Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Computer Science and Computational Intelligence 2025 (ICCSCI 2025)

Comparative Analysis of AI Models for Air Quality Index Prediction in Central Jakarta

Stevie Odie Patric Mulyono[a,*], Filbert Alfredo Saputro[a], Christian Gavriel Emanuel Hariyadi[a]

[a]*Bina Nusantara University, Jakarta, 11480, Indonesia*

Abstract

The percentage of the urban population is increasing due to urban migration, which presents a challenge especially regarding air quality in the city. Air quality needs to be monitored properly. A higher AQI brought on by these chemicals, such as PM10, PM2.5, CO, CO2, SOX, NOX, NH3, and O3, can result in serious health risks from highly contaminated air. AQI monitoring and forecasting have turned into a crucial tool. Machine learning methodologies produce highly accurate AQI predictions. This research employed dataset from Kaggle entitled "Air Quality Index in Jakarta," implemented Random Forest, CatBoost, and SVR, focusing on Bundaran HI with data preprocessing, model training, and evaluation. The models were evaluated based on several metrics, such as: MAE, RMSE, and R^2 score. This research found that CatBoost with MAE value of 0.0350, RMSE of 0.1041, and R^2 Score of 0.9559. Random Forest with MAE value of 0.0293, RMSE of 0.1153, and R^2 Score of 0.9459. SVR model with MAE of 0.0929, RMSE of 0.1791, and lower R^2 Score of 0.8694. So CatBoost model was the best performing choice and SVR model has the lowest accuracy.

Keywords: SVR; CatBoost; Random Forest; Air Quality Index (AQI)

[*] Corresponding author.
E-mail address: stevie.mulyono@binus.ac.id

1. Introduction

The percentage of the urban population is increasing due to urban migration. By 2050 the UN estimates that 68 percent of the global population will live in metropolitan areas [1]. Meanwhile, in Indonesia it is estimated that there will be an increase of around 66.6 percent by 2035, which means that only 33.4 percent are rural residents [2]. Urbanization is influenced by several factors such as: limited employment opportunities, low living standards, and minimal educational facilities in villages. The increase in individuals in the city presents a challenge, especially regarding air quality in the city.

Air quality needs to be monitored properly, so that it provides welfare for us. Some substances that are important contributors to Air Quality Index (AQI), such as: PM10, Particulate 2.5 (PM2.5), Carbon Monoxide (CO), Carbon Dioxide (CO2), Sulfur Oxides (SOX), Nitrogen Oxides (NOX), Ammonia (NH3), Ozone (O3), and other substances. A higher AQI brought on by these chemicals harms the environment in many ways, such as acid rain, global warming, smog and aerosols, reduced visibility, and climate change [3]. Serious health risks can result from highly contaminated air, as indicated by a high AQI number. This cause has the potential to induce a wide array of severe pathological conditions in humans, ranging from bronchitis to cardiovascular disease, from pneumonia to pulmonary malignancies, among others [4].

High AQI indicated that our environment fills with tons of hazardous pollution, these things significantly can endanger human health. Consequently, AQI monitoring and forecasting have turned into a crucial tool for international sustainable health development [5]. Various models for the AQI prediction have been developed based on statistics, machine learning, and deep learning methodologies. These AI-Based predictive models work by analyzing data to predict pollution and help to improve the air quality of the surrounding area. AI-powered prediction tools are reliable, affordable, straightforward, and quick to deploy [6]. This sensor development made it to simply identify different air pollution levels, and it will automatically calculate the AQI. Machine learning methodologies produce highly accurate AQI predictions across a wide range of environmental health, making it an effective alternative to standard statistic approaches in data analysis.

The deterioration of ambient air quality, particularly in metropolitan areas, necessitates substantial enhancements. In Jakarta, air pollution constitutes one of the numerous grave environmental challenges arising from population expansion, heightened economic activity, and the concomitant transportation and industrial endeavors. This research employed an open-source dataset available on Kaggle entitled "Air Quality Index in Jakarta," authored by Taufiq Pohan, and implemented three machine learning algorithms, specifically Random Forest, CatBoost, and Support Vector Regression (SVR). These machine learning techniques exhibit robust efficacy across a variety of dataset characteristics, problem specifications, and anticipated results. Random Forest is adept at uncovering significant interactions and non-linear effects among the predictors. CatBoost represents an application of gradient boosting, utilizing binary decision trees as foundational predictors. With its relative simplicity and versatility in addressing a spectrum of classification challenges, SVR uniquely provide minimizing the generalized error bound rather than the observed training error to get generalized performance [7]. In light of these benefits, the current study adopts these models.

2. Background Study

Air pollution has been widely studied due to its significant impacts on human health and environmental sustainability. Air quality monitoring and prediction techniques have been the subject of several studies, with an increasing focus on models based on artificial intelligence (AI). This section examines pertinent research on predicting air quality, the use of AI in environmental monitoring, and the relative effectiveness of various machine learning models.

2.1 Air Quality Index and Its Significance

From year-by-year attention to risk factor of air quality pollution more increasing, such as respiratory and cardiovascular morbity disorders become more serious. This ambient could cause 254,000 deaths globally in 2015. Tropospheric O3, PM2.5, and NO2 are the primary air pollutants posing the most significant risk to human health, particularly in urban areas [8].

Because of this serious factor USEPA developed a tool named Air Quality Index to give accurately information about air pollution levels. The AQI compare of the actual levels by measuring carbon monoxide, lead, nitrogen dioxide, ozone, particulate matter, and sulfur dioxide all used as the basis for the AQI [9]. USEPA used a standard method measurement for calculating the Air Quality Index, by collected data through Federal Reference Methods and Federal Equivalent Methods, to gain more accuracy data. In Modern method, AQI used various machine learning techniques to analyze this air pollution it use algorithms such as Deep Learning and Machine Learning.

The Main function of the AQI is its analysis in air health monitoring, but beside its useful, AQI has faced an issue where its limitation in monitoring the effect of pollutants and failing to reflect non-threshold concentration-response relationships. To prevent this issue, AQI has been integrating multiple pollutant into a single metric health to provide more comprehensive risk assessment.

2.2 Previous Approaches to AQI Prediction

Autoregressive Integrated Moving Average (ARIMA) models and Multiple linear regression (MLR) are two statistical techniques commonly used in conventional AQI prediction methods for air pollution forecasting. Previous studies have used ARIMA time series models to forecast air pollution levels in Nanded city. The prediction results from the ARIMA model show that air pollution levels in Nanded city are increasing. This study emphasizes the importance of tracking air pollution levels in the environment within permissible limits [10]. On the other hand, there is a study by creating a seasonal MLR model in Chiang Rai Province to forecast PM10 levels according to climate variables. This study emphasizes the importance of taking seasonal fluctuations into account when using MLR to anticipate air quality [11].

However, this previous approach has limitations. This ARIMA's inability to handle nonlinear patterns could result in inaccurate forecasts, particularly when unexpected events or major policy changes cause rapid changes in air quality [12]. Limitation of the MLR model is the linear representation of MLR to nonlinear systems, such as air pollution. Furthermore, modeling and forecasting are difficult and complex due to the complexity of natural air pollution, which includes its creation, transport, and release into the atmosphere [13]. Consequently, there is now a greater need for increasingly complex AI-driven prediction models.

2.3 AI-Based Models for AQI Prediction

Traditional method of AQI prediction has several limitations. It can't handle large datasets, can't capture complex patterns, and has less accuracy. Deep Learning and Machine Learning have become strong substitutes for AQI prediction due to the availability of large datasets and advancements in processing capacity. For this task, researchers train models using the following regression techniques:

1) Random Forest: In order to increase accuracy, the Random Forest model constructs with multiple decision trees by recursively dividing the provided dataset into two groups according to a certain criterion until a predefined stopping condition is satisfied [14]. The random forest algorithm determines the error rate more accurately than decision trees. More specifically, it has been shown analytically that as the number of trees increases, the error rate always converges. Random Forests do not gradually alter the training set, yet they produce results that are competitive with boosting and adaptive bagging [15]. When using data from urban sensing, Random Forest produces more accurate predictions [16].
2) Catboost: CatBoost is a Decision Tree-based method that works well for machine learning problems that use heterogeneous, categorical data. CatBoost's hyper-parameter sensitivity and the significance of hyper-parameter adjustment [17]. CatBoost regression had the lowest RMSE, MSE, and MAE values in this investigation, as well as the greatest R-SQUARE [18]. In both applications (or analysis), the CatBoost classifier performed admirably, earning the best score (accuracy).
3) Support Vector Regression: Support Vector Regression: SVR is based of high dimensional feature space from calculation of linear regression function, in which the nonlinear function maps the input data. SVR possess benefits in spaces with high dimensionality since SVR optimization does not depend on the dimensionality of the input space. SVR is characterized by its primary goal of minimizing the generalized error bound rather than the observed training error in order to get generalized performance [19].

2.4 Comparative Analysis of AI Models in AQI

Every AQI prediction used several statistical and machine learning. From the comparative analysis of AI models in CatBoost, RandomForest, and SVR each have advantages and limitations [20]. CatBoost is optimized for categorical data and gains high accuracy in AQI forecasting. This happens because CatBoost able to handle complex and overfitting data. CatBoost outperforms traditional machine learning models in dealing with small datasets and missing data. Research done in India proved that CatBoost regression provides the highest accuracies for New Delhi (85.0847%) and Bangalore (90.3071%). This demonstrated that datasets that had the SMOTE algorithm applied to them produced higher accuracy [18]. The random forest models that used ensemble-based decision tress, has faced a challenge with overfitting, especially when dealing with small datasets. This deficiency provides that the random forest gains the lowest root mean square error values in Bangalore (0.5674), Kolkata (0.1403), and Hyderabad (0.3826) [18]. The SVR models that used to predict AQI by using the concentrations of NO2, CO, O3, PM2.5, PM10, and SO2. SVR exhibited high performance in terms of the investigated measures of quality [20]. This model supervised machine learning algorithm that is used for regression problems. SVR shows accuracy in AQI prediction, particularly for short-term forecasting but it struggles with facing the large datasets. SVR works better because of the Mathematical models, learning, and regression techniques. Overall, the models for prediction like SVR, Random Forest, and CatBoost is suitable and the best fit line for predicting and monitoring the accuration of AQI.

2.5 Research Gap and Motivation

There are still several research gaps in AI-driven AQI prediction despite notable progress. First, there is still an opportunity for investigation in quickly urbanizing areas like Bundaran HI in Central Jakarta. Its strategic location is in the heart of the administrative and commercial area, surrounded by government institutions, large office buildings, shopping centers, and hotels. Second, not much research has directly compared the performance of CatBoost, Random Forest, and SVR under Jakarta-specific environmental circumstances and pollution sources, which are different from those in other cities that have been studied, such as India. This study aims to close these gaps by comparing the predicted accuracy, effectiveness, and interpretability of different models using an open-source AQI dataset from Jakarta.

3. Methodology

3.1 Data Collection and Preprocessing

This dataset will be utilized in this investigation from Kaggle with the author Taufiq Pohan. This data was collected from 5 areas in Jakarta, namely Kelapa Gading (North Jakarta), Bundaran HI (Central Jakarta), Lubang Buaya (East Jakarta), Kebon Jeruk (West Jakarta) , and Jagakarsa (South Jakarta). This study only focuses on Bundaran HI, therefore the data taken is only 4384 records. This dataset was compiled from 2010 to 2023. To provide in-depth knowledge of pollution patterns in Central Jakarta, this study views pollutant concentration values as an important aspect of the dataset. Pollutant concentrations consist of PM10, SO2, CO, O3, and NO2.

Data preprocessing is a crucial step in this study to improve the quality of datasets that will be used. This step can improve quality and reliability by reducing and inconsistency. The first stage of data preprocessing is handling missing value. Some of these data collection can have null or missing values, which will be handled to delete this row. These steps helped maintain the datasets by ensure no crucial value was missing value. If the data obtained is not subjected to data cleaning stages, this will cause changes in the results of the model, so this will have a significant impact on the results of the model predictions.

3.2 Utilized AI Models

This study will develop three machine learning models to predict the health categories within the scope of the air quality index. This model is used to compare which model is superior in predicting health and evaluate all performance differences. So that if you have obtained results that are faster, more efficient, and more accurate, it will be very helpful in maintaining air quality in Bundaran HI which is part of Central Jakarta. Here are some models that will be compared in this study:

1) Random forest is a variant of the ensemble model Begging. It is a supervised tree-based machine learning approach that is primarily utilized for multidimensional classification and regression problems. It may train independent decision tree (or forest) collections in parallel. Random Forest was part of Traditional decision trees, uses decision trees as weak learners. It has concept to increasing model diversity by randomly selecting features based on random sampling of samples. For instance, at each decision tree node, a subset of k features from the feature set is randomly selected, and an optimal feature is then chosen for splitting from this subset.
2) Support Vector Machines (SVM) are employed to address regression-related challenges within the framework of SVR. The Support Vector Machine represents a supervised learning methodology that delineates hyperplanes, serving as demarcations between diverse data points from which predictions can be derived. This methodology is applicable in the realms of classification, regression, and the identification of outliers. The SVR paradigm does not calculate the loss for each instance that resides within the hyperplane; rather, it formulates a hyperplane characterized by a margin of ϵ on either side of the linear function.
3) The Catboost algorithm is utilized for the construction of decision trees, derived from conventional gradient boosting methodologies. Boosting is founded on the premise that a collection of multiple, comparatively less robust models can be aggregated to form a singular, highly proficient predictive model that surpasses random guessing by a marginally significant extent. Gradient boosting mitigates inaccuracies through the sequential fitting of decision trees, wherein each tree acquires insights from the errors identified in the prior iteration. This methodology of incorporating additional functions into the ensemble is perpetuated until the specified loss function ceases to be minimized.

3.3 Model Training

The model training process is carried out on a Jupyter Notebook which serves as the development environment. This training is performed on a machine with Python 3.x installed along with Scikit-learn, NumPy, Matplotlib, CatBoost, and Pandas. The air pollutant feature dataset (PM10, SO2, NO2, CO, O3) and target variables (health) is preprocessed to balance the class distribution. The data is split in the ratio of 80% training and 20% testing to allow for adequate model validation while ensuring that all cases are represented in the sample. This splitting is done to prevent overfitting on test data and keep the statistical integrity of the data.

3.4 Model Implementation

Inside training the model, tuning hyperparameter should be done. This process needed so researchers can evaluate to optimize the model. Each hyperparameters combination is tested using cross-validation to find the best results. this will ensure the model performs on unseen data. The results of the tuning required using our model to gain accuracy in Machine Learning such as random forest, Catboost, and SVR are shown in Table 1 and Table 2.

Table 1. Hyperparameter Settings Random Forest and CatBoost

Category	Random Forest & CatBoost
Random Seed	55
Number Estimators	500
Kernel	N/A
epsilon	N/A
Gamma	N/A
Tree Depth	5

Table 2. Hyperparameter Settings SVR

Category	SVR
Random Seed	N/A
Number Estimators	N/A
Kernel	rbf
epsilon	0.05
Gamma	0.01
Tree Depth	N/A

4. Result Analysis

4.1 Data Preparation and Handling

In this research, the dataset focuses on the Air Quality Index in Central Jakarta, comprising several column attributes relevant for pollution levels. The dataset includes the "Tanggal" column which specific the date of data taken, and "Stasiun" column that monitoring the location where the data taken. It also contains specific pollutant concentration for “PM10”, “SO2”, “CO”, “O3”, “NO2”, and “PM2.5”, Each concentration representing the specific air pollutants measured in the air. Then the “Max” column shows the maximum value recorded among the pollutants and the “Critical” column identifies which pollutant had the most critical pollutant. Last the “Categori” column specific to classifies the air quality based on the pollutant levels into categories such as “baik” (good), “sedang” (moderate), “tidak sehat” (unhealthy), and “sangat tidak sehat” (very unhealthy). This dataset is used for evaluating the accuracy of air pollution, providing comprehensive temporal dominant causes in Central Jakarta.

In this training, the preprocessing does prepare the dataset by dropping the column "Stasiun" because each column provides duplication data. The "Max" column does not add new independent information so not needed for analysis. The pm25 was also dropped because it contain irrelevant data for analysis. Lastly the duplicate rows and NULL row were removed, and it has removed 433 rows. This preprocessing step optimizes the dataset analysis, ensures data integrity, and prevents redundancy.

This research step is continued by doing encode the target variable by converting these categorical into numeric. Next define the feature variables (x) and (y). These features (x) use 'pm10', 'so2', 'co', 'o3', and 'no2' which represent different air pollutants. The target variable (y) contains the encoded air quality categories for the model training.

4.2 Model Evaluation Analysis

The models were evaluated based on several metrics, such as: MAE, RMSE, and R^2 score. These metrics provide readers with insight into the overall predictive ability of the model. Comparison between models helps in determining which algorithm has the best predictive performance for AQI levels. The results are shown in Table 3.

Table 3. Models Result

Models	Evaluation Metrics	Score
Random Forest	MAE	**0,0293**
	RMSE	0,1153

	R^2 Score	0,9459
CatBoost	MAE	0,0350
	RMSE	**0,1041**
	R^2 Score	**0,9559**
Support Vector Regression	MAE	0,0929
	RMSE	0,1791
	R^2 Score	0,8694

This study compares the performance of each machine learning model (Random Forest, CatBoost, and SVR) to determine AQI in Central Jakarta. A comparative analysis of the three models revealed quite significant differences in performance. The evaluation results showed that CatBoost has superior ability in predicting with MAE value of 0,0350, RMSE of 0,1041, and R^2 Score of 0,9559. This value proves that the model is able to capture complex patterns that affect the determination of air quality. Then, Random Forest is in the next order with MAE value of 0,0293, RMSE of 0,1153, and R^2 Score of 0,9459.

In contrast, SVR showed much weaker performance with high MAE of 0.0929, RMSE of 0,1791, and lower R^2 Score of 0.8694. This result proves that SVR has a high error rate compared to the other two models. The overall results highlight that all three models are able to predict AQI to some extent, but the ensemble method approach has advantages over SVR in modelling the relationship between air pollution factors and AQI levels of health concern in the Bundaran HI as one of representative Central Jakarta environment.

5. Conclusion and Recommendations

This research found that the CatBoost model was the best performing choice because it was able to learn the dynamics of data from various levels of air pollutants. This was shown through the evaluation comparison process, which turned out that the CatBoost model had the lowest error rate compared to the other two models (SVR and Random Forest). The results we found were in line with previous research [18] where the CatBoost model they used proved to be superior in predicting AQI. The SVR model turned out to have the lowest accuracy compared to the other two models (Random Forest and CatBoost).

For future inquire about, it's suggested to undertake other models like XGBoost, LightGBM, or indeed profound learning strategies to conceivably get superior expectation comes about. Future thoughts can moreover see into including more important highlights, progressing include designing, or using gathering methods to create the show more exact and dependable.

References

[1] "68% of the world population projected to live in urban areas by 2050, says UN." Accessed: Mar. 10, 2025. [Online]. Available: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

[2] A. O. Victoria, "Kemendagri: Jakarta tak akan mampu tolak urbanisasi meski jadi DKJ." [Online]. Available: https://www.antaranews.com/berita/4069359/kemendagri-jakarta-tak-akan-mampu-tolak-urbanisasi-meski-jadi-dkj

[3] K. Balakrishnan *et al.*, "The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017," *Lancet Planet. Heal.*, vol. 3, no. 1, pp. e26–e39, Jan. 2019, doi: 10.1016/S2542-5196(18)30261-4.

[4] K. Kumar and B. P. Pande, "Air pollution prediction with machine learning: a case study of Indian cities," *Int. J. Environ. Sci. Technol.*, vol. 20, no. 5, pp. 5333–5348, May 2023, doi: 10.1007/s13762-022-04241-5.

[5] Y. Rybarczyk and R. Zalakeviciute, "Assessing the COVID-19 Impact on Air Quality: A Machine Learning Approach," *Geophys. Res. Lett.*, vol. 48, no. 4, Feb. 2021, doi: 10.1029/2020GL091202.

[6] U. Rahardja, Q. Aini, D. Manongga, I. Sembiring, and I. D. Girinzio, "Implementation of Tensor Flow in Air Quality Monitoring Based on Artificial Intelligence," *Int. J. Artif. Intell. Res.*, vol. 6, no. 1, pp. 2579–7298, 2023, doi: 10.29099/ijair.v6i1.430.

[7] F. Zhang and L. J. O'Donnell, "Support vector regression," *Mach. Learn. Methods Appl. to Brain Disord.*, vol. 11, no. 10, pp. 123–140, 2019, doi: 10.1016/B978-0-12-815739-8.00007-9.

[8] A. J. Cohen *et al.*, "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015," *Lancet*, vol. 389, no. 10082, pp. 1907–1918, 2017, doi: https://doi.org/10.1016/S0140-6736(17)30505-6.

[9] S. A. Horn and P. K. Dasgupta, "The Air Quality Index (AQI) in historical and analytical perspective a tutorial review," *Talanta*, vol. 267, p. 125260, 2024, doi: https://doi.org/10.1016/j.talanta.2023.125260.

[10] G. E. Kulkarni, A. A. Muley, N. K. Deshmukh, and P. U. Bhalchandra, "Autoregressive integrated moving average time series model for forecasting air pollution in Nanded city, Maharashtra, India," *Model. Earth Syst. Environ.*, vol. 4, no. 4, pp. 1435–1444, 2018, doi: 10.1007/s40808-018-0493-2.

[11] W. Kliengchuay *et al.*, "Particulate matter (PM10) prediction based on multiple linear regression: a case study in Chiang Rai Province, Thailand," *BMC Public Health*, vol. 21, no. 1, p. 2149, 2021, doi: 10.1186/s12889-021-12217-2.

[12] M. S. Ramadan, A. Abuelgasim, and N. Al Hosani, "Advancing air quality forecasting in Abu Dhabi, UAE using time series models," *Front. Environ. Sci.*, vol. 12, no. May, pp. 1–18, 2024, doi: 10.3389/fenvs.2024.1393878.

[13] S. R. Shams *et al.*, "Assessing the effectiveness of artificial neural networks (ANN) and multiple linear regressions (MLR) in forcasting AQI and PM10 and evaluating health impacts through AirQ+ (case study: Tehran)," *Environ. Pollut.*, vol. 338, p. 122623, 2023, doi: https://doi.org/10.1016/j.envpol.2023.122623.

[14] Matthias Schonlau and Rosie Yuyan Zou, "The random forest algorithm for statistical learning," *Stata J.*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[15] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[16] R. Yu, Y. Yang, L. Yang, G. Han, and O. A. Move, "RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems," 2016. doi: 10.3390/s16010086.

[17] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, p. 94, 2020, doi: 10.1186/s40537-020-00369-8.

[18] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis," *J. Environ. Public Health*, vol. 2023, no. 1, p. 4916267, Jan. 2023, doi: https://doi.org/10.1155/2023/4916267.

[19] S. P. and D. C. P. Debasish Basak, "Support Vector Regression," *Neural Inf. Process. – Lett. Rev.*, vol. 11, no. 10, pp. 203–204, 2007.

[20] S. Aram *et al.*, "Machine learning-based prediction of air quality index and air quality grade: a comparative analysis," *Int. J. Environ. Sci. Technol.*, Jun. 2023, doi: 10.1007/s13762-023-05016-2.