Improving Air Quality Prediction Models for Banting: A Performance Evaluation of Lasso, mRMR, and ReliefF

Siti Khadijah Arafin[1], Suvodeep Mazumdar[2], Nurain Ibrahim[1, 3]*
School of Mathematical Sciences, College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[1]
Information School, University of Sheffield, The Wave, 2 Whitham Road, Sheffield S10 2AH, United Kingdom[2]
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[3]

***Abstract*—This study explores the effectiveness of various feature selection methods in forecasting next-day PM2.5 levels in Banting, Malaysia. The accurate prediction of PM2.5 concentrations is crucial for public health, enabling authorities to take timely actions to mitigate exposure to harmful pollutants. This study compares three feature selection methods: Lasso, mRMR, and ReliefF using a dataset consisting of 43,824 data points collected from Banting air quality monitoring stations (CA22B). The dataset includes ten variables, including pollutant concentrations such as O3, CO, NO2, SO2, PM10, and PM2.5, along with meteorological parameters such as temperature, humidity, wind direction and wind speed. The results revealed that Lasso outperformed both mRMR and ReliefF in terms of various performance metrics, including accuracy, sensitivity, precision, F1 score, and AUROC. Lasso demonstrated superior ability to handle multicollinearity, significantly improving the interpretability of the model by retaining only the most important variables. This suggests that the effectiveness of feature selection methods is highly dependent on the characteristics of the dataset, such as correlations among features. Thus, the top eight features to predict PM2.5 levels in Banting selected by Lasso method are relative humidity, PM2.5, wind direction, ambient temperature, PM10, NO2, wind speed, and O3. The findings from this study contribute to the growing body of knowledge on air quality prediction models, highlighting the importance of selecting the appropriate feature selection method to achieve the best model performance. Future research should explore the application of Lasso method in other geographical regions, including urban, suburban and rural areas, to assess the generalizability of the results.**

***Keywords*—PM2.5 concentration; feature selection; Lasso; mRmR; RBFNN; ReliefF**

I. Introduction

There is an increasing emphasis on air quality research, such as air quality prediction and the health effects of contaminants. This surge in attention is driven by rising pollution levels, growing health concerns, and advancements in technology that make monitoring and prediction more accessible. According to [1], aside from Kuala Selangor, all monitoring stations in the Klang Valley reported poor air quality days, with Banting having the second-highest occurrence at five days, eventhough Banting has less traffic compared to other urban area such as Shah Alam.

Prediction of PM2.5 levels (particulate matter with a diameter of less than 2.5 microns) has attracted considerable interest because of its significant impact on human health and its role as a key indicator of air quality. Previous studies such as in study [2] highlight the association between prenatal and postnatal PM2.5 exposure and a high incidence of tic disorders in children. Additionally, the study in [3] used machine learning techniques to predict PM2.5 concentrations using historical air quality data from Kuala Lumpur, demonstrating that advanced modelling approaches can significantly improve the accuracy of air quality predictions, which is critical for public health advisories and environmental management.

More recently, neural networks (a type of machine learning) have gained prominence in predicting air quality due to their ability to model complex, nonlinear relationships inherent in air quality data. One significant study by [4] on artificial neural networks (ANNs) in Peninsular Malaysia demonstrated their effectiveness in accurately predicting pollutants and the Air Pollutant Index. Radial Basis Function Neural Network (RBFNN), a machine learning method that has a simple architecture consisting of an input, hidden, and output layers, demonstrated faster convergence during training compared to more complex architectures like Multi-Layer Perceptrons (MLP). Moreover, the accuracy of the RBFNN model in predicting the occurrence of haloketones in tap water, as shown in [5], exhibited high performance, effectively capturing the complex relationships between input parameters and the predicted outcomes.

To further improve the accuracy of PM2.5 prediction models, researchers have increasingly focused on feature selection methods. According to a study by [6], the findings indicated that feature selection methods improved prediction accuracy by at least 13.7% compared to models that did not employ feature selection. ReliefF is a filter-based feature selection method that is effectively used to train models for classification due to its ability to identify relevant features by ranking them based on their capacity to distinguish between instances [7]. In addition, the study in [8] noted that many

previous studies address multicollinearity issues in feature selection methods by removing redundant and irrelevant features from high-dimensional data, which can be effective in preventing deterioration in model performance. However, removing redundant features based solely on the correlation between variables might not provide accurate predictions, as the removed variables may have unique characteristics.

Therefore, a feature selection method that focuses on the correlation between variables is needed. For example, the study in [9] discussed using correlation-embedded attention modules to reduce multicollinearity can improve the model performance and interpretability. Minimum Redundancy Maximum Relevance (mRMR) is a filter-based, correlation-based feature selection method that focuses on selecting features that maximize relevance to the target variable while minimizing redundancy among them, thereby extracting the most representative features closely related to the target variable [10]. In addition, Least Absolute Shrinkage and Selection Operator (Lasso) is another feature selection method that effectively addresses multicollinearity by applying regularization to shrink less important feature coefficients to zero, thus enhancing model interpretability and reducing overfitting [11].

There are numerous feature selection techniques, such as Recursive Feature Elimination (RFE) and XGBoost Feature Importance. However, our study focuses on Lasso, mRMR, and ReliefF due to their distinct selection mechanisms and prior applications in air quality prediction research. The research in [12] points out that while RFE can achieve high performance, it suffers from computational inefficiency due to its reliance on iterative model training. This issue is particularly relevant when dealing with large datasets, as the time required for multiple iterations can become prohibitive. Meanwhile, XGBoost Feature Importance ranks features based on their contribution to decision trees but does not explicitly account for multicollinearity. Unlike Lasso, which penalizes correlated predictors to improve interpretability, XGBoost may distribute importance among correlated variables, making it less effective for identifying the most relevant individual predictors in highly correlated datasets.

Additionally, a study by [13] compared the ReliefF, mRMR, and Lasso methods using air quality data from Shah Alam. The study concluded that ReliefF outperformed the other two methods and recommended that future researchers compare these three feature selection methods in other urban air quality datasets. Based on this recommendation, our study applies these methods to air quality data from Banting, providing insights into how these methods perform in a different urban setting with distinct environmental and meteorological conditions. As an industrial area, Shah Alam is influenced by emissions from factories, vehicles, and urban activities, leading to higher concentrations of pollutants such as PM2.5, PM10, SO2, NO2, and CO. In contrast, Banting, an agricultural area, experiences pollution primarily from agriculture activities and occasional biomass burning, which may result in different pollutant levels and patterns compared to Shah Alam.

Therefore, this study will compare the performance of the RBFNN model combined with ReliefF, mRMR, and Lasso feature selection methods to determine the most effective approach to predict the next day concentration of PM2.5 in Banting. The study's findings will benefit policymakers and other relevant parties by providing evidence-based insights into the most effective feature selection methods for improving air quality prediction models.

II. Research Methods

A. Dataset

This study used the Banting air quality dataset provided by the Department of Environment Malaysia (DOE). Ten variables (PM2.5, PM10, SO2, NO2, O3, CO, wind direction, wind speed, relative humidity, and ambient temperature) used as independent variables, were extracted from hourly data spanning four years (2018 to 2022). However, there are missing values in the dataset as shown in Table I, with NO2 having the highest percentage of missing data at 8.68%, while relative humidity is the lowest at 1.39%. In order to address the missing data points, we employed linear interpolation method, as was suggested in study [14].

TABLE I. Percentage of Missing Values

Variable	N	Missing Value
PM2.5	43207	617 (1.41%)
PM10	43075	749 (1.71%)
SO2	40938	2886 (6.59%)
NO2	40021	3803 (8.68%)
O3	41182	2642 (6.03%)
CO	40735	3089 (7.05%)
WD	43135	689 (1.57%)
WS	42927	897 (2.05%)
Humidity	43217	607 (1.39%)
Temperature	42443	1381 (3.15%)
$PM2.5_{D+1}$	43207	617 (1.41%)

The target variable in this study is the PM2.5 levels of the next day. Hence, the hourly dataset was transformed to daily categorical data by averaging values over 24 hours. The PM2.5 breakpoint of air quality categories is shown in Table II based on guidelines by DOE. Furthermore, this study employed classification task, hence the target variable, $PM2.5_{D+1}$ was transformed into binary classification system, not polluted (0) and polluted (1) as suggested in [15]. The "good" and "moderate" category represent not polluted (0) class, while other than that are represent polluted (1) class [15].

TABLE II. 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

B. Research Framework

Research framework of this study is shown in Fig. 1. The process begins with data extraction as mentioned previously. Then, followed by extensive data pre-processing steps which include imputation using linear interpolation, converting hourly data into daily averages, binary categorization of PM2.5 levels, min-max normalization, and balancing the dataset with SMOTE technique. Subsequently, the three feature selection methods: ReliefF, mRMR, and Lasso are applied to rank and select the top eight variables most relevant to PM2.5 prediction as suggested in study [13] and study [16]. The selected features are used to train a Radial Basis Function Neural Network (RBFNN). A comparative evaluation of accuracy, specificity, precision, F1 score, and AUROC was finally used to identify the most effective model.

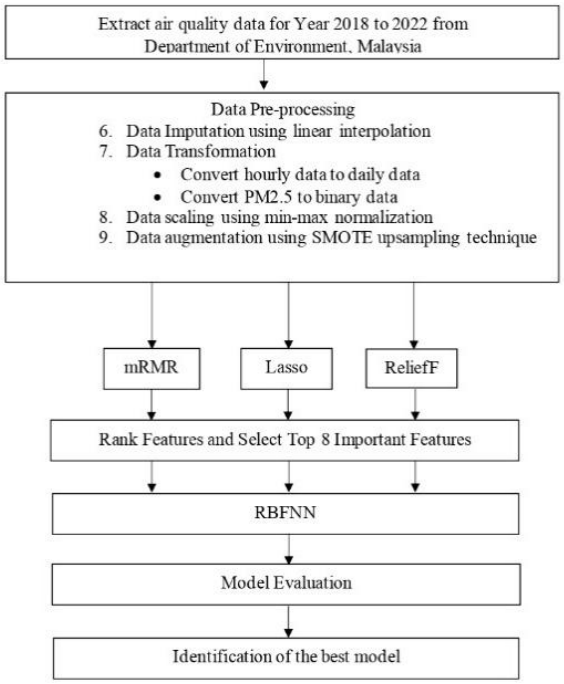

Fig. 1. Research flowchart.

C. Feature Selection Method

1) ReliefF: The ReliefF algorithm is a filter-based feature selection method known for its effectiveness in high-dimensional and noisy datasets [17]. It evaluates feature importance by assessing their ability to distinguish between instances of different classes. Unlike traditional methods relying on statistical correlations, ReliefF employs a distance-based approach, sampling data points and identifying the k-nearest neighbors within the same class (nearest hits) and from other classes (nearest misses). By comparing feature values between these neighbors, ReliefF assigns higher importance to features with substantial variation across classes and minimal variation within the same class.

The weight of a feature W[A] is updated iteratively using the formula in Eq. (1) [18], where H_i is the nearest hit (same class), while M_i is the nearest miss (different class). Moreover, the difference is calculate as shown in Eq. (2), where $diff(A, X, Y)$ are the difference in feature A values between instances X and Y.

$$W[A] = W[A] - \frac{1}{m}\sum_{i=1}^{m}(diff(A, H_i) - diff(A, M_i)) \quad (1)$$

$$diff(A, X, Y) = \begin{cases} 0 \; if \; X[A] = Y[A] \\ 1 \; if \; X[A] \neq Y[A] \end{cases} \quad (2)$$

2) Maximum Relevance Minimum Redundancy (mRMR): Maximum Relevance Minimum Redundancy (mRMR) is a feature selection method that aims to choose the most relevant features while minimizing redundancy among them. It is particularly useful in high-dimensional datasets where feature selection is crucial for improving model performance. The approach maximizes the relevance of selected features to the target variable and minimizes the redundancy between them. The relevance of a feature x_i to the target variable, c are calculated using mutual information, while redundancy between features is determined by the pairwise mutual information between features x_i and x_j. Eq. (3) and Eq. (4) shows the formulas to calculate maximum relevance and minimum redundancy, respectively.

$$maxD(S, c), D = \frac{1}{|s|}\sum_{x_i \in s} I(x_i, c) \quad (3)$$

$$minR(S), R = \frac{1}{|s|^2}\sum_{x_i, x_j \in s} I(x_i; x_j) \quad (4)$$

3) Lasso: The Least Absolute Shrinkage and Selection Operator (Lasso) is a method that helps improve model interpretability and performance by performing feature selection and regularization. It is particularly effective when working with datasets that contain many features, as it can reduce overfitting by penalizing less relevant variables [19]. Lasso works by adding a penalty term to the loss function, specifically the L1 penalty, which forces some of the feature coefficients to become exactly zero [19]. The L1 penalty, controlled by a tuning parameter λ, directly influences how many coefficients are driven to zero, with larger values of λ resulting in more features being eliminated. This regularization technique ensures that the model is both efficient and less likely to overfit the data. The mathematical formulation of Lasso is represented in Eq. (5). Where the first term represents the residual sum of squares, and the second term is the L1 penalty.

$$\hat{\beta} = min\,\beta\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (5)$$

4) Radial basis function neural network: Radial Basis Function Neural Networks (RBFNN) are a specialized type of artificial neural network that utilize radial basis functions for activation. These networks are well-suited for tasks such as function approximation, classification, and regression, owing to

their capacity to model involved nonlinear relationships. RBFNN typically consist of three essential layers: the input layer, the hidden layer, and the output layer, with weights connecting each layer.

The input layer receives the source node, representing the independent variable, and links it to the surrounding network. The hidden layer then applies a nonlinear transformation, mapping the input space to a higher-dimensional hidden space. The output layer generates the final predicted result based on the transformed inputs from the hidden layer. In the hidden layer, each unit corresponds to a transfer function, often a Gaussian function. The radial basis function (RBF), which has a symmetric shape, acts as the transfer function in this case. The number of hidden units is directly related to the number of RBF employed in the network. The Gaussian RBF is mathematically expressed as shown in Eq. (6), meanwhile the output is computed by summing the weighted contributions of the RBF as shown in Eq. (7).

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

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

where w_j are the weights assigned to each radial basis function, C_j represents the centers of the RBF, and r is a scaling factor. In the output layer, a logistic (sigmoid) activation function is commonly used for binary classification. This function transforms the weighted sum of the hidden layer outputs into a probability between 0 and 1, and is defined as in Eq. (8). Thus, the final output of the RBFNN for binary classification is calculated using Eq. (9), where w_0 is the bias term.

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

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

D. Model Performances

The developed model's performance will be evaluated using six metrics: accuracy, sensitivity, specificity, precision, F1 score, and AUROC. To evaluate the classification performance of the developed model, a confusion matrix is first constructed. The confusion matrix, shown in Table III provides a detailed summary of the model's predictions by categorizing them into four groups: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Meanwhile, the total number of instances is the sum of all four categories: TP, TN, FP, and FN.

TABLE III. CONFUSION MATRIX

	Actual Positive (1)	**Actual Negative (0)**
Predicted Positive (1)	TP	FP
Predictive Negative (0)	FN	TN

Accuracy reflects the percentage of correct predictions and is calculated as shown in Eq. (10). Sensitivity, or the true positive rate, measures the proportion of actual positive cases correctly identified in Eq. (11), while specificity, or the true negative rate, quantifies the proportion of actual negatives accurately calculated using in Eq. (12).

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

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

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

Precision determines the percentage of correct positive predictions among all predicted positives in Eq. (13). The F1 score, a harmonic mean of precision and sensitivity in Eq. (14), assesses the model's ability to maintain balance between these metrics.

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

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

The Area Under the ROC Curve (AUROC) curve illustrates the relationship between sensitivity and the False Positive Rate, with AUROC measuring the model's ability to differentiate between classes. A higher AUROC value, nearing 1.0, signifies better classification performance. Collectively, high scores across these metrics indicate a reliable and effective model.

III. RESULTS AND DISCUSSION

This section discusses the results of the study. Table IV summarizes the descriptive statistics of the independent variables, revealing a substantial range in standard deviations, from 0.001 to 38.15. This wide variability reflects the diverse scales of the variables, necessitating data normalization. Based on Fig. 2, the histogram reveals PM2.5, PM10, and CO have higher variability, suggesting potential impact from external factors. The distribution of these variables over the four years may have been affected by the COVID-19 period, with changes in human activity significantly impacting pollution levels during this time.

To address this, min-max normalization was employed to transform all variables to a consistent scale, ensuring comparability across features. Min-Max normalization is one of the most commonly used normalization methods in various applications, including air quality datasets [20]. Additionally, Fig. 3 illustrates the distribution of the $PM2.5_{D+1}$ category, highlighting a significant class imbalance in the dataset, where one category is disproportionately represented. Such imbalances can bias predictive models, undermining their reliability and generalizability. To resolve this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, a proven method for addressing class imbalance by generating synthetic samples for underrepresented categories. By balancing the dataset, SMOTE enhances the model's ability to learn from all

categories effectively, thereby improving prediction accuracy and ensuring robust, reliable outcomes.

TABLE IV. DESCRIPTIVE STATISTICS OF BEFORE DATA PRE-PROCESSING

Variable	N	Mean	Median	Std. Dev.	Skewness
PM2.5	1825	21.236	18.648	12.157	3.653
PM10	1825	30.433	28.016	14.137	3.165
SO2	1825	0.001	0.001	0.001	1.377
NO2	1825	0.009	0.009	0.003	0.335
O3	1825	0.019	0.018	0.006	0.881
CO	1825	0.596	0.575	0.201	0.883
WD	1825	163.226	159.162	38.15	0.787
WS	1825	1.026	0.985	0.329	1.967
Humidity	1825	83.366	83.827	6.023	-2.678
Temperature	1825	27.123	27.138	1.14	-0.265

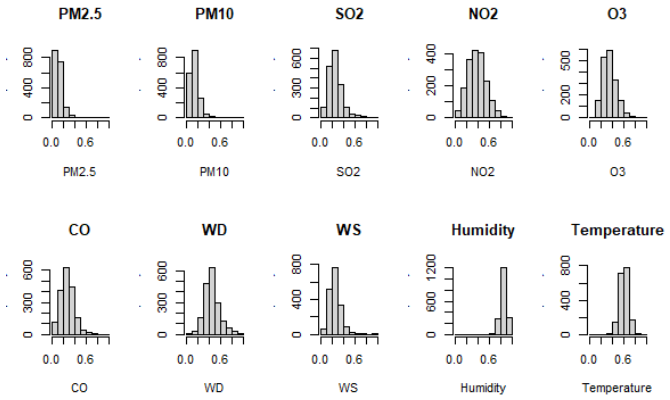

Fig. 2. $PM2.5_{Dt1}$ Distribution of variables.

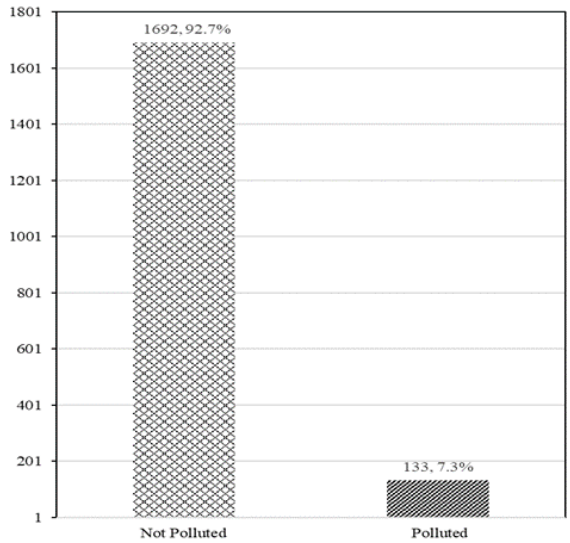

Fig. 3. $PM2.5_{Dt1}$ Distribution of (Before SMOTE).

Table V presents the descriptive statistics after data preprocessing. The application of min-max normalization successfully scaled the data, as evidenced by all mean and median values now falling within the standard range of 0 to 1. The data scaling ensures that the feature scales are consistent, which makes model training more successful. Furthermore, the skewness values have moved closer to zero, indicating a more symmetric and balanced distribution throughout the sample.

TABLE V. DESCRIPTIVE STATISTICS OF AFTER DATA PRE-PROCESSING

Variable	N	Mean	Median	Std. Dev
PM2.5	3288	0.172	0.136	0.143
PM10	3288	0.201	0.167	0.152
SO2	3288	0.278	0.255	0.137
NO2	3288	0.433	0.435	0.17
O3	3288	0.359	0.34	0.123
CO	3288	0.319	0.307	0.139
WD	3288	0.431	0.418	0.122
WS	3288	0.24	0.229	0.094
Humidity	3288	0.832	0.834	0.06
Temperature	3288	0.62	0.626	0.076

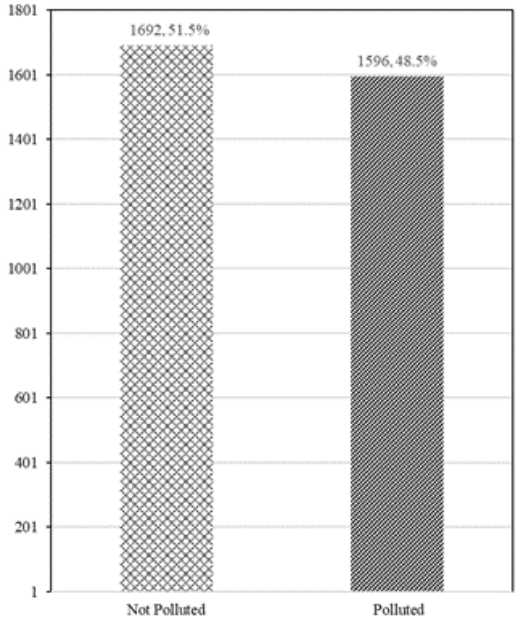

Fig. 4. $PM2.5_{Dt1}$ Distribution of (After SMOTE).

Fig. 4 illustrates the distribution of the $PM2.5_{D+1}$ category following the application of SMOTE for upsampling. This technique effectively addressed the class imbalance, resulting in a nearly equal representation of both groups: 1692 samples (51.5%) classified as "not polluted" and 1596 samples (48.5%) as "polluted." This balance dataset can enhance the reliability of the dataset for subsequent analyses and model predictions [21].

Table VI ranks independent variables according to three feature selection methods: Lasso, mRMR, and ReliefF, for predicting next-day PM2.5 levels. The rankings of independent variables by Lasso, mRMR, and ReliefF reveal both similarities and key differences. Lasso identifies relative humidity as the most important variable, emphasizing its critical role in PM2.5 prediction. In contrast, mRMR assigns the top 1 rank is SO2,

highlighting its unique contribution to the dataset. ReliefF, however, ranks NO2 as the most critical feature, suggesting its strong influence on pollutant dispersion and PM2.5 formation. These differences in top-ranking features underscore the variability in feature importance prioritization across the methods. Furthermore, the least important variables and exclusions also vary. CO is consistently ranked eighth by both Lasso and mRMR. However, ReliefF determine SO2 as the least important feature to predict PM2.5 in Banting [22].

Additionally, certain features were not selected by the three methods. Both particulate matter (PM10 and PM2.5) were excluded by mRMR method. Additionally, Lasso did not select SO2 and O3 as important feature. On other hand, ReliefF exclude both relative humidity and ambient temperature from the model.

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

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

Next, the three feature selection methods will be evaluated using the performance of the RBFNN model to predict $PM2.5_{D+1}$ in Banting is shown in Table VII. According to the table, the Lasso method yields higher values in accuracy (0.771), sensitivity (0.765), precision (0.778), F1 score (0.771), and AUROC (0.769), when compared to the mRMR and ReliefF methods. These findings contradict the results of [13], which concluded that ReliefF outperforms Lasso. The discrepancy arises from the contrasting characteristics of the study areas. Shah Alam, as an industrial area, is heavily influenced by traffic and industrial emissions, resulting in more consistent pollution sources throughout the year. In contrast, Banting, an agricultural region, experiences pollution patterns influenced by seasonal meteorological factors. For instance, according to a study in [23], during the northeast monsoon, precipitation in Banting have influence on air pollution levels. This difference in pollution patterns might explain why Lasso outperformed ReliefF in Banting, as ReliefF did not select two key meteorological factors, humidity and temperature, which are more significant in this region.

However, this study aligns with [23], which found that the Lasso algorithm enhances model performance more effectively than the ReliefF method. According to study [19], Lasso is known for its ability to handle multicollinearity in predictors, reducing the impact of correlated features, and enhancing model interpretability by retaining only significant variables. This explains why Lasso performs better than the ReliefF method. Moreover, the mRMR method also demonstrates better performance compared to ReliefF. This may be attributed to the high correlations among variables in the Banting air quality dataset.

TABLE VII. COMPARISON MODEL PERFORMANCE

Model	Lasso	mRMR	ReliefF
Accuracy	**0.771**	0.725	0.568
Sensitivity	**0.731**	0.623	0.544
Specificity	0.807	**0.819**	0.591
Precision	**0.778**	0.761	0.551
F1 Score	**0.754**	0.685	0.548
AUROC	**0.829**	0.777	0.608

Table VIII shows the confusion matrix provides a detailed breakdown of the Lasso-based RBFNN model's classification performance in predicting PM2.5 levels. In this case, the model correctly classified 231 polluted instances (TP) and 276 non-polluted instances (TN)**,** while misclassifying 66 non-polluted instances as polluted (FP) and 85 polluted instances as non-polluted (FN)**.** The total number of instances used for evaluation is 658, calculated as the sum of TP, TN, FP, and FN.

TABLE VIII. CONFUSION MATRIX

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	231	66
Predictive Negative (0)	85	276

IV. CONCLUSION

In conclusion, this study highlights the significant role of feature selection methods in enhancing the predictive performance of air quality models in Banting, Malaysia. Among the three methods evaluated, Lasso emerged as the most effective for predicting next-day PM2.5 levels, consistently outperforming both mRMR and ReliefF in key performance metrics such as accuracy, sensitivity, precision, F1 score, and AUROC. Thus, Thus, the top eight features to predict PM2.5 levels in Banting selected by Lasso method is relative humidity, PM2.5, wind direction, ambient temperature, PM10, NO2, wind speed, and O3. While previous studies have recognized the strengths of ReliefF in detecting relevant features, this research reinforces the advantages of Lasso, particularly in its ability to improve model performance by addressing feature redundancy and focusing on the most impactful variables.

The superior performance of Lasso can be attributed to its ability to handle multicollinearity, reduce the impact of correlated features, and enhance model interpretability by retaining only the most significant predictors. Given the success of Lasso, future research should further explore its application in different settings, including suburban and rural areas, to evaluate its generalizability across diverse environments, especially dataset that contains high correlation between variables.

This study uses data from 2018 to 2022, a period that includes the COVID-19 years. During this time, air quality is likely to have been positively affected due to lower traffic and congestion, which may have influenced the results. Future studies might compare the results by differentiating between the COVID-19 and non-COVID-19 years to better understand how these factors impact air quality predictions. The results from this study are specific to the Banting dataset and may not be directly applicable to other countries due to regional differences in air quality patterns. This study benefits researchers, policymakers, public health authorities, and technology developers by improving air quality prediction models. Policymakers can use these predictions to implement real-time air quality monitoring systems, establish dynamic traffic control measures to reduce vehicular emissions during high-pollution periods, and enforce stricter industrial regulations to limit pollutant discharge. Additionally, predictions can guide the development of targeted public health advisories, such as issuing alerts for vulnerable populations and adjusting outdoor activity recommendations during hazardous air quality events. This study also supports NGOs in advocating for better air quality regulations and raising public awareness. Moreover, future researchers can apply the methods explored in this study to other areas, such as water quality prediction or environmental monitoring in different ecosystems. By adapting the feature selection techniques to new domains, researchers can validate the approach's versatility and effectiveness in improving prediction models across various environmental factors.

ACKNOWLEDGMENT

The authors would like to acknowledge the Ministry of Higher Education (MOHE) for support this work under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2023/STG06/UITM/02/8). This research was financially supported by Institute of Postgraduate Studies Universiti Teknologi MARA (UiTM).

REFERENCES

[1] Ministry of Environment and Water Malaysia. (2021). Laporan Kualiti Alam Sekeliling 2021 [Environmental Quality Report 2021]. Department of Environment, Malaysia. https://enviro2.doe.gov.my/ekmc/wp-content/uploads/2022/10/Laporan-Kualiti-Alam-Sekeliling-2021.pdf

[2] Chang, Y., Jung, C., Chang, Y., Chuang, B., Chen, M., & Hwang, B. (2022). Prenatal and postnatal exposure to PM2.5 and the risk of tic disorders. Paediatric and Perinatal Epidemiology, 37(3), 191–200. https://doi.org/10.1111/ppe.12943

[3] Zaini, N., Ean, L. W., Ahmed, A. N., Malek, M. A., & Chow, M. F. (2022). PM2.5 forecasting for an urban area based on deep learning and decomposition method. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-21769-1

[4] Shafi, M. S. M., & Juahir, H. (2024). Forecasting Air Quality in Peninsular Malaysia: Unveiling the Power of Artificial Neural Networks.

[5] Deng, Y., Zhou, X., Shen, J., Xiao, G., Hong, H., Lin, H., ... & Liao, B. Q. (2021). New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water. Science of The Total Environment, 772, 145534.

[6] Nguyen, M. H., Nguyen, P. L., Nguyen, K., Le, V. A., Nguyen, T., & Ji, Y. (2021). PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and Encoder-Decoder Model. IEEE Access, 9, 57338–57350. https://doi.org/10.1109/access.2021.3072280

[7] Wu, Y., Liu, G., Li, Y., & Jiang, M. (2021). Filter - based feature ranking technique for target recognition by radar infrared combined sensors. IET Radar Sonar & Navigation, 16(1), 182 - 192. https://doi.org/10.1049/rsn2.12175

[8] Hikichi, S., Sugimoto, M., & Tomita, M. (2020). Correlation-centred variable selection of a gene expression signature to predict breast cancer metastasis. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-64870-z

[9] Chan, J. Y., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z., Lin, J., & Chen, Y. (2022). A Correlation-Embedded Attention Module to Mitigate Multicollinearity: an algorithmic trading application. Mathematics, 10(8), 1231. https://doi.org/10.3390/math10081231

[10] Huo, X., Su, H., Yang, P., Jia, C., Liu, Y., Wang, J., ... & Li, J. (2024). Research of Short-Term Wind Power Generation Forecasting Based on mRMR-PSO-LSTM Algorithm. Electronics, 13(13), 2469.

[11] Yan, Q., Wang, R., Dong, Y., Lv, X., Tang, X., Li, X., & Niu, Y. (2022). Logistic LASSO Regression for Dietary Intakes and Obesity: NHANES (2007-2016).

[12] Abdelwahed, N. M., El-Tawel, G. S., & Makhlouf, M. A. (2022). Effective hybrid feature selection using different bootstrap enhances cancers classification performance. BioData Mining, 15(1), 24.

[13] Arafin, S. K., Ul-Saufie, A. Z., Ghani, N. A. M., & Ibrahim, N. (2024). Feature Selection Methods Using RBFNN Based on Enhance Air Quality Prediction: Insights from Shah Alam. International Journal of Advanced Computer Science & Applications, 15(11).

[14] Ul-Saufie, A. Z., Hamzan, N. H., Zahari, Z., Shaziayani, W. N., Noor, N. M., Zainol, M. R. R. M. A., ... & Vizureanu, P. (2022). Improving air pollution prediction modelling using wrapper feature selection. Sustainability, 14(18), 11403.

[15] 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.

[16] Arafin, S. K., Ul-Saufie, A. Z., Ghani, N. A. M., & Ibrahim, N. (2024). A Two-Stage Feature Selection Method to Enhance Prediction of Daily PM2. 5 Concentration Air Pollution: 10.32526/ennrj/22/20240049. Environment and Natural Resources Journal, 22(6), 500-509.

[17] Desiani, A., Andriani, Y., Irmeilyana, I., Primartha, R., Arhami, M., Fitrianti, D., & Syafitri, H. N. (2023). The comparison of ReliefF and C.45 for feature selection on heart disease classification using backpropagation. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 17(2). https://doi.org/10.22146/ijccs.82948

[18] Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53, 23-69.

[19] Ibrahim, N. B. (2020). Variable selection methods for classification: application to metabolomics data. The University of Liverpool (United Kingdom).

[20] Umar, M. A., Chen, Z., Shuaib, K., & Liu, Y. (2024). Effects of feature selection and normalization on network intrusion detection. Data Science and Management. https://doi.org/10.1016/j.dsm.2024.08.001

[21] Koc, K., Ekmekcioğlu, Ö., & Gurgun, A. P. (2022). Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods. Engineering Construction & Architectural Management, 30(9), 4486–4517. https://doi.org/10.1108/ecam-04-2022-0305

[22] Rahman, M. N. A., Ismail, M. S., Wakid, S. A., Syazwan, W. M., Abd Aziz, N. A., Mustafa, M., ... & Yap, C. K. (2023). A Preliminary Checklist of Molluscs in the Kelanang Coast at Banting: An Observational Study. Journal ISSN, 2766, 2276.

[23] Hammad, M. S., Ghoneim, V. F., Mabrouk, M. S., & Al-Atabany, W. I. (2023). A hybrid deep learning approach for COVID-19 detection based on genomic image processing techniques. Scientific Reports, 13(1), 4003.