anomalies improved the performance of LSTM forecasting, reducing the MAPE and RMSE of the test data to 7.494% and 99.233, and increasing the correlation between actual and predicted data to 0.864, meaning it resulted in better model generalization. These findings suggest that accurate anomaly identification can increase forecast reliability by reducing the impact of extreme observations on model learning. However, the results still depend on the selected macroeconomic variables, anomaly threshold specifications, and smoothing procedures, which can affect detection sensitivity and potentially dampen extreme events. From a practical perspective, the proposed framework can support tourism forecasting systems and provide policymakers with a tool for distinguishing exceptional disruptions from normal fluctuations, thereby improving tourism planning, policy evaluation, and crisis-response decision making.

Suggestions

Future research is expected to incorporate additional relevant variables that may influence international tourist arrivals, such as transportation indicators, economic growth, mobility data, or tourism policy variables, in order to capture more complex relationships within multivariate time series data. Furthermore, the development of temporal feature engineering, including adaptive lag structures, rolling statistics, and dynamic seasonal representations, is recommended to better represent temporal dependency and structural changes in anomaly detection.

BIBLIOGRAPHY

Ahmar, A. S., Idrus, S. A., & Asmar. (2024). Analyzing Rupiah-USD Exchange Rate Dynamics: A Study with ARCH and GARCH Models. *JOIV International Journal on Informatics Visualization*, 8(3–2), 1802. https://doi.org/10.62527/joiv.8.3-2.3251

Aldi, F. A. R., & Nathasia, N. D. (2025). Prediksi harga dan kinerja aset Bitcoin menggunakan algoritma Long Short-Term Memory. *JURNAL FASILKOM*, 15(1), 68–76. https://doi.org/10.37859/jf.v15i1.8902

Arimie, C. O., Biu, E. O., & Ijomah, M. A. (2020). Outlier detection and effects on modeling. *OALib*, 07(09), 1–30. https://doi.org/10.4236/oalib.1106619

Atkinson, A. C., Riani, M., & Corbellini, A. (2021). The Box–Cox Transformation: Review and extensions. *Statistical Science*, 36(2). https://doi.org/10.1214/20-sts778

Choi, K., Yi, J., Park, C., & Yoon, S. (2021). Deep

Learning for Anomaly Detection in Time-Series Data: Review, analysis, and guidelines. *IEEE Access*, 9, 120043–120065. https://doi.org/10.1109/access.2021.3107975

Fitri, A., Purnamasari, I., & Siringoringo, M. (2019). Peramalan Jumlah Wisatawan Mancanegara Menggunakan Model ARIMA. Jurnal Statistika Universitas Muhammadiyah Semarang, 7(1), 14–19. https://jurnal.unimus.ac.id/index.php/statistik/article/view/4781/

Genest, C., Okhrin, O., & Bodnar, T. (2023). Copula modeling from Abe Sklar to the present day. *Journal of Multivariate Analysis*, 201, 105278. https://doi.org/10.1016/j.jmva.2023.105278

Horváth, G., Kovács, E., Molontay, R., & Nováczki, S. (2020). Copula-Based anomaly scoring and localization for Large-Scale, High-Dimensional continuous data. *ACM Transactions on Intelligent Systems and Technology*, 11(3), 1–26. https://doi.org/10.1145/3372274

Iwueze, I. S., Nwogu, E. C., Nlebedim, V. U., Nwosu, U. I., & Chinyem, U. E. (2018). Comparison of methods of estimating missing values in time series. *Open Journal of Statistics*, 08(02), 390–399. https://doi.org/10.4236/ojs.2018.82025

Karnick, S., Lakshminarayanan, S., Palati, M., & R, P. (2025). Impact of outlier detection techniques on time-series forecasting accuracy for multi-country energy demand prediction. *International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering*, 15(6), 5067. https://doi.org/10.11591/ijece.v15i6.pp5067-5079

Khoiriyah, N. S., Silfiani, M., Novelinda, R., & Rezki, S. M. (2023). Peramalan Jumlah Penumpang Kapal di Pelabuhan Balikpapan dengan SARIMA. *Jurnal Statistika Dan Komputasi*, 2(2), 76–82. https://doi.org/10.32665/statkom.v2i2.2303

Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020). COPOD: Copula-Based Outlier Detection. *IEEE International Conference on Data Mining*, 2020-November, 1118–1123. https://doi.org/10.1109/ICDM50108.2020.00135

Lisnawati, A., Alam, S., & Kurniawan, I. (2025). Aplikasi Prediksi Kedatangan Wisatawan Mancanegara Ke Indonesia Dengan Algoritma Long Short-Term Memory (LSTM). *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(5), 8343–8349. https://doi.org/10.36040/jati.v9i5.15055

Maharani, N. S., Angraini, Y., Rahmawan, M. A., Putri, O. A., Kurniawan, S., Safitri, T. A., Rizki, A., Ningsih, W. a. L., Hidayatulloh, N. G. T., & Ratnasari, A. P. (2023). Aplikasi Model Arima Garch Dalam Peramalan Data Nilai Tukar Rupiah Terhadap Dolar Tahun 2017-2022. *Jurnal Matematika Sains Dan Teknologi*, 24(1), 37–50. https://doi.org/10.33830/jmst.v24i1.4875.2023

Nabila, J., & Munifatussa'idah, A. (2022). How Do Trade, Inflation, Exchange Rates, And Information Technology Influence International Tourist Visits In Indonesia? *Airlangga Journal of Innovation Management*, 3(1), 1–17. https://doi.org/10.20473/ajim.v3i1.36922

Nugraha, Y. R. P., & Naylah, M. (2023). Indonesian Tourism Demand by ASEAN Tourist: A Panel Data analysis.

Signifikan Jurnal Ilmu Ekonomi, 12(1), 45–56. https://doi.org/10.15408/sjie.v12i1.29894

Nurhambali, M. R., Angraini, Y., & Fitrianto, A. (2024). Implementation of Long Short-Term Memory for Gold Prices Forecasting. *Malaysian Journal of Mathematical Sciences*, 18(2), 399–422. https://doi.org/10.47836/mjms.18.2.11

Pranolo, A., Setyaputri, F. U., Paramarta, A. K. I., Triono, A. P. P., Fadhilla, A. F., Akbari, A. K. G., Utama, A. B. P., Wibawa, A. P., & Uriu, W. (2024). Enhanced multivariate time series analysis using LSTM: A comparative study of Min-Max and Z-Score normalization techniques. *ILKOM Jurnal Ilmiah*, 16(2), 210–220. https://doi.org/10.33096/ilkom.v16i2.2333.210-220

Rasyidi, M. A. (2024). Resilience and Recovery: Clustering international tourism trends in Indonesia amidst COVID-19. *Sustainability and Social Impact*, 1(2), 29–38. https://doi.org/10.64849/ssi.v1i2.52

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series. *Proceedings of the VLDB Endowment*, 15(9), 1779–1797. https://doi.org/10.14778/3538598.3538602

Shahrestani, S., Carranza, E. J. M., & Sanislav, I. (2026). Geochemical anomaly delineation utilizing copula-based outlier detection method. *Applied Computing and Geosciences*, 29, 100325. https://doi.org/10.1016/j.acags.2026.100325

Sutisna, E., Dafitri, R., & Daryani, D. (2026). Forecasting foreign tourist arrivals in Indonesia using Google Trends Index as exogenous variable: A comparative study of SARIMAX and LSTM models. *Journal of Applied Informatics and Computing*, 10(2), 1864–1871. https://doi.org/10.30871/jaic.v10i2.12384

Syahzaqi, I., Sediono, S., Oktavia, S. S., Anggakusuma, A. C., & Wieldyanisa, E. E. (2025). Peramalan Jumlah Barang Kereta Api di Indonesia Menggunakan Metode Seasonal Autoregressive Integrated Moving Average (SARIMA). *Jurnal Statistika Dan Komputasi*, 4(1), 13–22. https://doi.org/10.32665/statkom.v4i1.4424

Syamad, & Handoyo, R. D. (2023). Analysis of symmetric and asymmetric effects of exchange Rate Pass-Through in Inflation-Targeting Countries. *Jurnal Ilmu Ekonomi Terapan*, 8(2), 312–337. https://doi.org/10.20473/jiet.v8i2.45150

Taylor, S. J., & Letham, B. (2017). Forecasting at scale. *The American Statistician*, 72(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

West, R. M. (2021). Best practice in statistics: The use of log transformation. *Annals of Clinical Biochemistry International Journal of Laboratory Medicine*, 59(3), 162–165. https://doi.org/10.1177/00045632211050531

Wicaksono, B. B. (2022). Sebuah analisis Pengaruh Inflasi, Suku Bunga dan Nilai Tukar Terhadap Perkembangan Pariwisata di Indonesia. *Journal Economics And Strategy*, 3(1), 13–23. https://doi.org/10.36490/jes.v2i2.277

Yidan, X., Shaolin, H., & Guotao, Y. (2023). Analysis and Improvement Approach of the impact of data disturbance on LSTM Prediction Algorithm. *Transactions on*

Machine Learning and Artificial Intelligence, 11(5). https://doi.org/10.14738/tecs.115.15411

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PYOD: A Python toolbox for scalable outlier detection. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1901.01588