Feature Selection Methods Using RBFNN Based on Enhance Air Quality Prediction: Insights from Shah Alam

Siti Khadijah Arafin[1], Ahmad Zia Ul-Saufie[2], Nor Azura Md Ghani[3], Nurain Ibrahim[4]*
School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[1, 2, 3, 4]
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[4]

***Abstract*—This study examines the predictive efficiency of several feature selection approaches in air quality models aimed to predict next-day PM2.5 concentrations in Shah Alam, Malaysia. Air pollution in urban areas is a significant public health concern, and accurate prediction models are essential for timely interventions. However, determining the most important parameters to include in these models remains difficult, especially in complex urban areas with several pollution sources. To address this, we employed three different feature selection methods and applied them to a dataset comprising 43,824 air quality data points provided by the Department of Environmental Malaysia. The data set contained ten variables, such as gas pollutants and meteorological indicators. Each feature selection approach determined top eight variables to include in a Radial Basis Function Neural Network (RBFNN) model. The results showed that ReliefF outperformed Lasso and mRMR in terms of accuracy, specificity, precision, F1 Score, and AUROC, making it the most effective feature selection method for this study. This study contributes to the body of knowledge on air quality modelling by emphasising the relevance of using proper feature selection techniques that are suited to the specific characteristics of the dataset and urban area. Furthermore, it proposes that future study should look into the use of ReliefF-RBFNN in other settings, such as suburban and rural areas, as well as hybrid feature selection approaches to improve prediction performance across several context.**

***Keywords*—Lasso; mRMR; PM2.5 concentration; RBFNN; ReliefF**

I. Introduction

Globally, air quality has grown to be a major environmental and public health concern, especially in urban areas such as Shah Alam, Malaysia where pollution levels can have a substantial negative influence on people's quality of life. According to study [1], the rapid urbanization in Shah Alam has worsened environmental problems, including air quality deterioration. Predicting air quality, particularly the concentration of dangerous pollutants like PM2.5, is essential for mitigating these risks and providing guidance for public health initiatives. Although various pollutants contribute to air pollution, [2] suggest that PM 2.5 is the most significant affect air pollution. Numerous studies have explored various methods to forecast air quality by utilizing a variety of meteorological and environmental data such as PM10, PM2.5, CO, O3, relative humidity, ambient temperatures and wind speed.

In recent years, there has been a lot of focus on improving the accuracy of air quality predictions using powerful machine learning algorithms. Among these, feature selection approaches help to improve model performance by finding the most relevant variables while minimizing data dimensionality. [3] stated in their study that feature selection can improve model generalization by avoiding overfitting and mitigating the effects of the curse of dimensionality. According to study [4], filter technique, wrapper technique and embedded technique are three technique of feature selection. Various studies have applied different feature selection techniques on air quality data including Lasso, mRMR and reliefF. For instance, [5] used Lasso to find key features that influenced ozone (O3) levels during China's COVID-19 lockdown. Their findings revealed that Lasso efficiently identified key variables, such as O3 and meteorological conditions, which improved the model's interpretability. Moreover, the study in [6] presented a novel method that combines mRMR with Random Forest (RF) and Long Short-Term Memory (LSTM) networks to estimate the Air Quality Index (AQI). Their research showed that mRMR successfully identified key variables influencing AQI, greatly improving the model's predictive capability. Besides, ReliefF was used as a feature selection method for air pollution analysis in the Zonguldak region of Turkey in a study by [7]. They compared the performance of ReliefF with a firefly-based feature selection algorithm and found that even though ReliefF was effective, the firefly-based method outperformed it in classification tasks using Random Forest classifiers. However, most researchers prefer filter approaches because they have a straightforward algorithmic framework and are thus simple to apply [7]. However, the effectiveness of filter and embedded feature selection methods especially in predicting air quality data is still questionable.

Hence, this study aims to determine which feature selection method provides better performance in predicting air quality. This study compares two filter feature selection method and one embedded feature selection method which are Maximum Relevance Minimum Redundancy (mRMR), ReliefF and Least Absolute Shrinkage and Selection Operator (Lasso). The findings of this study will help to build more reliable air quality forecast models, with consequences for public health and environmental management. This paper is organized as follows: I. Introduction, II. Method, III. Results and Discussion. Then, it followed by the conclusion in Section IV and the reference lists.

II. METHOD

A. Research Flowchart

Fig. 1 below shows the flowchart outlines of this study to predicting next-day PM2.5 levels in Shah Alam, Malaysia, using air quality data from 2018 to 2022 provided by the Department of Environment, Malaysia. Data extraction is the first step in the process, which is then followed by extensive data pre-processing, such as imputation using linear interpolation, converting hourly data to daily figures, binary categorisation of PM2.5 levels, min-max normalisation, and dataset balancing using SMOTE. Next, three feature selection methods which are mRMR, Lasso, and ReliefF are applied to rank and select the top eight variables most relevant to predicting PM2.5. These selected features are then used to train a Radial Basis Function Neural Network (RBFNN), with the model's performance evaluated based on accuracy, specificity, precision, F1 Score, and AUROC. Finally, the best-performing model is identified by combining effective feature selection with the RBFNN.

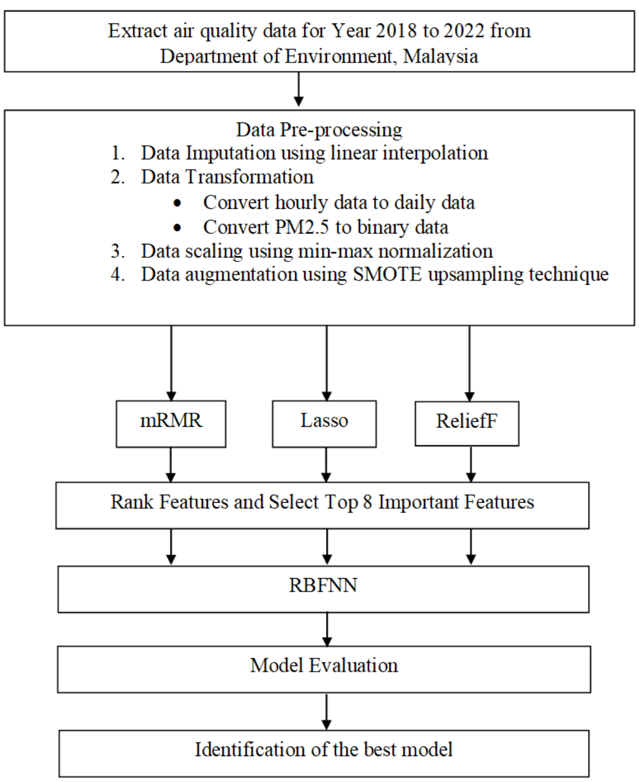

Fig. 1. Research flowchart.

B. Data Description

The Department of Environmental Malaysia provided 43,824 air quality data points for 10 variables, such as gas pollutants and meteorological parameters, collected in Shah Alam. Table I below shows the percentage of missing values for each variable. Based on Table I, all variables have missing values below 10% except NO2 with 12.65% of missing data. In contrast, PM2.5 has a low percentage of missing values with just 1.29% of missing data. Missing values can arise from various sources, including sensor malfunctions, environmental conditions, or data transmission errors. High levels of missing data, particularly in gas pollutants, could introduce biases or reduce the statistical power of the analysis if not properly addressed. Therefore, we applied linear imputation methods to ensure that these gaps do not compromise the accuracy of the predictive models.

TABLE I. PERCENTAGE OF MISSING VALUES

Variable	N	Missing Value
PM2.5	43257	567 (1.29%)
PM10	43151	673 (1.54%)
SO2	41242	2582 (5.89%)
NO2	38280	5544 (12.65%)
O3	41198	2626 (5.99%)
CO	40629	3195 (7.29%)
WD	42925	899 (2.05%)
WS	42868	956 (2.18%)
Humidity	42907	917 (2.09%)
Temperature	42926	898 (2.05%)

To predict the next day's air quality based on PM2.5 levels, we used a binary classification system where 0 represents “not polluted” and 1 represents “polluted”. We followed the methodology of [8], wherein the Air Quality Index (AQI) categories “Good” and “Moderate” were combined into the “not polluted” class, while the other categories were grouped into the “polluted” class. Table II shows the PM2.5 breakpoints (24-hour average) as defined by the U.S. Environmental Protection Agency (EPA).

TABLE II. BINARY LABELS FOR THE RESPECTIVE PM2.5 BREAKPOINT AND AQI CATEGORIES

AQI Category	PM2.5 Breakpoints
Good	0.0-12.0
Moderate	12.1-35.4
Unhealthy for Sensitive Groups	35.5-55.4
Unhealthy	55.5-150.4
Very Unhealthy	150.5-250.4
Hazardous	250.5 and above

C. Feature Selection Method

1) Lasso: The Least Absolute Shrinkage and Selection Operator (Lasso) is an embedded feature selection method that improves model performance by selecting and regularizing variables at the same time. Linear regression assigns weight to each feature, while LASSO regression proposed by Robert Tibshirani removes less significant ones from the subset [9]. Lasso effectively eliminates less significant features from the model by forcing some of the coefficients to be exactly zero by adding a penalty to the loss function that is equal to the absolute value of the magnitude of the coefficients. Furthermore, the study in [10], stated that Lasso method goals are to lower the variance in models with a high number of unnecessary variables.

The formula to calculate Lasso is shown in Eq. (1) below. Where, $\sum_{j=1}^{p}|\beta_j|$ is known as L_1 penalty term and the value is must below or equal to t, which is the upper bound of the summation of the absolute coefficients. While λ is the tuning parameter which controls the strength of the penalty (Ibrahim, 2020).

$$\hat{\beta} = \min \beta \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (1)$$

2) Maximum relevance minimum redundancy (mRMR): Maximum Relevance Minimum Redundancy (mRMR) is a widely used feature selection technique that seeks to reduce redundancy among the independent variables while identifying a subset of features that are most significant to a target variable. This method works especially well with high-dimensional datasets, where the number of features can significantly exceed the number of observations, making traditional modelling techniques less effective. The formula to calculates the maximum relevance shown in Eq. (2):

$$maxD(S,c), D = \frac{1}{|s|} \sum_{x_i \in s} I(x_i, c) \quad (2)$$

Based on the Eq. (2), x_i represents the i-th feature, while $c = \{c_0, c_1\}$ represents the class variables which is not polluted and polluted. L is 2 which denotes the total number of classes, and S indicates the feature subset. Moreover, to calculates the minimum redundancy shown in Eq. (3) below.

$$minR(S), R = \frac{1}{|s|^2} \sum_{x_i, x_J \in s} I(x_i; x_J) \quad (3)$$

3) ReliefF: ReliefF algorithm is an extended version of the Relief algorithm. It is a filter-based feature selection method that used to identify a feature's contribution to a target variable's prediction in order to find relevant features in high-dimensional data [11]. Unlike conventional methods, which just rely on statistical correlations or regression analysis, ReliefF operates by evaluating each feature's ability to differentiate between instances that belong to distinct classes. It has been demonstrated that ReliefF can handle noisy data and identify reelevant features in high-dimensional datasets [12].

The algorithm finds the k-nearest neighbours by repeatedly choosing random examples from training datasets. Finding the k-nearest in distinct classes, M_j (C), (j=1,2,…,k), where Euclidean distance is employed to determine the k-nearest neighbours, and H_j, (j=1,2,…,k) of R inside the same class [13]. The significance of each feature is estimated using the variation in feature values between these neighbours. Features that show significant variation across classes and small variations within the same class are more important. The weight of each characteristic is determined using the Eq. (4) below. While Eq. (5) is how the $diff$ is calculated [13].

$$w(f_i) = w(f_i) - \sum_{j=1}^{k} \frac{diff\ (f_i, R, H_j)}{mk} + \sum_{c \neq class(R)} \frac{p(C)}{1 - p(class(R))}\ X \sum_{j=1}^{k} \frac{diff\ \left(f_i, R, M_j(C)\right)}{mk}, \quad (i = 0,1, \dots, d) \quad (4)$$

$$diff(A, R_1, R_2) = \begin{cases} \frac{|R_1[A] - R_2[A]|}{\max A - \min A}, & if\ A\ is\ continous \\ 0, & if\ A\ is\ discrete\ and\ R_1[A] = R_2[A] \\ 1, & if\ A\ is\ discrete\ and\ R_1[A] \neq R_2[A] \end{cases} \quad (5)$$

where $diff(A, R_1, R_2)$ denotes the difference between samples R_1 and R_2 on feature A. While $R_1[A]$ and $R_2[A]$ indicate the values of sample R_1 and R_2 on feature A (Zhang et al., 2022).

4) Radial basis function neural network: Radial Basis Function Neural Networks (RBFNNs) are a subset of artificial neural networks that used radial basis functions as activation functions. The abilities of the model to describe complex nonlinear interactions makes them especially useful for problems involving function approximation, classification, and regression. There are three important layers in RBFNN which are the input layer, hidden layer, and output layer that are generally connected by weights. Firstly, a source node, also known as the independent variable is connected to the network to its surroundings in the input layer, meanwhile, the hidden layer involves a nonlinear transformation from input space to a high-dimension hidden space. Lastly, the output layer is the outcome of the network that applied to the input layer or called the predicted output. Fig. 2 shows the general framework of RBFNN [14].

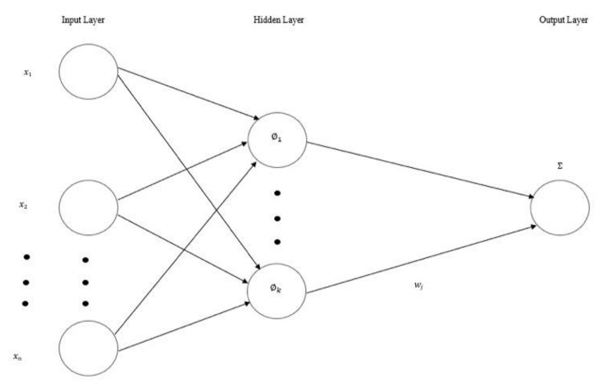

Fig. 2. General framework of a RBFNN.

The hidden layer comprises individual units, each corresponding to a transfer function $\emptyset_j$, which is generally a Gaussian function. The radial basis function (RBF), characterized by its radially symmetric shape, serves as the transfer function in this context. The number of hidden layer units directly corresponds to the number of RBFs used. The Gaussian radial basis function is formally defined in Eq. (6).

$$\emptyset(x) = \exp(-\frac{||x-c||^2}{2\sigma^2}) \quad (6)$$

where x is the input vector, c is the center of the RBF, and σ is the spread (width) parameter. The output of the hidden layer is calculated by taking a weighted sum of these radial basis functions. Specifically, for an input vector $X = \{x_1, x_2, \ldots, x_n\}$, the output of the hidden layer is given in Eq. (7).

$$g(X) = \sum_{j=1}^{k} w_j \emptyset_j (r\|X - C_j\|) \quad (7)$$

where w_j are the weights associated with the radial basis functions, C_j are the centers of the RBFs, while r is a scaling factor. In RBFNN, the output layer typically utilizes a logistic (sigmoid) activation function for binary classification tasks. The logistic function, as represented in Eq. (8), is employed to convert the weighted sum of the hidden layer outputs into a probability value within the interval of 0 to 1.

$$\sigma(\mathrm{z}) = \frac{1}{1+e^{-\mathrm{z}}} \quad (8)$$

Hence, the final output calculation of the RBFNN for binary classification is shown in Eq. (9), where w_0 is a bias term.

$$\hat{y} = \sigma\,(w_0 + \sum_{j=1}^{k} w_j \emptyset_j (r\|X - C_j\|)) \quad (9)$$

D. Model Performances

The performance of developed model in this study will be evaluated based on accuracy, sensitivity, specificity, precision, F1 score and AUROC. The accuracy is the number of correct predictions in a given number of predictions [15]. The formula to calculates accuracy is shown in Eq. (10):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (10)$$

Sensitivity and specificity are used to explain the relationship between the system's input and output variables and to evaluate the model's resilience in the face of uncertainty. The percentage of real positives that are correctly identified is known as sensitivity, or the True Positive (TP) rate, while the percentage of real negatives that are correctly recognised is known as specificity, or the True Negative (TN) rate. The following Eq. (11) and Eq. (12) provide the formulas for determining specificity and sensitivity, respectively.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (11)$$

$$Specificity = \frac{TN}{(TN+FP)} \quad (12)$$

Precision quantifies the proportion of correctly classified positive samples out of all samples classified as positive. On the other hand, the F1 score represents the harmonic mean of precision and sensitivity, providing an indication of whether the model's performance is well-balanced. Eq. (13) and Eq. (14) below present the formulas for calculating precision and the F1 score, respectively.

$$Precision = \frac{TP}{(TP+FP)} \quad (13)$$

$$F1\ Score = \frac{2\ x\ Precision\ x\ Sensitivity}{(Precision+Sensitivity)} \quad (14)$$

ROC (Receiver Operating Characteristic) curve examines the relationship between the true positive rate (sensitivity) on the y-axis and the false positive rate (1-specificity) on the x-axis, serving as a tool to assess the performance of the classifier. The AUROC (Area Under the ROC Curve) quantifies the model's ability to distinguish between classes. The AUC value ranges from 0 to 1, where an AUC of 0.0 signifies a model with entirely incorrect predictions, and an AUC of 1.0 indicates a model with perfectly accurate predictions. Hence, high values of accuracy, sensitivity, specificity, F1 score and AUROC indicates that the performance model is good.

III. Results and Discussion

This section displays the result and discussion of this study. Table III presents the descriptive statistics for the independent variables, which shows a broad range of standard deviations from 0 to 48.507, showing different scales for each variable. To address this, the data was standardized using min-max normalization, as described by [16] in their study on prediction of air pollutants for air quality using deep learning methods in a metropolitan city. Furthermore, the histogram for the $PM2.5_{Dt1}$ category in Fig. 3 shows an imbalance in the distribution. Thus, the Synthetic Minority Over-sampling Technique (SMOTE) was used to ensure a balanced dataset, improving the dependability of future results.

The descriptive statistics after data pre-processing are shown in Table IV. Following min-max normalization, all mean and median values now fall within the range of 0 to 1, indicating that the data has been successfully scaled to a standard range. Additionally, the skewness values are now closer to 0, reflecting a more balanced distribution across the dataset. Moreover, Fig. 4 shows the distribution of $PM2.5_{Dt1}$ after the application of SMOTE up sampling to the dataset. It demonstrates that there is a consistent amount of sample sizes in both groups, with 1676 (50.6%) not polluted and 1639 (49.4%) are polluted.

TABLE III. Descriptive Statistics of before Data Pre-Processing

Variable	N	Mean	Median	Std. Dev.	Skewness
PM2.5	1825	23.321	21.187	11.712	4.142
PM10	1825	32.554	30.244	13.739	3.086
SO2	1825	0.001	0.001	0.000	1.330
NO2	1825	0.015	0.015	0.005	0.308
O3	1825	0.020	0.019	0.007	0.800
CO	1825	0.770	0.754	0.266	0.447
WD	1825	206.597	205.583	48.507	0.106
WS	1825	0.820	0.783	0.240	1.468
Humidity	1825	80.138	80.109	6.603	-0.151
Temperature	1825	27.552	27.573	1.247	-0.155

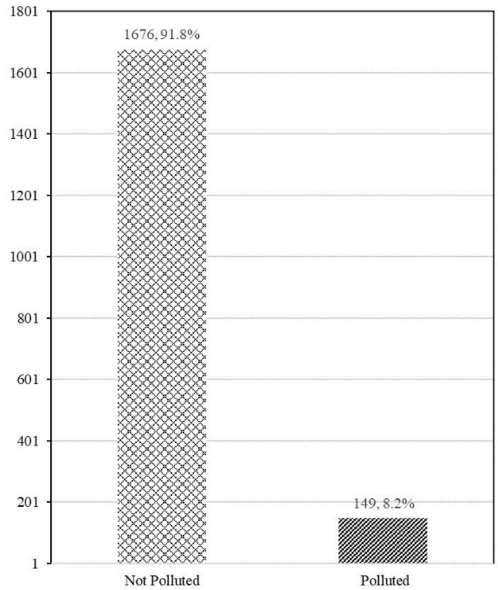

Fig. 3. PM2.5$_{Dt1}$ distribution of (Before SMOTE).

TABLE IV. DESCRIPTIVE STATISTICS OF AFTER DATA PRE-PROCESSING

Variable	N	Mean	Median	Std. Dev
PM2.5	3315	0.176	0.144	0.142
PM10	3315	0.219	0.190	0.146
SO2	3315	0.227	0.215	0.100
NO2	3315	0.426	0.423	0.165
O3	3315	0.333	0.315	0.130
CO	3315	0.395	0.388	0.165
WD	3315	0.522	0.510	0.183
WS	3315	0.214	0.201	0.100
Humidity	3315	0.520	0.517	0.146
Temperature	3315	0.559	0.565	0.118

Table V shows the rankings independent variables according to three feature selection methods which are Lasso, mRMR, and ReliefF on predicting PM2.5$_{D+1}$ in Shah Alam. Each features methods identifies top 8 independent variables to includes in the RBFNN model. The Lasso method ranks PM2.5 as the most critical feature to predict PM2.5 levels of the next day, a result that contrasts with the mRMR and ReliefF methods, where PM2.5 is not included among the top 8 features. However, the mRMR method select PM10 as the most significant feature while Lasso method rank PM10 as second most important features.

ReliefF, on the other hand, identify wind direction as the most important feature. This implies that variations in wind direction can result in significant variations in the dispersion or concentration of pollutants in urban environments. Moreover, [17] investigated the relationship between wind direction and air quality, specifically focusing on fine particulate matter (PM2.5) and its precursor gases. The investigation shows that the distribution and concentration of these contaminants are strongly influenced by wind direction. Moreover, this study shows ReliefF ranks both humidity and temperature as the least important features which is 7 and 8, in contrast to the Lasso and mRMR methods, which assign higher importance to these variables in predicting PM2.5 levels.

Furthermore, all three methods did not select SO2 as top 8 important variables to predict PM2.5 levels of the next day. According to a study by [18] on analysis of air pollution levels in a settlement area using passive sampling methods. Despite of SO2 presence, their results showed that SO2 had a small impact on the area under study's overall air quality index (AQI). This study supports the findings that SO2 may not have a significant impact on air quality projections, especially in areas where other pollutants predominate.

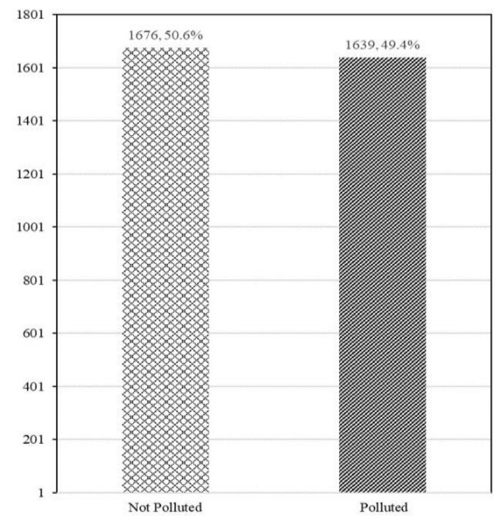

Fig. 4. PM2.5$_{Dt1}$ distribution of (After SMOTE).

TABLE V. FEATURE RANKING ACROSS LASSO, MRMR, AND RELIEFF METHODS

Variable	Lasso	mRMR	reliefF
PM2.5	1	-	5
PM10	2	1	6
SO2	-	-	-
NO2	4	6	2
O3	-	3	4
CO	6	7	3
WD	7	8	1
WS	8	2	-
Humidity	3	5	7
Temperature	5	4	8

Table VI shows the comparison of RBFNN model performance by using different feature selection methods to predict air quality of the next day based on PM2.5 level. According to Table VI, ReliefF outperformed Lasso and mRMR method with higher accuracy, specificity, precision, F1 Score and AUROC value which are 0.757, 0.719, 0.719, 0.758 and 0.759 respectively. This findings contrast with a study by

[9], who found that the Lasso method outperformed the ReliefF method in terms of performance metrics. However, it is important to note that [9] utilized a different classifier, specifically KNN, whereas this study employs RBFNN. Besides, a comparative study by [19], ReliefF was evaluated alongside Lasso and mRMR for survival prediction models. The findings showed that ReliefF was able to consistently find more relevant features than Lasso and mRMR, which had difficulty to maintain stability in their selections.

TABLE VI. COMPARISON MODEL PERFORMANCE

Model	Lasso	mRMR	ReliefF
Accuracy	0.735	0.753	**0.757**
Sensitivity	0.771	**0.818**	0.801
Specificity	0.702	0.703	**0.719**
Precision	0.702	0.703	**0.719**
F1 Score	0.735	0.756	**0.758**
AUROC	0.736	0.756	**0.759**

IV. CONCLUSION

In conclusion, this study's findings highlight the variety in feature selection approaches and their impact on the predictive effectiveness of air quality models in urban area which is Shah Alam, Malaysia. Among the three methods evaluated, ReliefF emerged as the most successful feature selection method for predicting next-day PM2.5 levels, outperforming both Lasso and mRMR in terms of accuracy, specificity, precision, F1 Score, and AUROC. This outcome aligns with some research that underscores ReliefF's ability to reliably detect relevant features, although other studies have favored the Lasso method. This study recommends that future research to explore the application of reliefF-RBFNN method to other type of areas such as sub-urban and rural area. This study also suggests that future studies should be conducted in other urban regions with varying climatic and pollutant features to confirm the generalisability of the findings and to develop more robust air quality prediction models that can be tailored to various situations. Moreover, it is recommended future researcher to explore on hybrid feature selection method such as ReliefF (filter) integrated with Lasso (embedded) feature selection method that might improve prediction performance across a variety of urban settings. However, the output of this study is not generalised to other country as the air quality patterns are differed across country.

ACKNOWLEDGMENT

Authors acknowledge acknowledge the Ministry of Higher Education (MOHE) for funding under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2023/STG06/UITM/02/8).

REFERENCES

[1] Y. A. Abdullah et al., "Urban Governance Approaches for Low Carbon Cities. The Case of Shah Alam Logal Government, Malaysia," Plan. MALAYSIA, vol. 20, no. 23 SE-Article, Nov. 2022, doi: 10.21837/pm.v20i23.1169.

[2] J. Angelin Jebamalar and A. Sasi Kumar, "PM2.5 prediction using machine learning hybrid model for smart health," Int. J. Eng. Adv. Technol., vol. 9, no. 1, pp. 6500–6504, 2019, doi: 10.35940/ijeat.A1187.109119.

[3] M. Hosni, A. Idri, and A. Abran, "On the value of filter feature selection techniques in homogeneous ensembles effort estimation," J. Softw. Evol. Process, vol. 33, no. 6, pp. 1–38, 2021, doi: 10.1002/smr.2343.

[4] A. Z. Ul-Saufie et al., "Improving Air Pollution Prediction Modelling Using Wrapper Feature Selection," Sustainability, vol. 14, no. 18. 2022, doi: 10.3390/su141811403.

[5] S. Liu et al., "Distinct regimes of O3 response to covid-19 lockdown in China," Atmosphere (Basel)., vol. 12, no. 2, pp. 1–10, 2021, doi: 10.3390/atmos12020184.

[6] H. Wu, T. Yang, H. Li, and Z. Zhou, "Air quality prediction model based on mRMR–RF feature selection and ISSA–LSTM," Sci. Rep., vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-39838-4.

[7] E. S. Eşsiz, V. N. Kılıç, and M. Oturakçı, "Firefly-Based feature selection algorithm method for air pollution analysis for Zonguldak region in Turkey," Turkish J. Eng., vol. 7, no. 1, pp. 17–24, 2023, doi: 10.31127/tuje.1005514.

[8] J. Kalajdjieski et al., "Air Pollution Prediction with Multi-Modal Data and Deep Neural Networks," Remote Sensing, vol. 12, no. 24. 2020, doi: 10.3390/rs12244142.

[9] M. S. Hammad, V. F. Ghoneim, M. S. Mabrouk, and W. I. Al-atabany, "A hybrid deep learning approach for COVID-19 detection based on genomic image processing techniques," Sci. Rep., vol. 13, no. 1, pp. 1–22, 2023, doi: 10.1038/s41598-023-30941-0.

[10] N. Ibrahim, "Variable Selection Methods for Classification : Application To Metabolomics Data," University of Liverpool, 2020.

[11] S. S. Md Noh, N. Ibrahim, M. M. Mansor, N. A. Md Ghani, and M. Yusoff, "Hybrid embedded and filter feature selection methods in big-dimension mammary cancer and prostatic cancer data," IAES Int. J. Artif. Intell., vol. 13, no. 3, pp. 3101–3110, 2024, doi: 10.11591/ijai.v13.i3.pp3101-3110.

[12] A. Desiani et al., "The Comparison of ReliefF and C.45 for Feature Selection on Heart Disease Classification Using Backpropagation," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 17, no. 2, pp. 183–194, 2023, doi: 10.22146/ijccs.82948.

[13] B. Zhang, Y. Li, and Z. Chai, "A novel random multi-subspace based ReliefF for feature selection," Knowledge-Based Syst., vol. 252, p. 109400, 2022, doi: https://doi.org/10.1016/j.knosys.2022.109400.

[14] H. Adeli and M. Wu, "Regularization neural network for construction cost estimation," J. Constr. Eng. Manag., vol. 124, no. 1, pp. 18–24, 1998.

[15] S. S. Md Noh, N. Ibrahim, M. M. Mansor, and M. Yusoff, "Hybrid filtering methods for feature selection in high-dimensional cancer data," Int. J. Electr. Comput. Eng., vol. 13, no. 6, p. 6862, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6862-6871.

[16] B. Das, Ö. O. Dursun, and S. Toraman, "Prediction of air pollutants for air quality using deep learning methods in a metropolitan city," Urban Clim., vol. 46, p. 101291, 2022, doi: https://doi.org/10.1016/j.uclim.2022.101291.

[17] H. Karimian et al., "Spatio-temporal variation of wind influence on distribution of fine particulate matter and its precursor gases," Atmos. Pollut. Res., vol. 10, no. 1, pp. 53–64, 2019, doi: https://doi.org/10.1016/j.apr.2018.06.005.

[18] N. Djamal, A. Suryanto, A. Artiningsih, N. Nasrullah, R. D. T. Manningtyas, and Y. Velina, "Analysis of Air Pollution Level In Settlement Area Using Passive Sampler Method," Int. J. Hydrol. Environ. Sustain., vol. 1, no. 2, pp. 97–107, 2022, doi: 10.58524/ijhes.v1i2.77.

[19] J. Hu et al., "Development of survival predictors for high-grade serous ovarian cancer based on stable radiomic features from computed tomography images," iScience, vol. 25, no. 7, p. 104628, 2022, doi: https://doi.org/10.1016/j.isci.2022.104628.