Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

Air Pollution Classification Prediction Model with Deep Neural Network based on Time-Based Feature Expansion and Temporal Spatial Analysis

Muhamad Dika Muldani*, Sri Suryani Prasetiyowati, Yuliant Sibaroni

School of Informatics, Informatics, Telkom University, Bandung, Indonesia
Email: dikamuldani@student.telkomuniversity.ac.id, srisuryani@telkomuniversity.ac.id, yuliant@telkomuniversity.ac.id
Correspondence Author Email: dikamuldani@student.telkomuniversity.ac.id
Submitted: **24/07/2024**; Accepted: **09/09/2024**; Published**: 09/09/2024**

Abstract−Air pollution is one of the most significant global challenges, with serious impacts on the health of living beings. In Indonesia, particularly in major cities such as Jakarta and Surabaya, the increase in the Air Quality Index (AQI) over the past few years indicates worsening air quality conditions. This decline in air quality is caused by increased industrial activities, motor vehicle emissions, and deforestation. Rising AQI levels pose severe health risks, including respiratory and cardiovascular diseases, and present major challenges for urban planning, public health management and environmental policy. Addressing this issue requires concerted efforts to implement sustainable practices, reduce emissions, and improve air quality management. The increasing air pollution level indicate the need for a more effective approach to identify and classify air quality index results with relevant success rates without using relatively expensive air quality index detection tools. This research aims to classify the air quality index using a Deep Neural Network model based on time-based feature expansion and spatial-temporal analysis. The Deep Neural Network model is used to extract complex patterns and hidden features in the data and help generate more accurate air pollution classifications. Meanwhile, time-based feature expansion is useful for extending the time representation in the data. The results of this research are expected to make a significant contribution in improving the global understanding of air pollution with accuracies up to 86.89% and 84.99%, respectively. By providing a cost-effective and efficient method for air quality monitoring, this study can lead to better pollution control measures. Furthermore, the insight gained from this research can help policymakers develop strategies to mitigate the adverse effects of air pollution on public health and the environment.

Keywords: Air Quality Index; Classification; Deep Neural Network; Time-Based Feature Expansion; healthier;

1. INTRODUCTION

Air is one of the primary elements that maintains the balance of the Earth's environment. According to the World Health Organization (WHO), air pollution is one of the most serious environmental risks caused by natural scenarios such as climate change, volcanic eruptions, emissions from living organisms, and salt sprays. However, human activities over time have become the primary cause of air pollution. The most common human activities include burning fossil fuels and producing chemicals [1]. According to the latest data from WHO, every year 4,200,000 people die due to air pollution, and 91% of the world's population still resides in areas where air quality exceeds WHO standards [2]. These statistics underscore the urgent need for comprehensive action to address air quality issues worldwide. Effective strategies and advanced technologies are essential to reduce the health risks associated with poor air quality, ensuring a healthier environment for all individuals.

In Indonesia, one of the regions severely affected by air pollution is Jakarta. The air quality in DKI Jakarta has an average PM2.5 concentration that continues to increase to 49.4 µg/m³ in 2019, which is almost five times higher than the annual average PM2.5 guideline set by the World Health Organization (WHO) [3]. In 2020, there was a decrease in the average PM2.5 concentration, which became 52.98 µg/m³[4]. This decrease in PM2.5 leveles can be attributed to several factors, including tighter environmental regulations, increase public awareness of air pollution, and the reduction in industrial activity and vehicle emissions due to the blocking and curbing of the COVID-19 pandemic played an an important role in improving air quality during this period. Joint efforts by governmental and non-governmental organizations in promoting cleaner technologies and practices also contributed to the reduction in PM2.5 concentrations observed, demonstrating the effectiveness of coordinated action in reducing air pollution. In 2021, there was another decrease in the average PM2.5 concentration to around 17-40 µg/m³ [5]. This decrease was due to government efforts to limit outdoor activities during 2021 and 2022. In 2022, there was an increase in concentration, but the highest concentration did not exceed 50 µg/m³ [6], as activities have returned to normal and can be carried out as usual. This return to normalcy marks a shift from previous government-imposed restrictions that aimed to limit outdoor activities during 2021 and 2022. These efforts were instrumental in temporarily reducing pollution levels. However, the subsequent increase in PM2.5 concentrations underscores the ongoing challenge of maintaining air quality standards that protect public health.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

In previous research [2] conducted by D. A. Kristiyanti et al., the collected data used to train a classification prediction model utilizing the Recurrent Neural Network algorithm. Although previous studies have shown quite significant success using attributes like APSI such as NO2, CO, CO3, PM2.5, PM10, obtaining an 88.86% accuracy rate and a 0.320 error rate, these studies have limitations in terms of the number of data samples. In another research by M. A. Faishol et al, utilizing the RNN-LSTM algorithm showed that out of several scenarios of data split between training and testing data using APSI attributes like NO2, CO, CO3, PM2.5, PM10 showed the best results with a 95%:5% data composition for training and testing data, with an RMSE error calculation yielding a value of 1.880 [7].

In another study by, a high accuracy rate of 96.41% was achieved by utilizing APSI attributes such as NO2, CO, CO3, PM2.5, PM10 using the Novel Binary Chimp Optimization Algorithm with LSTM or BChOA-LSTM [8]. Another previous research utilizing the CNN-BiLSTM-IDW algorithm with APSI attributes such as NO2, CO, CO3, PM2.5, PM10 showed a 16% improvement in prediction performance compared to using the regular IDW method [9].

With the influence of attributes determined by APSI related to NO2, CO, CO3, PM2.5, PM.10, the determination of air quality using tools which need to spend quite expensive, so this research will focus on the air quality classification model with the Deep Neural Network classification method with time-based feature expansion in predicting air quality to improve the processing results while maintaining the accuracy of the predictions used, which increases flexibility and reduces area overhead [10], especially in the Indonesian region, as previous studies have recommended Deep Neural Network due to their superior performance in handling complex prediction tasks [11], in other previous studies have recommended Deep Network cause Deep Network can perform better even with less data by having deep knowledge of the dataset and quality adjustments to the model [12]. especially in the Java area with attributes other than those determined by APSI previously such as the use of rainfall data, the number of trees or the number of vehicles in the Java area. This effort will include providing stimulus to the model to trigger more accurate results. Therefore, this research will try to overcome this limitation by using more dynamic and representative data.

With this approach, this study aims to better understand air pollution patterns in Indonesia, particularly in the Java region. This will assist the government and relevant agencies in taking more timely and effective mitigation measures, as well as developing better policies to protect public health. In addition, a more accurate analysis of air pollution will also provide a better understanding of the environmental and climate impacts that have important implications for environmental sustainability.

2. RESEARCH METHODOLOGY

In the context of this research, an overview of the system modeling flow is presented, which is represented through a flowchart in Figure 1.

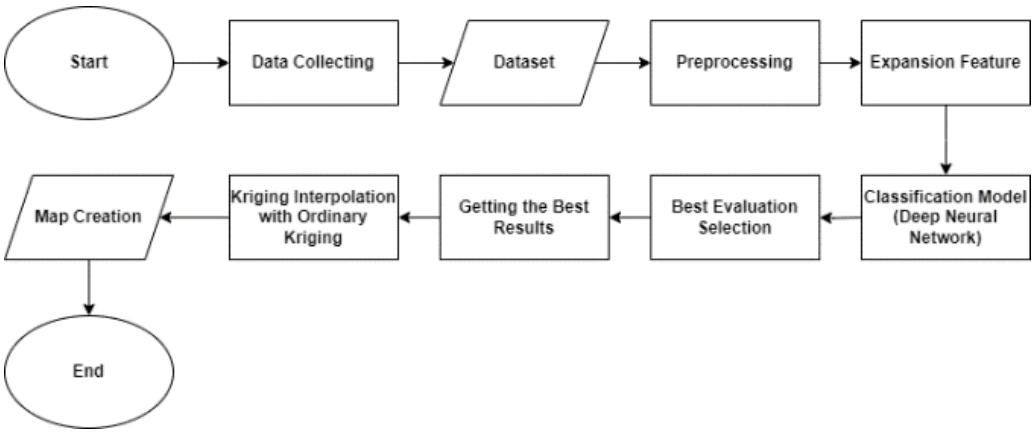

Figure 1. Flowchart of Overall System Design for Air Pollution Classification Prediction Model

This research system design uses the Deep Neural Network Time-Based algorithm to predict the classification of air pollution in Java Island, based on feature expansion from September 2020 to April 2022. The algorithm uses temporal and spatial features to improve prediction accuracy, taking into account factors such as meteorological data, pollutant levels, and seasonal variations. Performance evaluation, through metrics such as accuracy, F1 score, and RMSE, determined the best classification model, ensuring robust and reliable predictions. The model was then used for a geospatial visualization, which provided a comprehensive overview of the air pollution distribution on Java Island, highlighting areas with varying pollution levels. This visualization helps in identifying pollution hotspots, understanding trends over time, and supporting policy-making decisions aimed at improving air quality and public health on the island..

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

2.1 Data Collecting

The data used in this study comes from air quality values published by the Department of Environment and Forestry, which were collected during the period September 2020 to April 2022. Data acquisition was done by searching specific websites provided by the environment and forestry department and open data platforms available for regions in Java Island, which resulted in Comma Separated Value (CSV) files. This dataset consists of 29,233 entries, including various air quality indicators such as PM10, SO2, NO2, CO, and O3 levels. Each entry is tagged with date, temperature, humidity, rainfall, length of sunshine, wind speed, population, number of trees, number of vehicles, altitude, and location, making it possible to conduct a detailed study of how air pollution changes over time and in different places. The data covers many urban and rural areas, reflecting a wide range of environmental conditions, which helps the model learn from various situations, thereby improving the accuracy and reliability of predictions. Data preprocessing included handling missing values, standardizing features, and dividing the data into training and testing sets to properly assess model performance.

Table 1. Data Attributes

Attributes	Description
X_1	Minimum Temperature
X_2	Maximum Temperature
X_3	Average Temperature
X_4	Average Humidity
X_5	Rainfall
X_6	Length of Sunshine
X_7	Wind Speed
X_8	Wind Speed Direction
X_9	Most Wind Speed
X_{10}	Total Population
X_{11}	Number of Trees
X_{12}	Number of Vehicles
X_{13}	Altitude
Y	Number of predictions of each region's air pollution standard index each month

2.2 Preprocessing

The data obtained will go through a preprocessing stage first. This process aims to produce quality data for use in classification. In this research, the data preprocessing stage is carried out through stationarity tests and data normalization that ensure data quality before being used in a classification model that aims to mitigate potential problems that can affect model performance.

a. Augmented Dickey-Fuller test: Augmented Dickey-Fuller (ADF) is a method to test for stationary in time-series data. The ADF test works by calculating the time-series data model (t) and comparing it with a value from ADF distribution. This ADF test aims to detemine whether the time-series data is positively or negatively stationary [13]. In this hypothesis for testing stationary using the Dickey-Fuller unit root test was a follows.

$$Xt = \phi Xt - 1 + Wt \quad (1)$$

If -1 < ϕ 1 then Xt is stationary, and otherwise
If $\phi = 1$ then Xt is nonstationary

b. Min-Max Normalization: Normalization method that performs a linear transformation of the original data create balanced value comparison between the data before and after the process [14].

$$Xnew = \frac{X - min(X)}{max(X) - min(x)} \text{ [15]} \quad (2)$$

2.3 Feature Expansion

Feature expansion uses for selection feature and iteration to extend multiple feature [16]. This feature used to create a classification model to predict air pollution standard index value in Java Island. This expansion feature works based on pre-existing data, allowing the model to consider a broader range of variables and patterns. By iterating through and selecting relevant features, the process enhances the model's ability to capture intricate relationships within the data. Consequently, this leads to more accurate and reliable predictions of air pollution levels, contributing significantly to environtmental monitoring and decision-making in Java Island.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

Table 2. Data Class Label

Models	Combination	Training Data Attributes	Target
t-1	1	January 2022	February 2022
	2	December 2021	January 2022
	:	:	:
	17	September 2020	October 2020
t-2	1	December 2021 – January 2022	February 2022
	2	November 2021 – December 2021	January 2022
	:	:	:
	16	September 2020 – October 2020	November 2020
:	:	:	:
t-15	1	November 2020 – January 2022	February 2022
	2	October 2020 – December 2021	January 2022
	3	September 2020 – November 2021	December 2021

In this study, feature expansion is performed using Sklearn's SelectBest library, which improves the accuracy and performance of the prediction model. This method prioritizes the most relevant features, simplifies the model and reduces noise from less informative variables. By focusing on these key features, the model is better equipped to make precise predictions, resulting in more reliable results. [17]. This method selects the first k features with the highest values calculated using the f_classif function. The f_classif function evaluates the dependencies in the dataset by analyzing the variance based on the average of the characteristics, helping to identify the most significant features. This selection process reduces dimensionality while retaining key information, improving the accuracy and classification efficiency of the model [18].

2.4 Deep Neural Network

Deep neural network (DNN) is a product of artificial intelligence and is a collection of neurons organized in a sequence of many previous layers and performing simple calculations [19]. DNN uses connected layers to automatically understand and extract high-level features from data.

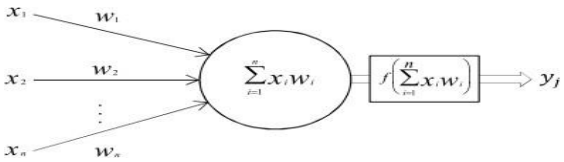

Figure 2. Mathematical model of a neuron

Figure 2 is a systematic model of neurons based on human brain neurons, where x is a neuron input that has a weight *Wi*. The sum of all inputs has a weight *XiWi*, then passed through a nonlinear activation function f, to concert the neuron's preactivation level into an output Yj [20].

Figure 3 illustrates the architecture of the Time-Based DNN model, which has more than one hidden layer to improve classification results. In the input layer, the variables t-1, t-2, and t-k represent the input vectors, which are obtained from the feature expansion process. The hidden layers, denoted by J and I, include the bias for the data to be predicted.

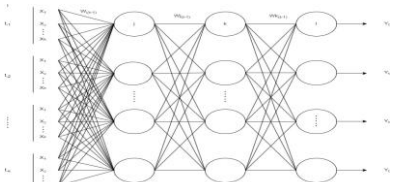

Figure 3. Deep Neural Network Time-Based Architecture

The variables i, j, k, and l serve as activation functions, introducing complexity and non-linearity to the model. The output layer produces Yt, which represents the best accuracy achieved in the classification process.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

2.5 Ordinary Kriging Interpolation

Interpolation is the process of estimating values in an unsmooted area to the produce a map of the distribution of values in the area. Kriging is a geostatistical interpolation method that uses the distance and direction between sample point to explain surface variations. The killing process involves adjusting mathematical functions at specific points to predict values at new locations, through steps such as statistical analysis, variogram modeling, and surface generation [21]. And Ordinary kriging is a method that utilizes spatial values at sampled locations and variograms that show the correlation between spatial points to predict values at other unsampled locations, where the predicted values depend on their proximity to the sampled locations [22].

$$\hat{y}_{(t+k)}(S_o) = \sum_{i=0}^{N} \lambda_i^{OK} y_{(t+k)}(S_i) \quad (3)$$

The formula explains how to estimate the prediction of the unknown class $\hat{y}_{(t+k)}(S_o)$ at the future time t+kt+k for a given location So. The estimation uses the class prediction $y_{(t+k)}(S_i)$ at nearby locations Si. with the NN representing the total number of such locations included in the interpolation. Each location Si contributes to the prediction in So based on the Ordinary Kriging weights λ_i^{OK}, which optimally combines the spatial information to produce an accurate estimate for $\hat{y}_{(t+k)}(S_o)$.

2.6 Air Pollution Standard Index

In the regulation of the Minister of Environtment and Foresty number 14 of 2020 article, the Air Pollution Standard Index (APSI) is a unitless number that describes the condition of ambient air quality at a certain location, which is based on the impact on human health and other living things [23]. based on research, the APSI table used is only good, medium, and unhealthy as in the table below.

Table 3. APSI parameter value conversion

APSI	24 Hour PM10 (µg/m^3)	24 Hour PM2.5 (µg/m^3)	24 Hour SO2 (µg/m^3)	24 Hour CO (µg/m^3)	24 Hour O3 (µg/m^3)	24 Hour NO2 (µg/m^3)	24 Hour HC (µg/m^3)
0-50	50	15,5	52	4000	120	80	25
51-100	150	55,4	180	8000	235	200	100
101-200	350	150,4	400	15000	400	1130	215
201-300	420	250,4	800	30000	800	2260	432
>300	500	500	1200	45000	1000	3000	648

Description:

a) Measurement data for 24 hours continuously.
b) APSI calculation results for particulate matter (PM2.5) are presented hourly for 24 hours
c) The APSI calculation results for particulate matter (PM10), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), and hydrocarbons (HC) are taken as the highest and lowest APSI values for each hour.

The APSI calculation is based on the upper limit APSI, lower limit APSI, upper limit ambient, lower limit ambient, and measured ambient concentration values. The mathematical equation for the calculation is as follows:

$$I = \frac{I_a - I_b}{X_a - X_b}(X_a - X_b) + I_b \quad (4)$$

Based on the extensive research conducted, the data results indicate that the APSI (Air Pollution Standard Index) tables utilized in the analysis are categorized solely into three distinct levels of air quality: good, moderate, and unhealthy. This categorization is crucial for understanding the varying degrees of air pollution on Java Island over the 15-month research period. The classification into these specific categories allows for a more detailed and precise analysis of air quality trends and patterns, aiding in the development of effective mitigation strategies. The table below clearly delineates these three levels, providing a comprehensive overview of the air quality status as determined by the APSI standards used in this study.

Table 4. Air Pollution Standard Index Category

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

Label	Category	Color Status	Range
0	Good		1-50
1	Medium		51-100
2	Unhealthy		101 – 200

2.6 Evaluation

The evaluation stage measures the effectiveness of Deep Neural Network classification method using a confussion matrix. The predicted value resulting from dividing the total of True Positive and True Negative by the sum of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) values of a confussion matrix [24].

Table 5. Confussion Matrix

Data		Actually Positive	Actually Negative	Actually Neutral
Prediction	Positive	TP	FN	FN
	Negative	FP	TN	TN
	Neutral	FP	TN	TN

Based on confussion matrix, several evaluation metrics can be calculated, including accuracy, precision, recall, and F1-score. Accuracy measures how well the model classifies the predicted results. Accuracy measures the percentage of true positive predictions among all positive predictions. Recall measures the percentage of positive cases that are correctly identified. And F1-score combines the precision and recall calculated.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \text{ x } 100 \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \text{ x } 100 \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \text{ x } 100 \quad (7)$$

$$F1 - Score = \frac{2\ x\ (Recall\ x\ Precision)}{Recall + Precision} \text{ x } 100 \quad (8)$$

3. RESULT AND DISCUSSION

3.1 Result

a. Best Performance of t-k Deep Neural Network Model
The Table below shows the confussion matrix value of each best t-k Deep Neural Network Time-Based Model.

Table 6. Best T-K Deep Neural Network Time-Based Model

t-k	Accuracy	F1-score	Recall	Precision
t-1	79.35%	75.84%	70.00%	82.41%
t-2	78.44%	74.58%	70.00%	78.69%
t-3	79.87%	74.89%	70.00%	79.42%
t-4	79.93%	77.34%	80.00%	83.25%
t-5	80.00%	79.17%	80.00%	85.71%
t-6	70.00%	67.03%	70.00%	81.25%
t-7	70.00%	67.03%	70.00%	81.25%
t-8	80.00%	79.16%	80.00%	85.71%
t-9	86.89%	84.89%	80.00%	87.16%
t-10	84.99%	84.89%	80.00%	88.32%
t-11	85.65%	85.75%	80.00%	85.45%
t-12	85.84%	86.13%	70.00%	86.13%
t-13	86.43%	83.87%	80.00%	85.53%
t-14	80.19%	79.65%	80.00%	85.71%
t-15	85.35%	83.94%	80.00%	86.32%

The results obtained from the Deep Neural Network (DNN) Time-Based model show varying percentages of accuracy across different time steps. For example, at t-1, the model achieved accuracies of 79.35%, 75.84%,

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

70.00%, and 82.41% across various metrics. As the time steps increased, the accuracy fluctuated, with a notable improvement observed at t-5, where the accuracy reached 80.00%, 79.17%, 80.00%, and 85.71%. The model shows significant performance peaks at t-9 and t-10, with accuracies up to 86.89% and 84.99%, respectively. On t-11 and t-12, the model maintained a high level of accuracy, with t-12 achieving the highest accuracy of 85.84% across its metrics. Overall, the results reflect the model's ability to adapt and improve over time, with variations in performance across different time steps indicating its sensitivity to temporal changes in air quality data.

b. Best Perfomance of t+k Deep Neural Network Model

The table below shows the confussion matrix value of each best t+k Deep Neural Network Time-Based Model.

Table 7. T+1 and T+2 Deep Neural Network with RMSE

t+k	RMSE
t+1	0.4297
t+2	0.4634

The results of the Deep Neural Network (DNN) Time-Based model, evaluated using Root Mean Square Error (RMSE), show performance across future time steps. For example, at t+1, the RMSE was 0.574355, and at t+2, the RMSE increased slightly to 0.618013. These values reflect the prediction accuracy of the model over time, with the error gradually increasing as the forecast horizon expands. The RMSE values give an idea of the model's ability to generalize and maintain accuracy when predicting air quality further into the future.

Table 8. Best T+K Deep Neural Network Time-Based Model

Location	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
Jendral Sudirman Tangerang	2	1	2	1	2	1	1	2	2	2	2	2	2
Sudimara Ciledug Tangerang	1	2	2	1	2	2	1	2	2	2	2	2	2
Ahmad Yani Semarang	0	0	1	1	1	1	1	0	0	0	0	1	0
Balai Kota Depok	0	0	1	2	1	1	0	1	1	0	0	1	1
Bandung	0	0	0	1	1	0	0	0	1	0	1	1	1
Banyuwangi	0	0	0	0	1	0	0	0	0	1	0	0	0
Budiarto	1	1	1	1	1	1	1	1	0	0	0	1	0
Cilacap	0	1	1	1	1	1	1	0	0	0	0	1	0
Citeko Bogor	1	0	0	1	1	1	1	0	0	0	0	0	1
DK1 Bunderan HI	1	2	2	1	2	2	2	1	0	1	0	1	1
DK2 Kelapa Gading	1	1	1	2	1	2	2	0	0	0	0	2	0
DK3 Jagakarsa	2	2	2	2	2	2	2	1	1	1	1	1	1
DK4 Lubang Buaya	1	1	2	2	1	1	1	2	2	2	2	2	2
DK5 Kebun Jeruk	2	1	2	2	2	2	2	1	1	1	1	2	1
Kebun Sari Surabaya	1	1	1	1	1	1	1	0	1	0	0	1	1
Malang	0	0	0	0	1	1	0	0	0	0	0	0	0
Meikarta Bekasi	1	1	1	1	2	1	1	0	0	0	1	1	1
Nganjuk	0	0	0	1	1	0	0	0	0	1	0	0	0
Pasuruan	0	0	0	0	1	1	1	0	0	0	1	0	0
Pendopo Banten	1	1	1	1	1	2	1	1	0	2	1	1	1
Samsat Balaraja Banten	1	1	1	2	1	1	1	0	1	0	2	1	0
Samsat Cikokol Banten	2	1	1	1	2	2	1	0	0	2	0	2	2
Samsat Serpong Banten	2	1	2	1	1	1	1	1	2	0	1	1	2
Semarang	0	0	1	1	0	1	1	1	0	1	0	1	0
Serang	0	0	0	1	0	1	1	0	1	0	0	0	1
Sleman	1	1	0	1	0	1	1	1	0	1	0	1	1
Tegal	0	1	0	0	0	0	0	0	1	0	0	0	0
Wonorejo Surabaya	1	0	1	0	1	1	1	1	1	0	0	1	1

c. Visualization of Air Pollution distribution classification

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

The visualization of the maps below shows the distribution of air pollution based on the classification result of predictions using a Deep Neural Network combined with Time-Based Feature Expansion and spatial temporal analysis.

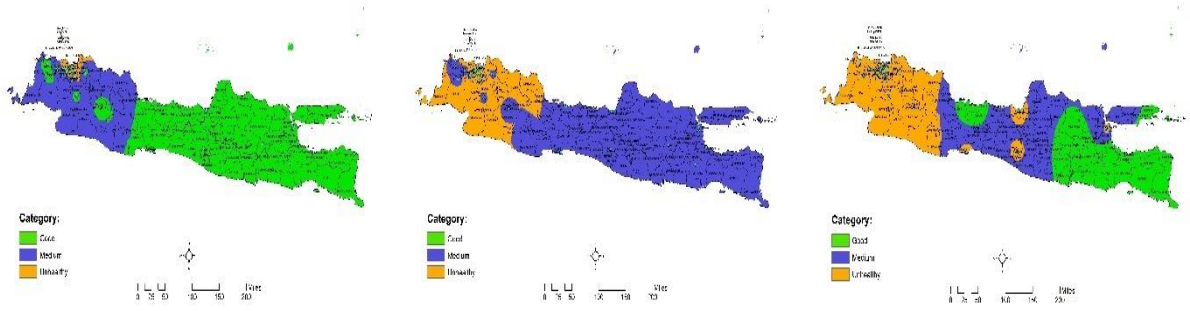

Figure 4. Prediction Map of Air Pollution Distibution May 2022 to July 2022

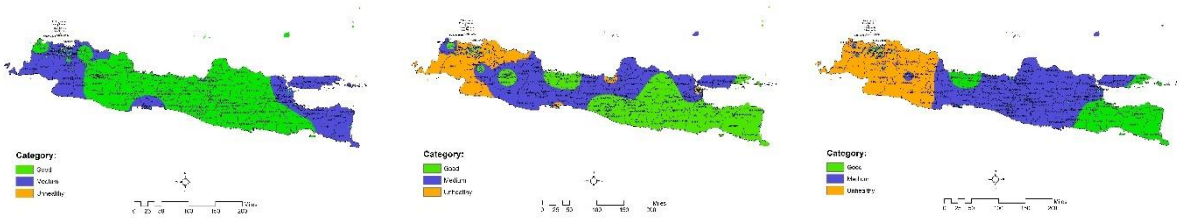

Figure 5. Prediction Map of Air Pollution Distibution August 2022 to October 2022

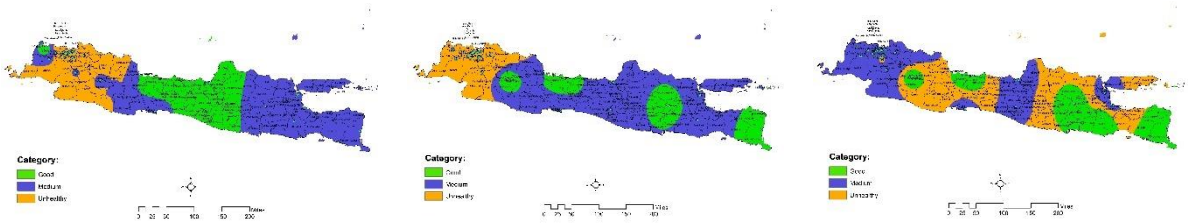

Figure 6. Prediction Map of Air Pollution Distibution November 2022 to January 2023

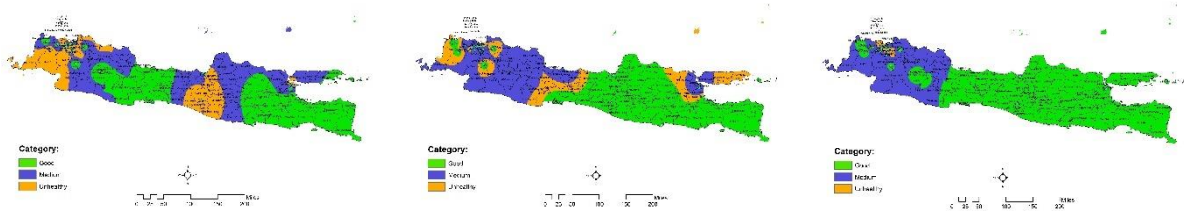

Figure 7. Prediction Map of Air Pollution Distibution February 2023 to April 2023

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

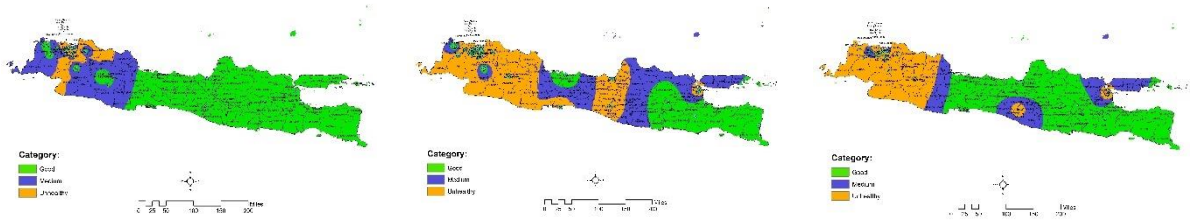

Figure 8. Prediction Map of Air Pollution Distribution May 2023 to July 2023

d. Discussion

Analysis of Table 6 shows that the Time-Based Deep Neural Network (DNN) model exhibits a general trend of improved performance metrics up to time step t-10, where accuracy peaks at 86.89% and F1 score reaches 89.00%. After t-10, the accuracy and F1 score tend to stabilize or decrease, indicating that t-10 is the most optimal time to make predictions. This trend shows that the model performs best for short- to medium-term forecasts, with accuracy decreasing for longer time horizons, reflecting the increasing complexity of predictions further into the future.

The model shows high accuracy and F1 scores, with significant improvement over time. The peak performance at t-10 indicates strong predictive ability, but the decline in performance at subsequent time steps emphasizes the need for improvements in long-term forecasting. The dominant features PM10, SO2, CO, O3, and NO2 are critical for accurate air quality prediction, highlighting their importance in the model and the need to focus on these pollutants for effective air quality management.

Overall, the time-based Deep Neural Network (DNN) with feature expansion demonstrated strong performance in predicting air pollution levels across Java Island, surpassing traditional methods. The model achieved high accuracy and minimal prediction error over the 15-month study period, highlighting its reliability in predicting air pollution trends. The VAR model proved effective with significant parameter estimates and smaller MAPE and MAE values, despite not having the smallest RMSE [25]. In contrast, the RNN-LSTM model improves with larger data volumes, reducing prediction error and improving accuracy [7] [26]. Overall, While VAR models provide a solid foundation for time series analysis, integrating RNN-LSTM neural networks can improve air quality forecasts, especially with larger data sets. This approach effectively identifies key features that affect pollution and overcomes the limitations of traditional models by considering factors such as vehicle density and population. However, the increased RMSE in long-term forecasts suggests that the model is more suitable for short- to medium-term predictions and needs to be refined for long-term accuracy. The Time-Based DNN with feature expansion remains robust, even with moderate data volumes, overcoming the common problems of data set size and overfitting.

The air pollution distribution maps in Figures 4 to 6 illustrate a pattern where high and moderate air quality categories are concentrated in urban areas such as DKI Jakarta and its surroundings. This pattern underscores the need for targeted air quality management strategies, especially in densely populated areas. The optimal classification prediction period for the Time-Based DNN model is about 15 months, with the most frequently selected features providing valuable insights for future air quality assessment. The findings can guide the government and communities in Java Island in effectively addressing air pollution and implementing measures to improve air quality.

4. CONCLUSION

Based on the results, it can be concluded that combining feature expansion with time series data to build a classification prediction model using Deep Neural Network significantly improves the performance of the model. Specifically, the Deep Neural Network models demonstrated accuracy, precision, recall, and F1-score metrics. For example, model t-5 among models t-1 to t-5 achieved a maximum accuracy of 80.00%, with precision values reaching 85.71%. The recall values for these models ranged from 70.00% to 80.00%. Models t-6 to t-10 reached a maximum accuracy of 86.89% on model t-9 with a precision value of 87.16%. The recall values for these models ranged from 70.00% to 80.00%. Models t-11 to t-15 showed further improvement, with some combinations exceeding 80.00% accuracy. The maximum accuracy was 86.43%, with precision values reaching 85.53% and average recall reaching 80.00%, resulting in a high F1-score. These results show that the use of Deep Neural Network with better feature expansion is effective for accurately predicting and categorizing data. This method has the potential to improve decision-making with data and enhance environmental monitoring. Overall, the results of this study can be used as a reference in handling air pollution in Java, so for future research, it is recommended to enrich the dataset with additional attributes that can significantly affect air pollution levels and explore other machine learning algorithms to obtain better results. In

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

addition, incorporating real-time data collection methods can improve the model's responsiveness to air quality changes, thus providing more accurate and timely predictions. This approach can assist the government in Java in identifying and implementing effective solutions to manage and mitigate air pollution.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to all those who have helped me during this research. This research would not have been achieved and completed without the support and assistance of various parties. The dedication, effort, and support given are very meaningful. Hopefully the result of this research can provide benefits for developing science and the welfare of society.

REFERENCES

[1] L. Mampitiya *et al.*, "Machine Learning Techniques to Predict the Air Quality Using Meteorological Data in Two Urban Areas in Sri Lanka," *Environ. - MDPI*, vol. 10, no. 8, pp. 1–18, 2023, doi: 10.3390/environments10080141.
[2] D. A. Kristiyanti, E. Purwaningsih, E. Nurelasari, A. Al Kaafi, and A. H. Umam, "Implementation of Neural Network Method for Air Quality Forecasting in Jakarta Region," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012037.
[3] Z. Zulkarnain and A. Ghiffary, "Impact of Odd-Even Driving Restrictions on Air Quality in Jakarta," *Int. J. Technol.*, vol. 12, no. 5, p. 925, 2021, doi: 10.14716/ijtech.v12i5.5227.
[4] K. I. Solihah, D. N. Martono, and B. Haryanto, "Analysis of Spatial Distribution of PM2.5and Human Behavior on Air Pollution in Jakarta," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 940, no. 1, 2021, doi: 10.1088/1755-1315/940/1/012018.
[5] A. S. Yuwono, A. V. A. Pinem, Supandi, K. Nisa, and C. Arif, "Evaluation of Air Pollution Standard Index for NO2 Parameter in Jakarta and Bogor," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics, 2023. doi: 10.1088/1755-1315/1134/1/012023.
[6] N. F. Prih Waryatno, N. P. Kinanti, and Taryono, "Kondisi Pencemaran Udara pada Saat Periode Lebaran 2022 di Wilayah Jakarta," *Bul. GAW Bariri*, vol. 3, no. 2, pp. 25–31, 2022, doi: 10.31172/bgb.v3i2.68.
[7] M. A. Faishol, E. Endroyono, and A. N. Irfansyah, "Predict Urban Air Pollution in Surabaya Using Recurrent Neural Network – Long Short Term Memory," *JUTI J. Ilm. Teknol. Inf.*, vol. 18, no. 2, p. 102, 2020, doi: 10.12962/j24068535.v18i2.a988.
[8] S. Baniasadi, R. Salehi, S. Soltani, D. Martín, P. Pourmand, and E. Ghafourian, "Optimizing Long Short-Term Memory Network for Air Pollution Prediction Using a Novel Binary Chimp Optimization Algorithm," *Electron.*, vol. 12, no. 18, 2023, doi: 10.3390/electronics12183985.
[9] K. Samal, K. Babu, and S. Das, "Spatio-temporal Prediction of Air Quality using Distance Based Interpolation and Deep Learning Techniques," *EAI Endorsed Trans. Smart Cities*, p. 168139, 2018, doi: 10.4108/eai.15-1-2021.168139.
[10] M. Song, J. Zhao, Y. Hu, J. Zhang, and T. Li, "Prediction based execution on deep neural networks," *Proc. - Int. Symp. Comput. Archit.*, pp. 752–763, 2018, doi: 10.1109/ISCA.2018.00068.
[11] P. Singh, T. L. Narasimhan, and C. S. Lakshminarayanan, "DeepAir: Air Quality Prediction using Deep Neural Network," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2019-Octob, pp. 869–873, 2019, doi: 10.1109/TENCON.2019.8929470.
[12] D. Lokhande, "Deep neural network in prediction of student performance," no. February, 2023.
[13] Ajewole KP, Adejuwon SO, and Jemilohun VG, "Test for Stationarity on Inflation Rates in Nigeria using Augmented Dickey Fuller Test and Phillips-Persons Test," *IOSR J. Math.*, vol. 16, no. 3, pp. 11–14, 2020, doi: 10.9790/5728-1603031114.
[14] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 4, no. 1, pp. 13–20, 2021, doi: 10.47738/ijiis.v4i1.73.
[15] N. Nofriani, "Machine Learning Application for Classification Prediction of Household's Welfare Status," *JITCE (Journal Inf. Technol. Comput. Eng.*, vol. 4, no. 02, pp. 72–82, 2020, doi: 10.25077/jitce.4.02.72-82.2020.
[16] M. A. Fauzi, R. F. N. Firmansyah, and T. Afirianto, "Improving sentiment analysis of short informal Indonesian product reviews using synonym based feature expansion," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 16, no. 3, pp. 1345–1350, 2018, doi: 10.12928/TELKOMNIKA.v16i3.7751.
[17] T. Desyani, A. Saifudin, and Y. Yulianti, "Feature Selection Based on Naive Bayes for Caesarean Section Prediction," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 879, no. 1, 2020, doi: 10.1088/1757-899X/879/1/012091.
[18] Elqi Ashok, Sri Suryani Prasetiyowati, and Yuliant Sibaroni, "DHF Incidence Rate Prediction Based on Spatial-Time with Random Forest Extended Features," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 4, pp. 612–623, 2022, doi: 10.29207/resti.v6i4.4268.
[19] M. Z. Alom *et al.*, "A state-of-the-art survey on deep learning theory and architectures," *Electron.*, vol. 8, no. 3, 2019, doi: 10.3390/electronics8030292.
[20] Z. Hu, "Estimation and application of matrix eigenvalues based on deep neural network," *J. Intell. Syst.*, vol. 31, no. 1, pp. 1246–1261, 2022, doi: 10.1515/jisys-2022-0126.
[21] G. H. Pramono, "Akurasi Metode IDW dan Kriging untuk Interpolasi Sebaran Sedimen Tersuspensi di Maros, Sulawesi Selatan," *Forum Geogr.*, vol. 22, no. 2, p. 145, 2008, doi: 10.23917/forgeo.v22i2.4988.
[22] N. Nur Rohma, "Pendugaan Metode Ordinary Kriging," *J. Penelit. Ilmu Sos. dan Eksakta*, vol. 2, no. 1, pp. 21–29, 2022, doi: 10.47134/trilogi.v2i1.33.
[23] B. K. Hidayatullah, M. Kallista, C. Setianingsih, P. S1, and T. Komputer, "Prediksi Indeks Standar Pencemar Udara Menggunakan Metode Long Short-Term Memory Berbasis Web (Studi Kasus Pada Kota Jakarta)," *e-Proceeding Eng.* , vol.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page: 867−877
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5675

9, no. 3, pp. 1247–1255, 2022, [Online]. Available: https://data.jakarta.go.id/
[24] M. Imam, S. Adam, S. Dev, and N. Nesa, "Air quality monitoring using statistical learning models for sustainable environment," *Intell. Syst. with Appl.*, vol. 22, no. March, p. 200333, 2024, doi: 10.1016/j.iswa.2024.200333.
[25] K. N. Sh, I. Irfani, and U. Mukhaiyar, "Predicting Air Pollution Levels in Jakarta Using Vector Autoregressive Analysis," vol. 2023, no. Icsmtr, pp. 14–22, 2023, doi: 10.2991/978-94-6463-332-0_3.
[26] Wihayati and F. W. Wibowo, "Prediction of air quality in Jakarta during the COVID-19 outbreak using long short-term memory machine learning," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 704, no. 1, 2021, doi: 10.1088/1755-1315/704/1/012046.