Environment and Natural Resources Journal

Volume 23
Issue 3
2025

AdjcorT-RBFNN for Air Quality Classification: Mitigating Multicollinearity with Real and Simulated Data

Siti Khadijah Arafin[1], Nor Azura Md Ghani[1], Marshima Mohd Rosli[2,3], and Nurain Ibrahim[1,4*]

[1]*School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia*
[2]*School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia*
[3]*Cardiovascular Advancement and Research Excellence (CARE Institute), Sungai Buloh Campus, 47000 Sungai Buloh, Selangor, Malaysia*
[4]*Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia*

ARTICLE INFO

Received: 3 Jan 2025
Received in revised: 4 Mar 2025
Accepted: 10 Mar 2025
Published online: 24 Apr 2025
DOI: 10.32526/ennrj/23/20250006

Keywords:
AdjcorT/ AdjcorT-RBFNN/ Air pollution/ Air quality prediction/ Particulate matter PM2.5/ RBFNN

*** Corresponding author:**
E-mail:
nurain@tmsk.uitm.edu.my

ABSTRACT

Air pollution levels have remained a significant issue worldwide despite advancements in technology, primarily due to rapid industrialization and urbanization. Among the various pollutants, PM2.5 significantly impacts air quality, posing health risks such as respiratory and cardiovascular diseases. Accurate prediction of PM2.5 levels is essential for effective air quality management. However, multicollinearity in air quality data can hinder model performance. To address this issue, this study introduces the AdjcorT-RBFNN, a two-stage feature selection method, to classify air quality in Klang, Selangor. The AdjcorT-RBFNN model selects the optimal combination of 9 feature combinations from 10 variables and outperforms the RBFNN model, which uses all 10 variables. With 7 hidden nodes and a learning rate of 0.01 for both models, AdjcorT-RBFNN achieves higher accuracy (0.62), sensitivity (0.64), specificity (0.60), precision (0.60), F1 score (0.62), and AUROC (0.62), confirming its effectiveness in classification tasks. The optimal features for predicting air quality in Klang are identified as PM2.5, PM10, relative humidity, SO_2, wind direction, O_3, CO, ambient temperature, and NO_2. Monte Carlo simulations validate the model's effectiveness, showing that AdjcorT-RBFNN consistently outperforms RBFNN, especially with strong negative correlations (ρ=-0.8) and larger sample sizes (N=150 and 200) further enhance classification accuracy. Compared to RBFNN, AdjcorT-RBFNN enhances class discrimination and reduces false positives, improving its reliability in detecting true classifications. These findings highlight the importance of feature selection in improving model performance, particularly in datasets with multicollinearity. Researchers, and health organizations can leverage AdjcorT-RBFNN for more accurate air quality predictions, supporting informed pollution control strategies.

1. INTRODUCTION

Air quality prediction has emerged as a significant issue in recent years, due to the increasing effects of air pollution on human health, climate change, and ecosystems. PM2.5, fine particulate matter with a diameter of 2.5 micrometers or smaller, is a major air pollutant often associated with severe health risks. These particles are small enough to be inhaled deeply into the lungs, and because of their size, they can also enter the bloodstream. PM2.5 is primarily generated from industrial emissions, vehicle exhaust, biomass burning, and other sources of combustion, as well as natural sources such as dust storms and wildfires. Due to its fine nature, PM2.5 can

Citation: Arafin SK, Ghani NAM, Rosli MM, Ibrahim N. AdjcorT-RBFNN for air quality classification: Mitigating multicollinearity with real and simulated data. Environ. Nat. Resour. J. 2025;23(3):242-255. (https://doi.org/10.32526/ennrj/23/20250006)

carry a variety of toxic compounds, including heavy metals, organic chemicals, and acids, which contribute to its toxicity. Pregnancy complications, lung cancer, cardiovascular and respiratory disorders, and other health problems have all been connected to exposure to PM2.5. Notably, Jalali et al. (2021) found a substantial correlation between elevated PM2.5 levels and higher mortality rates in a country in the Eastern Mediterranean. Furthermore, prolonged exposure to high levels of PM2.5 is known to cause oxidative stress, inflammation, and damage to cells and tissues, which exacerbate these health effects.

As a result, PM2.5 pollution has become a serious public health concern worldwide, including in Malaysia, where the Department of Environment Malaysia (DOE) began measuring PM2.5 in April 2017. Machine learning, particularly neural networks, is increasingly applied in classification and regression tasks. One of the primary advantages of using machine learning for air quality classification is its ability to process complex datasets and identify non-linear relationships among variables. A study by Zhang et al. (2023) on air quality index prediction in six Chinese urban areas found that ensemble approaches enhanced prediction accuracy by including numerous data sources, such as pollution and meteorological data. Additionally, Liu (2024) highlighted the limitations of traditional empirical models, stating that machine learning offers more accurate predictions for air quality management. Radial Basis Function Neural Network (RBFNN) is one of the well-known neural networks that offer fast convergence compared to others, such as Multi-Layer Perceptrons (Zhou et al., 2019a), primarily due to its simpler structure and fewer training parameters. A study by Li et al. (2022) on predicting water quality parameters also found that the RBFNN model demonstrated promising performance with high accuracy in individual indicator predictions, though it still exhibited a significant accuracy gap in some cases.

While neural networks offer advantages in processing complex datasets, effective feature selection is crucial in improving the reliability and performance of machine learning models. Suresh et al. (2022) emphasize that reducing the complexity of the dataset through feature selection can lead to improved model performance. Nazari et al. (2023) also highlight that feature selection helps classification algorithms focus on the most relevant features, reducing computational burden and improving accuracy. Similarly, Ul-Saufie et al. (2022) demonstrate that wrapper feature selection approaches can improve air pollution prediction by selecting important features, with brute-force being particularly effective.

Nevertheless, a major challenge in predictive modelling is addressing multicollinearity among air pollutants and meteorological factors, which can lead to inaccurate predictions. Many air pollutants are strongly correlated, meaning their concentrations tend to change together or also know has multicollinearity exist. Zhou et al. (2019b) found a high correlation between PM2.5 and PM10 levels in their study on air pollution and respiratory diseases. Their findings indicate that PM2.5 often constitutes a significant portion of PM10 concentrations, contributing to shared health impacts on respiratory health. Similarly, gases like carbon monoxide (CO) and nitrogen dioxide (NO_2) also demonstrate correlated increases, primarily as a result of vehicle emissions and combustion processes inherent to urban environments (Wang et al., 2022). When these relationships are not properly addressed, machine learning models may struggle to determine which factors are truly important, thus reducing the reliability of predictions.

This study focuses on Klang, Selangor, due to its severe pollution levels, heavily influenced by industrial emissions and port-related activities. Klang's air quality is largely affected by both local industrial zones and transboundary pollution from shipping activities, making it a crucial case study. Mohtar et al. (2022) highlighted in their study that the Klang station in Klang Valley, near Malaysia's busiest shipping port, Port Klang, often experiences the highest concentrations of particulate matter, making it a significant contributor to the country's air pollution issues. Despite its importance, few studies have examined feature selection techniques focused on multicollinearity issues for air quality classification in Klang.

Furthermore, this study considers ten key input variables influencing PM2.5 levels, including six pollutants (PM10, SO_2, NO_2, O_3, and CO). These pollutants were selected due to their significant impact on air pollution, as supported by previous studies on air quality classification, such as the study by Sapari et al. (2023). Moreover, Liu et al. (2020) found that changes in wind speed and temperature directly influence pollutant concentrations in China. Specifically, their study reported that an increase in wind speed generally improves air quality by dispersing pollutants away from densely populated areas. In addition, Wattimena et al. (2022) highlighted the importance of meteorological factors such as wind

speed, cloud volume, air pressure, temperature, relative humidity, and precipitation in forecasting air quality indices and improving prediction accuracy. However, only four meteorological parameters (wind direction, wind speed, relative humidity, and temperature) were included in this study, as these are the key parameters consistently recorded by the DOE.

The AdjcorT two-stage feature selection method has been shown to enhance RBFNN classification by mitigating multicollinearity, as demonstrated in a study by Arafin et al. (2024). In their study, they applied the method to Shah Alam's air quality dataset, improving predictive accuracy. However, their study did not explore its applicability in different urban environments with distinct pollution sources, such as Klang. This research aims to fill this gap by applying the AdjcorT-RBFNN method to Klang's air quality dataset and verifying its performance through Monte Carlo simulations.

By systematically evaluating the effect of multicollinearity and comparing AdjcorT-RBFNN with a RBFNN model, this study seeks to enhance air quality classification for improved environmental decision-making.

2. METHODOLOGY

2.1 Data description

The air quality dataset for the Klang, Selangor (CA21B) area was obtained from the Department of Environment (DOE) for the years 2018 to 2022. The extracted variables include Particulate Matter (PM2.5 and PM10), Sulphur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Ground-Level Ozone (O_3), Carbon Monoxide (CO), wind direction, wind speed, relative humidity and ambient temperature, all in hourly format, comprising a total of 43,824 samples. However, the PM2.5 data for 1st January 2018, is not available for every hour. Therefore, we removed 24 data points from 1st January 2018, reducing the dataset to 43,800 samples. Moreover, the dataset contains missing values as shown in Table 1. According to the table, the percentage of missing values of all variables are below 10%, with NO_2 has the highest missing value (7.7%). Meanwhile, data of ambient temperature has lowest missing value which is 0.4% of dataset. According to Chen and Li (2024), failure of monitoring instruments is one of the most common causes of missing data. The instruments malfunction might happen due to the extreme weather, power outages, or periodic maintenance, resulting in gaps in data collection. (Ghazali et al., 2021). Additionally, missing values in air quality datasets can adversely affect the performance of analytical models, leading to misleading (Ghazali et al., 2021). Thus, this study employed a widely used imputation method to impute the missing values which is linear interpolation. According to Van Rossum et al. (2023), linear interpolation is a simple method yet it has been shown to yield higher imputation accuracy.

Table 1. Percentage of missing values

Variable	N	Missing value
PM2.5	43,572	228 (0.5%)
PM10	43,462	338 (0.8%)
SO_2	40,648	3,152 (7.2%)
NO_2	40,414	3,386 (7.7%)
O_3	41,450	2,350 (5.4%)
CO	41,219	2,581 (5.9%)
WD	43,596	204 (0.5%)
WS	43,594	206 (0.5%)
Humidity	43,511	289 (0.7%)
Temperature	43,606	194 (0.4%)

2.2 Research framework

Figure 1 shows the research framework employed in this study to predict PM2.5 levels in Klang, Malaysia. The process begins with the collection of air quality data from the Department of Environment, Malaysia, for the years 2018 to 2022. Then, the data undergoes pre-processing, involving linear interpolation for imputing missing values, conversion of hourly data to daily averages, binary classification of PM2.5 levels, min-max normalization for feature scaling, and the application of the Synthetic Minority Oversampling Technique (SMOTE) to handle class imbalance.

Next, we use Spearman Correlation to explore the correlation between independent variables within the dataset to understand its extent and impact on model performance. To address this, the AdjcorT feature selection method is employed to rank variables based on their correlation and importance, mitigating the influence of multicollinearity. Subsequently, feature subsets are evaluated using an Artificial Neural Network (ANN) model to identify the best combinations for PM2.5 prediction. Additionally, two models are then developed and compared: a standard RBFNN and a Two-Stage AdjcorT-RBFNN, which integrates AdjcorT feature selection with RBFNN. The models are evaluated based on classification metrics such as accuracy, sensitivity, specificity, and

AUROC to identify the best-performing model. Finally, the robustness of the selected model is validated using simulation data with varying sample sizes and correlations, ensuring its reliability in identifying important predictors under various conditions.

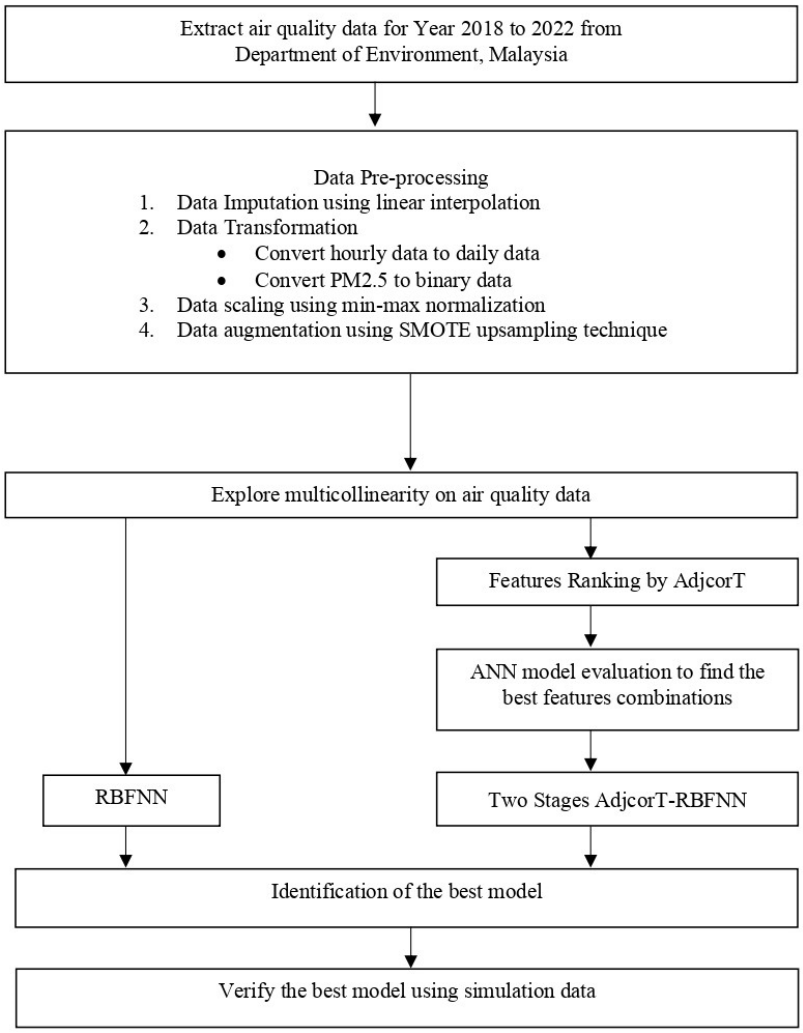

Figure 1. Research framework

2.3 Adjusted correlation sharing t-test

Ibrahim (2020) extended the variable selection method which is correlation sharing t-statistics (corT), by developing an adjusted version namely adjusted correlation sharing t-test (AdjcorT). Both methods rank the importance of features while considering the high correlation between variables. However, the algorithm of AdjcorT allowed both positive and negative correlation between variables, meanwhile corT considers only positive correlations. The standard t-statistics were calculated first using equation shows in (1). S_i is the pooled standard deviation within the group for the i-th variable. $\bar{x}_{ij}$ represents the average of the i-th variable for the j-th class or target variable (where j=0 or j=1). The equation of AdjcorT is displayed in (2):

$$T_i = \frac{\bar{x}_{i1} - \bar{x}_{i0}}{S_i} \quad (1)$$

$$r_i = \text{sign}\left(\frac{\bar{x}_{i1} - \bar{x}_{i0}}{S_i}\right) \times \left[\max_{(0 \le \rho \le 1)} \frac{1}{w} \sum_{j \in C_\rho(i)} |T_j|\right] \quad (2)$$

A score, or an AdjcorT value, r_i, is assigned to each variable. This value represents the average of all t-statistics for variables that have a correlation (in absolute value) of at least ρ with variable i. Additionally, w denotes the cardinality of $C_\rho(i)$, where $C_\rho(i)$ is the set of indices of variables whose correlation (in absolute value) with variable x_i is greater than or equal to ρ. According to Ibrahim (2020), the optimal value of ρ should be chosen to maximize the average. Moreover, since the t-scores for each variable are calculated to determine the correlation, this method is applicable only to continuous variables.

2.4 Artificial neural network

Artificial Neural Network (ANN) is a machine learning that designed inspired of biological neural networks. An ANN has three layers: the input layer (which consists of input variables), the hidden layer (which applies activation functions), and the output layer. They are designed to identify patterns and address complex problems by learning from data. ANNs consist of interconnected artificial neurons, also known as nodes, that work collaboratively to process information. Each neuron receives input, processes it, and generates output that can be transmitted to successive neurons in the network. This design enables artificial neural networks to perform tasks such as classification, regression, and pattern recognition. In our study, we used R Studio to implement an ANN model using the nnet package. To effectively train an ANN, several key hyperparameters must be set. These include the number of hidden neuron and learning rate, where the number of hidden neurons we set is 7 and 0.01 for learning rate as suggested by Ul-Saufie et al. (2022). The activation function used in the hidden layer is the logistic sigmoid, which maps inputs to a range between 0 and 1.

2.5 Radial basis function neural network (RBFNN)

RBFNN is a subset of ANN that share the same theoretical framework, consisting of three layers which is input, hidden and output layer. The difference is that RBFNNs use a radial basis function, typically Gaussian functions, as the activation function. RBFNN are particularly effective in addressing nonlinear relationships in data, such as air quality measurements. Their ability to approximate complex functions with relatively few parameters makes them highly suitable for tasks like pattern recognition and function approximation. The RSNNS package in RStudio was used to implement the RBFNN.

2.6 Performance metrics

Different performance metrics are essential for evaluating and comparing machine learning classification models, enabling informed decisions about model selection and improvement (Akshay et al., 2022). According to Ibrahim (2020), accuracy, sensitivity, specificity, and Area Under the Receiver Operating Characteristic (AUROC) are common key metrics used in previous study to assess classification model. For instance, Alalwany and Mahgoub (2024) employed accuracy, precision, recall, F1 score, and ROC metrics for evaluating the performance of their model on an ensemble learning-based real-time intrusion detection scheme for in-vehicle networks. Moreover, Chandra et al. (2022) used performance metrics including accuracy, specificity, F1 score, sensitivity, and precision to predict Jakarta's air quality.

In this study, accuracy, sensitivity, specificity, precision, F1 score, and AUROC are used to evaluate model performance of using real-world data and simulated data. Both real-world air quality data and simulated data were used to assess the classification effectiveness of the RBFNN and adjcorT-RBFNN models. The real data evaluation examines model performance based on actual air quality measurements from Klang, while the simulated data evaluation helps analyze the models' behavior under controlled conditions, particularly in handling multicollinearity. The combination of these two evaluations ensures a comprehensive assessment of the models' predictive capabilities. The specific settings for generating the simulated datasets, including sample sizes, correlation structures, and iteration processes, are detailed in Section 2.7.

2.7 Monte carlo simulation

Monte Carlo Simulation is a computational technique used to model the probability of different outcomes in systems influenced by randomness. It relies on repeated random sampling to approximate results, making it useful for solving problems involving uncertainty and complex decision-making (Liu, 2024). This method is widely applied in finance, engineering, healthcare, and machine learning, where it helps optimize strategies by evaluating potential scenarios. Monte Carlo Simulation is particularly valuable when analytical solutions are infeasible due to system complexity. In this study, Monte Carlo Simulation was employed to systematically generate datasets with varying correlation structures and sample sizes. This process allowed for the evaluation of how different levels of correlation affect the classification performance of RBFNN and adjcorT-RBFNN.

To achieve this, we use RStudio to run both the RBFNN and AdjcorT-RBFNN models using simulated data with varying sample sizes (n=50, 100, 150, 200) and correlation values (ρ=-0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8) to evaluate the impact of correlation and sample size on classification accuracy. Each scenario was simulated over 100 iterations to ensure statistical robustness and reduce variability in performance estimates. During each iteration, the models were trained and evaluated on a newly generated dataset, allowing us to analyse how correlation and sample size

influence classification performance. By incorporating multiple iterations, we mitigated the effects of random variations in individual datasets, ensuring more reliable comparisons between the models. This simulation-based validation helps confirm whether AdjcorT effectively addresses multicollinearity and enhances the classification capability of RBFNN, particularly under different correlation strengths.

3. RESULTS AND DISCUSSION

3.1 Data pre-processing

Data transformation was employed in this study, where the hourly dataset was converted to daily data by aggregating the values over 24 hours. Moreover, the dependent variable, $PM2.5_{D+1}$ values was transformed into binary data. Kalajdjieski et al. (2020) suggested separating the air quality category into two groups only, polluted and not polluted. Table 2 shows the PM2.5 breakpoints (24-hour average) according to DOE guidelines.

The descriptive statistics of the dataset after data transformation are shown in Table 3. The total number of samples (N) was reduced to 1,824 because the hourly data was transformed into daily data. Based on the table, the average PM2.5 levels is 26 µg/m^3, while the maximum value is 154 µg/m^3. Moreover, the standard deviation of SO_2 is the lowest (0.001), while the highest standard deviation is 35, which corresponds to wind direction. This wide difference in scale can affect the accuracy of classification. Hence, we employed min-max normalization to standardize these measurements, thereby enhancing the interpretability of the results, following a study by Aarthi et al. (2023). In addition, the distribution of PM2.5 categories is not balanced as shown in Figure 2, where not polluted (86%) category is more than polluted (14%). To address this issue, the applied Synthetic Minority Over-sampling Technique (SMOTE) was applied to the dataset.

Table 2. Binary labels for the respective PM2.5 breakpoint and AQI categories

AQI category	PM2.5 breakpoints	Binary labels
Good	0.0-12.0	Not polluted
Moderate	12.1-35.4	Not polluted
Unhealthy for sensitive groups	35.5-55.4	Polluted
Unhealthy	55.5-150.4	Polluted
Very unhealthy	150.5-250.4	Polluted
Hazardous	250.5 and above	Polluted

Table 3. Descriptive statistics before data pre-processing

Variable	N	Mean	Median	Std. Dev.	Skewness	Min	Max
PM2.5	1,824	26.309	24.206	12.341	3.720	9.134	154.845
PM10	1,824	35.890	33.103	15.529	2.880	10.774	180.227
SO_2	1,824	0.002	0.001	0.001	2.512	0.000	0.009
NO_2	1,824	0.017	0.016	0.005	0.453	0.003	0.040
O_3	1,824	0.015	0.015	0.005	0.662	0.002	0.040
CO	1,824	0.871	0.852	0.264	0.305	0.122	1.835
WD	1,824	169.988	161.410	35.632	1.093	72.039	327.733
WS	1,824	1.375	1.319	0.340	0.933	0.556	3.500
Humidity	1,824	80.518	80.413	5.916	-0.092	58.659	100.000
Temperature	1,824	28.350	28.442	1.157	-0.367	23.204	31.281

The descriptive statistics after data pre-processing were recomputed, as shown in Table 4. This table presents the descriptive statistics of the variables after data normalization and standardization. The total number of samples increased due to the application of the SMOTE up-sampling technique. The dataset now consists of 3,089 samples, with 80% used for training and 20% for testing. According to the table, the minimum and maximum values of all variables are 0 and 1, respectively. This result indicates that all variables has successfully scaled into a standard range between 0 and 1. Furthermore,

Figure 3 shows the distribution of PM2.5 category after employed SMOTE technique are more balanced now with 50.9% of dataset are not polluted category while 49.1% of dataset are polluted category. Hence, the dataset is more suitable to train for classification task because SMOTE method SMOTE can significantly enhance the classification performance of machine learning models, particularly in scenarios where the minority class is critical as highlighted by Ariansyah et al. (2023).

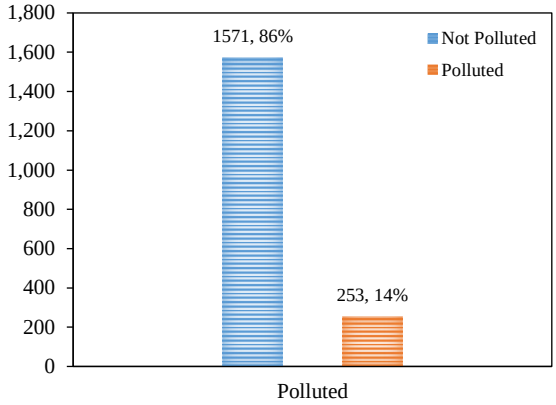

Figure 2. $PM2.5_{D+1}$ distribution (Before SMOTE)

Table 4. Descriptive statistics after data pre-processing

Variable	N	Mean	Median	Std. Dev	Skewness	Min	Max
PM2.5	3,089	0.127	0.108	0.087	2.908	0	1
PM10	3,089	0.158	0.136	0.094	2.259	0	1
SO_2	3,089	0.172	0.147	0.120	2.691	0	1
NO_2	3,089	0.381	0.368	0.133	0.621	0	1
O_3	3,089	0.349	0.329	0.137	0.694	0	1
CO	3,089	0.349	0.329	0.137	0.694	0	1
WD	3,089	0.373	0.337	0.130	1.288	0	1
WS	3,089	0.284	0.266	0.113	0.718	0	1
Humidity	3,089	0.515	0.506	0.140	0.044	0	1
Temperature	3,089	0.644	0.660	0.137	-0.474	0	1

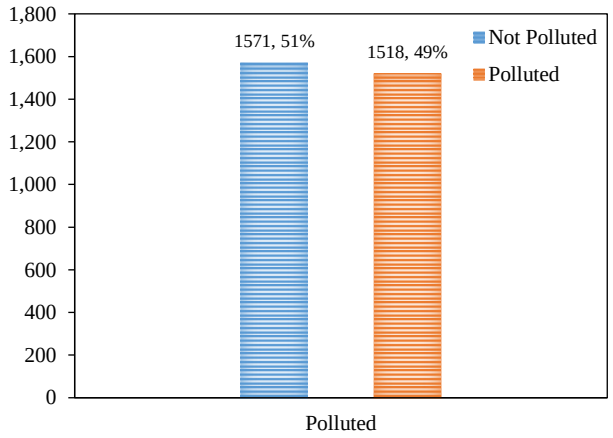

Figure 3. $PM2.5_{D+1}$ distribution (After SMOTE)

3.2 Correlation between features

The high correlation between features or also known as multicollinearity might distort the accuracy of predictive model because the highly correlated variables may share similar characteristics. For instance, a study by Kılıçoğlu and Yerlikaya-Özkurt (2024) highlighted that the high correlation among independent variables may reduce the reliability of regression coefficients, making it difficult to draw meaningful inference from the model. Spearman correlation is less sensitive to outliers than pearson correlation due to its ranking of data rather than raw values, reducing extreme values' influence on the correlation coefficient (Hou et al., 2022). Moreover, the spearman correlation is a non-parametric measure, and hence it does not assume a specific distribution for the data making it more reliable than pearson correlation (Hou et al., 2022).

Therefore, Spearman correlation matrix was computed to examine the correlation between variables in Klang's air quality dataset as shown in Table 5. Spearman correlation values range from -1 to 1, where values closer to 1 indicate a strong positive correlation, and values closer to -1 indicate a strong negative correlation. According to the analysis, the Spearman correlation value between PM2.5 and PM10 is 0.94, suggesting a very strong positive correlation between these variables. Furthermore, relative humidity and ambient temperature show a strong negative correlation, with a value of -0.84. Additionally, both particulate matters (PM10 and PM2.5) have moderate positive correlations with CO, with a value of 0.55 and 0.59 respectively. In addition, NO_2 also have moderate positive correlations with

CO, with a value of 0.57. Notably, wind speed also shows moderate negative correlation with CO with spearman correlation of -0.5. Other variables exhibit correlation values below±0.5.

Although certain variables, such as PM10 and PM2.5, exhibit high correlation, they should not be removed solely based on this criterion. Despite their strong raltionship, each variable may capture unique characteristics that contribute to air quality classification. For instance, PM2.5 and PM10, while strongly correlated, represent different particle size fractions with distinct health and environmental implications. Removing one could result in the loss of valuable information that enhances model performance. While multicollinearity can pose challenges in linear models by inflating variance and reducing interpretability, its impact on non-linear models like RBFNN is less pronounced. RBFNN can effectively learn complex relationships even when inputs are correlated. However, a high degree of multicollinearity may introduce redundancy, which is why the AdjcorT feature selection method was applied in Section 3.4. AdjcorT identifies the most informative features while preserving key variables that contribute to classification accuracy, ensuring that the model benefits from a diverse yet relevant set of inputs.

Table 5. Spearman correlation matrix

Variables	PM2.5	PM10	SO_2	NO_2	O_3	CO	WD	WS	Humidity	Temperature
PM2.5	1.00	0.94	0.09	0.37	0.16	0.59	-0.20	-0.07	-0.33	0.33
PM10	0.94	1.00	0.18	0.39	0.16	0.55	-0.16	-0.05	-0.37	0.33
SO_2	0.09	0.18	1.00	0.03	0.06	0.03	0.09	0.18	-0.20	0.09
NO_2	0.37	0.39	0.03	1.00	-0.05	0.57	-0.08	-0.50	0.14	-0.22
O_3	0.16	0.16	0.06	-0.05	1.00	-0.01	0.02	0.05	-0.39	0.36
CO	0.59	0.55	0.03	0.57	-0.01	1.00	-0.10	-0.23	-0.14	0.07
WD	-0.20	-0.16	0.09	-0.08	0.02	-0.10	1.00	0.03	0.07	-0.09
WS	-0.07	-0.05	0.18	-0.50	0.05	-0.23	0.03	1.00	-0.47	0.42
Humidity	-0.33	-0.37	-0.20	0.14	-0.39	-0.14	0.07	-0.47	1.00	-0.84
Temperature	0.33	0.33	0.09	-0.22	0.36	0.07	-0.09	0.42	-0.84	1.00

3.3 Feature combinations

The first stage involves finding the best feature combinations using the AdjcorT feature selection method. Figure 4 shows the ranking of feature importance, where higher values indicate greater significance of the variable to the target variable ($PM2.5_{D+1}$). According to the table, particulate matter (PM2.5 and PM10) are the most important variables for predicting PM2.5 in Klang, with AdjcorT values of 7.7 and 7.5, respectively, followed by relative humidity, SO_2, wind direction, O_3, CO, ambient temperature and NO_2. Additionally, wind speed is the least important feature for classifying PM2.5 in Klang. The features were then added to the ANN model one by one according to their ranking as shown in Figure 4 to determine the best feature combinations, as suggested by Arafin et al. (2024). The learning rate is set at 0.01 and the number of hidden nodes is determined by summing the number of variables and classes, dividing the result by two, and then adding one (Ul Saufie et al., 2022). Thus, this study use number of hidden nodes is 7. Table 6 presents the ANN model performances with varying numbers of features, based on accuracy, sensitivity, specificity, precision, F1 score, and AUROC. The highest value for each performance metric is typed in bold font. According to the table, the model with nine features achieves the best performance, with higher accuracy (0.67), sensitivity (0.85), F1 score (0.7), and AUROC (0.68). Arafin et al. (2024) concluded that eight features are sufficient to predict next-day PM2.5 concentrations in the urban area of Shah Alam. In contrast, this study found that nine features are needed to predict PM2.5 concentrations. The optimal feature combination for classifying $PM2.5_{D+1}$ in Klang includes PM2.5, PM10, relative humidity, SO_2, wind direction, O_3, CO, ambient temperature, and NO_2, based on AdjcorT value ranking.

3.4 Best model identification

In this section, the performance of the RBFNN model with all 10 variables is compared to the two-stage AdjcorT-RBFNN model, which utilizes the 9 best feature combinations. The RBFNN model with all

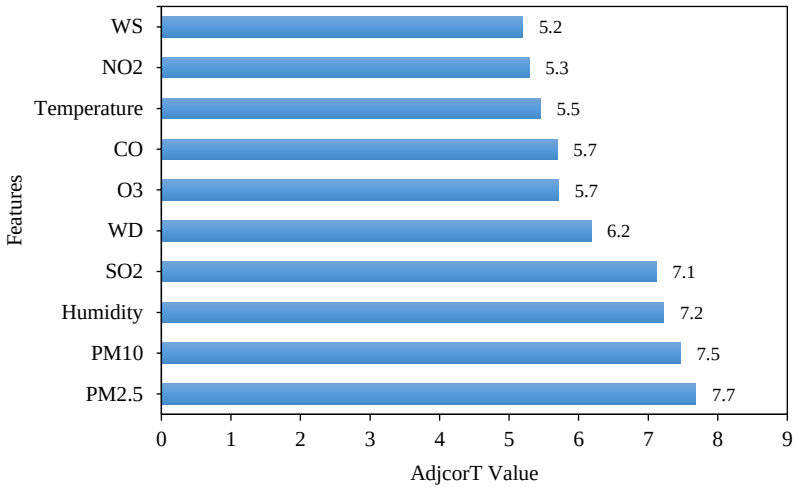

Figure 4. Feature ranking by AdjcorT value

Table 6. Model Performances for different numbers of features combination

No of features	1	2	3	4	5	6	7	8	9	10
Accuracy	0.58	0.55	0.61	0.59	0.63	0.66	0.67	0.65	**0.67**	0.67
Sensitivity	0.60	0.66	0.63	0.60	0.61	0.67	0.72	0.69	**0.85**	0.69
Specificity	0.55	0.51	0.59	0.57	**0.66**	0.65	0.62	0.62	0.60	0.64
Precision	0.55	0.51	0.59	0.57	**0.66**	0.65	0.62	0.62	0.60	0.64
F1 score	0.58	0.58	0.61	0.58	0.63	0.66	0.67	0.65	**0.70**	0.66
AUROC	0.58	0.56	0.61	0.58	0.62	0.66	0.67	0.65	**0.68**	0.67

10 variables represents the standard approach, while the AdjcorT-RBFNN model applies the AdjcorT method to mitigate multicollinearity by selecting the 9 most relevant features. Based on the results in Table 6, wind speed was excluded in the AdjcorT-RBFNN model. The number of hidden nodes is 7 for both models, RBFNN and AdjcorT-RBFNN. In addition, the learning rate for both models are set to 0.01. Table 7 presents a comparison of performance metrics for both models, with the highest values highlighted in bold font. According to the table, the two-stage AdjcorT-RBFNN model outperforms the RBFNN model, achieving higher accuracy (0.62), sensitivity (0.64), specificity (0.60), precision (0.60), F1 score (0.62), and AUROC (0.62). This finding is consistent with the research conducted by Arafin et al. (2024), which demonstrates that the AdjcorT-RBFNN model can enhance the performance of the RBFNN model. However, their study excludes relative humidity and ambient temperature as an important feature for classifying $PM2.5_{D+1}$ in Shah Alam. In contrast, our study found that both meteorological parameter, which is relative humidity and ambient temperature are important factor to predict PM2.5. The differences in the selection of the best features may be due to the different study areas, as our study was conducted in a Klang, while theirs was conducted in Shah Alam.

Table 7. RBFNN and AdjcorT-RBFNN model performances

Model	RBFNN	AdjcorT-RBFNN
Accuracy	0.59	**0.62**
Sensitivity	0.61	**0.64**
Specificity	0.56	**0.60**
Precision	0.56	**0.60**
F1 Score	0.58	**0.62**
AUROC	0.59	**0.62**

3.5 Monte Carlo simulations

A Monte Carlo simulation was applied to verify the best model, AdjcorT-RBFNN, using simulated data. The simulations of both models were run using various scenarios, with different sample sizes (N=50, 100, 150, 200) and correlations (ρ=-0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8). The line charts of accuracy, sensitivity, specificity, precision, F1 score, and AUROC for both models are shown in Figures 5, 6, 7, 8, 9, and 10, respectively. Based on Figure 5, the accuracy of the AdjcorT-RBFNN model is highest with strong negative correlation (ρ=-0.8) across all sample sizes. The accuracy decreases as correlation weakens, but it

gradually improves with positive correlation. However, it doesn't reach the high levels seen with negative correlations. Additionally, larger sample sizes, particularly N=150 and N=200, result in better accuracy. The RBFNN model exhibits less variation across correlation levels, consistently achieving lower accuracy than AdjcorT-RBFNN. Under strong positive correlation (ρ=0.8) and N=200, RBFNN attains 50.9% accuracy, whereas AdjcorT-RBFNN reaches 60.3%, demonstrating its superior ability to select important features. Figure 6 demonstrates that the AdjcorT-RBFNN model's sensitivity is high with strong negative correlation and decreases as correlation weakens, improving with positive correlation but not returning to initial levels. Larger sample sizes improve sensitivity, especially for N=150 and N=200. In contrast, the RBFNN model's sensitivity remains consistent and high across all correlation levels, while AdjcorT-RBFNN's sensitivity fluctuates with changes in correlation and sample size.

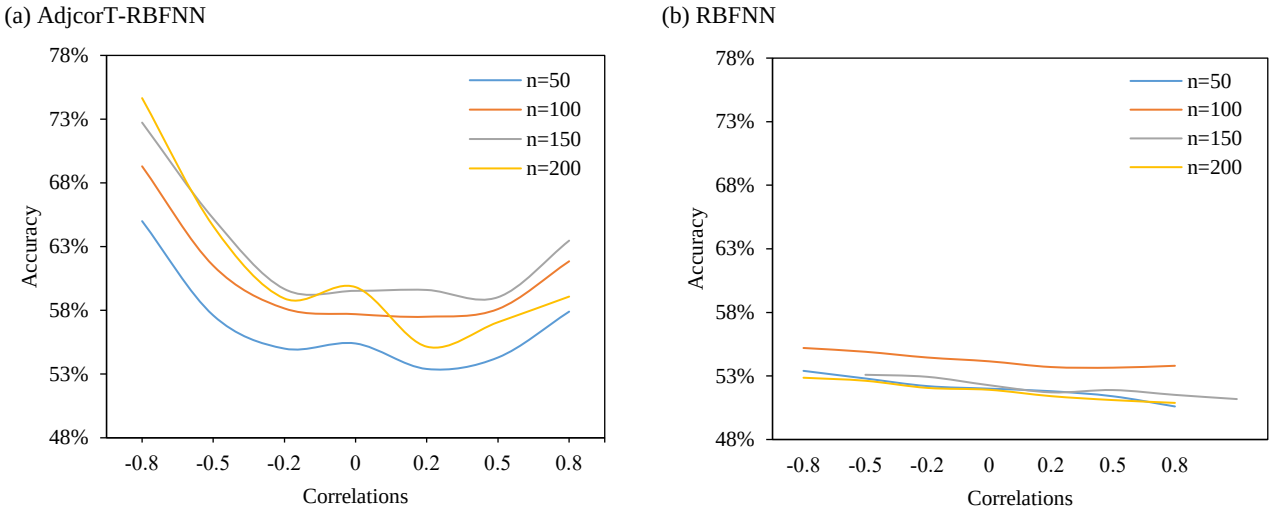

Figure 5. Accuracy of simulation AdjcorT-RBFNN and RBFNN model

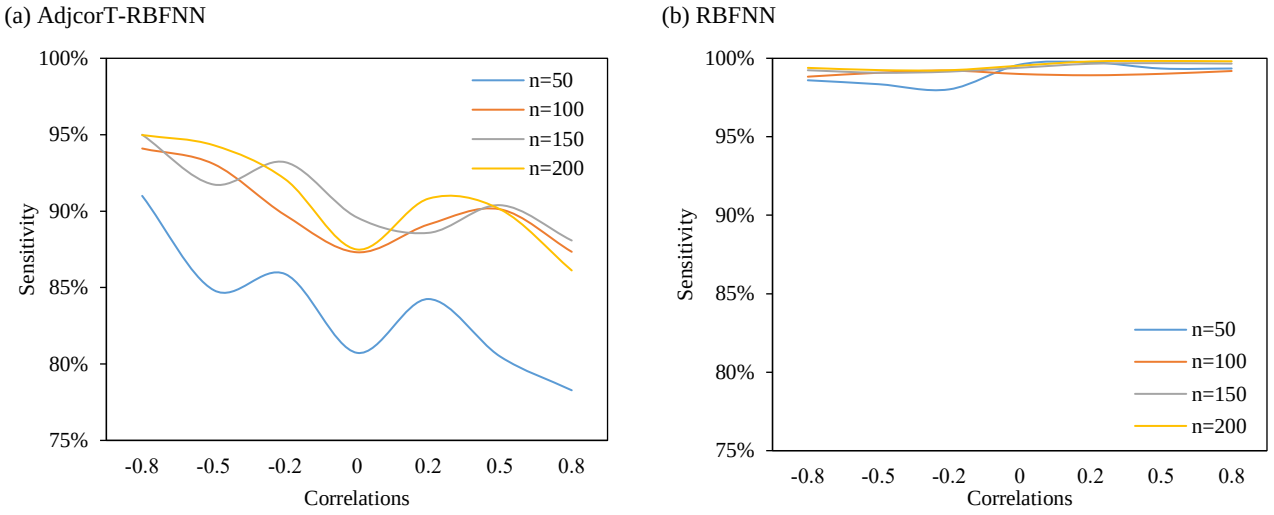

Figure 6. Sensitivity of simulation AdjcorT-RBFNN and RBFNN model

Figure 7 illustrates specificity, showing low values for both models, especially for RBFNN (below 11%) and AdjcorT-RBFNN (21-52%). The AdjcorT-RBFNN model performs better at correctly identifying negative cases, particularly with negative correlations and larger sample sizes. Figure 8 shows the precision of the simulations for both models. According to the line charts, AdjcorT-RBFNN performs better,

especially with strong correlations, while RBFNN consistently shows lower precision with minimal variation. The AdjcorT-RBFNN model is particularly effective at reducing false positives among predicted positive cases, especially with stronger correlations and larger sample sizes. Figure 9 compares F1 scores, showing that AdjcorT-RBFNN performs better with larger sample sizes and stronger correlations, especially for negative correlations. In contrast, RBFNN maintain a stable F1 scores between 64% and 70% across all conditions. Lastly, Figure 10 compares AUROC values. The AdjcorT-RBFNN model shows varying AUROC depending on the correlation, with the highest value (74%) for strong negative correlation (ρ=-0.8) and N=200. As correlation weakens, AUROC decreases. RBFNN's AUROC remains between 50% and 55%, suggesting that the model has limited ability to distinguish between classes. To sum up, the AdjcorT-RBFNN model outperforms RBFNN in discriminating between classes, particularly with strong correlations, as demonstrated by simulated data. Similarly, Ibrahim (2020) highlighted that the AdjcorT provides a flexible variable selection approach for classification, particularly in medium to large datasets with negative correlations.

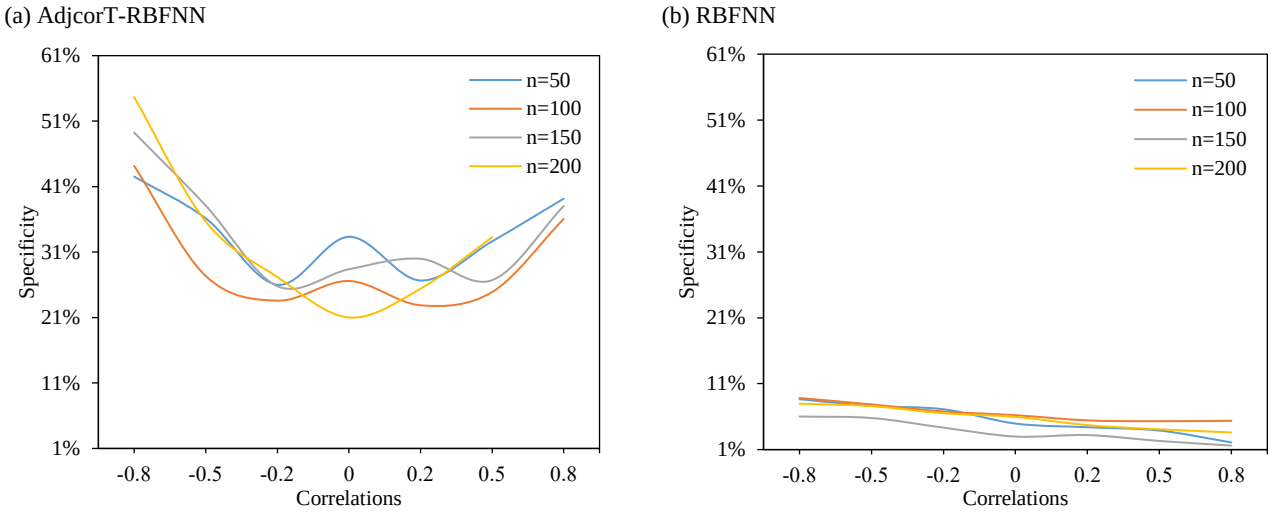

Figure 7. Specificity of simulation AdjcorT-RBFNN and RBFNN model

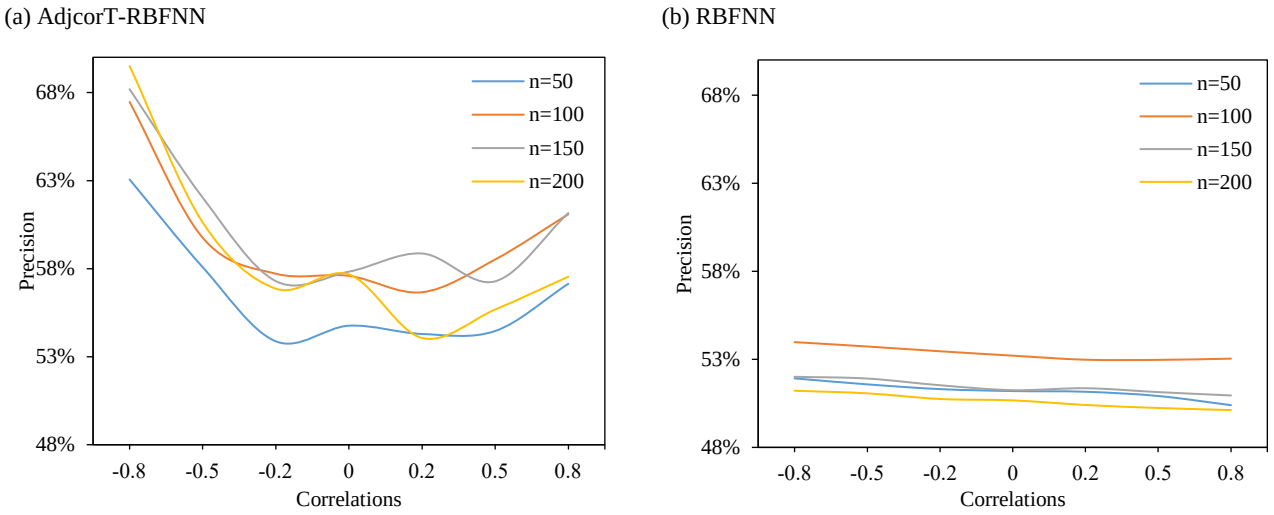

Figure 8. Precision of simulation AdjcorT-RBFNN and RBFNN model

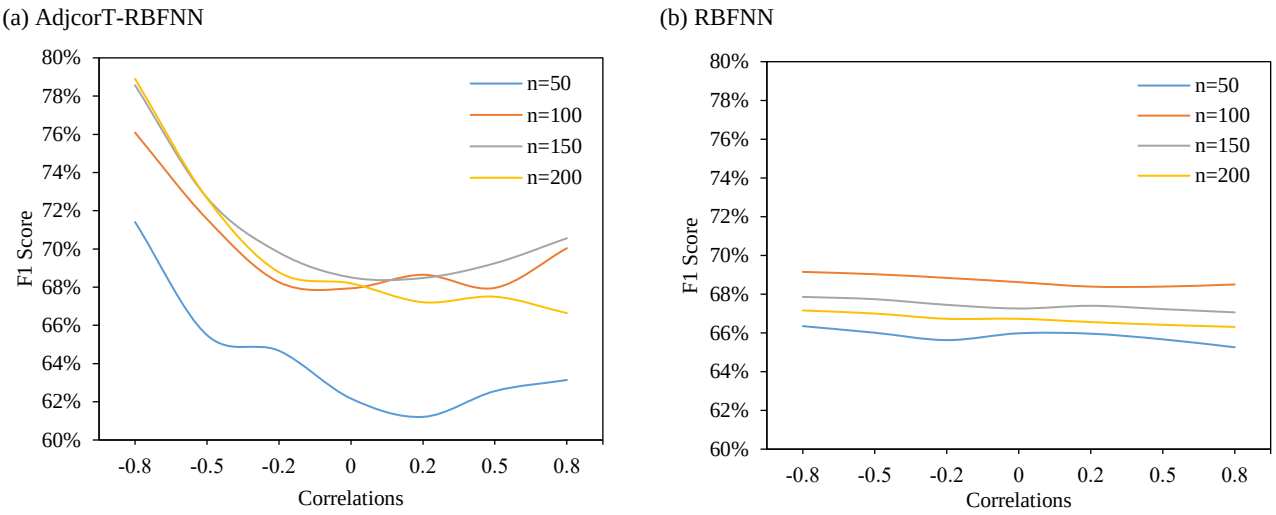

Figure 9. F1 Score of simulation AdjcorT-RBFNN and RBFNN model

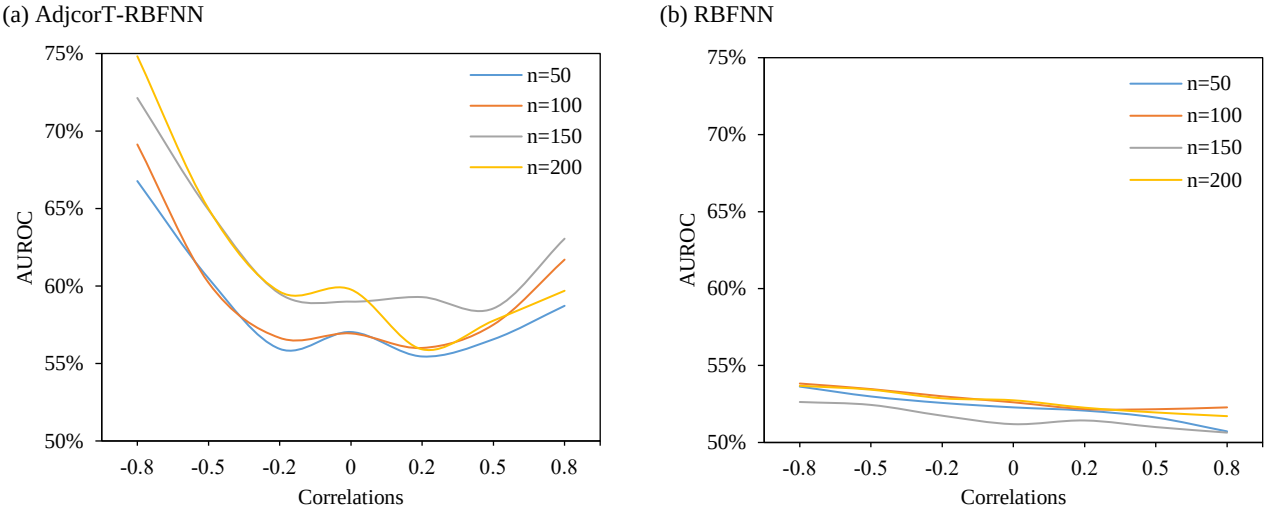

Figure 10. AUROC of Simulation AdjcorT-RBFNN and RBFNN Model

4. CONCLUSION

This study aims to classify air quality in Klang, Selangor while considering the high correlation between features using the two-stages feature selection method, AdjcorT-RBFNN. This study found that the AdjcorT-RBFNN model outperformed the RBFNN model, achieving higher performance metrics, including accuracy, sensitivity, specificity, precision, F1 score, and AUROC. Specifically, the AdjcorT-RBFNN model achieved an accuracy of 0.62, a sensitivity of 0.64, a specificity of 0.60, a precision of 0.60, an F1 score of 0.62, and an AUROC of 0.62, which were consistently higher than those of the standard RBFNN model. Based on the AdjcorT method, 9 features were identified as the best feature combination to predict air quality in Klang, namely PM2.5, PM10, relative humidity, SO_2, wind direction, O_3, CO, ambient temperature and NO_2. These features were selected based on their importance ranking, ensuring that only the most relevant predictors were retained while reducing redundancy caused by multicollinearity.

Moreover, this study verified the AdjcorT-RBFNN model's performance using simulated data. Based on the simulation results, the findings demonstrate that the AdjcorT-RBFNN model consistently outperforms the RBFNN model in distinguishing between classes, particularly when there are strong positive and negative correlations between the variables. When the correlation was strong (ρ=-0.8), the AdjcorT-RBFNN model achieved the highest accuracy, especially for larger sample sizes (N=150 and N=200). In contrast, the RBFNN model exhibited lower accuracy across all correlation levels, with less variation. Specifically, when ρ=0.8 and N=200, AdjcorT-RBFNN achieved an accuracy of

60.3%, outperforming RBFNN, which only reached 50.9%. Furthermore, AUROC values showed that AdjcorT-RBFNN was most effective under strong negative correlations, reaching 74% when ρ=-0.8 and N=200, whereas RBFNN's AUROC remained between 50% and 55% across all conditions. The AdjcorT-RBFNN model's ability to select relevant features based on correlation strength allows it to better handle complex data relationships, resulting in improved performance in terms of accuracy, sensitivity, and other key metrics. In contrast, the RBFNN model shows a more limited ability to differentiate between classes, as it lacks a dedicated feature selection mechanism to address multicollinearity. These results further highlight the importance of an effective feature selection method in improving model performance, especially when the dataset exhibits high multicollinearity.

The two-stage feature selection method, AdjcorT-RBFNN, has been shown to enhance RBFNN classification by considering the high correlation between features, using both real and simulated datasets. However, this study limited to air quality data in Klang an urban area. Therefore, we suggest future researchers apply this method to air quality data in other urban, suburban or rural areas to confirm its effectiveness. Moreover, due to the compromised results of the simulation, we also suggest future researchers apply this two-stage feature selection method in other areas where multicollinearity issues exist.

ACKNOWLEDGEMENTS

The authors would like to specially acknowledge the Ministry of Higher Education (MOHE) for funding under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2023/STG06/UITM/02/8). We are also thankful to the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Research Nexus UiTM (ReNeU) and College of Computing, Informatics and Mathematics, UiTM.

AUTHOR CONTRIBUTIONS

Author 1 carried out the experiment and prepared the manuscript content, Author 2, Author 3 and Author 4 verified the manuscript content and Author 4 also supervised the project.

DECLARATION OF COMPETING INTERESTS

No conflict of interest concerning this manuscript.

REFERENCES

Aarthi C, Ramya VJ, Falkowski-Gilski P, Divakarachari PB. Balanced spider monkey optimization with Bi-LSTM for sustainable air quality prediction. Sustainability 2023;15: Article No. 1637.

Akshay A, Abedi M, Shekarchizadeh N, Burkhard FC, Katoch M, Bigger-Allen A, et al. MLcps: Machine learning cumulative performance score for classification problems. GigaScience 2022;12:Article No. 108.

Alalwany E, Mahgoub I. An effective ensemble learning-based real-time intrusion detection scheme for an In-Vehicle network. Electronics 2024;13:Article No. 919.

Arafin SK, Ul-Saufie AZ, Ghani NA, Ibrahim N. A two-stage feature selection method to enhance prediction of daily PM2.5 concentration air pollution. Environment and Natural Resources Journal 2024;22(6):500-9.

Ariansyah MH, Winarno S, Fitri EN, Retha HMA. Multi-Layer perceptron for diagnosing stroke with the SMOTE method in overcoming data imbalances. Innovation in Research of Informatics 2023;5(1):1-8.

Chandra W, Resti Y, Suprihatin B. Implementation of a breakpoint halfway discretization to predict Jakarta's air quality. INOMATIKA 2022;4(1):1-10.

Chen C, Li K. Spatiotemporal stacking method with daily-cycle restrictions for reconstructing missing hourly PM2.5 records. Transactions in GIS 2024;28:349-67.

Ghazali SM, Shaadan N, Idrus Z. A comparative study of several EOF based imputation methods for long gap missing values in a Single-Site Temporal Time Dependent (SSTTD) Air Quality (PM10) dataset. Pertanika Journal of Science and Technology 2021;29(4):2625-43.

Hou J, Ye X, Feng W, Zhang Q, Han Y, Liu Y, et al. Distance correlation application to gene co-expression network analysis. BMC Bioinformatics 2022;23(1):Article No. 81.

Ibrahim NB. Variable Selection Methods for Classification: Application to Metabolomics Data [dissertation]. United Kingdom: The University of Liverpool; 2020.

Jalali S, Karbakhsh M, Momeni M, Taheri M, Amini S, Mansourian M, et al. Long-term exposure to PM2.5 and cardiovascular disease incidence and mortality in an Eastern Mediterranean country: Findings based on a 15-year cohort study. Environmental Health 2021;20(1):Article No. 112.

Kalajdjieski J, Zdravevski E, Corizzo R, Lameski P, Kalajdziski S, Pires IM, et al. Air pollution prediction with multi-modal data and deep neural networks. Remote Sensing 2020;12:Article No. 4142.

Kılıçoğlu Şevval, Yerlikaya-Özkurt F. A novel comparison of shrinkage methods based on multi criteria decision making in case of multicollinearity. Journal of Industrial and Management Optimization 2024;20:3816-42.

Li T, Lu J, Wu J, Zhang Z, Chen L. Predicting aquaculture water quality using machine learning approaches. Water 2022; 14:Article No. 2836.

Liu R. Monte-Carlo Simulations and applications in machine learning, option pricing, and quantum processes. Highlights in Science Engineering and Technology 2024;88:1132-7.

Liu Y, Zhou Y, Lu J. Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. Scientific Reports 2020;10:Article No. 14518.

Mohtar AAA, Latif MT, Dominick D, Ooi MCG, Azhari A, Baharudin NH, et al. Spatiotemporal variations of particulate

matter and their association with criteria pollutants and meteorology in Malaysia. Aerosol and Air Quality Research 2022;22:Article No. 220124.

Nazari L, Aslan MF, Sabanci K, Ropelewska E. Integrated transcriptomic meta-analysis and comparative artificial intelligence models in maize under biotic stress. Scientific Reports 2023;13(1):Article No. 15899.

Sapari AM, Hadiana AI, Umbara FR. Air quality classification using extreme gradient boosting (XGBOOST) algorithm. Innovation in Research of Informatics 2023;5(2):44-51.

Suresh S, Newton DT, Everett TH, Lin G, Duerstock BS. Feature selection techniques for a machine learning model to detect autonomic dysreflexia. Frontiers in Neuroinformatics 2022;16:Article No. 901428.

Ul-Saufie AZ, Hamzan NH, Zahari Z, Shaziayani WN, Noor NM, Zainol MRRMA, et al. Improving air pollution prediction modelling using wrapper feature selection. Sustainability 2022;14:Article No. 11403.

Van Rossum MC, Da Silva PMA, Wang Y, Kouwenhoven EA, Hermens HJ. Missing data imputation techniques for wireless continuous vital signs monitoring. Journal of Clinical Monitoring and Computing 2023;37:1387-400.

Wang W, Yang S, Yin K, Zhao Z, Ying N, Fan J. Network approach reveals the spatiotemporal influence of traffic on air pollution under COVID-19. Chaos an Interdisciplinary Journal of Nonlinear Science 2022;32:Article No. 041106.

Wattimena EMC, Annisa A, Sitanggang IS. CO and PM10 prediction model based on air quality index considering meteorological factors in DKI Jakarta using LSTM. Scientific Journal of Informatics 2022;9:123-32.

Zhang B, Duan M, Sun Y, Lyu Y, Hou Y, Tan T. Air Quality Index Prediction in six major Chinese urban agglomerations: A comparative study of single Machine Learning Model, ensemble Model, and Hybrid Model. Atmosphere 2023; 14:Article No. 1478.

Zhou Y, Mu T, Pang Z-H, Zheng C. A survey on hyper basis function neural networks. Systems Science and Control Engineering 2019a;7:495-507.

Zhou H, Wang T, Zhou F, Liu Y, Zhao W, Wang X, et al. Ambient air pollution and daily hospital admissions for respiratory disease in children in Guiyang, China. Frontiers in Pediatrics 2019b;7:Article No. 00400.