IJoICT
ISSN 2356-5462
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 10, No. 2, Dec 2024. pp.270-286
doi: 10.21108/ijoict.v10i2.1035

Prediction of Classification of Air Quality Distribution in Java Island using ANN with Time-Based Feature Expansion and Spatial Analysis

Soni Andika Gutama [1*], Sri Suryani Prasetiyowati [2], Yuliant Sibaroni [3]

[1,2,3]*School of Informatics, Informatics, Telkom University, Bandung, Indonesia*
[*]soniandika@student.telkomuniversity.ac.id

Abstract

Air pollution has a significant impact on human health and the environment, especially in densely populated areas such as Java Island in Indonesia. Air pollution is caused by high air quality indexes originating from concentrations of hazardous pollutants such as sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), and hydrocarbons (HC), and particles (PM10, PM2.5). This study uses Artificial Neural Network (ANN) with time-based feature expansion to predict the classification of air quality indexes in Java Island for the next few months. While LSTM is used as a baseline for performance comparison with the proposed method. The results obtained show that the performance of the ANN model with time-based feature expansion can match the performance of LSTM with an accuracy of 92.30% and an F1 Score of 92.19%. This shows that the time-based feature expansion scenario in ANN is able to capture the spatial dynamics of time in the distribution of air quality in Java Island. The contribution of this study is to support the creation of effective policies and strategies in preventing and handling the impacts of air pollution as early as possible.

Keywords: Air quality index, Artificial Neural Network, Prediction, Time-based feature expansion, Spatial analysis, Java Island

I. Introduction

Air quality is one of the main environmental issues that requires serious attention, both in urban and rural areas. Increased transportation mobility and uncontrolled industrial activities have become a major source of air pollution, leading to a significant increase in pollutant emissions [1][2]. The decline in air quality has adverse effects on human health, including an increased risk of respiratory illnesses, and also harms the environment, contributing to issues like global warming, acid rain, and climate change [3].

The island of Java, as the region with the largest population and the most economically active in Indonesia, faces a high risk of air pollution. Transportation accounts for about 70% of the region's total air pollution, with emissions such as carbon monoxide (CO), nitrogen oxides (NOx), and dust particles (SPM10) frequently exceeding safe thresholds [4][5]. In addition to transportation, industrial activities also worsen air quality in the

Received on 01 Jan 2025. Revised on 8 Jan 2020. Accepted and Published on 11 Mar 2025.

region. Natural factors such as wind direction, seasonality, and topography also affect the distribution of pollutants, making the distribution pattern of air quality very dynamic [4].

Various efforts have been made to reduce the impact of air pollution, such as planting trees in urban areas. Trees play an important role in absorbing pollutants such as carbon dioxide (CO2) through photosynthesis, while also capturing dust particles in the atmosphere [6]. However, this step alone is not enough to significantly address the complexity of the air pollution problem. To understand and manage pollution more effectively, a technology-based approach that is able to accurately predict the air quality index is needed.

Traditionally, statistical methods such as linear regression have been widely used for air quality analysis. These methods are valued for their interpretability and effectiveness in modeling simple relationships between variables. However, they often struggle to capture complex, nonlinear interactions inherent in environmental data, especially when dealing with large and dynamic datasets [7]. In contrast, machine learning approaches like Artificial Neural Networks (ANNs) excel in identifying intricate patterns within such data, providing more robust and adaptive solutions for air quality prediction [8].

Air quality prediction plays an important role in supporting data-driven decision-making, environmental policy planning, evaluating the effectiveness of interventions, and mitigating pollution impacts. In this context, machine learning methods such as Artificial Neural Networks (ANNs) have become an attractive option. ANNs, with *backpropagation* algorithms, are able to identify patterns in complex historical data, even on limited and poorly structured datasets [9], [10]. The model offers a balance between simplicity, flexibility, and effectiveness in a variety of studies, including air quality prediction [11], [12].

This study employs the Artificial Neural Network model to predict the classification of air quality across Java Island. The model incorporates time-based features and spatial analysis to offer a more comprehensive understanding of air pollution distribution patterns. With this approach, the research aims to support more effective and strategic environmental management on the island of Java.

II. LITERATURE REVIEW

A. *Prediction*

Predictive modeling is crucial in many sectors, including healthcare, business, climate change, and transportation. [13]. In air quality research, the ability to predict future conditions is essential for understanding the dynamics of air pollution in a specific area. Previously, predictive approaches often relied on statistical methods that utilized temporal and spatial patterns. However, with advancements in technology, machine learning has increasingly become a widely used method. This technique excels not only in handling structured and unstructured data but can also be applied for various purposes, including regression, classification, clustering, and prediction [14].

B. *Artificial Neural Network*

An Artificial Neural Network (ANN) is a computational framework inspired by the human brain's structure, consisting of layers of interconnected neurons. Each connection between neurons is assigned a weight that signifies the strength of the relationship. The Perceptron, a basic ANN model, is a single-layer network that can only process linear data. To overcome this limitation, the Multi-Layer Perceptron (MLP) was developed, incorporating an input layer, one or more hidden layers, and an output layer. The addition of multiple hidden layers allows MLPs to solve non-linear problems and model more complex patterns in the data [15].

The MLP architecture is made up of multiple layers, where every neuron in one layer is entirely connected to the neurons in the subsequent layer. At each hidden layer, neurons output z_j calculated by the equation 1, where x_i is the input value, w_{ij} is the weight, b_j is the bias, and $\emptyset_h$ is the activation function on the hidden layer, usually using the ReLU function [16].

$$z_j = \emptyset_h \left(\sum_{i=1}^{n} w_{ij} x_i + b_j \right) \tag{1}$$

At the output layer, the output of neurons y_k calculated using equations 2, where represents the weight from the hidden layer to the output layer, is the bias in the output layer, and is the activation function in the output layer, usually the Softmax function, which is used to generate probabilities $v_{jk} b_k$ $\emptyset_o$[17].

$$y_k = \emptyset_\text{o} \left(\sum_{j=1}^{m} v_{jk} z_j + b_k \right) \tag{2}$$

The ReLU (Rectified Linear Unit) activation function transforms negative values into zero while preserving positive values, calculated using equation 3. The Softmax function, applied to the output layer, transforms the output value into probabilities using the equation 4, z_k where the output of the neuron is and represents the number of classes present. Softmax guarantees that the total of the probabilities across all classes equals 1, making it ideal for multi-class classification tasks.kK [18].

$$\text{ReLU}(x) = \max(0, x) \tag{3}$$

$$\text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_{i=1}^{K} e^{z_i}} \tag{4}$$

The training process of the MLP model utilizes a backpropagation algorithm to adjust the weights by minimizing prediction errors. The sparse categorical cross entropy loss function is used for multi-class classification tasks. To enhance convergence, the Adam optimizer is commonly used due to its efficiency in managing large and complex datasets [19]. In air quality prediction, the application of ANNs and MLPs is particularly effective, as they excel at capturing complex interactions between various environmental factors and air quality. These models deliver more accurate, efficient, and reliable predictions by accounting for the intricate relationships among multiple input variables [20].

C. Feature Expansion

Feature engineering methods, such as feature selection, feature expansion, and oversampling, are crucial for building classification models with Artificial Neural Networks (ANNs). Feature selection aims to identify the most important features from the dataset, while feature expansion involves creating new features from the existing data to reveal additional patterns and relationships [19]. This methodical process seeks to identify the best combination of features that most effectively capture the patterns in the data. The selected features are then employed to train the ANN model, enabling it to learn intricate relationships between input variables by assessing different feature combinations [19]. The process also involves dividing the data into training and testing sets, along with using the SMOTE (Synthetic Minority Over-sampling Technique) approach to tackle class imbalance, thereby improving the model's capacity to generalize [19]. These feature engineering techniques greatly enhance prediction accuracy and provide more reliable results, especially in cases with underrepresented data or complex non-linear relationships between features.

D. Data Collecting

This study uses air quality data provided by the Ministry of Environment and Forestry, covering the period from June 2019 to April 2022. The data, accessible in CSV format on the Ministry's official website and the open data platform for the Java Island region, consists of 11,114 entries, each containing various air quality indicators as detailed in Table 1. These indicators provide a detailed overview of pollution trends across different times and locations. The dataset spans several regions on Java Island, reflecting diverse environmental conditions, allowing the model to learn from a variety of scenarios, thus improving the accuracy and reliability of predictions. The data preprocessing involves addressing missing values, standardizing features, dividing the dataset into training and testing sets, and thoroughly assessing model performance.

TABLE I
DATA ATTRIBUTES DESCRIPTION

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	Air quality indexes

E. Air Pollution Standard Index

Based on the Regulation of the Minister of Environment and Forestry Number 14 of 2020, the Air Pollutant Standard Index (ISPU) is a dimensionless value that indicates the quality of the air in a particular area, considering its impact on human health and other living organisms [19]. This study concentrated on three ISPU categories: good, moderate, and unhealthy.

TABLE II
APSI PARAMETER VALUE CONVERSION

APSI	**24Hour PM10** $(\mu g/m^3)$	**24Hour PM2.5** $(\mu g/m^3)$	**24Hour SO2** $(\mu g/m^3)$	**24Hour CO** $(\mu g/m^3)$	**24Hour O3** $(\mu g/m^3)$	**24Hour NO2** $(\mu g/m^3)$	**24Hour HC** $(\mu g/m^3)$
0-50	50	15,5`	52	4000	120	80	25
51-100	150	55,4	108	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) Data from continuous measurements conducted over a 24-hour period.
b) Hourly results of APSI calculations for particulates (PM2.5) recorded over 24 hours.
c) The maximum and minimum APSI values for each hour are chosen for particulates (PM10), sulfur dioxide (SO2), carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), and hydrocarbons (HC) to be used as the calculation outcomes.

The ISPU calculation relies on several key parameters, including the upper and lower limits of ISPU, the upper and lower limits of ambient, and the measurable ambient concentration value. The mathematical formulation for this calculation is given by the equation 5.

$$I = \frac{I_a - I_b}{X_a - X_b}(X_a - X_b) + I_b \tag{5}$$

The results of data analysis show that the ISPU classification in this study is divided into three levels: good, moderate, and unhealthy, by adopting AQI color standards to facilitate the visualization of air quality status. Good categories are marked with green, medium with yellow, and unhealthy with red. This categorization allows for a more structured analysis of air quality trends in Java over a 15-month study period. This division into three specific categories with color markers provides a clear framework for analyzing patterns of air quality change and designing effective treatment strategies.

TABLE III
AIR QUALITY INDEX CATEGORY

Label	Category	Color Status	Range
0	Good	🟩	1-50
1	Medium	🟨	51-100
2	Unhealty	🟧	101-200

III. RESEARCH METHOD

The research process for predicting air quality distribution classifications on Java Island using a time-based Artificial Neural Network (ANN) began with designing a feature extension based on historical time data. This was followed by the development of an ANN model incorporating time-based feature expansion. Subsequently, an optimization process was conducted to identify the most effective combination of features for each dataset.

A. Data Matrix Design with Feature Expansion

Feature expansion is employed for selecting and iterating through various combinations of relevant features to enhance model performance [19]. This approach supports the development of a classification model that predicts the standard air pollution index in Java. The process involves randomly selecting feature combinations from the dataset, enabling the model to explore a broader range of variables and patterns. By assessing these combinations based on the model's performance, it identifies the optimal subset of features that improve prediction accuracy. This method results in more precise and dependable air pollution predictions, which are crucial for environmental monitoring and effective decision-making in Java.

TABLE IV
DATA CLASS LABEL

Models	Combination	Training Data Attributes	Target
t-1	1	June 2019	July 2019
	2	July 2019	August 2019
	:	:	:
	34	March 2022	April 2022
t-2	1	June 2019 – July 2019	August 2019
	2	July 2019 – August 2019	September 2019
	:	:	:
	33	February 2022 – March 2022	April 2022
:	:	:	:
t-33	1	June 2019 – February 2022	March 2022
	2	July 2019 – March 2022	April 2022
t-34	1	June 2019 – March 2022	April 2022

The features selected for expansion include critical air pollution indicators such as PM10, CO, and NO2, which are directly related to human health risks and environmental impacts. Studies have shown that PM10 is a significant contributor to respiratory diseases, while CO and NO2 play a major role in the development of air quality-related health problems, such as asthma and cardiovascular conditions [21]. These pollutants are key components of the air quality index and their inclusion in the feature expansion process enhances the model's ability to predict pollution levels accurately.

In this study, the exploration of feature combinations was carried out randomly as much as possible until the best combination was found based on the F1 score value generated by the classification model. This approach refers to iterative exploratory methods that are often used to identify relevant features in data-driven models [22]. This process guarantees that the chosen features are the most relevant for enhancing the model's accuracy and efficiency, while preserving key information from the dataset. By focusing on the best combination of features, the model becomes better equipped to generate accurate and reliable predictions on air pollution data [23].

B. Design of Feature Selection and Expansion for Artificial Neural Network-based Classification

Artificial Neural Network (ANN) is used to build a classification prediction model in the $t - k$ and predict the classification on $t + k$. The implementation of ANN is carried out through two main stages, namely the identification of the best combination of features in the $t - k$ and the application of ANN to prediction $t + k$.

1) Implementation of Artificial Neural Network in Phase $t - k$

In the $t - k$, ANNs are used to evaluate different combinations of features to determine the best subset of features that provide the highest classification performance. This process begins with data pre-processing, where the SMOTE technique is applied to address the class imbalance. SMOTE creates a synthetic sample by calculating the nearest k-neighbors for each minority sample and generating a new sample using the equation 6.

$$x_{new} = x_{minority} + \text{rand}(0{,}1) \cdot (x_{neighbor} - x_{minority}) \quad (6)$$

After pre-processing, feature combinations are evaluated using an ANN, with the data divided into training and test sets in an 80:20 ratio. The ANN model is trained on the training data and assessed using a weighted F1-score metric, calculated using equation 7.

$$F1_{weighted} = \sum_{k=1}^{C} w_k \cdot F1_k \quad (7)$$

The ANN architecture consists of three hidden layers, with a ReLU activation function applied to the hidden layers and a Softmax activation function used in the output layer. The output of each hidden layer is calculated through linear operations, which are subsequently processed by their respective activation function.

2) Implementation of Artificial Neural Network in Phase $t + k$

In the t + k, Artificial Neural Network (ANN) model is used to predict classification by using the features that have been selected in the $t - k$. The ANN model applied in this phase has a similar architecture to the previous model, with inputs in the form of identified features and outputs calculated using the Softmax layer for class prediction. The output of the model is calculated using equation 8, $\widehat{y_\iota}$is the prediction probability for the ith class, z_i is the output of the previous layer for class I, and C is the number of classes.

$$\widehat{y_\iota} = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (8)$$

The model is trained using Adam's optimization algorithm with a sparse_categorical_crossentropy loss function, which is defined in equation 10, where N is the number of samples, C is the number of classes, y_{ik} is the actual label for the ith sample and the class ke-k, and $\hat{y}_{i,k}$ is the predicted probability for the kth class. The training process was carried out by oversampling techniques using SMOTE to handle class imbalances and normalize features using MinMaxScaler.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} y_{ik}\log\left(\hat{y}_{i,k}\right) \tag{10}$$

C. Interpolation kriging ordinary

Interpolation is a technique of structuring values in areas without data to describe the distribution of values in an area. Among the various interpolation methods, Kriging exists as a geostatistical approach that analyzes spatial relationships based on distances and orientations between sample points. This method applies a series of mathematical functions through systematic stages including statistical analysis, variogram construction, and surface generation [24].

Specifically, Ordinary Kriging implements the spatial concept by utilizing sample values and variograms to estimate values at points that have not yet been measured. The accuracy of the prediction depends largely on the degree of proximity to the locations that already have measurement data [25].

$$\hat{y}_{(t+k)}(S_o) = \sum\nolimits_{i=0}^{N} \lambda_i^{0K}\, y_{t+k}(S_i) \tag{11}$$

The equation 11 shows the mechanism of value estimation $\hat{y}_{(t+k)}(S_o)$ at t+k time for a location S_o. This process involves values $y_{t+k}(S_i)$ from N nearby locations S_i , with Ordinary Kriging weight λ_i^{0K} optimized to integrate spatial data, resulting in accurate predictions at the $\hat{y}_{(t+k)}(S_o)$.

D. Evaluation

The confusion matrix, which includes True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN), will be utilized to evaluate the performance of the Artificial Neural Network (ANN) classification method. From this matrix, metrics like accuracy, precision, recall, and F1 score are derived. Accuracy indicates the proportion of correct predictions, precision assesses the accuracy of positive predictions, recall gauges the model's ability to identify positive cases, and the F1 score combines precision and recall for a comprehensive evaluation, particularly for imbalanced datasets. Based on the confusion in Table V, the matrix will be used to calculate some of the evaluation matrices in equation 12 - 15.

TABLE V
CONFUSION MATRIX

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

$$\text{Accuracy} \quad = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100 \tag{12}$$

$$Precision \quad = \frac{TP}{\text{TP} + \text{FP}} \times 100 \tag{13}$$

$$Recall = \frac{TP}{\text{TP} + \text{FN}} \times 100 \tag{14}$$

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \times 100 \tag{15}$$

E. Experiments

Experiments were conducted to assess and compare the performance of the Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) classification methods using an identical dataset. In the ANN method, classification is done by applying feature expansion to explore relevant feature combinations that enhance prediction accuracy. On the other hand, the LSTM experiment solely uses the target class without incorporating additional features, allowing the LSTM to generate the feature representation automatically.

IV. Results And Discussion

A. Result

This section presents the results of implementing and assessing the ANN classification model with the Time-Based Feature Expansion approach, which incorporates time-based features t − k, based on a dataset of 34 available entries.

1) Best Perfomance of t-k Artificial Neural Network

The t-k Artificial Neural Network (ANN) Model exhibited remarkable performance, as evidenced by the confusion matrix values presented in the Table VI. These values highlight the model's effectiveness in accurately

TABLE VI
BEST PERFORMANCE T-K ANN TIME-BASED MODEL

Scenario	**Optimal**	**Perfomance**			
	Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
t-1	1o	90,00%	91,67%	90,00%	89,90%
t-2	2f	80,00%	85,71%	80,00%	79,17%
t-3	3n	80,00%	85,71%	80,00%	79,17%
t-4	4z	84,61%	87,44%	84,62%	82,83%
t-5	5x	85,71%	86,90%	85,71%	85,53%
t-6	6y	92,30%	93,59%	92,31%	92,11%
:	:	:	:	:	:
:	:	:	:	:	:
t-33	33a	63,49%	78,57%	63,54%	62,37%
t-34	34	62,38%	63,39%	62,59%	61,93%

The analysis reveals that the highest performance was achieved in scenarios t-6, t-8, t-14, t-17, t-21, t-24, and t-25, with accuracy ranging from 92.30% to 92.85% and F1-scores between 92.11% and 92.67%, demonstrating high stability. Models in scenarios t-5, t-7, t-9, and t-10 also exhibited consistent performance with accuracy above 85%, despite minor fluctuations. Conversely, significant performance declines were observed in scenarios t-33 and t-34, with accuracy values of 63.49% and 62.38%, and F1-scores of 62.37% and 61.93%, likely due to suboptimal feature combinations or model configurations. Overall, the pattern indicates that models with high accuracy tend to exhibit greater stability, while those with lower performance show sharper declines, providing critical insights for selecting optimal models in similar applications.

2) Comparison of ANN and LSTM Model Performance

The experimental results presented in Table 1 demonstrate that ANNs surpass LSTMs, especially in terms of accuracy, precision, recall, and F1-score. Based on these outcomes, this study chooses ANN as the main classification method. This decision is based on ANN's excellence in producing more accurate predictions as well as higher training efficiency. In contrast to LSTM which requires sequential data analysis, ANN is faster and more effective in handling non-sequential datasets, making it more suitable for air quality monitoring applications in Java.

TABLE VII
COMPARISON OF ANN AND LSTM MODEL PERFORMANCE BASED ON EVALUATION METRICES

	Artificial Neural Network				**Long Short Term Memory**			
Model	**accuracy**	**precision**	**recall**	**f1_score**	**accuracy**	**precision**	**recall**	**f1_score**
t-1	90,00%	91,67%	90,00%	89,90%	92,86%	94,05%	92,86%	92,67%
t-2	80,00%	85,71%	80,00%	79,17%	77,78%	83,33%	77,78%	73,81%
t-3	80,00%	85,71%	80,00%	79,17%	87,54%	93,71%	92,41%	93,06%
t-4	84,61%	87,44%	84,62%	82,83%	76,60%	72,69%	79,32%	75,86%
t-5	85,71%	86,90%	85,71%	85,53%	69,93%	61,88%	66,92%	64,30%
:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:
t-34	62,30%	63,30%	62,50%	61,90%	50,00%	25,00%	50,00%	33,33%

TABLE VIII
OPTIMAL FEATURE COMBINATION FOR PREDICTION MODELS

Scenario	**Optimal Model**	**Optimal Features**
t-1	1o	x113,x17,x16,x12,x15
t-2	2f	x210,x28,x23,x212,x211,x25,x16,x26,x18,x24,x111,x14
t-3	3n	x11,x110,x313,x35,x23,x15,x38,x14,x113,x212,x24,x311,x310,x36,x18,x34,x27,x29,x12, x31,x21,x22,x32,x210,x211,x37
t-4	4z	x13,x12,x33,x22,x27,x49,x110,x410,x34,x24,x311,x112,x310,x38,x17,x42,x26,x43,x212, x36,x113,x32,x39,x15,x211,x11,x46,x14,x21,x210,x16,x312,x35
t-5	5x	x42,x24,x34,x39,x57,x28,x12,x210,x33,x112,x410,x29,x212,x411,x38,x311,x512,x510,x51,x13,x36,x49,x48,x513,x19,x15,x113,x310,x37,x213,x313,x31,x110,x47,x35,x11,x211,x16,x412,x22,x58,x26,x43,x55,x41,x312,x21,x45,x27
t-6	6y	x21,x612,x13,x511,x49,x63,x611,x54,x112,x48,x27,x312,x610,x613,x46,x38,x44,x412,x410,x18,x41,x17,x67,x11,x28,x110,x210,x29,x51,x211,x12,x513,x69,x52,x26,x25,x313,x61,x45,x212,x42,x310,x58,x55
:	:	:
:	:	:
:	:	:

3) Optimal Feature Combination

The table below presents the optimal feature combinations for the prediction models. These combinations are designed to improve the models' predictive accuracy. Additionally, they contribute to the overall reliability of the predictions. The results indicate that the best-performing combinations do not always utilize all features. In scenario t-34, excluding certain features improved performance, emphasizing the importance of automated feature selection in optimizing model accuracy and efficiency.

TABLE IX
CLASS PREDICTION OF AIR POLLUTION DISTRIBUTION FROM MAY 2022 TO APRIL 2023

Location	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Des**	**Jan**	**Feb**	**Mar**	**Apr**
	2022	**2022**	**2022**	**2022**	**2022**	**2022**	**2022**	**2022**	**2023**	**2023**	**2023**	**2023**
Jend.Sudirman Tangerang	1	0	0	2	2	2	2	2	2	0	0	0
Sudimara Ciledug	0	1	0	2	2	2	2	2	2	0	0	0
A. Yani Semarang	1	1	0	1	0	1	0	0	0	0	0	0
Balai Kota Depok	1	0	0	0	0	0	0	2	0	0	0	0
Bandung	0	1	0	1	0	0	1	0	0	0	1	0
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
Wonorejo Surabaya	1	0	0	1	0	1	0	0	0	0	0	0

TABLE X
CLASS PREDICTION OF AIR POLLUTION DISTRIBUTION FROM MAY 2023 TO APRIL 2024

Location	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Des**	**Jan**	**Feb**	**Mar**	**Apr**
	2023	**2023**	**2023**	**2023**	**2023**	**2023**	**2023**	**2024**	**2024**	**2024**	**2024**	**2024**
Jend.Sudirman Tangerang	2	2	2	1	2	2	2	2	2	2	2	2
Sudimara Ciledug	2	2	2	0	2	2	2	2	2	2	2	2
A. Yani Semarang	0	0	1	1	0	1	0	0	0	0	0	0
Balai Kota Depok	0	0	2	0	2	2	0	0	0	0	0	0
Bandung	0	0	0	1	0	0	0	0	0	0	0	0
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
Wonorejo Surabaya	0	0	1	1	0	0	0	0	0	1	0	0

4) Best T+K Artificial Neural Network Time-Based Model

The results from t-k, processed using the same algorithm to generate predictions for t+k, demonstrate the algorithm's consistency and reliability. These predictions underscore the model's capability to generalize effectively across different scenarios, highlighting its potential for robust forecasting in similar contexts.

TABLE XI
CLASS PREDICTION OF AIR POLLUTION DISTRIBUTION FROM MAY 2024 TO FEBRUARY 2025

Location	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Des**	**Jan**	**Feb**
	2024	**2024**	**2024**	**2024**	**2024**	**2024**	**2024**	**2024**	**2025**	**2025**
Jend.Sudirman Tangerang	2	0	2	1	2	1	1	0	0	0
Sudimara Ciledug	2	0	2	0	2	0	0	0	0	0
A. Yani Semarang	0	1	0	0	0	0	0	1	1	0
Balai Kota Depok	0	0	0	0	0	0	0	0	0	0
Bandung	0	0	0	1	0	0	1	1	0	1
:	:	:	:	:	:	:	:	:	:	:
Wonorejo Surabaya	0	1	0	1	0	1	1	1	1	1

5) Visualization of Air Pollution distribution classification Map

The air quality index prediction results generated by the Artificial Neural Network model can be utilized to create a map depicting the distribution of air pollution across Java. This map serves to visually and informatively represent the air quality variations in various regions.

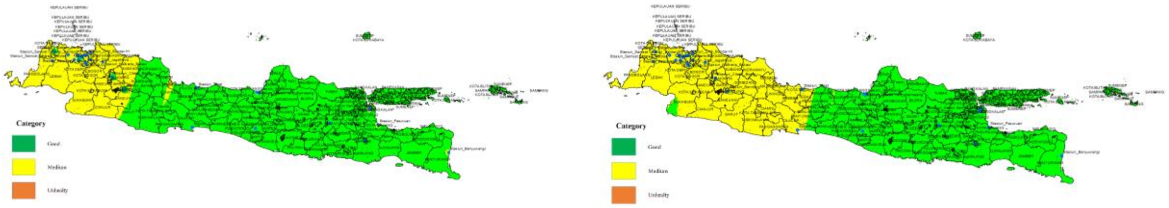

Fig. 1. Prediction Map of Air Pollution Distribution May 2022 to June 2022

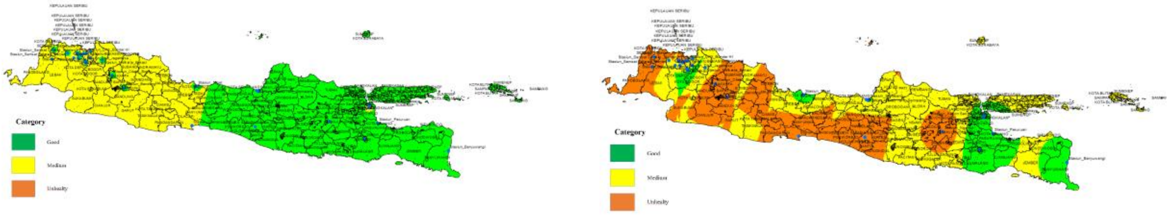

Fig. 2. Prediction Map of Air Pollution Distribution July 2022 to August 2022

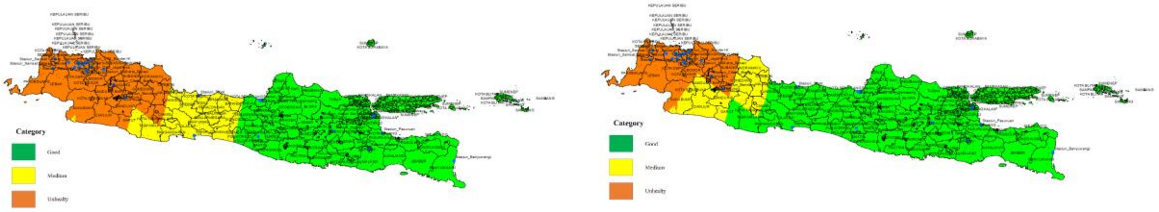

Fig. 3. Prediction Map of Air Pollution Distribution September 2022 to October 2022

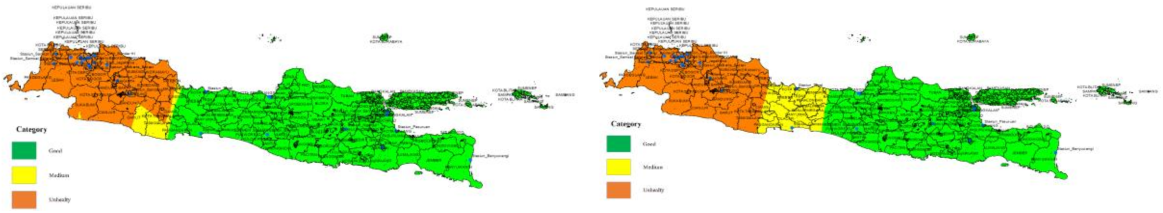

Fig. 4. Prediction Map of Air Pollution Distribution November 2022 to December 2022

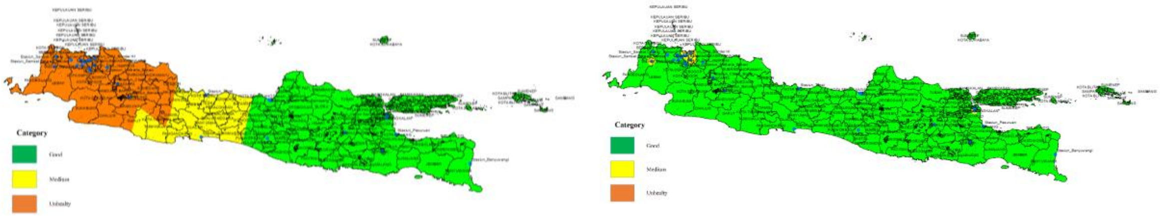

Fig. 5. Prediction Map of Air Pollution Distribution January 2023 to February 2023

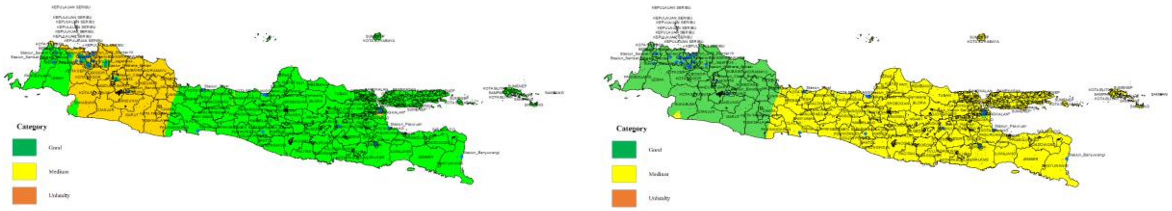

Fig. 5. Prediction Map of Air Pollution Distribution March 2023 to April 2023

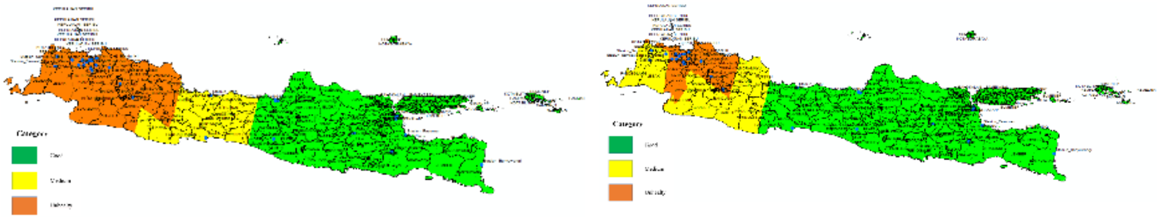

Fig. 6. Prediction Map of Air Pollution Distribution May 2023 to June 2023

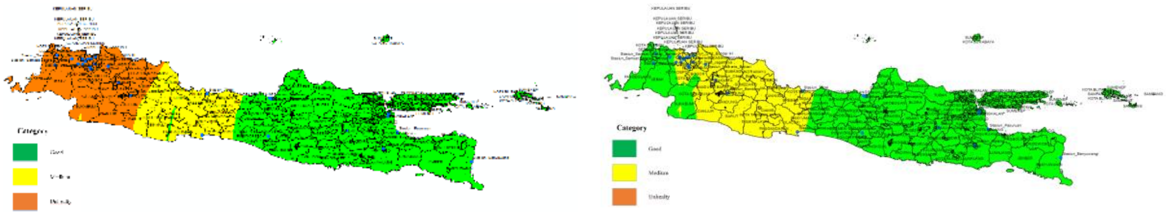

Fig. 7. Prediction Map of Air Pollution Distribution July 2023 to August 2023

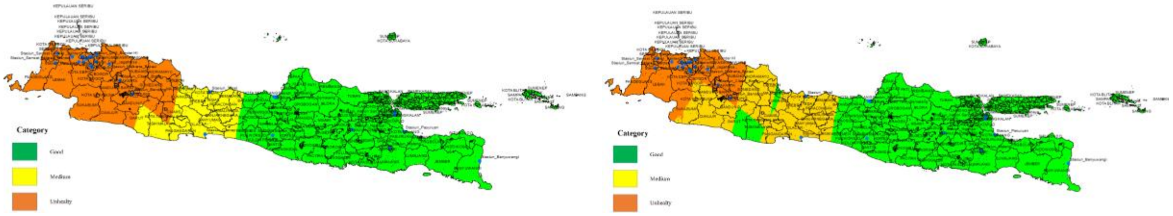

Fig. 8. Prediction Map of Air Pollution Distribution September 2023 to October 2023

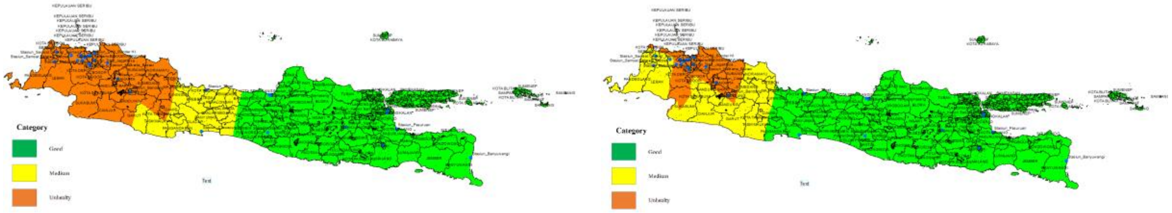

Fig. 9. Prediction Map of Air Pollution Distribution November 2023 to December 2023

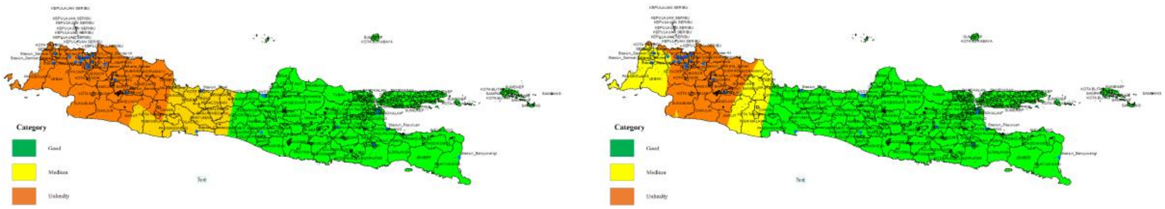

Fig. 10. Prediction Map of Air Pollution Distribution January 2024 to February 2024

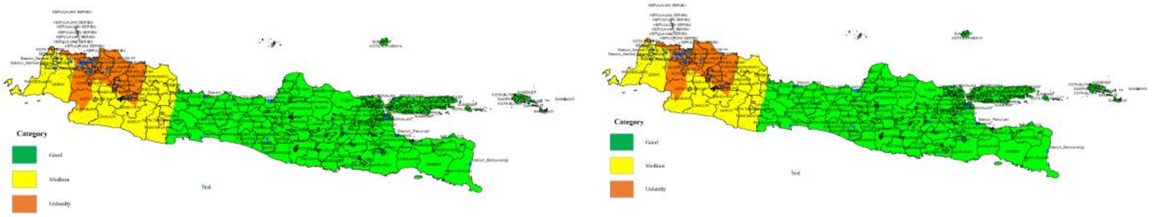

Fig. 11. Prediction Map of Air Pollution Distribution March 2024 to April 2024

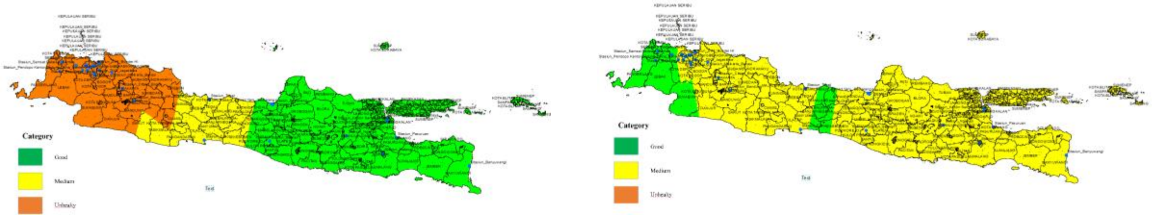

Fig. 12. Prediction Map of Air Pollution Distribution May 2024 to June 2024

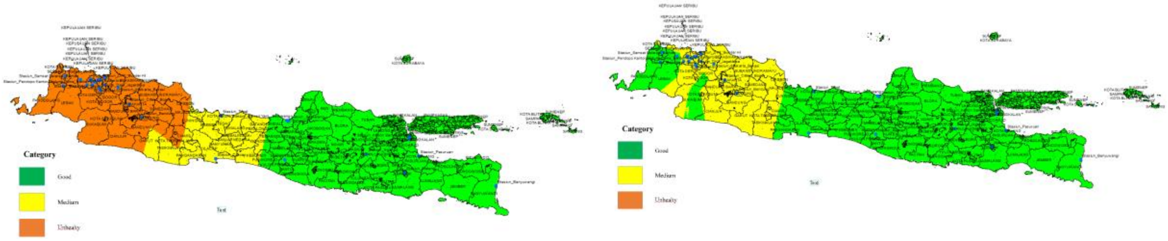

Fig. 13. Prediction Map of Air Pollution Distribution July 2024 to August 2024

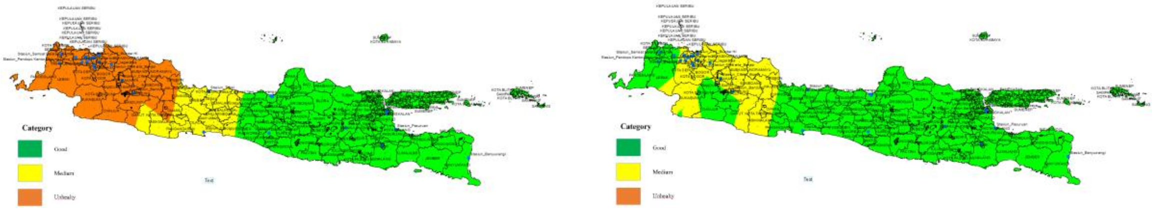

Fig. 14. Prediction Map of Air Pollution Distribution September 2024 to October 2024

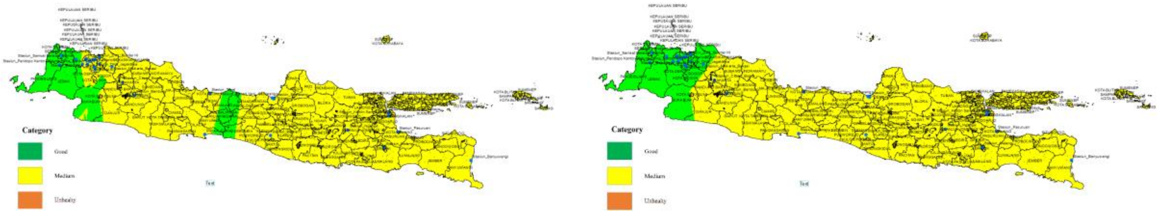

Fig. 15. Prediction Map of Air Pollution Distribution November 2024 to Desember 2024

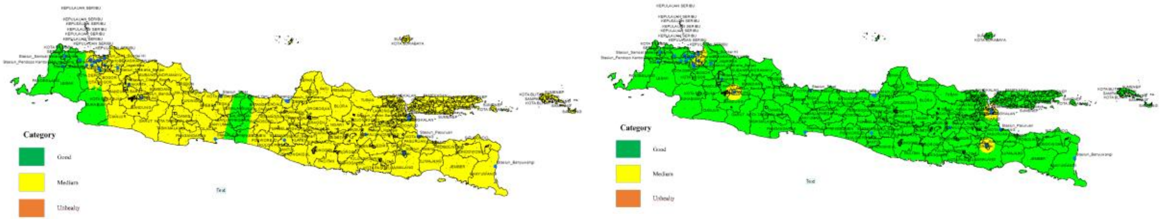

Fig. 17 Prediction Map of Air Pollution Distribution January 2025 to February 2025

B. Discussion

This research demonstrates that the Artificial Neural Network (ANN) model, combined with a time-based feature expansion approach, performs effectively in predicting air quality on Java. Experimental results indicate that the ANN achieves up to 92.30% accuracy and a 92.19% F1-score in the t-6 scenario, highlighting its capability to manage non-linear variable relationships. The steady improvement in accuracy and F1-score up to this scenario underscores the ANN's proficiency in delivering reliable predictions for short- to medium-term forecasts. However, its performance begins to decline beyond the t-10 scenario, likely due to the increased complexity of longer-term predictions.

The identification of dominant features such as PM10, SO2, CO, O3, and NO2 is a key factor in the success of the model in generating accurate predictions. These features contribute significantly to helping Artificial Neural Network (ANN) recognize air quality distribution patterns that are often difficult to handle by conventional methods or models such as Long Short-Term Memory (LSTM). To provide additional insight, correlation analysis showed that PM10 demonstrated the strongest relationship with the air quality index, followed by SO2 and NO2. In contrast, features such as wind speed showed a much weaker relationship, indicating their limited impact on predictions. This reinforces the selection of relevant features in improving model performance. In addition, time-based feature expansion methods allow for the exploration of various combinations of input variables, providing the model with the ability to capture complex temporal patterns [26]. The iterative approach to feature selection also revealed that certain combinations of features, such as PM10, NO2, and population density, consistently improved prediction accuracy. This highlights the importance of targeted feature engineering for specific scenarios.

Spatial visualization through kriging interpolation adds to the understanding of air quality distribution, especially in urban areas such as DKI Jakarta and Surabaya, which have higher concentrations of air pollution due to high vehicle and industrial activity. Jakarta's dense urbanization and limited airflow exacerbate pollutant concentration, while seasonal factors, such as lower rainfall during dry months, contribute to temporarily elevated PM10 and CO levels. This underscores the importance of seasonal adjustments in predictive models. This pattern emphasizes the need for mitigation policies that focus on areas with high population density and economic activity. In terms of policy applications, the results suggest several actionable strategies. For example, predictive data can be used to enforce vehicle restrictions in high-pollution zones during peak times or optimize the placement of air quality sensors in areas predicted to exhibit high pollution variability. Such measures can enhance the effectiveness of pollution control efforts while prioritizing resource allocation.

However, this study has several limitations that need to be considered. Reliance on the quality and availability of sensor data is a major challenge, especially in regions with inadequate data coverage. In addition, the Artificial Neural Network (ANN) model is still not able to capture sudden changes in air quality due to external factors[27], such as natural disasters or extreme weather conditions. These limitations indicate the need for additional data integration and exploration of hybrid methods to strengthen model performance, especially in long-term predictions.

This study offers significant insights for environmental management efforts on the island of Java. The ANN models have the potential to facilitate data-informed decision-making, including the strategic placement of air quality sensors and the formulation of more effective air pollution control policies. Future research could investigate the integration of real-time data and spatial modeling enhancements, such as dynamic interpolation methods, to improve prediction accuracy in rapidly changing environments. Additionally, exploring hybrid models combining ANN with deep learning architectures like RNN-LSTM or transformers may further optimize long-term forecasting capabilities. These approaches are anticipated to contribute sustainably to improving air quality management practices.

V. CONCLUSION

This study demonstrates that the ANN model with time-based feature expansion significantly enhances the accuracy of air quality predictions across Java, achieving a maximum accuracy of 92.30% and an F1-score of 92.19% in the t-6 scenario, proving its effectiveness for short- to medium-term predictions. The identification of dominant features, including PM10, SO2, CO, O3, and NO2, validates the model's capacity to accurately capture air pollution distribution patterns. These findings provide valuable insights for air quality management, such as optimized sensor placement strategies and evidence-based mitigation policy development. Future research should focus on enriching the dataset with additional attributes, integrating real-time data streams, and exploring advanced machine learning models, such as hybrid or transformer-based architectures, to further enhance prediction accuracy and adaptability to sudden environmental changes.

ACKNOWLEDGMENT

The authors wish to thank Telkom University for the support of funds and infrastructure, so that this research can be completed.

REFERENCES

[1] S. Aswatha *et al.*, “Smart Air Pollution Monitoring System,” *Global Nest Journal*, vol. 25, no. 3, pp. 125–129, Mar. 2023, doi: 10.30955/gnj.004396.

[2] M. A. Fath, “Pengaruh Kualitas Udara Dan Kondisi Iklim Terhadap Perekonomian Masyarakat (Literature Review),” *Media Gizi Kesmas*, vol. 10, no. 2, pp. 2021–329, 2021.

[3] K. Kumar and B. P. Pande, “Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities,” *International Journal of Environmental Science and Technology*, vol. 20, no. 5, pp. 5333–5348, May 2023, doi: 10.1007/s13762-022-04241-5.

[4] S. ’Mulyono, “KLHK Sebut Risiko Pencemaran Udara di Jawa Tinggi, Tidak Hanya di Jakarta,” JawaPos.com. Accessed: Aug. 11, 2024. [Online]. Available: https://www.jawapos.com/nasional/012704898/klhk-sebut-risiko-pencemaran-udara-di-jawa-tinggi-tidak-hanya-di-jakarta

[5] N. Kusminingrum and G. Gunawan, “Polusi Udara Akibat Aktivitas Kendaraan Bermotor di Jalan Perkotaan Pulau Jawa dan Bali,” *Jurnal Jalan Jembatan*, vol. 25, no. 3, pp. 13–13, 2008.

[6] N. K. T. Martuti, “The role of plants against air pollution in the protocol street of Semarang city,” *Journal Biosantifika*, vol. 5, no. 1, pp. 36–42, 2013. [Online]. Available: http://journal.unnes.ac.id/nju/index.php/biosaintifika

[7] M. Reza Akbar, M. Ihsan Akbar, dan Rizki Achmad Darajatun, U. Singaperbangsa Karawang Jl HSRonggo Waluyo, K. Karawang, and J. Barat, “Analisis Regulasi Uji Emisi Gas Buang Kendaraan Berdasarkan Pengaruhnya Terhadap Indeks Kualitas Udara di DKI Jakarta Menggunakan Metode Korelasi Pearson dan Regresi Linear,” *J. Statistika*, vol. 15, no. 1, 2022. [Online]. Available: www.unipasby.ac.id

[8] O. Dwi *et al.*, “Perbandingan hasil peramalan dengan metode double exponential smoothing Holt dan metode jaringan syaraf tiruan,” [Online]. Available: www.dinkesjatim.go.id/imunisasi. [Accessed: 25-November-2024].

[9] A. Singh Bharatpur, “A Literature Review on Time Series Forecasting Methods,” 2022.

[10] H. Liu, G. Yan, Z. Duan, and C. Chen, "Intelligent modeling strategies for forecasting air quality time series: A review," *Appl. Soft Comput.*, vol. 102, p. 106957, 2021. doi: 10.1016/j.asoc.2020.106957.

[11] R. Qamar and B. A. Zardari, "Artificial neural networks: An overview," *Mesopotamian J. Comput. Sci.*, vol. 2023, pp. 124-133, 2023. doi: 10.58496/mjcsc/2023/015.

[12] J. T. Hardinata, M. Zarlis, E. B. Nababan, D. Hartama, and R. W. Sembiring, "Modification of Learning Rate with Lvq Model Improvement in Learning Backpropagation," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Dec. 2017. doi: 10.1088/1742-6596/930/1/012025.

[13] M. Robnik, "Explanation of Prediction Models With Explain Prediction," 2018. [Online]. Available: https://fri.uni-lj.si/en/employees/marko-robnik-sikonja. [Accessed: Nov. 25, 2024].

[14] B. Nikparvar and J. C. Thill, "Machine learning of spatial data," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 9, p. 600, 2021. doi: 10.3390/ijgi10090600.

[15] K. L. Du, C. S. Leung, W. H. Mow, and M. N. S. Swamy, "Perceptron: Learning, generalization, model selection, fault tolerance, and role in the deep learning era," *Mathematics*, vol. 10, no. 24, p. 4730, 2022. doi: 10.3390/math10244730.

[16] Y. S. Park and S. Lek, "Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling," in *Developments in Environmental Modelling*, vol. 28, Elsevier B.V., 2016, pp. 123–140. doi: 10.1016/B978-0-444-63623-2.00007-4.

[17] S. S. Prasetiyowati, A. Yahya, and A. A. Rohmawati, "Performance of time-based feature expansion in developing ANN classification prediction models on time series data," *International Journal on Information and Communication Technology (IJoICT)*, vol. 9, no. 2, pp. 162-176, 2023, doi: 10.21108/ijoict.v9i2.868.

[18] R. D. Ramadhani, A. N. A. Thohari, C. Kartiko, A. Junaidi, T. G. Laksana, and N. A. S. Nugraha, "Optimasi akurasi metode convolutional neural network untuk identifikasi jenis sampah," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 312-318, 2021, doi: 10.29207/resti.v5i2.2754.

[19] M. Zare, H. R. Pourghasemi, M. Vafakhah, and B. Pradhan, "Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: A comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms," *Arabian Journal of Geosciences*, vol. 6, no. 8, pp. 2873–2888, Aug. 2013, doi: 10.1007/s12517-012-0610-x.

[20] H. Putra and N. Ulfa Walmi, "Penerapan Prediksi Produksi Padi Menggunakan Artificial Neural Network Algoritma Backpropagation," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 6, no. 2, pp. 100–107, Sep. 2020, doi: 10.25077/teknosi.v6i2.2020.100-107.

[21] N. P. D. Arwini, "Dampak pencemaran udara terhadap kualitas udara di Provinsi Bali," *Jurnal Ilmiah Vastuwidya*, vol. 2, no. 2, pp. 20-30, 2019.

[22] 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 Computing Electronics and Control)*, vol. 16, no. 3, pp. 1345–1350, Jun. 2018, doi: 10.12928/TELKOMNIKA.v16i3.7751.

[23] T. Desyani, A. Saifudin, and Y. Yulianti, “Feature Selection Based on Naive Bayes for Caesarean Section Prediction,” in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Aug. 2020. doi: 10.1088/1757-899X/879/1/012091.

[24] H. Purnomo and R. A. E. Wijaya, “Pemetaan Sebaran Kadar Al2O3 dan RSiO2 pada Endapan Laterit Bauksit Menggunakan Pendekatan Metode Interpolasi Ordinary Kriging Dan Inverse Distance Weighting,” *Angkasa: Jurnal Ilmiah Bidang Teknologi*, vol. 14, no. 1, May 2022, doi: 10.28989/angkasa.v14i1.1227.

[25] E. Respatti *et al.*, “Perbandingan Metode Ordinary Kriging dan Inverse Distance Weighted untuk Estimasi Elevasi Pada Data Topografi (Studi Kasus: Topografi Wilayah FMIPA Universitas Mulawarman) Comparison of Ordinary Kriging and Inverse Distance Weighted Methods for Estimation of Elevations Using Topographic Data (Case Study: FMIPA University of Mulawarman’s Topographic),” *Jurnal Eksponensial*, vol. 5, no. 2, 2014.

[26] A. H. Suhendar, A. A. Rohmawati, and S. S. Prasetyowati, “Performance of CART Time-Based Feature Expansion in Dengue Classification Index Rate,” *Sinkron Jurnal dan Penelitian Teknik Informatika*, vol. 9, no. 1, pp. 1–9, Jan. 2024, doi: 10.33395/sinkron.v9i1.13023.

[27] F. Rozi, F. Sukmana, J., "Penggunaan Moving Average Dengan Metode Hybrid Artificial Neural Network Dan Fuzzy Inference System Untuk Prediksi Cuaca," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 1, no 02, 2016.