ADI Journal on Recent Innovation (AJRI)
Vol. 5 No. 1 September 2023

p-ISSN : 2685-9106
e-ISSN : 2686-0384

Predicting Air Quality Index using Ensemble Machine Learning

Marviola Hardini[1], Richard Andre Sunarjo[2], Marsani Asfi[3], Mochamad Heru Riza Chakim[4], Yulia Putri Ayu Sanjaya[5]
University of Raharja[1, 4, 5], Pelita Harapan University[2], Catur Insan Cendekia University[3]
Jln. Jenderal Sudirman No.40, RT.002/RW.006, Cikokol, Kec. Tangerang, Kota Tangerang, Banten[1, 4, 5],
Jln. M.H. Thamrin Boulevard No.1100, Kelapa Dua, Tangerang Regency, Banten[2]
Jln. Kesambi No. 202, Drajat, Kec. Kesambi, Cirebon, West Java[3]
Indonesia[1, 2, 3, 4, 5]

e-mail: marviola.hardini@raharja.info, rasusi2019@gmail.com, marsani.asfi@cic.ac.id, heru.riza@raharja.info, yulia.putri@raharja.info

Author Notification
03 July 2023
Final Revised
01 August 2023
Published
22 August 2023

To cite this document:
Hardini, M., Sunarjo, R. A., Asfi, M., Riza Chakim, M. H., & Ayu Sanjaya, Y. P. (2023). Predicting Air Quality Index using Ensemble Machine Learning. ADI Journal on Recent Innovation, 5(1Sp), 78–86.
DOI: https://doi.org/10.34306/ajri.v5i1Sp.981

Abstract

This study focuses on utilizing Ensemble Machine Learning to predict and forecast the Air Quality Index (AQI). The research is motivated by the adverse effects of industrialization and population growth on air quality, leading to detrimental impacts on human health. While numerous methodologies exist for air quality prediction, it is crucial to anticipate future air conditions to minimize their larger consequences. Hence, this study proposes an air quality evaluation system to facilitate future predictions. The study comprises three primary modules: Data Preparation, AQI Forecasting, and Evaluating Air Quality. The Data Preparation Module involves real-time data collection and formatting to ensure compatibility with subsequent modules. In this research, the Sparse Spectrum GPR (SSGPR) method is employed for AQI forecasting, while the cloud model is adopted for air quality evaluation. The study's findings demonstrate the capability of the proposed model to account for the uncertainty and randomness inherent in air quality prediction. Performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) are employed to evaluate the models' effectiveness. Based on the evaluation results, it can be concluded that the Ensemble Machine Learning method utilized in this study effectively predicts and forecasts the Air Quality Index. These predictions play a crucial role in minimizing the adverse impact of air pollution on human health by providing insights into future air conditions. Overall, this research contributes significantly to comprehending and addressing the increasingly urgent challenges associated with air quality.

***Keywords**: Air Quality Index (AQI), Ensemble Machine Learning, Industrialization and Air pollution*

1. Introduction

In an era of ever-evolving industry, rapid population growth is a significant challenge. The increase in population not only has an impact on economic and social development, but also contributes to pressing environmental problems. The increased use of fossil fuels, high levels of transportation, and deforestation as a result of urbanization and industrial growth all play a role in increasing levels of environmental pollution[1].

Air pollution is one of the most detrimental results of human activity. The process of burning fossil fuels in motor vehicles, industry, and power plants results in the emission of various pollutants into the atmosphere. Harmful particles, toxic gases, and volatile organic compounds released into the air slowly settle in the environment around us. This air pollution has a serious impact on human health[2]. Exposure to air pollution in the long term can cause various diseases such as heart disease, respiratory problems, and even lung cancer.

To address air pollution problems and protect public health, a deep understanding of current air quality and the ability to predict future air conditions are needed[4]. In this context, this study proposes an innovative and effective evaluation system to measure and forecast the Air Quality Index (AQI) using the Ensemble Machine Learning approach[5].

AQI is a parameter used to evaluate air quality based on the concentration of certain pollutant substances contained in it[6]. In this study, the Ensemble Machine Learning approach will be used to integrate historical air quality data, weather data, and other environmental factors to obtain accurate AQI predictions in the future. The Ensemble Machine Learning approach is a method that combines several machine learning algorithms, where each algorithm makes its own contribution in obtaining better predictive results. By using this technique, it is hoped that the proposed evaluation system will be able to provide reliable and useful results for decision makers and related parties in taking steps to mitigate and control air pollution[7].

Through this research, it is hoped that concrete steps will be created in reducing the negative impact of air pollution. By understanding future patterns and trends in air quality, we can take appropriate and effective actions to protect our environment and safeguard public health. This research also contributes to the development of science and technology, especially in the field of air quality prediction using the approach[8].

2. Related Work

This research presents a hybrid air quality warning system that incorporates forecasting and evaluation. First, they offer a hybrid forecasting model based on "decomposition and ensemble" theory combined with a sophisticated data pre-processing approach, with the supporting vector engine playing an important role in this system. Furthermore, to supplement this research, they used fuzzy assessment, which is very significant in warning systems. Forecasting models and fuzzy evaluation methodologies complement each other. The experimental findings demonstrate that the proposed method is considerably superior in terms of accuracy and efficacy in assessing air quality. Furthermore, the application of forecasting and evaluation allows for precise and useful forecasts of future air quality, which provides considerable benefits[9].

This study creates a dynamic evaluation model based on a fuzzy synthetic evaluation approach in order to swiftly evaluate future air quality. Using a newly created computational intelligence optimization approach, they improved the least squares SVM, which estimates six air pollutant concentrations. The fuzzy synthetic assessment model with entropy weights is used to predict future air quality conditions. The findings and analysis of air quality demonstrate that estimations of urban air pollution concentrations and air quality conditions can be objectively and reliably measured. This demonstrates that the suggested dynamic scoring model, through simulation design, might be a valuable tool for air quality monitoring[10].

The proposed system integrates complexity analysis, data pre-processing, and the estimated-optimal module. The suggested system uses a modified least squares support vector engine to perform complexity analysis on the original time series, which is then optimized using a multi-objective multiverse optimization algorithm[11]. The system then forecasts the AQI series every hour with a modified least squares support vector engine and optimizes it with a multi-objective multiverse optimization method. Experiments using datasets from eight major

Chinese cities show that the suggested system can achieve high accuracy and stability for air quality monitoring in an efficient manner.

This research presents a new hybrid model that blends outlier detection and correction approaches with heuristic intelligent optimization tactics[12]. To begin, they employ a data pre-processing technique to discover and correct outliers while revealing the primary properties of the original time series. They then employ an algorithm.

This study describes a new hybrid air quality early warning system with three modules: data pre-processing, forecasting, and air quality evaluation[13]. To extract chaotic features from raw data, a new hybrid data pre-processing technique is utilized, resulting in a more stable collection of pollution data for forecasting. To increase the forecasting module's accuracy and stability, a multi-purpose locust optimization technique is applied. To offer thorough results, a fuzzy air quality evaluation module is also provided. This forecasting approach not only outperforms earlier models in terms of accuracy and stability, but the assessment module also delivers valid air quality data[14].

To extract critical elements from complicated spatiotemporal interactions and limit error accumulation and propagation, this study offers a Deep Multi-output LSTM (DM-LSTM) neural network model with three deep learning algorithms (mini-batch gradient descent, neuron dropout, and L2 regularization). in air quality predictions with multiple steps ahead. Three time series of PM2.5, PM10, and NOx were evaluated simultaneously at five air quality monitoring stations in Taipei City, Taiwan, to put the proposed DM-LSTM model to the test. The suggested DM-LSTM model increased the spatial-temporal stability and accuracy of multi-step-ahead air quality forecasting at the regional level by merging three deep learning algorithms[15].

This study introduces a novel Gaussian Process (GP) regression model that is rarely used. The key idea is to narrow the GP's spectral representation, resulting in a direct and realistic regression approach[16]. This study reveals that there is a trade-off between prediction accuracy and processing requirements, and that these models are typically better than today's more advanced sparse regression models. In the framework of weight space and function space representation, this novel design prioritizes functions that are always stationary and can approach various forms of covariance functions within their class[17].

3. Proposed Work

3.1 Data Preparation

This module's goal is to collect real-time data from sensors and calculate AQI in real time[18].

3.1.1 Data collection

The primary aim in this experiment is to configure the sensors. To get a location, this study used unique sensors for each air pollutant in conjunction with a GPRS connection. The Arduino MKR1000 board is used to create efficient IoT applications that include on-board wifi.

After configuring the IoT device, the sensor is connected to the Arduino web editor to be programmed. To activate the sensor, the sketch must be written in the Arduino web editor and uploaded to the MKR1000 board. When the upload is finished, the sensor begins reading the data. The following major step is to save the obtained real-time data. As a result, the Thingspeak platform is used as an efficient data storage in this study[19].

However, only the principal pollutant concentrations were included in the data repository, notably PM2.5, PM10, SO2, O3, NO2, NH3, and CO. The information is solely for the pollutants that are saved in the Thingspeak channel. After collecting and updating real-time data in Thingspeak, the next step is to compute AQI[20].

3.1.2 AQI calculation

The AQI is calculated by computing the sub-index for each pollutant and finding the greatest value between them. The sub-index is calculated according to a predetermined formula. The maximum value of all the sub-indices obtained is termed the AQI. This calculation's drawing will be implemented and combined with the previous stage. When the entire sketch is uploaded to the Arduino board, the sensor will read the real-time data and simultaneously

update the data to the Thingspeak channel, along with the real-time AQI calculations. As a result, the preliminary data for prediction will be prepared and presented as input for the forecasting module[21].

3.2 AQI calculation

The forecasting module in this study attempts to forecast present and future AQI trends. This module employs SSGPR, an efficient real-time learning approach. SSGPR is a probabilistic and non-parametric variation of GPR (Gaussian Process Regression). Because of its increased computing performance, SSGPR was chosen to analyse real-time data and perform real-time AQI predictions. The GPR model is improved by modifying the GPR power spectrum and utilizing a sparse approximation[22].

SSGPR's major purpose is to discover the best amount of frequencies. After determining the appropriate frequency, the initial mean and variance of the new data can be computed. The uncertainty in the new output can be assessed using the interval (UB, LB) based on the mean and variance.

3.3 AQI calculation

In this study, air quality is evaluated using cloud models that are based on predictive values. Using decision-making procedures, this model may convert quantitative data into qualitative data. The cloud model is based on three key parameters: Ex, En, and HE. These three characteristics are useful for simulating data ambiguity and randomness[23].

- Ex – The expected value of the qualitative concept.
- En – Represents ambiguity in a qualitative concept.
- He – Randomness in a qualitative concept.

The air quality evaluation process involves the following series of steps:

1. Select the pollutant to be considered for evaluating air quality.
2. Apply cloud models for each pollutant at each level.
3. Using the entropy-based Super scale weighting method to calculate the weight of each pollutant.
4. Based on the weights obtained in the previous step, it produces a degree of certainty for the pollutant to determine the level of air quality.
5. By calculating the average of the results from the previous steps, the final degree of certainty can be determined.
6. Observing the degree of certainty at each level, consider the level that has the maximum degree of certainty as the designated air quality level.

4. discussion result

4.1 Reliability Analysis

This study uses Cronbach's α to test the reliability of the results. Based on his exams, Cronbach's α values for the 5 constructs ranged from 0.8 to 0.9, indicating that the results were higher than the standard value, which was 0.6[24].

In this study, it has 5 variables consisting of population, vehicle use, air quality, ensemble machine learning and data preparation (PD) along with the framework images that will be examined in this study.

p-ISSN : 2685-9106
e-ISSN : 2686-0384

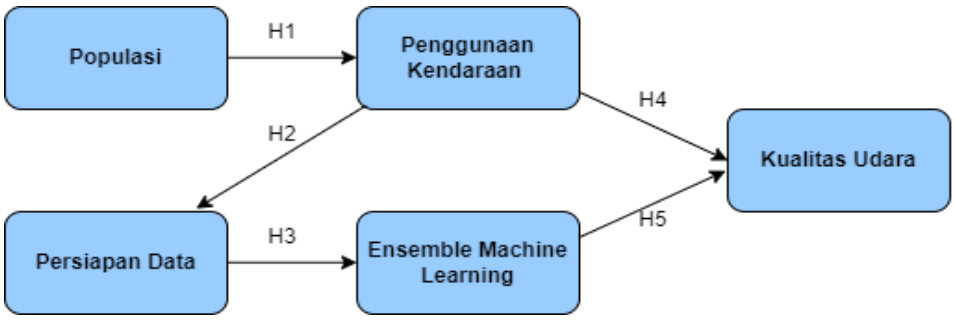

Figure 1. Research Framework

4.2 Measurement Model

The variables used in this study were obtained through indicators consisting of questionnaires. To ensure the validity of the resulting data, a truth or validity test was carried out on the two components to test the construct validity. The level of construct convergence reliability and validity will be higher if the AVE (Average Variance Extracted) value is also high. Overall, the measurement model shows adequate reliability, good convergent validity, and adequate discriminant validity[25].

4.3 Structural Model

The research variables used in this study were the results of a collection of indicators gathered through the distribution of questionnaires [26]. As a result, the generated data must be examined for the truth or validity of the two components in order to determine construct validity. The loading factor and AVE with a value of 0.5 determine the first component, convergent validity. In this study, two metrics were utilized to test dependability: composite reliability and Cronbach alpha [27]. Cronbach alpha must be greater than 0.6 and composite reliability must be greater than 0.7. If the reliability value (alpha) is more than the specified threshold of dependability, the computation results can be regarded as a measurement tool with high precision and consistency of thought [28].

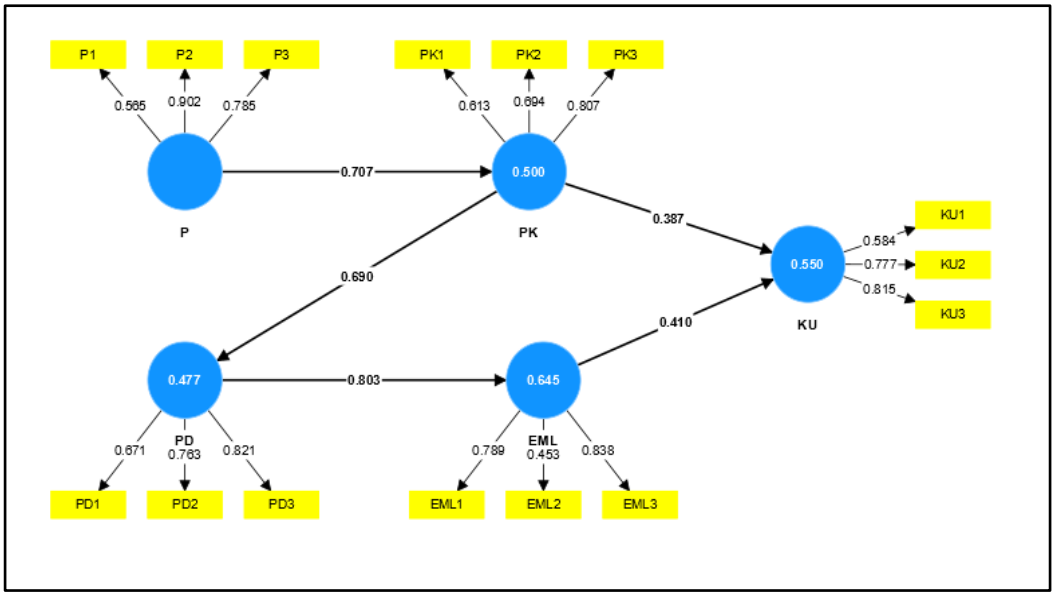

Figure 2. Structural Model Results

p-ISSN : 2685-9106
e-ISSN : 2686-0384

Tabel 1. AVE Result Display

Nama Variable	Average Variance Extracted
Population	0.583
Vehicle Use	0.503
Air Quality	0.536
Machine Learning Ensembles	0.510
Data Preparation	0.569

Based on table 1, the AVE results for each variable have met the value above 0.5

Table 2. Reliability Test Results

Nama Variable	Cronbach's Alpha	Composite Reliability (rho_a)	Composite Reliability (rho c)
Population	0.616	0.650	0.802
Vehicle Use	0.500	0.511	0.750
Air Quality	0.549	0.544	0.773
Machine Learning Ensembles	0.506	0.585	0.746
Data Preparation	0.616	0.622	0.798

Based on the data in table 2, it is clear that all variables meet the criteria for a Cronbach's alpha value greater than 0.6. Similarly, the composite dependability value for each variable fits the criteria for a value greater than 0.7. Overall, the measurement model (outer model) met the requirements, allowing this research to move on to the structural model stage (inner model) [29].

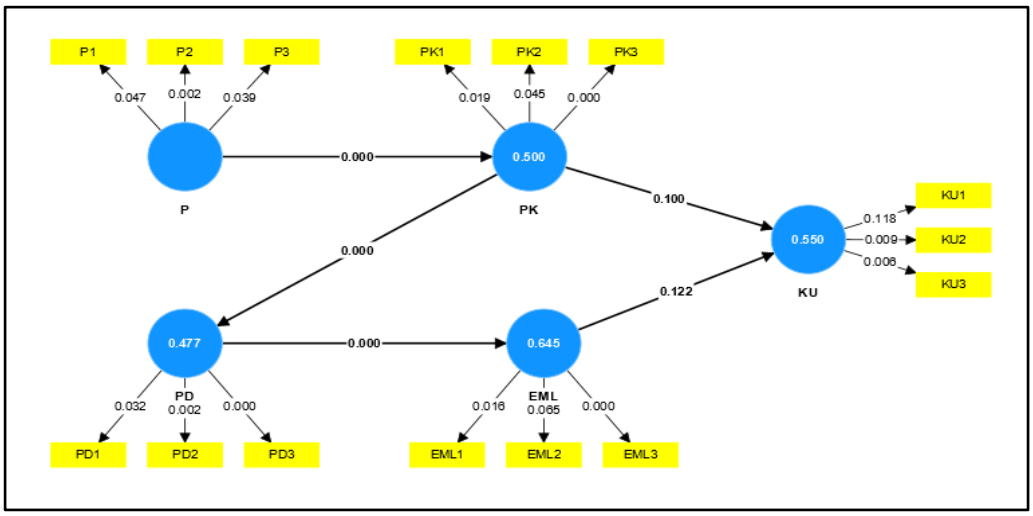

Figure 3. Path Coefficients results

p-ISSN : 2685-9106
e-ISSN : 2686-0384

Table 3. Path Coefficients Test Results

Nama Variable	Koefisien	T value	p value	Result
Ensemble Machine Learning -> Air Quality	0.265	1.546	0.122	Not Affect
Population -> Vehicle Users	0.114	6.201	0.000	Influence
Data Preparation -> Ensemble Machine Learning	0.119	6.762	0.000	Influence
Vehicle Users -> Air Quality	0.235	1.647	0.100	Not Affect
Vehicle User -> Data Preparation	0.176	3.934	0.000	Influence

From the results of the analysis using SmartPLS, it can be observed that the p-value is <0.01, which indicates that each variable has a significant level. However, based on existing data, it appears that the relationship between Social Norms and Behavior, as well as Self Monitoring and Self Efficiency in Air Pollutants does not show a significant effect [30].

5. Conclusion

The suggested system in this work is made up of three modules: data preparation, forecasting, and air quality evaluation. Sensors capture real-time data, and AQI is generated in real-time depending on the data. Furthermore, the Sparse Spectrum GPR model is trained to analyze real-time data while accounting for uncertainty, and it is utilized to forecast AQI. Furthermore, cloud models that can deal with ambiguity and randomization are employed to assess air quality.

As a result, the suggested model is capable of predicting and forecasting future AQI while accounting for the uncertainty, fuzziness, and randomness of real-time data. Metrics such as MAE, RMSE, and MAPE are used to assess model performance. More emphasis can be placed in future research on incorporating data normalization approaches to increase AQI prediction accuracy.

References

[1] Rahardja, U., Aini, Q., Manongga, D., Sembiring, I., & Girinzio, I. D. (2023). Implementation of Tensor Flow in Air Quality Monitoring Based on Artificial Intelligence. International Journal of Artificial Intelligence Research, 6(1).

[2] Rahardja, U., Aini, Q., Manongga, D., Sembiring, I., & Sanjaya, Y. P. A. (2023). Enhancing Machine Learning with Low-Cost P M2. 5 Air Quality Sensor Calibration using Image Processing. APTISI Transactions on Management (ATM), 7(3), 11-19.

[3] Lutfiani, N., Wijono, S., Rahardja, U., Iriani, A., Aini, Q., & Septian, R. A. D. (2023). A bibliometric study: Recommendation based on artificial intelligence for ilearning education. Aptisi Transactions on Technopreneurship (ATT), 5(2), 109-117.

[4] Aini, Q., Manongga, D., Sembiring, I., & Apriliasari, D. (2023). Transformation of payment in education use bitcoin with reduced confirmation times. Aptisi Transactions on Technopreneurship (ATT), 5(1), 1-8.

[5] Aini, Q., Harahap, E. P., Santoso, N. P. L., Sari, S. N., & Sunarya, P. A. (2023). Blockchain Based Certificate Verification System Management. APTISI Transactions on Management (ATM), 7(3), 1-10.

[6] Rahardja, U., Aini, Q., Sunarya, P. A., Manongga, D., & Julianingsih, D. (2022). The use of tensorflow in analyzing air quality artificial intelligence predictions pm2. 5. Aptisi Transactions on Technopreneurship (ATT), 4(3), 313-324.

[7] Rahardja, U. (2022). Camera Trap Approaches Using Artificial Intelligence and Citizen Science. International Transactions on Artificial Intelligence, 1(1), 71-83.

[8] Rahardja, U., Ngadi, M. A., Sutarman, A., Apriani, D., & Nabila, E. A. (2022, November). A Mapping Study Research on Blockchain Technology in Education 4.0. In 2022 IEEE Creative Communication and Innovative Technology (ICCIT) (pp. 1-5). IEEE.

[9] Rawat, B., Bist, A. S., Rahardja, U., Harahap, E. P., & Septian, R. A. D. (2022, October). Novel Framework to Define Extended & Mixed Reality for Online Learning. In 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1-4). IEEE.

[10] Agarwal, V., Lohani, M. C., Bist, A. S., Rahardja, U., Khoirunisa, A., & Octavyra, R. D. (2022, September). Analysis of Emerging Preprocessing Techniques Combined with Deep CNN for Lung Disease Detection. In 2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA) (pp. 1-6). IEEE.

[11] Suraya, P. A., Ramadhan, T., Lutfiani, N., Khoirunisa, A., & Rahardja, U. (2022, September). Blockchain, Information and Speculation Calculations in Indonesia: Recent Work. In 2022 10th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-8). IEEE.

[12] Gui, H., Rahardja, U., Yang, X., & Yan, Y. (2022). Ability Orientation or Good Character? Moderated Mediation Mechanism to Determine the Impact of Telepresence on Consumer Purchasing Intention in Cross-Border E-Commerce. Frontiers in Psychology, 13.

[13] Farida, I., Ningsih, W., Lutfiani, N., Aini, Q., & Harahap, E. P. (2023). Responsible Urban Innovation Working ith Local Authorities a Framework for Artificial Intelligence (AI). Scientific Journal of Informatics, 10(2), 121-126.

[14] Dewi, K. I. M., Narayana, I. W. G., & Rahardian, R. L. (2023). Application of Certification Management Information Systems at LSP Engineering Hospitality Indonesia. Aptisi Transactions on Technopreneurship (ATT), 5(3), 12-23.

[15] Manawar, A. (2023). An Innovative and Secure Platform for Leveraging the Blockchain Approach for Online Exams. Aptisi Transactions on Technopreneurship (ATT), 5(1), 99-108.

[16] Asmolov, A., & Ledentsov, A. (2023). Impact on Educational Effectiveness Using Digital Gamification. Startupreneur Business Digital (SABDA Journal), 2(1), 98-105.

[17] Septiani, N., Lutfiani, N., Oganda, F. P., Salam, R., & Devana, V. T. (2022, February). Blockchain technology in the public sector by leveraging the triumvirate of security. In 2022 International Conference on Science and Technology (ICOSTECH) (pp. 1-5). IEEE.

[18] Gunawan, I. K., Lutfiani, N., Aini, Q., Suryaman, F. M., & Sunarya, A. (2021). Smart Contract Innovation and Blockchain-Based Tokenization in Higher Education. Journal of Education Technology, 5(4), 636-644.

[19] Rahardja, U., Lutfiani, N., Aini, Q., & Annisa, I. Y. (2021). The potential utilization of blockchain technology. Blockchain Frontier Technology, 1(01), 57-67.

[20] Setyowati, W., Kurniawan, P. C., Mardiansyah, A., Harahap, E. P., & Lutfiani, N. (2021). The Role Of Duty Complexity As A Moderation Of The Influence Auditor's Professional Knowledge And Ethics On Audit Quality. Aptisi Transactions on Management (ATM), 5(1), 20-29.

[21] Janarthanan, R., Partheeban, P., Somasundaram, K., & Elamparithi, P. N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities and Society, 67, 102720.

[22] Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, *7*, 128325-128338.

[23] Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. Applied Sciences, 9(19), 4069.

[24] Jiao, Y., Wang, Z., & Zhang, Y. (2019, May). Prediction of air quality index based on LSTM. In 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) (pp. 17-20). IEEE.

[25] Manurung, E. B. (2023). Gantry Robot System Checkers Player. ADI Journal on Recent Innovation, 5(1Sp), 9-19.

[26] Wiwin, N., Sunarya, P. A., Azizah, N., & Saka, D. A. (2023). A Model for Determine Upgrades for MSMEs using Analitical Hyrarcy Process. ADI Journal on Recent Innovation, 5(1Sp), 20-32.

[27] Sulivyo, L., & Dewi, F. M. (2023). Strategy Management Analysis in the Face of Business Competition. ADI Journal on Recent Innovation, 5(1Sp), 1-8.

[28] Ginantra, N. L. W. S. R., Asana, I. M. D. P., Parwita, W. G. S., & Eriana, I. W. E. (2022). Mobile-based customers management system in ayunadi supermarket. ADI Journal on Recent Innovation, 4(1), 86-101.

[29] Kumgliang, O., Khamwon, A., & Buasri, R. (2022). Online Brand Community Components in Gadgets Industry, Thailand. ADI Journal on Recent Innovation, 4(1), 67-76.

[30] Basri, Y. Z., & Arafah, W. (2023). Determinant of Interest in Paying Zakat with Age as a Moderating Variable (Study on Minang Society). APTISI Transactions on Management (ATM), 7(2), 92-104.