Environment and Natural Resources Journal

Volume 22
Issue 6
2024

A Two-Stage Feature Selection Method to Enhance Prediction of Daily $PM_{2.5}$ Concentration Air Pollution

Siti Khadijah Arafin[1], Ahmad Zia Ul-Saufie[1], Nor Azura Md Ghani[1], and Nurain Ibrahim[1,2*]

[1]School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2]Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

ARTICLE INFO

Received: 28 Feb 2024
Received in revised: 9 Sep 2024
Accepted: 11 Sep 2024
Published online: 4 Nov 2024
DOI: 10.32526/ennrj/22/20240049

Keywords:
Adjusted correlation sharing t-test/ Air pollution/ Particulate matter $PM_{2.5}$/ Radial basis function neural network

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

ABSTRACT

In recent decades, air pollution has negatively affected human health and the environment. One of the important features contributing to air pollution is called $PM_{2.5}$. However, daily prediction of $PM_{2.5}$ is still lacking, especially using feature selection infused into the model. Hence, the main objective of this research is to utilize the feature selection procedures by proposing two stages feature selection methods namely adjusted correlation sharing t-test (adjcorT) and radial basis function neural network (RBFNN) in identifying the important features. This consequently also helps enhance the prediction of daily $PM_{2.5}$ concentrations. Secondary data were obtained from the Department of Environment Malaysia (DOE) from 2018 until 2022 that consists of 5 years of air pollutant daily data. The results found that adjcorT-RBFNN identified the NO_2, $PM_{2.5}$, PM_{10}, CO, O_3, wind speed and SO_2 as important features. The finding revealed that the accuracy, sensitivity, specificity, precision, F1 score and AUROC value, for a day-ahead prediction in Shah Alam are 0.756, 0.801, 0.717, 0.717, 0.757, and 0.758 respectively. Additionally, the predicted model may serve as an instrument for an early warning system, providing local authorities with information on air quality for formulation of strategies of air quality improvement.

1. INTRODUCTION

Air is the mixture of gases that surrounds the Earth and extends into its atmosphere. About 78% of it is nitrogen, 21% is oxygen, and the remaining other gases such as argon, carbon dioxide, and water vapor. Since air contains the oxygen that people, animals, and most other species require to breathe, it is essential for sustaining life on Earth. Air quality is important, where a low concentration of pollutants that endangers both human health and the environment is indicative of good air quality. Conversely, low air quality denotes high pollution concentrations. The air quality is typically assessed based on the concentration of various pollutants such as particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), and others.

The air quality is a growing concern worldwide. Air pollution has been shown to increase the risk and mortality of various pulmonary diseases, including lung cancer, chronic obstructive pulmonary disease (COPD), asthma, and infectious diseases such as pneumonia and tuberculosis, as evidenced by Ko and Kyung's (2022) research on its adverse effects on pulmonary diseases. There are serious risks to the environment and public health associated with air pollution, especially when fine particulate matter ($PM_{2.5}$) is present. $PM_{2.5}$ refers to particulate matter with a diameter of 2.5 micrometers or less, which is a major air pollutant and a significant component of air quality (Wang and Tian, 2018). Even at low concentrations, long-term exposure to PM2.5 has been shown to cause lung cancer (Ko and Kyung, 2022).

Citation: Arafin SK, Ul-Saufie AZ, Md Ghani NA, Ibrahim N. A two-stage feature selection method to enhance prediction of daily $PM_{2.5}$ concentration air pollution. Environ. Nat. Resour. J. 2024;22(6):500-509. (https://doi.org/10.32526/ennrj/22/20240049)

Accurate air quality prediction can channel strategically valuable information to the government for air pollution control and management. However, it is a challenging task to predict air quality movements as it is subject to big data elements, including many internal and external factors, and poses a big challenge to researchers who try to predict it (Sokhi et al., 2021; Zhao et al., 2022). Artificial Neural Network (ANN) has gained popularity in various fields due to its ability to solve non-linear problems and its potential for accurate predictions (Larasati et al., 2018). In the context of predicting air quality, ANN has been widely used due to its capability to build air quality models with comparable or better accuracy than other methods (Suleiman et al., 2019). Moreover, the Radial Basis Function Neural Network (RBFNN) has been utilized to predict air quality. The research conducted by Yadav and Nath (2018) comparing the predictive performance of RBFNN and Generalized Regression Neural Network (GRNN) models for the prediction of PM_{10} levels in Ardali Bazar, Varanasi, India, found that RBFNN performed better in measured and predicted PM_{10} values compared to GRNN. A study from Algeria claimed that RBFNN outperformed other models being compared in their study for electric consumption forecast (Kahoui and Chekouri, 2024). Another study also employed RBFNN with an optimization technique for classification purposes, and it obtained good classification performances (Ali et al., 2024).

Feature selection is crucial for predicting air quality due to the complex nature of the factors influencing air pollution. As there are numerous environmental factors involved in building a model, variable selection is used to effectively filter out the important variables that affect air quality. However, air quality data can indeed face multicollinearity issues due to the interrelationship of environmental variables. Several studies have highlighted the presence of multicollinearity in air quality data and its implications for statistical modelling. For instance, Farrell et al. (2019) highlighted the challenges posed by multicollinearity among covariates in statistical models for analyzing the effects of environmental factors on the spatial distribution of species. Hence, this study aims to predict the next day's air quality based on $PM_{2.5}$ concentrations while considering the multicollinearity issues between the features. Consequently, the important features will be identified, and these features can be used for the next day's air quality based on $PM_{2.5}$ prediction in the future, reducing the air pollution that exists in the air.

2. METHODOLOGY

2.1 Research framework

This section discusses the research framework as shown in Figure 1, the data, feature selection methods, classification modelling and model evaluation used in this research. Specifically, air quality data is extracted from the Department of Environment, Malaysia and data management was done on the respective dataset including data conversion, normalization and SMOTE. This research also aim to explore the collinearity among the features in air quality data followed by the feature selection procedures. Then, the proposed two stages adjcorT-RBFNN were compared to the adjcorT and RBFNN in the model evaluation. Finally, the best model are identified in this research.

Our research data consist of variables or pollutant factors involve in air quality data which is Particulate Matter (PM_{10} and $PM_{2.5}$), Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Ground-Level Ozone (O_3) and Sulphur Dioxide (SO_2) that due to the burning of natural gas, coal and wood, industries and vehicles. Besides pollutant factors, the meteorological parameters are wind direction, wind speed, relative humidity and ambient temperature taken as variables that might have an impact on the air quality. 43824 sample sizes of a dataset from air quality monitoring stations at Shah Alam, Selangor in hourly data format is obtained from the collaborators, which is the Department of Environmental (DOE) for duration of five-year period from 2018 to 2022. Table 1 shows the description of air quality data consisting of 10 variables obtained from the DOE.

2.2 Adjusted correlation sharing t-test

Adjusted correlation sharing t-test (adjcorT) is an extended variable selection method which is correlation sharing t-test (corT). The variable selection method, corT only considers positive relationships between the variables which might produce less accurate results (Ibrahim, 2020). However, the adjcorT method allows both positive and negative correlations between the variables from -1 to 1. The equation of adjcorT is displayed in (1):

$$\mathrm{r_i} = \mathrm{sign}\ (\frac{\bar{\mathrm{x}}_{\mathrm{i}1} - \bar{\mathrm{x}}_{\mathrm{i}0}}{\mathrm{s_i}}) \times \left[\max_{(0 \leq \rho \leq 1)} \frac{1}{\mathrm{w}} \textstyle\sum_{\mathrm{j} \in \mathrm{C}_{\rho}(\mathrm{i})} |\mathrm{T_j}|\right] \qquad (1)$$

Where; max is the maximum. In addition, each variable is assigned a score r_i, which equals to the average of all t-statistics for variables having correlation (absolute) at least ρ with variable i, choosing the best value of ρ to maximize the average. AdjcorT is a filter feature selection method, and it is easy to use. This method is only applicable for continuous variables as it calculates the t-score of each variable and assesses the correlation with the other independent variables.

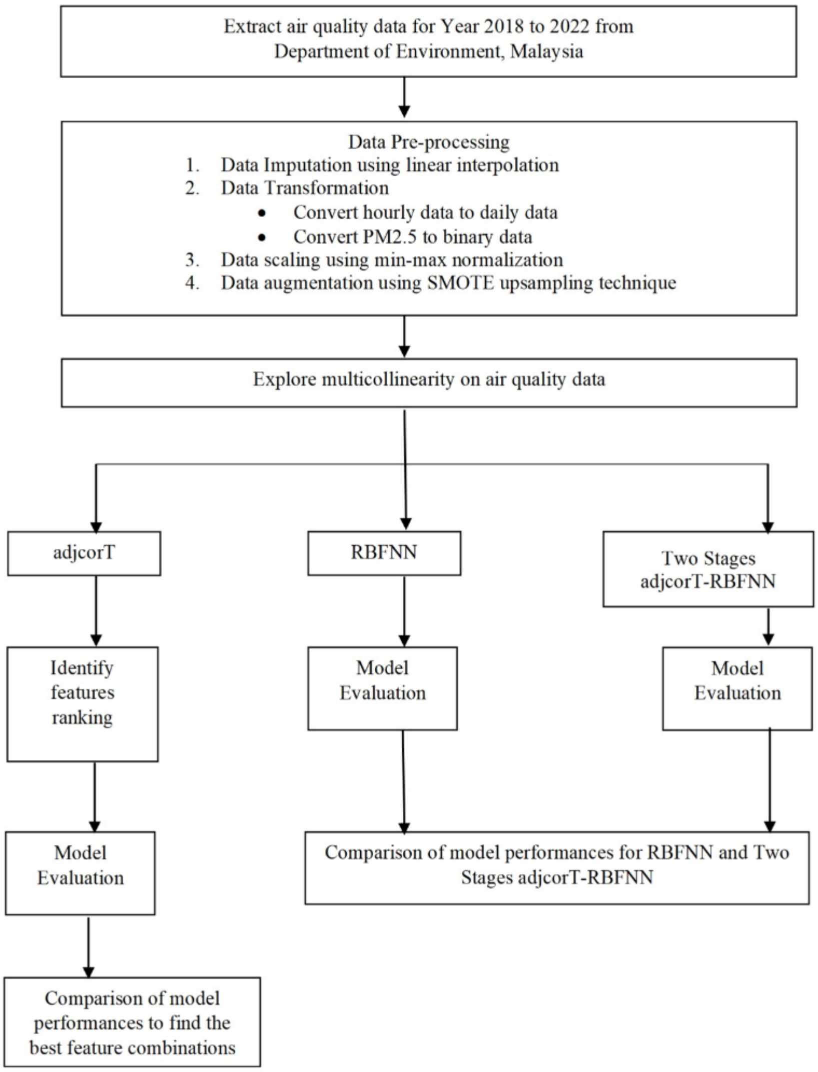
Extract air quality data for Year 2018 to 2022 from Department of Environment, Malaysia
Data Pre-processing
1. Data Imputation using linear interpolation
2. Data Transformation
• Convert hourly data to daily data
• Convert PM2.5 to binary data
3. Data scaling using min-max normalization
4. Data augmentation using SMOTE upsampling technique
Explore multicollinearity on air quality data
adjcorT
RBFNN
Two Stages adjcorT-RBFNN
Identify features ranking
Model Evaluation
Model Evaluation
Model Evaluation
Comparison of model performances for RBFNN and Two Stages adjcorT-RBFNN
Comparison of model performances to find the best feature combinations

Figure 1. Research flowchart

Table 1. Air quality data description

Variable	Unit	Description
$PM_{2.5}$	(µg/m^3)	Particulate matter 2.5 micrometres or less in diameters
PM_{10}	(µg/m^3)	Particulate matter 10 micrometres or less in diameters
SO_2	(ppm)	Sulphur dioxide
NO_2	(ppm)	Nitrogen dioxide
O_3	(ppm)	Ground-level ozone
CO	(ppm)	Carbon monoxide
WD	(°)	Wind direction
WS	(m/s)	Wind speed
Humidity	(%)	Relative humidity
Temperature	(°C)	Ambient temperature

2.3 Radial basis function neural network

Radial Basis Function Neural Network (RBFNN) is a type of artificial neural network that uses radial basis functions as activation functions in its hidden layer. The input layer, hidden layer, and output layer are the three main layers of an RBFNN, and they are typically connected by weights. First, a source node, also known as the independent variable, connects the network to its surroundings in the input layer. Meanwhile, a nonlinear transformation from input space to a high-dimension hidden space takes place in the hidden layer. The RBF neurons in the hidden layer use Gaussian or other radial basis functions to compute their activations based on the distance between their centres and the input data. The final layer is called the output layer, which is the result of the network applied to the input layer, also known as the predicted output. The network's output is then created by combining and weighting these activations. RBFNNs can approximate complex functions with comparatively few parameters which are very helpful for tasks involving function approximation and pattern recognition. This method can effectively handle high-dimensional input spaces and is particularly useful when handling nonlinear relationships in data such as air quality data. Figure 2 shows the theoretical framework for RBFNN.

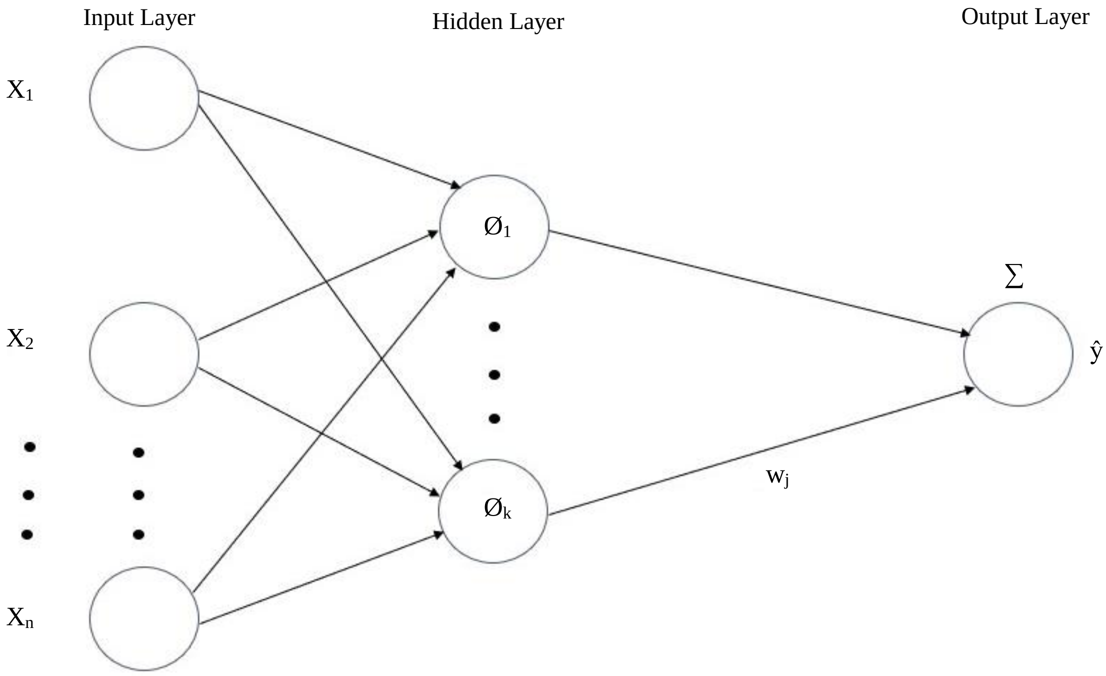

Figure 2. Theoretical framework for RBFNN (Wu, 1998)

2.4 Artificial neural network

An artificial neural network, or ANN, is a type of computational model that's design and operations are modelled after biological neural networks. It consists of networked nodes, sometimes known as "neurons," which work together to process and generate complex information. ANNs learn through by continuously modifying the connections among neurons, utilizing feedback to enhance predictions in comparison to intended results. This iterative learning process allows ANNs to gradually improve their capacity for prediction, also known as backpropagation.

2.5 Model evaluation

This research uses accuracy, sensitivity, specificity, precision, F1 score and Area Under the Receiver Operating Characteristic (AUROC) to evaluate the ANN classification model. Accuracy represents the proportion of correct predictions among the total predictions made by the model. To determine the accuracy of the model, the number of correct predictions is divided by the total number of predictions (Atif et al., 2023; Md Noh et al., 2023). Sensitivity, also known as the true positive rate, measures the proportion of actual positive cases that are correctly identified by the model (Zhang-James et al., 2023). While specificity, or the true negative rate, quantifies the proportion of actual negative cases that are correctly identified by the model (Oboya et al., 2023). Precision measures the proportion of true positive predictions among all positive predictions made by the model (Holakoei and Sajedi, 2023). The F1 score is the harmonic mean of precision and sensitivity, providing a balance between the two metrics (Peng et al., 2023). Lastly, the AUROC measures a classifier's ability to discriminate between groups (Weng et al., 2023).

3. RESULTS AND DISCUSSION

3.1 Results

This section presents and analyses the results obtained from our study, delving into the implications and significance of our findings. Table 2 shows missing values of variables in Shah Alam from the period 2018 to 2022. All variables have missing values below 8% except NO_2 with 12.65% of missing values. In this research, we use linear interpolation method to impute the missing values in the hourly dataset.

Table 2. Percentage of missing values

Variable	N	Missing value
$PM_{2.5}$	43,257	567 (1.29%)
PM_{10}	43,151	673 (1.54%)
SO_2	41,242	2582 (5.89%)
NO_2	38,280	5544 (12.65%)
O_3	41,198	2626 (5.99%)
CO	40,629	3195 (7.29%)
WD	42,925	899 (2.05%)
WS	42,868	956 (2.18%)
Humidity	42,907	917 (2.09%)
Temperature	42,926	898 (2.05%)

3.2 Data pre-processing

Data transformation and data cleaning are parts of the data pre-processing to improve the quality of the dataset. To facilitate the process of predicting $PM_{2.5}$ category for the following day ($PM_{2.5Dt1}$), the hourly dataset is converted to daily dataset. The $PM_{2.5}$ breakpoints (24-hour average) in the Table 3 below are based on the U.S. Environmental Protection Agency (EPA) that aims to protect public health from the harmful effects of fine particle pollution. Since our goal is to forecast the $PM_{2.5Dt1}$ category, we adopted a binary classification framework for air pollution prediction, where Air Quality Index (AQI) categories "good" and "moderate" were combined to represent the "not polluted" class, while the others AQI categories were grouped into the "polluted" class, following Kalajdjieski et al. (2020).

According to the descriptive statistics of independent variables in Table 4, the standard deviations range from 0 to 11.712. Data normalization was applied since it is clear that different scales were observed across variables in this study by using min-max normalization similar to a study by Du et al. (2019) on hybrid deep learning framework for air quality prediction. Besides, the histogram of the $PM_{2.5Dt1}$ category shows that the distribution of the category is not balanced as shown in Figure 3. Thus, this research also applied Synthetic Minority Over-sampling Technique (SMOTE) to balance the datasets.

Table 3. Binary labels for the respective $PM_{2.5}$ breakpoint and AQI categories

AQI category	$PM_{2.5}$ 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 4. Descriptive statistics before data pre-processing

Variable	N	Mean	Median	Std. Dev.	Skewness	Min	Max
$PM_{2.5}$	1,825	23.321	21.187	11.712	4.142	5.466	144.917
PM_{10}	1,825	32.554	30.244	13.739	3.086	7.184	156.553
SO_2	1,825	0.001	0.001	0.000	1.330	0.000	0.001
NO_2	1,825	0.015	0.015	0.005	0.308	0.002	0.037
O_3	1,825	0.020	0.019	0.007	0.800	0.002	0.060
CO	1,825	0.770	0.754	0.266	0.447	0.148	1.967
WD	1,825	206.597	205.583	48.507	0.106	86.023	317.928
WS	1,825	0.820	0.783	0.240	1.468	0.386	2.504
Humidity	1,825	80.138	80.109	6.603	-0.151	56.347	99.155
Temperature	1,825	27.552	27.573	1.247	-0.155	22.310	31.939

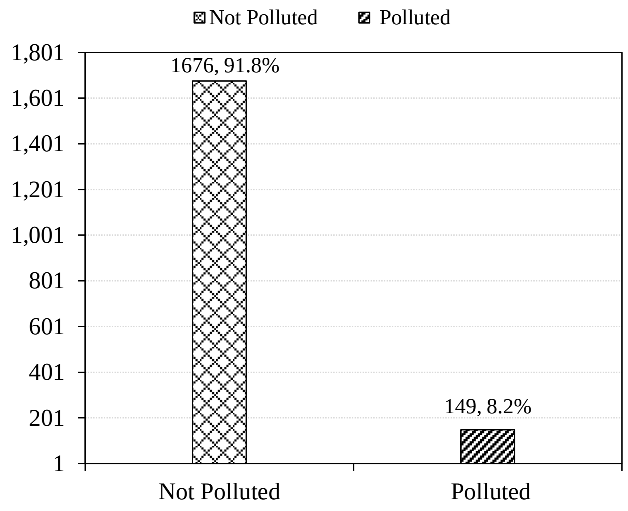

Figure 3. $PM_{2.5Dt1}$ distribution (Before SMOTE)

Descriptive statistics in Table 5 shows significant changes following the application of SMOTE and data normalization. Now, all mean, median values are within the range of 0 and 1 indicates that scaling to a standard range has been achieved. Following normalization, standard deviations have dropped, suggesting less variability among variables. In general, the values of skewness are closer to value 0, indicating a more balanced distribution. Figure 4 displayed the $PM_{2.5Dt1}$ distribution after SMOTE up sampling was applied to the dataset. It shows that both categories have a consistent number of sample sizes in which 1,676 (50.6%) are not polluted and 1,639 (49.4%) are polluted. Meanwhile, Figure 5 shows other features' distribution after the data pre-processing applied to the dataset. There are bell-shaped distributions for all features except for $PM_{2.5}$, PM_{10}, SO_2 and wind speed. By addressing issues of class imbalance and ensuring fair comparison across features, these transformations improve the dataset's suitability for developing precise predictive models.

Table 5. Descriptive statistics after data pre-processing

Variable	N	Mean	Median	Std. Dev.	Skewness	Min	Max
$PM_{2.5}$	3,315	0.176	0.144	0.142	3.239	0	1
PM_{10}	3,315	0.219	0.190	0.146	2.820	0	1
SO_2	3,315	0.227	0.215	0.100	1.332	0	1
NO_2	3,315	0.426	0.423	0.165	0.279	0	1
O_3	3,315	0.333	0.315	0.130	1.301	0	1
CO	3,315	0.395	0.388	0.165	0.512	0	1
WD	3,315	0.522	0.510	0.183	0.078	0	1
WS	3,315	0.214	0.201	0.100	1.174	0	1
Humidity	3,315	0.520	0.517	0.146	0.025	0	1
Temperature	3,315	0.559	0.565	0.118	-0.253	0	1

3.3 Correlation between features

The correlation between the features were examined in order to explore the correlation between features in the air quality data. The spearman correlation matrix provides insight into the relationships between the features. Values closer to 1 denote a strong positive correlation, while values closer to -1 denote a strong negative correlation, and values around 0 suggest no linear correlation between the independent variables. Based on the spearman correlation matrix output, there is a strong positive correlation between $PM_{2.5}$ and PM_{10} which is 0.97, indicating a significant association between these two pollutants. Similarly, NO_2 and O_3 have a fairly positive correlation (0.66), which illustrates moderate positive relationship between these two variables. Conversely, humidity shows a significant negative correlation (-0.8) with temperature, indicating a strong negative relationship between these meteorological variables. Other correlation values are not showing any multicollinearity issue exists seriously in the air quality data.

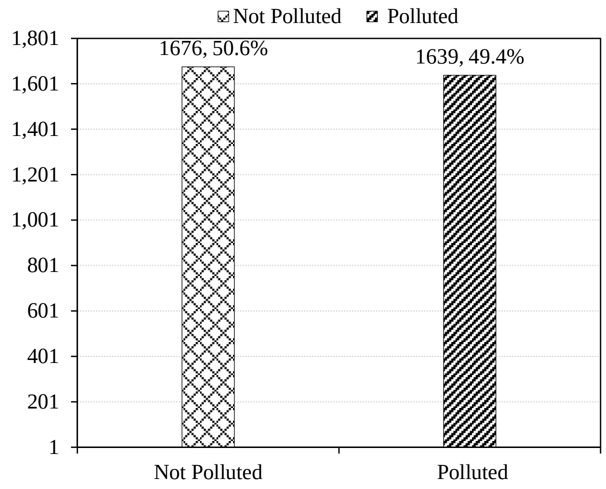

Figure 4. PM2.5$_{Dt1}$ distribution (After SMOTE)

3.4 Optimizing feature selection

The adjcorT values were ranked, the higher the value indicates the more significant the variable for developing a new model to predict the category of $PM_{2.5D+1}$ in Shah Alam, Selangor. As shown in Figure 5, the most important variable to predict $PM_{2.5D+1}$ is NO_2 while the least important variable is temperature with adjcorT value 23.81 and 14.99 respectively. By adding features one by one to the ANN model based on features ranking, we can determine the number of optimized features to predict $PM_{2.5D+1}$. The number of hidden nodes used for the ANN are followed as suggestion by Ul-Saufie et al. (2022), number of hidden nodes=(number of attributes + number of classes) / 2+1, while the learning rate values are set to 0.01. Based on the performance of ANN model in Table 6, the accuracy, sensitivity, specificity, precision and F1 score shows that the optimize features to predict $PM_{2.5D+1}$ is top 8 variables ranked by adjcorT method which is NO_2, $PM_{2.5}$, PM_{10}, CO, O_3, WS and SO_2. Figure 6 displays the various performances in a line charts for a clearer comparison view.

3.5 Model comparison

This research used RBFNN to verify the number of optimized features to predict $PM_{2.5D+1}$. The number of hidden nodes used for RBFNN and adjcorT-RBFNN model are 7 and 6 respectively as suggested by Ul-Saufie et al. (2022), number of hidden nodes = (number of attributes + number of classes) / 2 + 1, while the learning rate values are set to 0.01. Table 7 and Bar graph in Figure 7 shows the performance of the RBFNN model when using all 10 variables and using top 8 variables provided by adjcorT values ranking. We can conclude that the model using the top 8 variables in adjcorT-RBFNN outperformed the traditional RBFNN with higher accuracy, specificity, precision, F1 Score and AUROC which are 0.756, 0.801, 0.701, 0.757, and 0.758 respectively.

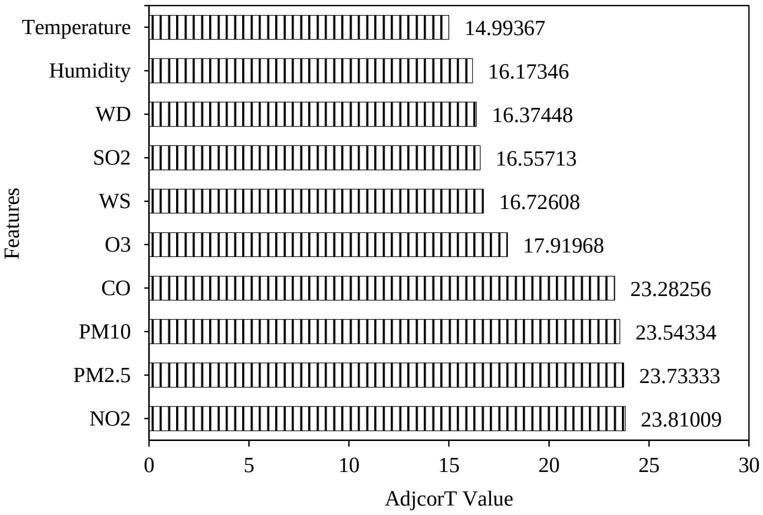

Figure 5. adjcorT values for each features

Table 6. Model performances for different numbers of features combination

No of features	1	2	3	4	5	6	7	8	9	10
Accuracy	0.67	0.75	0.76	0.76	0.75	0.80	0.78	0.81	0.80	0.80
Sensitivity	0.72	0.77	0.79	0.79	0.79	0.85	0.84	0.87	0.86	0.86
Specificity	0.63	0.73	0.73	0.74	0.71	0.75	0.73	0.76	0.75	0.76
Precision	0.63	0.73	0.73	0.74	0.71	0.75	0.73	0.76	0.75	0.76
F1 score	0.67	0.75	0.76	0.76	0.75	0.80	0.78	0.81	0.80	0.81
AUROC	0.67	0.75	0.76	0.76	0.75	0.80	0.78	0.81	0.80	0.81

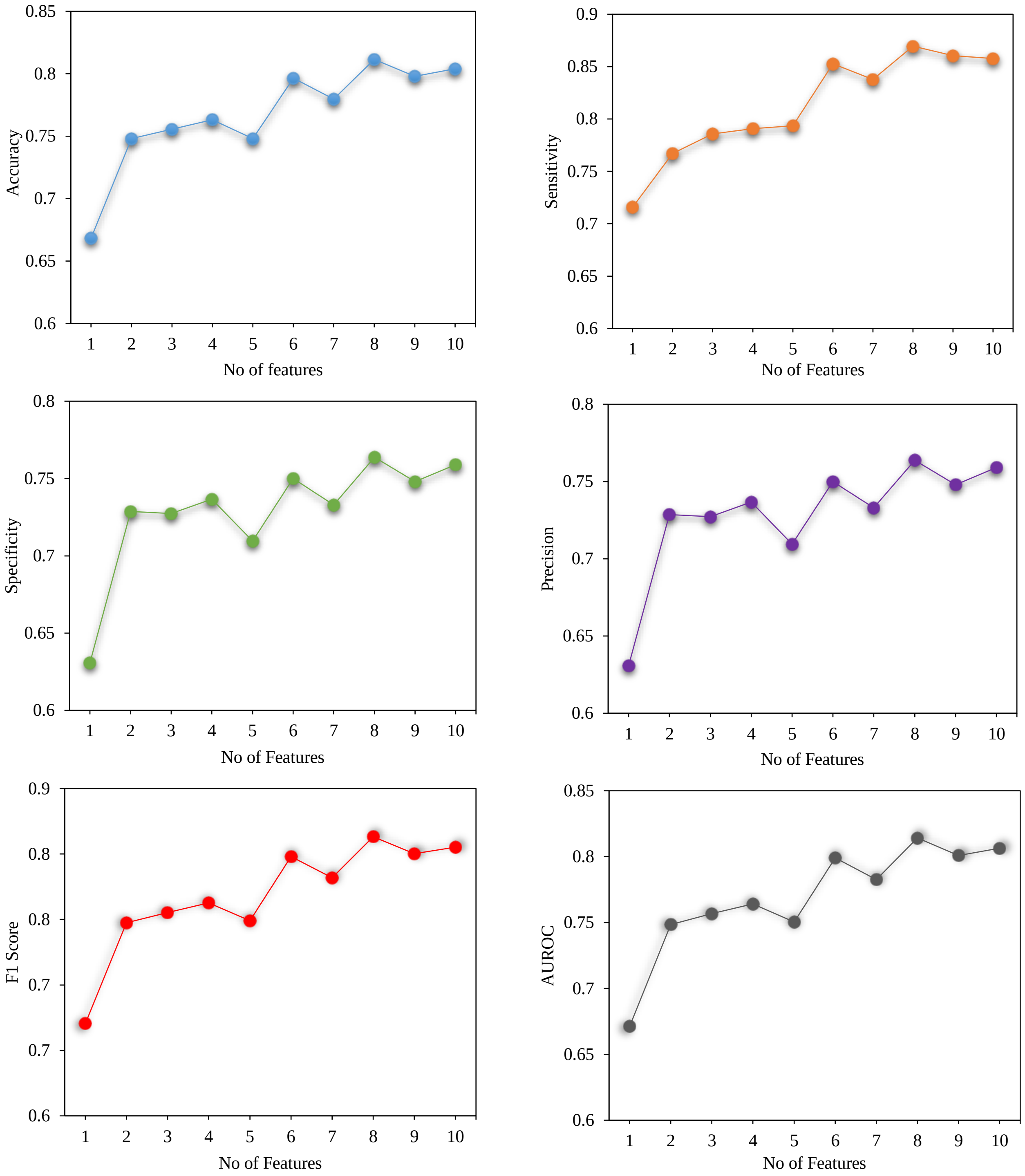

Figure 6. The model performances in different features combination included into the model

According to a study by Shaziayani et al. (2020), they found that combining SVM and BRT with feature selection techniques effectively reduced the prediction error and identified city-specific important variables for PM_{10} prediction. They found that the most important variables are PM_{10} followed by NO_2, CO, SO_2, relative humidity, temperature, and the least important is O_3 while wind speed is excluded from the model. Their findings have similarities with our research, in which all the important variables are the same with ours except the relative humidity and temperature. However, their research did not consider the multicollinearity issues in the air quality data since the researcher used other variable selection methods that did not consider the high correlation between the independent variables. Additionally, according to research by Afrin et al. (2021) has discovered wind speed and wind direction are the significant variables to predict $PM_{2.5}$ concentrations with 94% of total $PM_{2.5}$ variability explained by the model. Artificial neural networks have been used widely for prediction and classification including in air quality areas. A study by Hamami and Fithriyah (2020) focused on the classification of air pollution levels using artificial neural networks. They highlight the application of neural networks for air pollution classification into their model giving high accuracy, sensitivity and also specificity which is above 90%.

Table 7. RBFNN and adjcorT-RBFNN model performances

Model	RBFNN	adjcorT-RBFNN
Accuracy	0.753	0.756
Sensitivity	0.801	0.801
Specificity	0.712	0.717
Precision	0.712	0.717
F1 score	0.754	0.757
AUROC	0.755	0.758

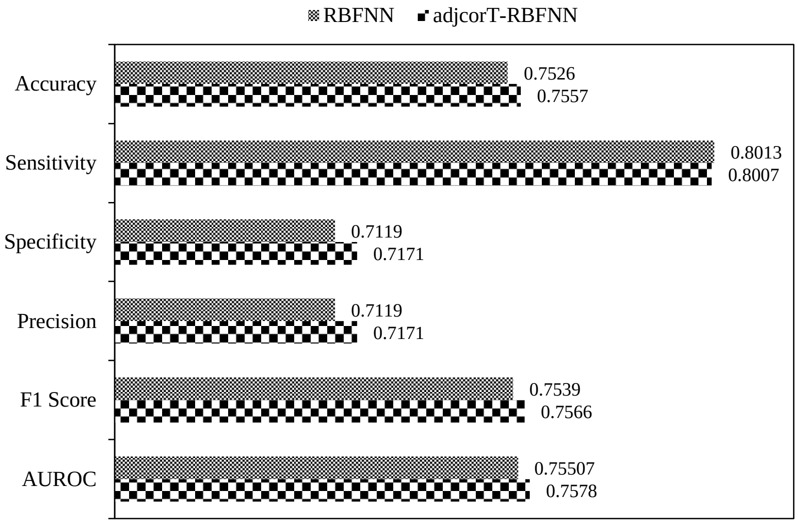

Figure 7. Horizontal Barchart of RBFNN and adjcorT-RBFNN model performances

4. CONCLUSION

This research utilizes the feature selection procedures especially on air quality data in enhancing prediction of $PM_{2.5}$. The result revealed that the model with two stages feature selection technique, adjcorT-RBFNN has proven to have better performance with higher accuracy, specificity, precision, F1 score and AUROC compared to RBFNN model. Ultimately, the daily $PM_{2.5}$ concentrations in Shah Alam may be anticipated using the proposed models. Additionally, the predicted model may serve as an instrument for an early warning system, providing local authorities with information on air quality for formulation of strategies of air quality improvement. In terms of the limitation, this study explored a comprehensive feature selection method applied on air quality data in Shah Alam, Selangor. However, further and in-depth studies may be needed to confirm its effectiveness in different monitoring stations in Klang Valley. Future study may study other feature selection methods and other machine learning classification techniques to apply on the air quality data. In addition, these methods also can be applied to other data in different fields.

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.

REFERENCES

Afrin S, Islam MM, Ahmed T. A meteorology based particulate matter prediction model for megacity Dhaka. Aerosol and Air Quality Research 2021;21(4):1-14.

Ali M, Khan F, Atta MN, Khan A, Khan A. Hybrid crow search and RBFNN: A novel approach to medical data classification. Journal of Informatics and Web Engineering 2024;3(1):252-64.

Atif M, Anwer F, Talib F, Alam R, Masood F. Analysis of machine learning classifiers for predicting diabetes mellitus in the preliminary stage. International Journal of Artificial Intelligence 2023;12(3):1302-11.

Du S, Li T, Yang Y, Horng SJ. Deep air quality forecasting using hybrid deep learning framework. IEEE Transactions on Knowledge and Data Engineering 2019;33(6):2412-24.

Farrell A, Wang G, Rush SA, Martin JA, Belant JL, Butler AB, et al. Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data. Ecology and Evolution 2019;9(10):5938-49.

Hamami F, Fithriyah I. Classification of air pollution levels using artificial neural network. Proceedings of the International Conference on Information Technology Systems and Innovation; 2020 Oct 19-23; Bandung-Padang: Indonesia; 2020.

Holakoei HR, Sajedi F. Compressive strength prediction of SLWC using RBFNN and LSSVM approaches. Neural Computing and Applications 2023;35(9):6685-97.

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

Kahoui H, Chekouri SM, Sahed A. A comparative study of ARIMA, RBFNN, and Hybrid RBFNN-ARIMA models for electricity net consumption forecasting in Algeria. Review of Socio-Economic Perspectives 2024;9(1):189-98.

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(24):1-9.

Ko UW, Kyung SY. Adverse effects of air pollution on pulmonary diseases. Tuberculosis and Respiratory Diseases 2022; 85(4):313-9.

Larasati A, Dwiastutik A, Ramadhanti D, Mahardika A. The effect of Kurtosis on the accuracy of artificial neural network predictive model. Proceedings of the MATEC Web of Conferences; 2018 Aug 30-31; Malang: Indonesia; 2018.

Noh SS, Ibrahim N, Mansor MM, Yusoff M. Hybrid filtering methods for feature selection in high-dimensional cancer data. International Journal of Electrical and Computer Engineering 2023;13(6):6862-71.

Oboya WM, Gichuhi AW, Wanjoya A. A Hybrid DNN-RBFNN model for intrusion detection system. Journal of Data Analysis and Information Processing 2023;11(04):371-87.

Peng HY, Duan SJ, Pan L, Wang MY, Chen JL, Wang YC, et al. Development and validation of machine learning models for nonalcoholic fatty liver disease. Hepatobiliary and Pancreatic Diseases International 2023;22(6):615-21.

Shaziayani WN, Ahmat H, Razak TR, Zainan Abidin AW, Warris SN, Asmat A, et al. A novel hybrid model combining the support vector machine (SVM) and boosted regression trees (BRT) technique in predicting PM_{10} concentration. Atmosphere 2022;13(12):1-17.

Sokhi RS, Moussiopoulos N, Baklanov A, Bartzis J, Coll I, Finardi S, et al. Advances in air quality research-current and emerging challenges. Atmospheric Chemistry and Physics Discussions 2021;22(7):4615-703.

Suleiman A, Tight MR, Quinn AD. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM_{10} and $PM_{2.5}$). Atmospheric Pollution Research 2019;10(1):134-44.

Ul-Saufie AZ, Hamzan NH, Zahari Z, Shaziayani WN, Noor NM, Zainol MR, et al. Improving air pollution prediction modelling using wrapper feature selection. Sustainability 2022;14(18):Article No. 11403.

Wang Z, Tian Z. Analysis of correlation between $PM_{2.5}$ and major pollutants by the method of path analysis. Proceedings of the International Symposium on Communication Engineering and Computer Science; 2018 Jul 28-29; Hohhot: China; 2018.

Weng S, Chen J, Ding C, Hu D, Liu W, Yang Y, et al. Utilizing machine learning algorithms for the prediction of carotid artery plaques in a Chinese population. Frontiers in Physiology 2023;14:1-12.

Yadav V, Nath S. Daily prediction of PM_{10} using radial basis function and generalized regression neural network. Proceedings of the Recent Advances on Engineering, Technology and Computational Sciences; 2018 Feb 6-8; Allahabad: India; 2018.

Zhang-James Y, Hoogman M, Franke B, Faraone SV. Machine learning and MRI-based diagnostic models for ADHD: Are we there yet? Journal of Attention Disorders 2023;27(4),335-53.

Zhao Z, Wu J, Cai F, Zhang S, Wang YG. A statistical learning framework for spatial-temporal feature selection and application to air quality index forecasting. Ecological Indicators 2022;144:1-16.