UNIVERSITI TEKNOLOGI MARA

A TWO STAGE FEATURE SELECTION METHOD USING ADJCORT AND RADIAL BASIS FUNCTION NEURAL NETWORK FOR ENHANCE CLASSIFICATION OF DAILY PM2.5 CONCENTRATION

SITI KHADIJAH BINTI ARAFIN

MSc

June 2026

UNIVERSITI TEKNOLOGI MARA

A TWO STAGE FEATURE SELECTION METHOD USING ADJCORT AND RADIAL BASIS FUNCTION NEURAL NETWORK FOR ENHANCE CLASSIFICATION OF DAILY PM2.5 CONCENTRATION

SITI KHADIJAH BINTI ARAFIN

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Statistics)

Faculty of Computer and Mathematical Sciences

June 2026

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 28th November 2025 to conduct the final examination of Siti Khadijah Binti Arafin on her Masters of Science thesis entitled "A Two Stage Feature Selection Method Using AdjcorT And Radial Basis Function Neural Network for Enhance Classification of Daily PM2.5 Concentration" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiners recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Norizan Mat Diah, PhD
Associate Professor
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
(Chairman)

Wan Fairos Wan Yaacob, PhD
Associate Professor
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
(Internal Examiner)

Shamshuritawati Sharif, PhD
Associate Professor
School of Quantitative Sciences
University Utara Malaysia
(External Examiner)

PROFESSOR DR HJH ZURAEDA IBRAHIM
Dean
Institute of Postgraduates Studies
Universiti Teknologi MARA

Date: 11 June 2026

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Postgraduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Siti Khadijah Binti Arafin
Student ID. No.	:	2023202692
Programme	:	Master of Science (Statistics) – CDCS753
Faculty	:	Computer and Mathematical Sciences
Thesis Title	:	A Two Stage Feature Selection Method Using AdjcorT and Radial Basis Function Neural Network for Enhance Classification of Daily PM2.5 Concentration
Signature of Student	:	……..
Date	:	June 2026

ABSTRACT

This study addresses the critical challenge of multicollinearity in air quality datasets by proposing a two-stage feature selection method, AdjcorT, integrated with the Radial Basis Function Neural Network (RBFNN) to improve the classification of daily PM2.5 concentrations. Multicollinearity, characterized by strong correlations among predictor variables, can distort model performance and obscure the true importance of features, resulting in less reliable classifications. The AdjcorT method explicitly accounts for both positive and negative correlations, effectively mitigating the adverse effects of multicollinearity during feature selection. The AdjcorT-RBFNN model was evaluated using datasets from Shah Alam and Banting. In Shah Alam, AdjcorT-RBFNN outperformed other models, including RBFNN, Lasso-RBFNN, mRMR-RBFNN, and ReliefF-RBFNN, while in Banting, Lasso-RBFNN showed better predictive performance. However, AdjcorT-RBFNN demonstrated competitive accuracy with significantly faster computational time. The study highlights the critical role of feature selection that considers high correlations among features to improve predictive accuracy. Key predictors for next-day PM2.5 in Shah Alam include NO2, PM2.5, PM10, CO, O3, wind speed, SO2, and wind direction. In Banting, significant predictors include humidity, PM2.5, wind direction, temperature, PM10, NO2, wind speed, and CO. The research underscores the importance of addressing multicollinearity, as both AdjcorT and Lasso consistently identified core predictors such as PM2.5, PM10, and NO2. Although Lasso-RBFNN achieved higher accuracy in Banting, AdjcorT-RBFNN was more computationally efficient while delivering similarly robust results. Furthermore, simulations demonstrate AdjcorT-RBFNN's effectiveness in managing highly correlated variables, especially under strong correlation conditions. This study contributes to the development of more accurate and computationally efficient models for air quality classification, emphasizing the need to select appropriate feature selection methods based on dataset characteristics, particularly when multicollinearity is present.

ACKNOWLEDGEMENT

First and foremost, I am profoundly grateful to Allah for granting me the strength, patience, and resilience to complete this journey. Without His divine guidance and blessings, this achievement would not have been possible.

I would like to extend my deepest appreciation to my supervisor, Dr. Nurain Ibrahim, for her steadfast support, insightful guidance, and continuous encouragement throughout the course of this research. Her patience and expertise have played a crucial role in shaping the direction and quality of this work. My sincere thanks also go to my co-supervisor, Prof. Madya Ts. Dr. Ahmad Zia Ul-Saufie Mohamad Japeri, for his valuable feedback, thoughtful suggestions, and ongoing support, all of which have been instrumental in improving this thesis.

My heartfelt thanks go to my beloved parents, Arafin Bin Jahar and
Nor, for their endless prayers, unconditional love, and unwavering belief in me. Your sacrifices and support have been the cornerstone of everything I have achieved.

I am also deeply thankful to my husband, Nik Ahmad Amaruddin Bin Nik Azman, for being my source of strength and comfort. Your emotional support, patience, and faith in me have helped carry me through the most difficult moments of this journey.

I would also like to take a moment to acknowledge myself for the perseverance, determination, and resilience throughout this journey. Balancing academic responsibilities alongside professional commitments and personal challenges was not always easy, but I remained committed to seeing this journey through. This experience has taught me the value of patience, growth, and believing in my own capabilities.

To my dear friends, thank you for your kindness, shared laughter, and consistent encouragement that helped me stay motivated along this journey.

This achievement is shared with all who stood beside me through every step. I am forever thankful. Alhamdulillah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Research Background

Malaysia has frequently experienced transboundary haze events, in which the amount of particulate matter in the air, exceptionally high and has a negative impact on both human health and the ecosystem. Hence, it is important to monitor the air quality to prevent the citizen expose to the harmful air. 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 & Kyung's (2022) research on its adverse effects on pulmonary diseases. Figure 1.1 shows the 68 of air quality monitoring stations in Malaysia to make sure the air quality of each location is in the range of good quality. Air quality index in Malaysia has five categories which are Good (0-50), Moderate (51-100), Unhealthy for Sensitive Groups (101-150), Unhealthy (151-200), Very Unhealthy (201- 300) and Hazardous (more than 300). The Malaysia Ambient Air Quality Guidelines (MAAQG) utilize Fine Particulate Matter (PM2.5) levels, along with other pollutants such as Particulate Matter (PM10), carbon monoxide (CO), sulphur dioxide (SO2), nitrogen dioxide (NO2), and ozone (O3), to assess and measure air quality across the country (Shaziayani et. al, 2021).

Figure 1.1 Air Quality Monitoring Station in Malaysia

PM2.5 refers to fine particulate matter with a diameter of 2.5 micrometers or less, capable of deeply penetrating the respiratory system and posing health risks. Recent research reveals that PM2.5 is primarily caused by anthropogenic and natural sources, with fossil fuel combustion accounting for 27.3% of global emissions, along with residential energy use, industrial processes, and agricultural practices (McDuffie et. al, 2021). Major pollutants which are PM10, SO2, NO2, CO and O3 were used as factors to predict PM2.5 concentrations (Wu et.al, 2023). They also used meteorological parameters such as relative humidity, barometric pressure, temperature and wind speed to build the predictive model. In addition, Ul Saufie et al. (2022) emphasize the need to prioritize PM2.5 prediction because it poses greater health risks compared to PM10. However, the Department of Environment (DOE) only began incorporating PM2.5 into the Air Pollutant Index (API) in 2018. As a result, fewer studies have focused on PM2.5 despite its harmful effects. As a result, there is a limited number of studies focusing on PM2.5 in Malaysia despite its significant health impact.

To address the growing need for accurate PM2.5 classification and to better capture the complex relationships between multiple pollutants and meteorological factors, recent studies have increasingly adopted machine learning approaches. Machine learning techniques are being increasingly utilized in air quality research to study the correlation between air pollutant levels, emissions, and meteorological changes over time (Gao et al., 2024; Kulkarni et al., 2022; Nouri, Lak & Valizadeh, 2021). Kulkarni et.al (2022) conduct a study to compare statistical model and machine learning model in PM2.5 prediction, evaluating several statistical approaches alongside machine learning models such as Random Forest, Deep Learning, and eXtreme Gradient Boosting. Similarly, Gao et al. (2024) used a variety of machine learning techniques to predict daily average PM2.5 levels, such as decision trees, random forests (RF), support vector machines (SVM), support vector regression (SVR), k-nearest neighbors, neural networks, and Gaussian process regression. Among these approaches, artificial neural network–based models have been widely applied due to their strong capability in capturing complex nonlinear relationships inherent in air quality data (Goudarzi et al., 2021). In addition, Nouri, Lak & Valizadeh (2021) combined Principal Component Analysis (PCA) with neural network models, including multi-layer perceptron (MLP) and radial basis function neural networks (RBFNN), to predict PM2.5 concentrations in Urmia, Iran.

However, many of the existing approaches for PM2.5 prediction have methodological limitations. Traditional statistical models, such as linear regression and generalized additive models, rely on strict assumptions and may not sufficiently capture the non-linear and complex interactions between variables (Kulkarni et al., 2022). While some machine learning models, such as deep neural networks, have shown good predictive performance, they often do not explicitly address multicollinearity among input variables (Radočaj et al., 2023). However, according to Chan et al. (2022), ignoring multicollinearity can reduce model interpretability and may compromise robustness when models are applied to different datasets, because correlated predictors can produce inconsistent or misleading importance measures and unstable predictive behaviour.

These limitations are particularly important because air quality data are often high-dimensional and subject to multicollinearity, which complicates modeling and increases the risk of overfitting. Feature selection is therefore essential to reduce dimensionality, eliminate redundant or irrelevant variables, and enhance model performance. In air quality prediction, effective feature selection helps mitigate the effects of multicollinearity and ensures that only the most informative predictors are included. In addition, some studies do not incorporate feature selection before model training, which can lead to the inclusion of irrelevant or redundant variables and increase the risk of overfitting (Ding et al., 2025). Moreover, Cheng & Tsai (2022) highlight the importance of variable selection in identifying key environmental factors that impact air quality. By selecting these key variables, researchers can improve the model's capacity to capture the intricate relationships between different parameters that influence air quality. Thus, these limitations highlight the need for a targeted feature selection strategy that accounts for inter-variable correlations while enhancing both predictive accuracy and generalizability.

In this context, artificial neural network (ANN) based models are well suited for PM2.5 classification due to their strong capability in capturing complex nonlinear relationships between air pollutants and meteorological variables. Among ANN architectures, the radial basis function neural network (RBFNN) has been widely applied in air quality studies because of its effective nonlinear approximation ability, fast convergence, and robustness when handling noisy and high-dimensional environmental data (Goudarzi et al., 2021; Nouri, Lak & Valizadeh, 2021). These

characteristics make RBFNN a suitable predictive model for PM2.5, particularly when combined with an appropriate feature selection strategy to address multicollinearity.

Feature selection methods can be grouped into correlation-based approaches such as Lasso, mRMR and AdjcorT, and non-correlation-based approaches like ReliefF. However, many of these commonly used methods, particularly non-correlation-based ones, evaluate variables independently and may overlook the strong correlations that typically exist among pollutants and meteorological factors. For instance, ReliefF, which is commonly employed as a filter-based feature selection method, can identify relevant features but does not remove redundancy among correlated predictors, limiting its ability to address multicollinearity (Pudjihartono et al., 2022). As a result, such methods may fail to adequately address multicollinearity in complex environmental datasets. Moreover, some correlation-based feature selection methods, such as mRMR, do not explicitly consider negative correlations between variables, which may lead to the overlooking of important predictors with inverse relationships. In contrast, AdjcorT addresses multicollinearity by ranking factors using t-scores while allowing both positive and negative correlations among variables, thereby reducing redundancy and overfitting.

Therefore, to address these limitations, this research proposes an integrated two-stage feature selection method that combines the Adjusted Correlation Sharing t-test (AdjcorT) and the Radial Basis Function Neural Network (RBFNN) model. In the first stage, AdjcorT is employed to select and rank significant predictors while accounting for inter-variable correlations, and in the second stage, the selected features are used as inputs to the RBFNN for PM2.5 classification. The proposed approach ranks factors using t-scores while allowing both positive and negative correlations among variables, with the aim of mitigating multicollinearity and overfitting issues. This method helps to identify significant underlying factors for future air quality classification. The performance of the proposed models is evaluated using accuracy, sensitivity, specificity, precision, F1 score, and the area under the ROC curve (AUROC), based on both simulated data and air quality data obtained from the Department of Environment (DOE), Malaysia.

1.2 Problem Statement

Nowadays, air pollution has become a major concern worldwide. Long-term exposure to hazardous air pollution can cause lung cancer, cardiovascular problems, and various respiratory diseases in humans. Fine particulate matter (PM2.5) is a particularly dangerous pollutant due to its small size, which allows it to penetrate deeply into the lungs. Long-term exposure to PM2.5, even at low concentrations, has been shown in previous studies to increase the risk of lung cancer. Hence, accurate classification of PM2.5 levels is essential for air quality assessment, effective pollution control, and the protection of public health.

However, selecting the right input features to predict PM2.5 concentrations is one of the critical challenges due to many factors that affect air quality. The presence of high-dimensional input data not only increases model complexity but also affects the robustness and stability of predictive performance if irrelevant or redundant variables are included. Moreover, high-dimensional datasets often suffer from multicollinearity, a condition where two or more predictor variables are highly correlated with each other. This typically occurs because pollutant and meteorological variables such as PM10, NO2, CO, O3, temperature, and humidity tend to vary together due to shared emission sources and atmospheric conditions. Multicollinearity can distort the estimation of model parameters, making it difficult to determine the individual effect of each variable. As a result, it can lead to unstable models, inflated variance in predictions, and a decline in the overall accuracy and interpretability of the predictive model.

Hence, feature selection is a crucial step in air quality modelling to reduce dimensionality, remove redundant variables, and mitigate multicollinearity. However, many feature selection methods applied in air quality studies focus primarily on identifying variables that are strongly associated with the target variable, without explicitly accounting for the correlation structure among predictors. When predictor variables are highly correlated, such approaches may retain redundant features or fail to distinguish between shared and unique contributions of correlated pollutants. This limitation is particularly important in air quality data, where pollutant and meteorological variables often vary together due to common emission sources and atmospheric processes. Therefore, feature selection methods that explicitly consider inter-variable correlations are required to effectively address multicollinearity and improve model stability and interpretability. Nevertheless, relying on a single feature

selection method may not be sufficient to ensure model robustness, as different methods capture different aspects of feature relevance. Additionally, Lasso and mRMR are popular correlation-based feature selection methods used in various fields including air quality. However, these methods may eliminate some important features, as they tend to select only one feature from groups of highly correlated variables. This limitation is particularly critical in air quality applications, where correlated pollutants may jointly influence PM2.5 concentrations. To address this issue, the Adjusted Correlation Sharing t-test (AdjcorT) has been proposed as a feature selection method that ranks variables using t-scores while explicitly considering the magnitude of correlations among predictors. By considering both correlated and uncorrelated features during the ranking process, AdjcorT reduces the risk of discarding informative variables solely due to high correlation. However, none of previous studies has apply AdjcorT method to air quality data, particularly in the Malaysian context.

In addition, air quality data is characterized by big data elements that are nonlinear, nonparametric, complex, and chaotic, making it difficult to model and predict its future movement. As a result, machine learning technique was widely used in air quality prediction due to its ability to handle complex, non-linear relationships and large datasets efficiently. Machine learning models, such Artificial Neural Networks (ANN), have shown strong performance in various fields, including air quality prediction by learning patterns from historical data. Within the ANN family, the Radial Basis Function Neural Network (RBFNN) is recognised for its relatively simple network architecture, fast training convergence, and strong capability in approximating nonlinear functions, making it suitable for environmental datasets with limited sample sizes. Although RBFNN has been applied globally in air quality prediction, its application in Malaysian air quality studies remains limited, and its integration with correlation-based feature selection strategies has not been sufficiently explored. However, a major limitation of RBFNN is its sensitivity to correlated input variables, as multicollinearity can negatively affect network training and feature importance interpretation.

Therefore, this study aims to enhance the RBFNN model by integrating AdjcorT as an initial feature ranking stage to identify significant predictors while explicitly accounting for the correlation structure among variables. By combining a correlation-based feature selection method with a nonlinear learning model, the proposed framework seeks to improve model stability, robustness, and interpretability in PM2.5

classification. Besides, the existing noncorrelation-based and correlation-based feature selection remains questionable, as they do not explicitly consider the direction (positive or negative) of correlations between variables. Therefore, the proposed AdjcorT–RBFNN framework addresses this limitation by incorporating both the strength and direction of correlations during feature ranking, thereby reducing prediction errors and improving the identification of key factors contributing to PM2.5 concentrations.

1.3 Research Questions

The research questions of interest based on background of study and problem statements are as follows:

1. What are the significant factors that significantly contribute to PM2.5 concentration classification of the next day?
2. Which is the best feature selection method to enhance air quality classification?
3. How to verify the performance of the proposed model and existing model on air quality classification?

1.4 Research Objectives

This study proposes one robust model which is two-stage feature selection method of AdjcorT and RBFNN model in dealing big data and correlated data in air quality data. The research objectives include the following:

1. To determine the significant factors that contribute to PM2.5 concentration classification of the next day.
2. To compare the performance of a non-correlation based model and a two-stage correlation based feature selection model for air quality classification using real data.
3. To verify the performance of existing model with the proposed model via simulation data.

1.5 Significance of Study

The proposed of two-stages feature selection method, AdjcorT and RBFNN, represents a significant advancement of feature selection in mitigating multicollinearity issues within air quality data, consequently enhancing the accuracy of PM2.5 concentration classifications for the next day. The application of two-stages feature selection method can improves classifications of PM2.5 concentrations' predictive accuracy and offers a strong framework for further study into air quality modelling.

Furthermore, this study aims to improves the accuracy of PM2.5 concentration classifications, which is advantageous to governments, environmental agencies, and policymakers. The determination the significant factors that contribute to PM2.5 concentration can offer a good information to the government and policy frameworks to take early precautions to sustains the air quality in this rapid urban development. By identifying the key factors influencing PM2.5 concentrations, this method enables targeted policy interventions, such as stricter emissions controls or urban planning regulations aimed at reducing pollution in high-risk areas. Predicting poor air quality in advance also can help government to take quick action to protect risky group such as issuing alerts or advising people to limit outdoor activities when the air quality is particularly bad.

For the community, especially in rapidly urbanizing area, better classifications of air pollution levels can improve public awareness and health outcomes. With accurate data on air quality, communities can make informed decisions about daily activities and health precautions. Additionally, schools, hospitals, and other public institutions can use air quality forecasts to safeguard the well-being of children, the elderly, and those with pre-existing respiratory conditions.

In the industrial sector, this method provides valuable insights into the environmental impact of industrial activities. Industries can use the predictive models to optimize their operational schedules, reducing emissions during periods of high pollution or adjusting production processes to comply with environmental standards. Additionally, this aligns with the government's emphasis on encouraging industries to follow to environmental, social, and governance (ESG) guidelines.

1.6 Scope and Limitations

The scope of this study includes the analysis and classification of air quality in the region of Shah Alam and Banting, Malaysia. This is because Shah Alam, the state capital of Selangor is an urban area which are significant contributors to air pollution due to increased levels of industrialization, human activity, and vehicular traffic. Moreover, Banting is a suburban area focus on agricultural as main activities. The agricultural practices in Banting, including the use of fertilizers and pesticides, can contribute to varying levels of particulate matter and other pollutants. Thus, the findings and results of this study are limited to this specific area.

In this study, PM2.5 concentration is the dependent variable used for classification purposes, where air quality is categorized as either polluted or not polluted. The independent variables under study include a combination of air pollutant measurements and meteorological parameters. These consist of PM10, PM2.5, NO2, SO2, CO, and O3, as well as temperature, relative humidity, wind speed, and wind direction. These variables were selected based on their known influence on PM2.5 levels, as supported by past studies (Liu et al., 2024; Ramli et al., 2023). The inclusion of these predictors allows the model to capture both pollutant interactions and environmental effects contributing to PM2.5 variation.

The study is further limited by data availability because it only uses air quality measurements that were taken from 2018 to 2022. This period corresponds with the Department of Environmental Malaysia's start monitoring the PM2.5 concentration on mid of year 2017. The feature selection method used in this study, AdjcorT have not been explore to other areas than metabolomics data especially in air quality, hence the literature reviews on this method has limited research. Another limitation is the lack of previous studies considering multicollinearity issues in predictive modelling within the context of air quality data.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

The previous chapter was explained and discussed regarding the background of the study, problem statement, research question, objective, hypothesis and scope and limitation of this study. Thus, relevant empirical studies would be discussed in this chapter to support this research paper by referring to the information stated above. Section 2.2 discusses air quality in general, emphasizing its importance for environmental and public health and explaining key pollutants that impact it. Section 2.3 focuses specifically on Particulate Matter 2.5 (PM2.5), describing its sources, characteristics, and harmful effects on human health and the environment. Subsequently, section 2.4 identifies the factors contributing to PM2.5 concentrations, including meteorological variables and other air pollutants, which are crucial for accurate classification.

Section 2.5 explores the concept of multicollinearity, highlighting its challenges in modeling air quality data and the need to address it to ensure reliable classifications. Section 2.6 delves into feature selection methods, such as ReliefF, mRMR, Lasso, and AdjcorT, explaining their role in selecting significant predictors and improving model performance. Next, section 2.7 reviews classification methods like Artificial Neural Networks (ANN) and Radial Basis Function Neural Networks (RBFNN), discussing their advantages and applications in PM2.5 classification. Section 2.8 outlines the metrics used to evaluate model performance, including accuracy, sensitivity, specificity, precision, F1 score, and AUROC, emphasizing their importance in validating the models. Finally, section 2.9 concludes the chapter by summarizing the key points discussed and their relevance to the study's objectives, setting the stage for the methodology presented in the subsequent chapter.

2.2 Air Quality

Air quality refers to the level of contamination and pollutants in the air that humans breathe, including a variety of contaminants such as particulate pollution, ozone, nitrogen dioxide, sulphur dioxide, and carbon monoxide. Poor air quality can have detrimental impacts on human health, including respiratory illnesses, cardiovascular diseases, and potentially even cancer. The inhalation of air pollutants like ozone and fine airborne particles is associated with increased hospital admissions and, tragically, even fatalities (Usmani et al., 2020). Additionally, poor air quality can harm the environment, affect plants and animals, and contribute to climate change.

Gibergans-Báguena et al. (2020) emphasize the significance of the Air Quality Index (AQI) as a critical tool for assessing air quality. Numerous factors affect air quality, including emissions from industrial and vehicular sources, agricultural activities, natural events such as wildfires and dust storms, and meteorological conditions. Air quality indices are frequently employed by governments and organizations to inform the public of potential health hazards associated with outdoor activities or exposure to polluted air.

As a crucial environmental issue, poor air quality poses serious health risks and contributes to climate change. Accurate air quality classification is essential for providing valuable information to governments for air pollution control and management. Previous research has extensively explored air quality prediction using various methods and target variables such as the air quality index, PM10, PM2.5, and carbon monoxide (CO). For instance, in Kolkata, India, Imam et al. (2024) divided the projected AQI into six classes to forecast air quality categories. Similarly, Cheng and Tsai (2022) employed four classifiers to categorize air quality based on datasets from Fengshan and Mailiao, Taiwan. Moreover, Saminathan & Malathy, (2023), Gao et.al, (2022) and Yang et.al, (2022) had conducted their studies on air quality prediction in China, specifically in Aotizhongxin to Wanshouxigong, Wuhan and Beijing respectively. Meanwhile, a study on classification of carbon monoxide for air pollution monitoring in Igoumenitsa, Greece was conducted by Spyrou et.al, (2023). There are also have previous studies on air quality prediction in Malaysia. Ramli et.al, (2023) conducted a study to predict PM10 concentration in peninsular Malaysia using air quality datasets from 9 states in peninsular Malaysia which are Perlis, Pulau Pinang, Selangor, Negeri Sembilan, Johor, Terengganu, Kelantan and Pahang. While a

study by Chinatamby & Jewaratnam (2023) used datasets from 13 air quality monitoring in Malaysia to compare PM2.5 concentrations at industrial areas.

However, predicting air quality movements is challenging due to the complexity of big data elements, which include numerous internal and external factors. For example, pollution patterns fluctuate between weekdays and weekends due to differences in traffic and industrial emissions (Hassan et al., 2022). Advanced modelling technique that take into account temporal changes are needed due to this variability, making air quality forecast more challenging. Furthermore, meteorological elements must be included in predictive models. Studies show that including variables like temperature, humidity, and wind speed improves air quality forecasts (Wattimena et al., 2022). Nevertheless, accurately representing the interactions among these variables and understanding their combined effects on air quality remains a complex and non-linear challenge (Mokhtari et al., 2021). The need for real-time data and the capability to adapt to rapidly changing conditions further complicates the predictive landscape, particularly during pollution events where immediate responses are essential (Mokhtari et al., 2021). Thus, this complexity poses a significant challenge to researchers (Sokhi et al., 2021). According to Sokhi et al. (2021), machine learning models are particularly effective in air quality prediction, as they can process the large volumes of data required to account for multiple factors influencing air quality, including traffic patterns, industrial output, and weather conditions. Hence, accurate predictive models necessitate advanced techniques and a deep understanding of air quality factors.

2.3 Particulate Matter 2.5 (PM2.5)

Particulate matter 2.5 as known as PM2.5 which is a fine particulate matter with a diameter of 2.5 micrometers or less. These tiny particles are from various sources, including vehicle emissions, industrial processes, burning of fossil fuels, and natural occurrences such as wildfires and volcanic eruptions. Saha et al. (2023) indicate that emissions from automobiles significantly contribute to overall PM2. 5 levels in their research evaluating the impact of mobile sources on air quality. Furthermore, factory and power plant emissions, including combustion and manufacturing operations, make a major contribution to urban air pollution by increasing PM2.5 emissions (Xu et al. 2020). The combustion of fossil fuels for energy production is another major source of

PM2.5. This includes emissions from coal-fired power plants and other industrial activities that rely on fossil fuels (Amann et al., 2020). Also, wildfires are a major natural source of PM2.5, especially in places where they happen often. Studies have shown that smoke from wildfires raises PM2.5 levels significantly, which harms the quality of the air in large places (Liang et al., 2021). For instance, Liang et al. (2021) conclude that wildfire smoke has a considerable impact on interior air quality, contributing to higher PM2.5 levels. Besides, PM2.5 can also form through secondary processes in the atmosphere. For example, chemical reactions involving gases such as sulfur dioxide (SO2) and nitrogen oxides (NOx) can lead to the formation of secondary particulate matter, which contributes to overall PM2.5 levels (Xu et al., 2020).

The range of PM2.5 concentrations to considered good and bad in Malaysia align with the guidelines set by the World Health Organization (WHO) and the Malaysian Ambient Air Quality Guidelines. Generally, an annual mean of 10 μg/m³ is considered good air quality, while a 24-hour mean not exceeding 25 μg/m³ is also deemed acceptable. PM2.5 is a major air pollutant that significantly affect public health. Although there are many pollutants that cause air pollution, Jebamalar & Kumar (2019) suggests that PM 2.5 is the main one.

To determine the impact of PM2.5 on the AQI value, this study refer to the study by (Aamir et.al, 2021). The research conducted a regression analysis and found a strong relationship between AQI and PM2.5 with $r2$ value equal to 0.93. This indicates PM2.5 have a substantial impact on the AQI value. The high correlation coefficients suggest that changes in PM2.5 levels are closely associated with variations in the AQI value. This evidence supports the notion that PM2.5 plays a major role in influencing the AQI value, highlighting the importance of monitoring and controlling PM2.5 levels to maintain air quality standards.

Additionally, Fan et. al, (2022) conduct a study on a spatial correlation between PM2.5 exposure and health outcomes, such as the incidence of lung cancer. The study suggesting a significant association between PM2.5 exposure and adverse health effects. Research has shown that long-term exposure to PM2.5 is associated with increased susceptibility to cardiovascular diseases, leading to higher incidence and mortality rates in regions (Jalali et.al, 2021). Furthermore, it has been discovered that PM2.5 exposure has a negative impact on the outcomes of births, with prenatal exposure linked to lower birth weight and other negative effects on fetal development (Balakrishnan et al., 2022). In fact, WHO has indicated that PM2.5 is responsible for

approximately 4.2 million premature deaths worldwide each year due to its role in exacerbating respiratory and cardiovascular diseases (Amann et al., 2020).

Hence, accurate classifications of PM2.5 levels are essential for effective environmental management. People can have early awareness for them to take preventive steps such as limiting traffic on specific day if they know when and where PM2.5 levels are most likely to occur. This is particularly important in urban areas where PM2.5 can accumulate due to local emissions and meteorological conditions (Rahardja et al., 2022). According to the studies by Rahardja et al. (2022), forecasting PM2.5 can help in assessing the effectiveness of air quality regulations and policies over time. However, a meta-analysis showed that the majority of research on long-term PM2.5 exposure and its health effects has been focused in North America and Europe, with only a limited number of studies taking place in other areas, such as Southeast Asia (Jalali et al., 2021). This indicates that although Malaysia is progressing in PM2.5 research, there is still a necessity for more thorough studies to understand more the effects and dynamics of PM2.5 within the Malaysian context.

In conclusion, PM2.5 is a harmful air pollutant with a diameter of less than 2.5 micrometers, known to have detrimental effects on human health, including increased mortality rates and the incidence of diseases like lung cancer. Understanding the factors that contribute to rising PM2.5 is crucial to predict as it has great impact on public health and the environment.

2.4 Factors Contributed to PM2.5 Concentration Classification

This study focuses on next-day air quality classification using PM2.5 concentration breakpoints corresponding to Air Quality Index categories, which is critical for air quality management and public health protection. Accurate PM2.5 classification requires an understanding of the key factors influencing its variability, as PM2.5 concentration is governed by complex interactions between pollutant emissions, meteorological conditions, and temporal persistence. Previous studies have therefore employed a wide range of pollutant-related and meteorological variables to improve PM2.5 prediction and classification performance. Table 2.1 summarizes previous studies on PM2.5 prediction and classification, including study location, variables used, and outcomes.

Table 2.1
Summary of Related Studies on PM2.5 Prediction Factors

Study	Location	Variables Used	Model / Focus	Outcome
Mohammadi et al. (2024)	Iran (Isfahan)	Wind direction, wind speed, temperature, humidity	PM2.5 prediction	Meteorological data useful for forecasting
Saminathan & Malathy (2023)	China	PM2.5, PM10, SO2, NO2, O3, CO, wind direction, ambient temperature	PM2.5 Classification	Effective classification using multiple pollutant and meteorological inputs
Eşsiz et al. (2023)	Turkey (Zonguldak)	COVID-19, pollutant data	Firefly-based feature selection	COVID-19 variable not important in air pollution prediction
Ramli et al. (2023)	Malaysia (Shah Alam, Klang)	PM10, CO, NO2, NO, SO2, humidity, temperature, wind speed, wind direction (cos/sin)	Feature selection for prediction	Identified significant pollutant and meteorological predictors
Yang et al. (2022)	China	Hourly PM2.5, PM10, SO2, temperature, humidity, pressure	AQI prediction	Meteorological factors improve prediction accuracy
Gao et al. (2022)	China	PM2.5, lockdown-related features	Spatiotemporal model	Lockdown policies reduced PM2.5 concentrations
Wang et al. (2022)	Not specified	SO2, NO2, O3, CO	AQI prediction	Pollutants are crucial predictors for AQI
Zaini et al. (2022)	Malaysia (Batu Muda, Cheras)	PM10, PM2.5, SO2, NO2, O3, CO	Hybrid Ensemble Empirical Mode Decomposition (EEMD) + Long short-term memory (LSTM)	Accurate short-term PM2.5 forecasts
Zaman et al. (2021)	Malaysia (nationwide)	AOD (Himawari-8), NO2, SO2, CO, O3, temperature, humidity, wind speed and direction	Satellite + ground data fusion	Satellite data enhanced PM2.5 estimation
Zhou et al. (2020)	China (Guilin)	Temperature, pressure, humidity, boundary layer height, wind speed, wind direction	Meteorological influence analysis	Meteorological factors strongly influence PM2.5 variations

Based on the studies summarised in Table 2.1, pollutant-related variables are consistently identified as important contributors to PM2.5 concentration. Particulate matter such as PM10 is frequently reported to have a strong association with PM2.5, as both particulate fractions often originate from similar emission sources including vehicular traffic, industrial activities, and biomass burning (Saminathan and Malathy, 2023; Ramli et al., 2023; Zaini et al., 2022). Several studies conducted in urban environments highlight that PM10 contributes substantially to PM2.5 variability, reflecting shared sources and atmospheric processes (Ramli et al., 2023). In addition, gaseous pollutants such as SO2 and NO2 are commonly linked to secondary aerosol formation through atmospheric chemical reactions, thereby indirectly increasing PM2.5 concentration (Wang et al., 2022; Yang et al., 2022). Carbon monoxide is often used as an indicator of combustion-related emissions and has been shown to exhibit strong associations with PM2.5 levels in densely populated areas (Saminathan and Malathy, 2023; Zaini et al., 2022). Although O3 is not directly emitted, it plays a role in photochemical processes that influence secondary particulate formation and subsequently affect PM2.5 concentration (Yang et al., 2022).

Meteorological conditions are also widely recognised as key factors influencing PM2.5 concentration by regulating pollutant dispersion, accumulation, and transformation. Relative humidity has been reported to enhance the hygroscopic growth of fine particles, leading to higher PM2.5 mass concentration, particularly under stagnant atmospheric conditions (Yang et al., 2022; Zhou et al., 2020). Temperature affects atmospheric stability and chemical reaction rates, which in turn influence secondary aerosol formation (Zhou et al., 2020). Wind speed and wind direction determine the transport and dispersion of air pollutants, where low wind speeds tend to favour pollutant accumulation, while certain wind directions may transport pollutants from surrounding urban or industrial regions into the study area (Mohammadi et al., 2024; Ramli et al., 2023).

A comparison across studies suggests that models incorporating both pollutant-related and meteorological variables generally achieve better predictive performance than those relying on a limited set of inputs (Yang et al., 2022; Zaini et al., 2022). However, the relative importance of individual variables varies across regions due to differences in emission sources, climatic conditions, and geographical characteristics. For example, meteorological variables tend to play a stronger role in regions with

complex atmospheric dynamics, whereas pollutant variables are often more dominant in highly urbanised and industrialised areas (Zhou et al., 2020; Ramli et al., 2023).

In addition to external predictors, PM2.5 concentration itself is frequently included in short-term air quality models to capture temporal dependence. Previous studies have demonstrated that PM2.5 exhibits strong persistence over time, meaning that current or previous-day PM2.5 levels provide valuable information for classifying air quality on the following day (Zaini et al., 2022; Gao et al., 2022). This temporal continuity reflects the cumulative effects of emission patterns and atmospheric processes that cannot be fully explained by other variables alone. Consequently, incorporating lagged PM2.5 values has been shown to improve next-day air quality classification performance (Gao et al., 2022).

Overall, the existing literature indicates that PM2.5 concentration is influenced by a combination of pollutant-related and meteorological factors, with their relative importance varying according to local environmental conditions. Guided by these findings, this study considers pollutant-related variables including PM10, SO2, NO2, CO, and O3, together with meteorological factors such as relative humidity, temperature, wind speed, and wind direction, to classify next-day air quality categories based on PM2.5 concentration breakpoints.

2.5 Multicollinearity

Multicollinearity is a common issue in air quality data that poses significant challenges for predictive modelling. It occurs when two or more predictor variables are highly correlated with each other, making it difficult to isolate the individual effect of each variable. This is especially important in air quality studies, where pollutants such as PM2.5, PM10, NO2, and CO often come from common sources or are influenced by the same atmospheric conditions. For example, Kao et al. (2023) found a strong positive correlation between CO and NO2, with a correlation coefficient of 0.904 and a VIF value of 11. They also reported a high correlation between PM2.5 and PM10 ($r = 0.908$). Similarly, Su et al. (2021) found a correlation coefficient of 0.99 between PM2.5 and PM10.

Researchers often use the Pearson correlation coefficient, Spearman's rank correlation, and the variance inflation factor (VIF) to identify multicollinearity. VIF quantifies how much the variance of a regression coefficient is inflated due to

multicollinearity. However, VIF measures the combined collinearity of one variable with all others rather than the pairwise relationship between specific variables. Spearman's correlation is sometimes preferred over Pearson and VIF, particularly because it is more robust to outliers and non-linear relationships. By using ranks instead of raw values, Spearman's method provides more reliable correlation estimates when data are not normally distributed (Wu et al., 2024).

Multicollinearity hinders the identification of significant predictors within a model. The variance of coefficient estimates may increase, resulting in unstable results and inflated standard errors. This inflation can result in Type II errors, where important predictors are incorrectly classified as insignificant (Mamouei et al., 2022; Kumar et al., 2024). When predictor variables are highly correlated, their individual effects on the response variable become difficult to distinguish, which weakens the model's interpretability and prediction accuracy (Cheng & Tsai, 2022).

While some studies have not specifically addressed multicollinearity, they do recognise the importance of correlation among variables. For example, Cheng & Tsai (2022) applied correlation-based feature selection with Pearson correlation coefficients in their study on air quality prediction. In addition, Sethi and Mittal (2021) created a Correlation-based Adaptive LASSO Regression method for adjusting weights based on correlations. Banga et al. (2021) used an ANOVA F-value test to evaluate the relationship between features and the target variable when predicting PM2.5 concentrations in smart cities in China.

Moreover, multicollinearity can lead to biased estimates in predictive models like logistic regression, thereby reducing their reliability (Guan & Fu, 2022). This distortion becomes more severe with stronger correlations among variables (Yoo & Cho, 2019). Ramli et al. (2023) noted multicollinearity in air quality data from 2009 to 2018 for Klang and Shah Alam, Malaysia, but did not address the issue in their modelling process. Therefore, this study aims to highlight the issue of multicollinearity in predicting air quality in Malaysia. It is crucial to identify important features that contribute to PM2.5 concentration while mitigating multicollinearity to improve classification performance.

2.6 Existing Feature Selection Method

Feature selection is an essential process in machine learning and data mining, aimed at identifying the most relevant features in a dataset to enhance model performance and interpretability. By reducing dimensionality, feature selection helps mitigate the curse of dimensionality, prevent overfitting, and enhance model generalisation capability (Hosni et al., 2021). Feature selection methods are typically categorized into three main groups: filter, wrapper, and embedded techniques. These categories differ in how features are evaluated and selected, particularly in terms of their dependency on learning algorithms and their ability to handle feature interactions and correlations.

Air quality datasets are typically characterised by high dimensionality and strong correlations among pollutant and meteorological variables. Such correlation structures can adversely affect model stability, predictive accuracy, and interpretability if not properly addressed. Consequently, the treatment of feature correlation and redundancy becomes a central consideration when selecting an appropriate feature selection strategy for air quality classification tasks. To address these challenges, this study focuses on four representative feature selection methods: ReliefF, mRMR, AdjcorT, and Lasso. These methods were selected to cover different feature selection concepts, with ReliefF representing a non-correlation-based filter, mRMR and AdjcorT representing correlation-based filters, and Lasso representing a correlation-based embedded method. This selection allows for a direct comparison of how different feature selection strategies handle feature relevance, redundancy, and correlation structure in the context of PM2.5-based air quality classification. Before discussing these methods within their respective methodological categories, it is necessary to first introduce correlation-based feature selection as a distinct conceptual approach.

2.6.1 Correlation-Based Feature Selection Methods

Correlation-based feature selection methods aim to identify relevant predictors by explicitly considering the statistical relationships among features and between features and the target variable. These methods are particularly well suited to datasets with strong inter-feature dependencies, such as air quality datasets, where pollutant concentrations and meteorological variables often exhibit high levels of

multicollinearity. By accounting for feature redundancy, correlation-based methods seek to retain informative variables while eliminating redundant ones, thereby improving model robustness, interpretability, and generalisation performance.

Unlike general filter methods that assess features independently, correlation-based approaches evaluate both feature relevance and feature redundancy. Relevance is typically measured by the degree of association between an individual feature and the target variable, while redundancy is assessed through correlation coefficients or dependency measures among features. Features that are highly correlated with each other and convey overlapping information are considered redundant and may be removed to mitigate multicollinearity effects. This characteristic is especially important in environmental and air quality studies, where correlated predictors frequently coexist.

One widely used correlation-based feature selection technique is the Maximum Relevance Minimum Redundancy (mRMR) method. mRMR is grounded in mutual information theory and aims to select features that maximise relevance to the target variable while simultaneously minimising redundancy among the selected features. By balancing these two criteria, mRMR retains informative predictors while reducing the adverse impact of correlated variables. Houndekindo and Ouarda (2023) demonstrated that mRMR successfully eliminated multicollinearity in 13 out of 14 experimental cases when compared with other feature selection methods, including Elastic Net and Random Forest Embedded Selection, highlighting its effectiveness in datasets characterised by strongly correlated environmental variables.

Another important correlation-based method is Adjusted Correlation Thresholding (AdjcorT), proposed by Ibrahim (2020). AdjcorT extends conventional correlation-based feature selection by incorporating both the magnitude and direction of correlations through an adjusted correlation and t-statistic framework. Unlike traditional approaches that rely solely on absolute correlation values, AdjcorT is capable of distinguishing between informative correlated predictors and redundant variables, particularly in the presence of strong negative correlations. Ibrahim (2020) demonstrated that AdjcorT consistently outperformed corT under negative correlation structures, while exhibiting comparable performance under positive correlation conditions. Given the complex and mixed correlation patterns commonly observed in air quality predictors, correlation-based feature selection methods such as mRMR and AdjcorT provide a principled foundation for feature selection in PM2.5 classification tasks.

While correlation-based methods offer clear advantages in managing redundancy and multicollinearity, they can be implemented using different methodological paradigms. The following subsections therefore discuss how these methods, along with non-correlation-based techniques, are positioned within filter, wrapper, and embedded feature selection frameworks.

2.6.2 Filter Feature Selection Methods

Filter-based feature selection methods evaluate the relevance of features independently of the learning algorithm. These methods are widely used due to their simplicity and efficiency. Filter techniques unlike wrapper and embedding approaches are not dependent on certain machine learning models. This independence makes them computationally efficient and suitable for large datasets, though they may overlook feature interactions that more complex methods capture. Filter methods typically rely on statistical metrics such as correlation coefficients, mutual information, chi-square tests, or ANOVA F-tests to evaluate the relevance of individual features with respect to the target variable.

Several studies have applied filter-based feature selection techniques to air quality prediction problems. A study by Eşsiz et al. (2023) developed a FireFly Optimization Algorithm (FOA)-based feature selection approach, comparing it with the ReliefF method for air pollution classification. The study concluded that FOA outperformed ReliefF when using the Random Forest classifier. Bhimavarapu et al. (2022) proposed a novel filter-based algorithm that combines symmetric uncertainty with kurtosis to identify the most significant features affecting air pollution. Similarly, Albraikan et al. (2022) utilized the empirical receiver operating characteristic curve (EROC) method, which helps rank features by their importance, with higher EROC values signifying more crucial features for understanding the nonlinear aspects of the data.

Within the filter-based category, mRMR and AdjcorT function as correlation-based filters. As a filter method, mRMR ranks features using mutual information criteria without involving any learning algorithm, which makes it computationally efficient while explicitly addressing redundancy among features. Similarly, AdjcorT operates as a filter by selecting features based on adjusted correlation thresholds and statistical significance. The inclusion of both correlation-based and non-correlation-based filter

methods enables a comprehensive evaluation of the role of correlation handling in feature selection for air quality datasets.

In contrast, ReliefF represents a non-correlation-based filter method. ReliefF evaluates feature importance based on how well features differentiate between neighbouring instances belonging to different classes, rather than relying on correlation measures. This instance-based mechanism allows ReliefF to capture local feature relevance and nonlinear patterns, making it complementary to correlation-based filter methods. By comparing ReliefF with mRMR and AdjcorT, this study investigates the impact of explicitly incorporating correlation information into filter-based feature selection.

2.6.3 Wrapper Feature Selection Methods

Wrapper-based feature selection methods evaluate feature subsets by measuring their impact on the performance of a specific predictive model, such as a decision tree, neural network, or support vector machine. Unlike filter methods, wrappers assess subsets of features by iteratively training and testing the model, allowing them to capture interactions among variables that filters may overlook (Adesunkanmi et al., 2023). This method can result in more accurate predictions, as it tailors the selected feature set to the specific learning algorithm.

The primary advantage of wrapper methods is their ability to produce highly accurate feature subsets, especially when model performance is a priority. Since the model is trained with various feature combinations, the method can effectively determine which variables contribute most to the model's predictive power. In the context of air quality prediction, Ul-Saufie et al. (2022) employed several wrapper techniques, including forward selection, backward selection, stepwise selection, brute-force feature selection, genetic algorithms (GA), and weight-guided methods to improve prediction models. They suggested that future studies combine wrapper methods with other approaches, such as filter and embedded techniques, to enhance model performance. Their study found that the backward selection method was the most effective for predicting PM10 concentrations in Klang.

Despite their effectiveness, wrapper methods are also known for being computationally inefficient, especially when dealing with high-dimensional data. The

method involves repeated model training and validation, which is both costly and time-consuming (Song et al., 2023). Additionally, wrapper methods are more prone to overfitting, as they closely align feature selection with the specific model and training data (Shin et al., 2023). Without proper cross-validation, selected features may not generalize well to unseen data. Moreover, unlike certain filter or embedded techniques, most wrapper methods do not explicitly account for correlations between variables. As a result, they may retain redundant features that are highly correlated with one another. To overcome this limitation, some studies have proposed hybrid approaches which combine wrapper methods with filter techniques that consider the correlation between variables to enhance selection robustness and interpretability.

2.6.4 Embedded Feature Selection Methods

Embedded feature selection methods integrate feature selection directly into the process of model training, making them more efficient than filter and wrapper methods while still considering model performance. These methods incorporate feature selection within the learning algorithm itself, allowing the model to automatically identify and retain the most relevant features as it learns from the data. This integrated nature enables embedded methods to balance predictive accuracy with computational efficiency.

One of the most widely used embedded methods is the Least Absolute Shrinkage and Selection Operator (Lasso). Lasso applies L1 regularization, which penalizes the absolute size of regression coefficients, forcing some of them to shrink to exactly zero. As a result, irrelevant or less important features are effectively removed from the model. This process helps in reducing overfitting and simplifying the model, particularly in the context of high-dimensional datasets. Lasso is beneficial for addressing multicollinearity by removing redundant features that exhibit high correlation with each other (Kayanan & Wijekoon, 2020). However, when faced with a group of highly correlated features, Lasso tends to retain only one and discard the rest, which may lead to the exclusion of other informative variables and a potential reducing interpretability.

In a study by Wang et al. (2021), Lasso was used to develop a risk score system for predicting the survival of patients with adenocarcinoma of the esophagogastric junction. The study demonstrated Lasso's ability to reduce the impact of collinearity and improve model stability, showing its effectiveness in producing a simplified and interpretable model. Hence, Lasso can similarly improve interpretability in air quality

prediction activities by reducing the feature set while preserving high predictive performance.

Additionally, neural network-based approaches have also been explored as embedded feature selection techniques. In these models, the importance of input features can be evaluated using the connection weights between the input and hidden layers. This approach considers how much influence each input node has on the network's predictions, making it possible to rank feature importance directly from the trained model. Ghanizadeh et al. (2020) demonstrated that the connection weight method provides a more accurate assessment of feature significance compared to traditional methods like Garson's algorithm. Furthermore, Da Costa et al. (2021) proposed a new methodology to evaluate the stability of the Weight-Based Feature Selection (WBFS) method using a voting-based approach, improving the robustness of feature rankings derived from neural networks.

2.7 Air Quality Classification Method

This study proposes the application of a classification method to predict PM2.5 level. Classification techniques are crucial in data analysis, enabling the categorization of data into distinct groups based on specific features which is particularly important for understanding the complex relationships in air quality datasets. Traditional methods for air quality classification often include statistical techniques such as linear regression, time series analysis, and empirical modelling. These methods are often based on predefined relationships between variables and may struggle to capture the complex, non- linear interactions found in environmental data. According to Ameer et al. (2019), traditional statistical linear methods frequently give poor air pollution estimates due to the complexity and variability of time-series data. In contrast, machine learning approaches, including neural networks, support vector machines, and decision tree-based models, have demonstrated strong performance in air quality classification tasks.

2.7.1 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are among the most widely used machine learning models for air quality classification due to their ability to model complex nonlinear relationships between input variables and target outcomes. In the context of PM2.5 classification, ANN-based models are particularly effective in handling interactions between pollutant concentrations and meteorological variables. For example, Liu and Zhang (2022) demonstrated that integrating neural network-based models significantly improved the accuracy of PM2.5 measurements obtained from low-cost air quality sensors. Their findings indicate that ANN models enhance data reliability, which is essential for public health monitoring and environmental decision-making.

Several studies have compared ANN with other machine learning classifiers for PM2.5 prediction. Meena et al. (2024) evaluated multiple classification algorithms, including Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbour, Support Vector Machine, and Multinomial Logistic Regression, in analysing the influence of air quality awareness on travel behaviour. Their results showed that Random Forest achieved the highest accuracy and F1-score. Similarly, Guo et al. (2023) compared multilayer perceptron (MLP), gradient boosting decision tree, logistic regression, decision tree, and KNN models with the CatBoost algorithm for indoor PM2.5 classification. The study concluded that CatBoost outperformed other models in terms of classification accuracy.

In the Malaysian context, Murugan and Palanichamy (2021) compared the performance of MLP and Random Forest classifiers for PM2.5 classification and found that Random Forest achieved higher accuracy than ANN-based MLP. Imam et al. (2024) further reported that Support Vector Machine models outperformed several classical machine learning algorithms, including Naive Bayes, Logistic Regression, Decision Tree, and Random Forest, achieving prediction accuracies of 97.98% and 93.29% for two different urban areas. These findings indicate that while ANN-based models are effective for PM2.5 classification, their performance can vary depending on data characteristics and model configuration.

2.7.2 Radial Basis Function Neural Network (RBFNN)

Radial Basis Function Neural Networks (RBFNNs) represent a specialised class of neural networks that have shown promising performance in classification tasks, particularly when dealing with nonlinear and high-dimensional data. RBFNNs are characterised by their simple network structure and fast learning capability, making them suitable for environmental monitoring applications. Ahmed and Hussein (2020) reported that RBFNNs have been widely applied in various classification problems due to their strong approximation capability. Similarly, Mansor et al. (2020) highlighted the effectiveness of RBFNNs in pattern recognition and image processing tasks, reinforcing their suitability for classification-oriented problems.

In environmental applications, RBFNNs have been successfully employed in water quality classification. Huang and Yang (2019) developed an optimised RBF neural network model for water quality evaluation and achieved higher classification accuracy compared to conventional methods. Their study demonstrated that RBFNNs possess strong generalisation capability, particularly when applied to small sample datasets, and are effective in capturing internal nonlinear relationships among environmental variables.

More recent studies have further emphasised the advantages of RBFNNs in classification tasks. Jia et al. (2020) reported that RBFNNs require shorter training time compared to traditional back-propagation neural networks due to their simpler learning process. Cano et al. (2024) also found that RBFNNs achieved competitive classification accuracy while maintaining lower computational complexity and faster convergence compared to SVM models. These characteristics make RBFNNs particularly attractive for real-time or near real-time air quality monitoring systems. Additionally, RBFNNs have demonstrated strong performance when applied to limited or imbalanced datasets, which is a common challenge in air quality studies (Razali et al., 2018).

Previous work has shown that feature ranking techniques can influence RBFNN prediction performance. However, existing RBFNN studies in air quality prediction, such as Yadav et al. (2025), have focused on regression-based modelling rather than classification tasks. Despite these advantages, the application of RBFNNs for air quality classification remains relatively limited, particularly in Malaysia, where studies predominantly focus on more commonly used classifiers such as Random Forest, SVM, and ANN-based models. Moreover, existing RBFNN-based studies rarely examine the

impact of multicollinearity among input variables, even when feature selection methods are employed. Therefore, this study aims to address this research gap by applying the RBFNN model for PM2.5 classification in Malaysia while explicitly incorporating a correlation-aware feature selection strategy to enhance model robustness and interpretability.

2.8 Model Performance

Model performance refers to the ability of a model to accurately predict outcomes based on the input data. Evaluating model performance is essential as it provides insights into how well the model is performing in making predictions or classifications. Understanding the gap between the model's classifications and the actual outcomes helps in assessing the reliability and effectiveness of the model (Yang et.al, 2022). A study on Spatial assessment of PM10 hotspots by Tella et.al (2021) found that the PM10 hotspot map produced by the Random Forest model indicates that urbanized and industrialized areas have high PM10 concentration with a good results sensitivity, specificity and accuracy value which is 0.99, 0.99 and 0.98 respectively. Guo et.al (2023) conduct a study of classification model of indoor PM2.5 concentration using CatBoost algorithm found that CatBoost method outperformed others classification method with good values of area under ROC (AUROC) which is 0.949. Meanwhile, Murugan & Palanichamy (2021) used accuravy, recall, precision and F1 score to compare the Multi-Layer Perceptron (MLP) and Random Forest classifier in predicting PM2.5 levels.

Moreover, a study by Hussain et.al (2020) on classification of indoor and outdoor atmospheric PM2.5 and PM10 mass concentration used sensitivity, specificity, precision and AUROC to evaluate the model performance. The study concludes that the highest AUC which is 1.00, was obtained using the fne Gaussian and cubic SVM along with the cubic and weighted KNN, while the highest accuracy (95.8%) was obtained using the cubic and coarse Gaussian SVM along with the cosine and cubic KNN. Cheng & Tsai (2022) applied four classification method, which are decision tree, random tree, random forest and extra tree to classify air quality. The results of their study show random forest classifier has the best performance metrics which is accuracy, AUROC, sensitivity, precision and F1 score compared to other methods (Cheng & Tsai, 2022). Meanwhile, Sethi & Mittal (2021) stated that they used weighted recall, weighted

precision, weighted F1 score and accuracy to validate the proposed features selection method which is Correlation based Adaptive LASSO (CbAL) Regression method to predict air quality index in Delhi, India. This highlights the importance of using multiple performance metrics when evaluating models that incorporate feature selection, particularly in the presence of correlated predictors.

Previous research studies on classification models have frequently used performance metrics such as accuracy, sensitivity, specificity, precision, F1 score, and AUROC. Feature selection techniques aim to improve classification performance by identifying the most relevant features from high-dimensional datasets. The effectiveness of feature selection is therefore commonly evaluated through classification performance metrics rather than feature ranking alone. Metrics such as accuracy, area under the receiver operating characteristic curve (AUROC), precision, recall, and F1 score have been widely used to assess the classification efficiency of models built using selected feature subsets (Çekik & Kaya, 2024; Cahyani & Irsyada, 2025). These metrics provide complementary perspectives on model performance, allowing researchers to evaluate whether feature reduction leads to improved predictive accuracy, class discrimination, and overall model robustness. Thus, to verify the effectiveness of the proposed model, this study employs these performance metrics.

2.9 Summary of Literature Review

To address the limitations identified in previous studies, this section synthesises the literature into three key aspects: (i) factors affecting PM2.5 concentrations, (ii) applications of machine learning models in air quality prediction, and (iii) applications of feature selection methods in air quality studies. This structured synthesis provides a systematic foundation for justifying the integration of feature selection with RBFNN in this study. Table 2.2 provides an overview of the air quality prediction studies reviewed in this chapter, while subsequent tables synthesise the literature according to specific thematic dimensions relevant to this study.

Table 2.2
Overview of Air Quality Prediction Studies Reviewed

Title	Authors & Years	Malaysia	Other Country	PM2.5 as DV	Classification	Multicollinearity	Feature Selection
Classification prediction model of indoor PM2. 5 concentrations using CatBoost algorithm.	Guo, Z., Wang, X., & Ge, L. (2023).		/	/	/		
Air Quality Monitoring Using Statistical Learning Models for Sustainable Environment.	Imam, M., Adam, S., Dev, S., & Nesa, N. (2024).		/		/		/
Classification of CO Environmental Parameter for Air Pollution Monitoring with Grammatical Evolution	Spyrou, E. D., Stylios, C., & Tsoulos, I. (2023).		/		/		
Ensemble-based classification approach for PM2.5 concentration forecasting using meteorological data	Saminathan, S., & Malathy, C. (2023).		/	/	/		
PM2. 5 prediction using machine learning hybrid model for smart health.	Jebamalar, J. A., & Kumar, A. S. (2019)		/	/			
PM2. 5 forecasting for an urban area based on deep	Zaini, N. A., Ean, L. W.,	/		/		/	

learning and decomposition method	Ahmed, A. N., Abdul Malek, M., & Chow, M. F. (2022).						
COVID-19 lockdowns and air quality: Evidence from grey spatiotemporal forecasts.	Gao, M., Yang, H., Xiao, Q., & Goh, M. (2022)		/	/			
Firefly-Based feature selection algorithm method for air pollution analysis for Zonguldak region in Turkey.	Eşsiz, E. S., Kilic, V. N., & Oturakçi, M. (2023).		/		/		/
Revealing influence of meteorological conditions on air quality prediction using explainable deep learning.	Yang, Y., Mei, G., & Izzo, S. (2022).		/	/			/
Performance of Bayesian Model Averaging (BMA) for Short-Term Prediction of PM10 Concentration in the Peninsular Malaysia.	Ramli, N., Abdul Hamid, H., Yahaya, A. S., Ul-Saufie, A. Z., Mohamed Noor, N., Abu Seman, N. A., ... & Deák, G. (2023).	/					
An Intelligent Time Series Model Based on Hybrid Methodology for Forecasting Concentrations of Significant Air Pollutants.	Cheng, C. H., & Tsai, M. C. (2022).		/		/	/	/
Applying hybrid feature selection methods for statistical modelling of roadside particle concentrations (PM 2.5 and PNC).	Suleiman, A., Tight, M. R., & Quinn, A. D. (2020).		/	/			/
A feature selection and multi- model fusion-based approach of predicting air quality.	Zhang, Y., Zhang, R., Ma, Q., Wang, Y.,		/	/		/	/

	Wang, Q., Huang, Z., & Huang, L. (2020).						
An efficient correlation based adaptive LASSO regression method for air quality index prediction	Sethi, J. K., & Mittal, M. (2021).		/		/	/	/
Performance analysis of regression algorithms and feature selection techniques to predict PM2.5 in smart cities.	Banga, A., Ahuja, R., & Sharma, S. C. (2021).		/	/		/	/
Improving Air Pollution Prediction Modelling Using Wrapper Feature Selection.	Ul-Saufie, A. Z., Hamzan, N. H., Zahari, Z., Shaziayani, W. N., Noor, N. M., Zainol, M. R. R. M. A., ... & Vizureanu, P. (2022).	/					/
Kurtosis-Based Feature Selection Method using Symmetric Uncertainty to Predict the Air Quality Index.	Bhimavarapu, U., & Sreedevi, M. (2022).		/		/		/

2.9.1 Factors Affecting PM2.5 Concentrations

Studies consistently report that PM2.5 concentrations are influenced by a combination of pollutant-related and meteorological factors. Common pollutant variables include PM10, CO, NO_2, SO_2, and O_3, while frequently examined meteorological parameters include temperature, relative humidity, wind speed, and wind direction. These variables are often strongly correlated due to shared emission sources and atmospheric processes. In urban and industrialised areas, PM10 and gaseous pollutants such as CO and NO_2 are frequently identified as significant contributors to PM2.5 variability. Meteorological factors further modulate PM2.5 levels by influencing pollutant dispersion, chemical transformation, and accumulation. As a result, air quality datasets typically exhibit strong positive and negative correlations among predictors, making multicollinearity an inherent characteristic rather than an anomaly. Table 2.3 summarises key factors affecting PM2.5 concentrations reported in recent studies.

Table 2.3
Factors Affecting PM2.5 Concentrations

Study	Location	Pollutant Factors	Meteorological Factors	Key Findings
Guo et al. (2023)	China	PM10, CO	Temperature, Humidity	PM10 strongly associated with PM2.5
Zaini et al. (2022)	Malaysia	PM10, NO_2, CO	Wind speed, Humidity	Meteorology moderates PM2.5 accumulation
Cheng & Tsai (2022)	Taiwan	PM10, SO_2	Temperature	Combined pollutant–meteorology effects
Ramli et al. (2023)	Malaysia	PM10	–	PM10 significant for particulate prediction
Yang et al. (2022)	China	Multiple pollutants	Multiple meteorological factors	Strong inter-variable dependence observed

2.9.2 Applications of Machine Learning Models in Air Quality Prediction

Machine learning models have been widely adopted to overcome the limitations of traditional statistical approaches in capturing nonlinear relationships in air quality data. Models such as Random Forest, Support Vector Machine, CatBoost, deep learning models, and Artificial Neural Networks have demonstrated strong predictive performance for PM2.5 forecasting and classification. RBFNN, as a subclass of ANN, has gained attention due to its simple structure, fast convergence, and ability to model nonlinear relationships effectively. However, most studies focus primarily on predictive accuracy and rarely examine how multicollinearity among input variables affects feature importance and model stability. Consequently, while machine learning models perform well numerically, their interpretability and robustness under correlated inputs remain insufficiently addressed. Table 2.4 summarises recent applications of machine learning models in air quality prediction.

Table 2.4
Applications of Machine Learning Models in Air Quality Prediction

Study	Model Used	Prediction Type	Limitation
Guo et al. (2023)	CatBoost	Classification	Correlation not discussed
Imam et al. (2024)	RF, SVM	Regression	Feature redundancy ignored
Jebamalar & Kumar (2019)	ANN	Regression	Sensitive to input redundancy
Zaini et al. (2022)	ANN / Deep Learning	Regression	Limited feature interpretability
Banga et al. (2021)	ANN, SVM, RF	Regression	Multicollinearity not analysed
Yang et al. (2022)	Deep Learning	Prediction	Correlation effects implicit
Ul-Saufie et al. (2022)	ANN-based models	Prediction	High computational cost
Yadav et al. (2025)	MLFFNN, RBFNN, GRNN, MLR	Prediction	Focused on regression-based prediction and multicollinearity among input variables not explicitly addressed

Although ANN and RBFNN-based models demonstrate strong predictive performance in air quality studies, most existing works focus primarily on accuracy and do not explicitly address the impact of multicollinearity among input variables. This limitation motivates the integration of correlation-based feature selection methods to enhance model robustness and interpretability.

2.9.3 Applications of Feature Selection Methods in Air Quality Studies

Feature selection is commonly employed to improve model performance and reduce dimensionality in air quality datasets. Methods such as mRMR, Lasso, ReliefF, and wrapper-based approaches have been applied to identify important predictors for PM2.5 and air quality indices. However, most feature selection methods either eliminate correlated variables or treat correlation as redundancy to be minimised. While this approach simplifies models, it may inadvertently remove correlated predictors that contain complementary information. Only a limited number of studies explicitly acknowledge multicollinearity, and even fewer adopt correlation-based strategies that allow correlated variables to be retained when they contribute meaningful discriminatory power. This limitation is particularly important in nonlinear classification frameworks, where correlated variables do not necessarily violate model assumptions but can still influence feature ranking and interpretation. Table 2.5 summarises feature selection methods applied in air quality studies.

Table 2.5
Applications of Feature Selection Methods in Air Quality Prediction

Study	Feature Selection Method	Correlation-Based	Limitation
Suleiman et al. (2020)	Hybrid FS	Yes	Redundant variables removed
Zhang et al. (2020)	Filter + Fusion	Yes	Correlation treated as redundancy
Sethi & Mittal (2021)	Adaptive Lasso	Yes	One correlated variable excluded
Banga et al. (2021)	Multiple FS	Mixed	Limited correlation discussion

2.10 Conclusion

The literature review indicates that although RBFNN has been applied in air quality prediction, its performance may be affected by multicollinearity when highly correlated predictors are present. Existing feature selection methods either remove correlated variables or do not explicitly consider correlation structure, which may limit their effectiveness in PM2.5 classification tasks. Notably, the AdjcorT feature selection method has not yet been applied to air quality data, including studies conducted in Malaysia. Given its ability to account for both positive and negative correlations while ranking feature importance, integrating AdjcorT with RBFNN provides a methodological advantage. This study therefore proposes a two-stage AdjcorT–RBFNN framework to enhance PM2.5 classification accuracy, improve feature interpretability, and address multicollinearity in a systematic and application-relevant manner.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the methodology used to predict PM2.5 concentrations in Shah Alam and Banting. These two locations were selected because they represent contrasting urban and sub-urban environments within the same region, allowing the proposed methodology to be evaluated under different air pollution characteristics and emission sources. A detailed justification of the urban and suburban classification of Shah Alam and Banting, supported by previous studies, is provided in Section 3.3.

To achieve the research objectives outlined, this study employs a two- stages methodology. The flow of this study and the method of two-stages feature selection are explained in section 3.2. Next, section 3.3 will describe the data sources, the variables involved, and the specific details of the datasets collected from monitoring stations in both Shah Alam and Banting between 2018 and 2022. This is followed by data pre-processing in Section 3.4, which explains the steps taken to clean and prepare the data for analysis. These steps include handling missing data, data aggregation, outlier detection, and normalisation techniques to ensure that the dataset is suitable for modelling. In addition, multicollinearity diagnostics, which examine the correlation among variables, are also explained in Section 3.4.

Section 3.5 discusses the feature selection methods applied in this study, namely ReliefF, mRMR, Lasso, and AdjcorT, which are used to identify significant predictors by ranking features. Meanwhile, the methodology of classification methods used in this study which are Artificial Neural Networks (ANN) and Radial Basis Function Neural Networks (RBFNN) are showed in section 3.6. Furthermore, the methods to assess the model performance such as accuracy, sensitivity, specificity, precision, F1 score and AUROC are explained in section 3.7. Moreover, section 3.8 discussed the methodology of connection weight approach to determine the feature importance based on the neural network's connection weight. The simulation setting of two-stages feature selection method are showed in section 3.9. Finally, section 3.10 concludes the methodology of the study.

3.2 Research Design

The flowchart of the proposed two-stage feature selection framework is presented in this research design and illustrated in Figure 3.1. The study begins with data extraction, where hourly air quality data recorded between 2018 and 2022 were obtained from the Department of Environment (DOE), Malaysia. Data from Shah Alam and Banting were selected to represent urban and suburban environments, respectively. Following data extraction, data pre-processing is performed to ensure data quality and suitability for modelling. This stage includes missing value imputation using linear interpolation, transformation of hourly observations into daily averages, conversion of PM2.5 concentrations into binary classification labels, outlier detection using Mahalanobis Distance, min–max normalisation, and data augmentation using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance

After pre-processing, multicollinearity among the independent variables is explored using Spearman's correlation coefficient. This analysis is conducted as a diagnostic step to examine the strength and direction of associations between pollutant and meteorological variables and to confirm the presence of correlated predictors commonly observed in air quality data. The mitigation of multicollinearity is performed during the feature ranking stage using the Adjusted Correlation Sharing T-test (AdjcorT). Unlike conventional feature selection methods that evaluate predictors independently, AdjcorT incorporates inter-feature correlations directly into its scoring mechanism by sharing t-statistics among correlated variables. This correlation-based ranking reduces the bias introduced by highly correlated predictors and mitigates multicollinearity effects prior to feature combination evaluation and model training.

Following AdjcorT ranking, an Artificial Neural Network (ANN) model is employed as an auxiliary tool to determine the optimal number of features. Features are added sequentially into the ANN model according to the AdjcorT ranking order, and classification performance is evaluated for increasing numbers of features. The ANN is used solely to identify the number of features that yields the best predictive performance. Once the optimal number of features is determined, feature selection methods are applied independently using the same feature count to enable fair comparison. AdjcorT, ReliefF, mRMR, and Lasso are each used as feature selection methods, where each method selects its own subset of features based on the optimal number of features.

Subsequently, classification models based on Radial Basis Function Neural Networks (RBFNN) are constructed. The baseline RBFNN model is trained using all available features, without feature selection. The two-stage classification models, including ReliefF-RBFNN, mRMR-RBFNN, Lasso-RBFNN, and the proposed AdjcorT-RBFNN model, are trained using feature subsets selected by their respective methods. Model performance is evaluated using standard classification metrics, including accuracy, sensitivity, specificity, precision, F1-score, and the area under the receiver operating characteristic curve (AUROC). The best-performing model is identified based on comparative evaluation results. To further interpret model behaviour, the connection weight approach is applied after model training to assess the relative importance of input variables based on neural network weights. This analysis is conducted to enable a comparison of feature importance rankings obtained from the baseline RBFNN model and the proposed two-stage AdjcorT-RBFNN model. The connection weight analysis provides insight into variable influence but does not form part of the feature selection process.

Finally, a Monte Carlo simulation study is conducted to verify the robustness of the proposed AdjcorT-RBFNN model under controlled multicollinearity conditions. In the simulation setting, datasets with predefined correlation structures and varying sample sizes are generated. AdjcorT is applied to rank the simulated variables, which are subsequently used as inputs to the RBFNN classifier. The ANN-based feature combination evaluation step applied in the real data analysis is not included in the simulation, as the true informative variables are known. The simulation results are used to assess variable selection accuracy and classification performance under different correlation strengths. Further methodological details are provided in Sections 3.3 to 3.13.

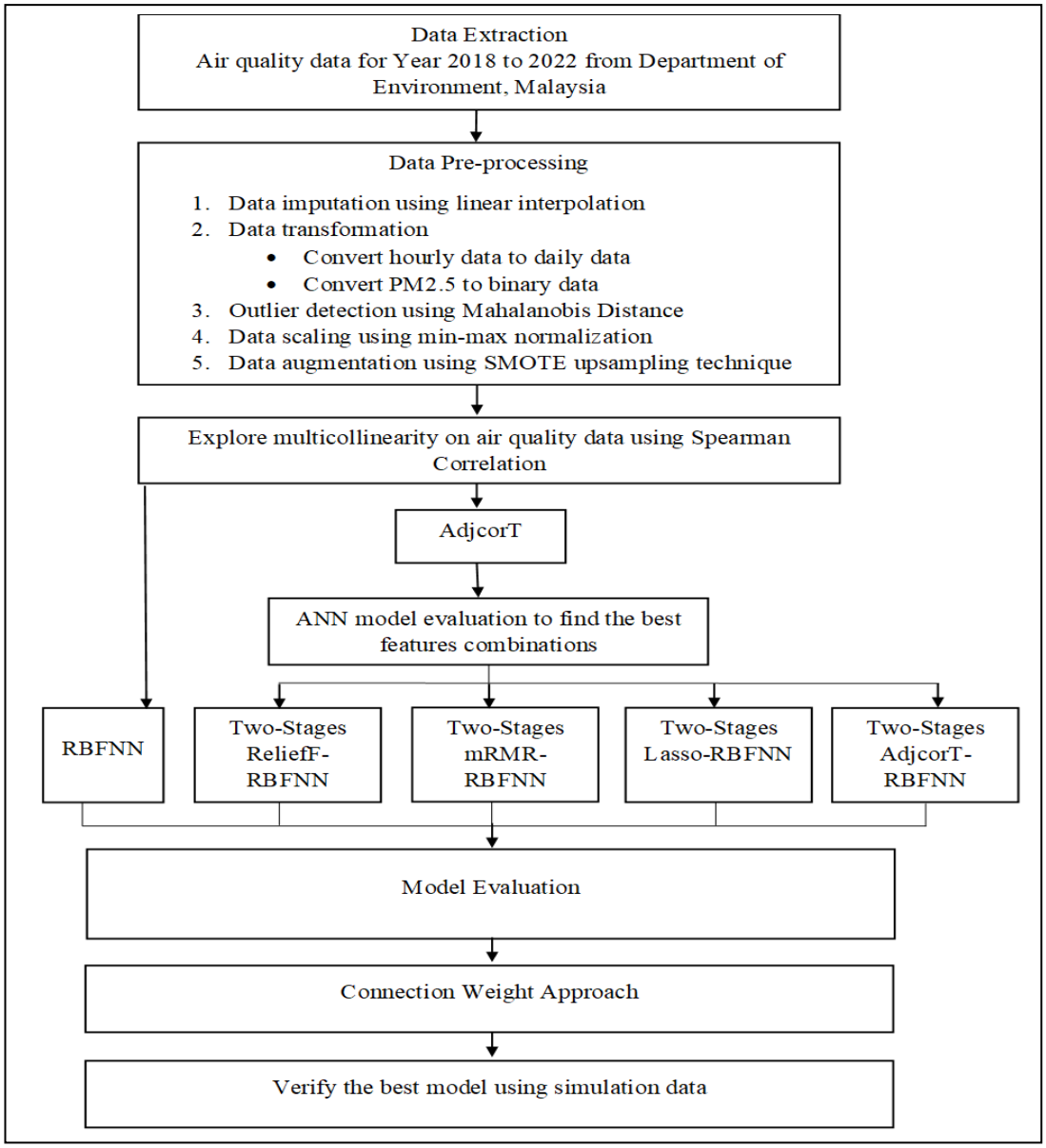

Figure 3.1 Research Design

The illustrations of the one-stage and two-stage feature selection frameworks are shown in Figures 3.2 and 3.3, respectively. In the one-stage approach, the RBFNN model identifies the relative importance of input features after model training through the adjustment of connection weights. The connection weight approach provides insight into feature importance based on the contribution of each input variable to the network output. However, RBFNN alone does not explicitly mitigate multicollinearity among predictors. Therefore, a two-stage feature selection framework is employed to enhance RBFNN performance in classifying next-day PM2.5 levels.

In the first stage of the two-stage framework, all ten independent variables are evaluated using the Adjusted Correlation Sharing T-test (AdjcorT). The features are ranked based on their AdjcorT values, where higher values indicate greater importance in classifying next-day PM2.5 levels while accounting for high correlations among variables. Following this ranking, features are added sequentially to an Artificial Neural Network (ANN) model according to the AdjcorT ranking order to evaluate different feature combinations. The optimal number of features is determined based on the classification performance of the ANN model.

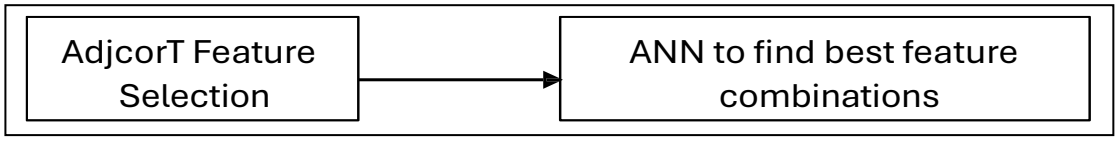

Figure 3.2 One-Stage of Feature Selection

Then, the optimal number of features is used as input to the RBFNN model to construct the proposed AdjcorT-RBFNN classifier. After model training, the weight matrix of the AdjcorT-RBFNN model is computed, and the connection weight approach is applied to re-rank features based on their influence on the trained network. Furthermore, the classification performance of the two-stage feature selection framework is compared with that of the baseline RBFNN model.

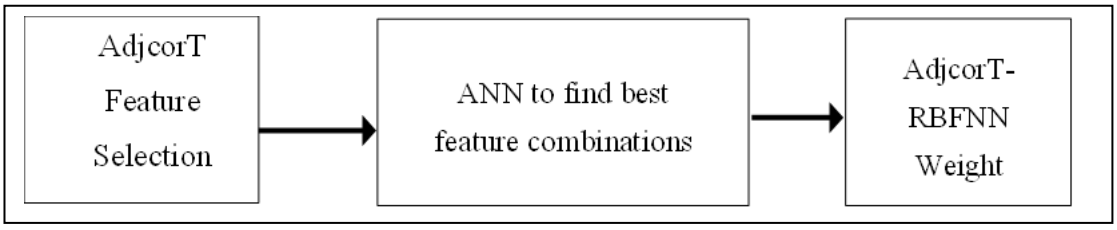

Figure 3.3 Two-Stages of Feature Selection

3.3 Data Description

This research data consists of variables or pollutant factors involved in air quality data such as Particulate Matter (PM10), Particulate Matter (PM2.5), Carbon Monoxide (CO), Nitrogen Dioxide (NO2), Ground-Level Ozone (O3) and Sulphur Dioxide (SO2) that are due to the burning of natural gas, coal and wood, industries and vehicles as shown in Table 3.1. Besides pollutant factors, the meteorological parameters which are wind direction, wind speed, relative humidity and ambient temperature are taken as variables that might have an impact on air quality.

In this study, PM2.5 is measured as a continuous variable ($\mu g/m^3$) at an hourly resolution. However, the objective is not to predict the continuous PM2.5 concentration directly. Instead, the continuous PM2.5 measurements are aggregated into 24-hour average values and subsequently transformed into a binary categorical target variable, referred to as the next-day PM2.5 category. This categorical target represents air quality status (polluted or not polluted) and is derived based on established PM2.5 breakpoints, as described in Section 3.4.2.2.

Shah Alam has been consistently chosen as a popular study location in various research events, further confirming its urban status (Shaadan & Din, 2022). Moreover, according to Raffee (2021), Banting is categorized as suburban area. However, Banting was more polluted compared to Shah Alam from 2014 to 2018 even though Banting has a high percentage of green land use and less industry or transportation land use (Ling et al., 2020). This study focuses on two monitoring locations to ensure a controlled evaluation of multicollinearity treatment in air quality data. As the methodological objective is to assess feature selection performance under correlated variables rather than spatial generalization, selecting one urban and one suburban location is sufficient. Including additional locations may introduce heterogeneous correlation patterns that complicate the interpretation of multicollinearity effects.

Hence, the data covers the air pollutant index in Malaysia from air quality monitoring stations at Shah Alam (CA20B) and Banting (CA22B). 43824 observations of dataset in hourly data format for each location are obtained from the collaborators, which is DOE for the duration of five years period from 2018 to 2022, and the data used for building proposed predictive model and performance evaluation purposes. This period was chosen because PM2.5 data in Malaysia only became officially available starting in 2018, making it the earliest reliable year for analysis.

Table 3.1
Data Description

Variable	Role	Unit	Description
Next-day PM2.5 category	Dependent Variable	0 (Not Polluted), 1 (Polluted)	Binary air quality category derived from next-day 24-hour average PM2.5 concentration
PM10	Independent Variable	(µg/m3)	Particulate matter 10 micrometres or less in diameters
PM2.5	Independent Variable	(µg/m3)	Particulate matter 2.5 micrometres or less in diameters
SO2	Independent Variable	(ppm)	Sulphur dioxide
NO2	Independent Variable	(ppm)	Nitrogen dioxide
O3	Independent Variable	(ppm)	Ground-level ozone
CO	Independent Variable	(ppm)	Carbon monoxide
WD	Independent Variable	(°)	Wind direction
WS	Independent Variable	(m/s)	Wind speed
Humidity	Independent Variable	(%)	Relative humidity
Temperature	Independent Variable	(°c)	Ambient temperature

3.4 Data Pre-Processing

Data pre-processing was conducted to improve data quality and ensure reliable model performance. This phase addresses missing values, data aggregation and transformation, outlier detection, data normalisation, class imbalance, and multicollinearity issues commonly present in air quality datasets. These steps are essential to enhance the robustness and predictive accuracy of the proposed classification models.

3.4.1 Missing Data Imputation

Missing values in the air quality time series were handled using linear interpolation. According to García et al. (2022), linear interpolation has been acknowledged as a straightforward yet efficient technique for managing missing data in time series, offering comparatively good performance when addressing isolated missing values in time series related to air quality. Formula of linear interpolation is shown in 3.1:

$$y = y_1 + \frac{(x - x_1)(y_2 - y_1)}{x_1 - x_1} \tag{3.1}$$

Where x_1 and x_2 represent the time indices of two known observations immediately before and after the missing value, y_1 and y_2 denote the corresponding observed air quality values at times x_1 and x_2, respectively, and x represents the time point at which the missing value is interpolated.

3.4.2 Data Transformation

Data transformation was performed to convert the raw air quality data into a suitable format for classification modelling. This process involved data aggregation and the conversion of the target variable into a binary category.

3.4.2.1 Data Aggregation

After imputing missing values, the hourly air quality data were aggregated into daily values by averaging the 24-hourly observations recorded within each day. The aggregation process is defined by Equation (3.2). Where x_i represents the PM2.5 concentration measured at hour i, and y denotes the aggregated daily PM2.5 concentration.

$$y = \frac{1}{24}\sum_{i=1}^{24} x_i \tag{3.2}$$

3.4.2.2 Conversion of Target Variable to Binary Category

Since the objective of this study is to predict the next-day PM2.5 category, the target variable was transformed into a binary classification outcome. The PM2.5 breakpoints (24-hour average) used in this study are based on the guidelines established by the United States Environmental Protection Agency (EPA), which aim to protect public health from the adverse effects of fine particulate matter pollution.

Following the Air Quality Index (AQI) classification, the AQI categories "Good" and "Moderate" were combined to represent the "not polluted" class, while the remaining AQI categories were grouped into the "polluted" class. This binary classification framework is consistent with previous air quality classification studies, such as Kalajdjieski et al. (2020). The corresponding PM2.5 breakpoints, AQI categories, and binary labels are summarised in Table 3.2.

Table 3.2
Binary Labels for The Respective PM2.5 Breakpoint and AQI Categories

AQI Category	PM2.5 Breakpoints	Binary Labels
Good	0.0-12.0	Not Polluted
Moderate	12.1-35.4	Not Polluted
Unhealthy for Sensitive Groups	35.5-55.4	Polluted
Unhealthy	55.5-150.4	Polluted
Very Unhealthy	150.5-250.4	Polluted
Hazardous	250.5 and above	Polluted

3.4.3 Outlier Detection

Outlier detection is essentials steps in data pre-processing. In addition, air quality data commonly exhibits extreme values due to extreme events, such as s evere haze pollution (Chen et.al, 2020). Additionally, Mahalanobis Distance (MD) are commonly applied on multivariate data to detect outliers. This technique measures the distance between a data point and the mean of the dataset (Mahajan et al., 2021). Mahalanobis distance is calculated using the formula in (3.3). Where O represents the vector of observed values or the data point for which you are calculating the Mahalanobis distance. While $\bar{O}$ is the mean vector, representing the average value for each feature and (S) is the covariance matrix describing the variance and covariance among the variables.

$$MD(O,\bar{O}) = \sqrt{(O-\bar{O})^T \mathrm{S}^{-1}(O-\bar{O})} \tag{3.3}$$

The calculated Mahalanobis Distance values were compared to a critical value to detect outlier. It was determined based on the chi-squared distribution with p degrees of freedom (where p is the number of variables). A significance level of 0.05 was used to obtain the critical value from the chi-squared distribution table as shown in equation (3.4). Any observation with a Mahalanobis Distance larger than this critical value was classified as an outlier.

$$X^2_{0.95}(10) = 18.307 \tag{3.4}$$

3.4.4 Data Normalization

To guarantee that the data is on a consistent scale and distribution, normalization techniques must be used for precise analysis and model performance. Since min-max normalization is frequently used in machine learning techniques for air quality data, this study employs this normalization technique (Aarthi et al., 2023). The min–max normalisation formula is given in Equation (3.5), where, x represents the original value of a variable, x_{min} and x_{max} denote the minimum and maximum values of the variable, respectively, and x_{new} is the normalize value.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.5}$$

3.4.5 Data Augmentation for Imbalanced Data (SMOTE)

In air quality datasets, the distribution of pollution categories is often imbalanced, where non-polluted observations substantially outnumber polluted observations. Such class imbalance can bias classification models towards the majority class and negatively affect prediction performance, particularly for the minority (polluted) class.

To address this issue, this study employs a data augmentation technique, namely the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates synthetic samples of the minority class by interpolating between existing minority class observations in the feature space, thereby increasing the representation of the minority class without simply duplicating existing samples. This approach helps improve the model's ability to learn decision boundaries associated with polluted conditions.

The application of SMOTE has been widely adopted in air quality prediction studies to handle class imbalance and enhance classification performance, particularly in machine learning-based models (Barid et al., 2024). Therefore, SMOTE is applied in this study as a data-level augmentation method prior to model training.

3.4.6 Multicollinearity Analysis

Spearman's correlation coefficient among the variables or factors was calculated to examine the presence of multicollinearity in the air quality data. To measure the spearman's correlation, ρ as shown in Equation (3.6), the two variables were first ranked. Let variable A has the rank (R_A) and variable B has the rank (R_B). Then, the difference, d between the two ranks were measure ($d = R_A - R_B$). While n is the number of ordered pairs.

$$\rho = 1 - 6 \sum d^2/n(n^2 - 1) \tag{3.6}$$

Spearman correlation are ranges from -1 to 1. According to War and Barlis (2023), if the spearman correlation value is greater than 0.6 between the two variables, it means that there is multicollinearity issue in the dataset. Therefore, multicollinearity identified in this study was addressed using the proposed two-stage feature selection method combining AdjcorT and RBFNN, which is explained in Phase 2 of the methodology.

3.5 Feature Selection Methods

Feature selection is a method to improve the accuracy of prediction by reducing the dimensionality of the dataset, removing irrelevant or redundant features. This study wants to emphasize the use of feature selection to enhance PM2.5 levels classification while mitigate the multicollinearity issues in air quality data. Thus, feature selection is applied in this study to enhance the classificaton of PM2.5 levels while addressing multicollinearity issues commonly found in environmental datasets. Four feature selection methods are employed: ReliefF, mRMR, Lasso, and AdjcorT.

The selection of these methods is based on their underlying selection principles and their relevance to the study's aim. Since AdjcorT is a filter-based and correlation-based method that accounts for both positive and negative correlations, it is chosen as the primary method to be evaluated. To assess its comparative performance, mRMR is employed because it is also a filter-based method that considers correlation but operates

by maximizing relevance and minimizing redundancy. This allows for a direct comparison between two correlation-based filter approaches.

ReliefF, on the other hand, is employed to represent a filter-based feature selection method that does not rely on correlation between features. It evaluates features based on their ability to distinguish between classes, offering a noncorrelation-based perspective for comparison with AdjcorT. Lastly, Lasso is chosen as it represents an embedded method that integrates feature selection into model training and applies regularization to reduce multicollinearity by shrinking some coefficients to zero. Unlike filter methods, Lasso incorporates model performance during selection, making it a valuable benchmark for assessing the embedded correlation-based approach against AdjcorT.

By selecting these three feature selection methods, which is mRMR, ReliefF, and Lasso, this study aims to provide a comprehensive evaluation of AdjcorT in comparison to other filter and embedded based feature selection methods. Each method is integrated with the RBFNN model to observe their influence on classification performance. The methodology and implementation of ReliefF, mRMR, Lasso, and AdjcorT are described in detail in Sections 3.5.1 through 3.5.4.

3.5.1 ReliefF as Noncorrelation- Based Feature Selection

The ReliefF feature selection method is an improvement to the algorithm of Relief method which makes it suitable for feature weight computations involving multiple samples (Chen et.al, 2020). It is a filter-based and noncorrelation-based feature selection method that used to identify relevant features in high-dimensional data by assessing their contribution to the prediction of a target variable. In contrast to traditional techniques that exclusively depend on statistical correlations or regression analysis, ReliefF functions by assessing the degree to which every feature may discriminate between examples belonging to different classes. ReliefF is particularly effective in handling noisy and high-dimensional data, making it a valuable tool in many machine learning and data mining applications.

The ReliefF algorithm operates through the following steps:

1. A random instance R is selected from the training dataset.
2. For the selected instance R, the algorithm identifies k nearest neighbours H_j $(j = 1,3, \dots, k)$ belonging to the same class as R, referred to as the nearest hits.
3. The algorithm then identifies k nearest neighbours $M_j(C)$ $(j = 1,3, \dots, k)$ from each different class C, referred to as the nearest misses. Euclidean distance is used to determine the nearest neighbours (Zhang et al., 2022).
4. The differences in feature values between the selected instance R and its nearest hits and misses are computed for each feature.
5. Feature weights are updated iteratively by decreasing the weight when a feature differs between R and its nearest hits and increasing the weight when a feature differs between R and its nearest misses.
6. Steps 1 to 5 are repeated for multiple randomly selected instances to obtain stable estimates of feature importance.

Based on these steps, the weight of each feature f_i is updated using Equation (3.7), which incorporates the contributions from both nearest hits and nearest misses across different classes. The feature weight update mechanism ensures that features with small within-class differences and large between-class differences receive higher importance scores.

$$\begin{aligned} & w(f_i) \\ & = w(f_i) \\ & - \sum_{j=1}^{k} \frac{diff\left(f_i, R, H_j\right)}{mk} \\ & + \sum_{c \neq class(R)} \frac{p(C)}{1 - p\left(class(R)\right)} X \sum_{j=1}^{k} \frac{diff\left(f_i, R, M_j(C)\right)}{mk}, \\ & \qquad (i = 0,1, \dots, d) \end{aligned} \tag{3.7}$$

The difference function used in Equation (3.7) is defined in Equation (3.8), which measures the normalized difference between two instances for a given feature:

$$diff(A, R_1, R_2) = \begin{cases} \dfrac{|R_1[A] - R_2[A]|}{\max A - \min A}, if\ A\ is\ continuous \\ 0, if\ A\ is\ discrete\ and\ R_1[A] = R_2[A] \\ 1, if\ A\ is\ discrete\ and\ R_1[A] \neq R_2[A] \end{cases} \quad (3.8)$$

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 $R_2[A]$ indicates the values of sample R_1 and R_2 on feature A (Zhang et.al, 2022).

3.5.2 Maximum Relevance Minimum Redundancy (mRMR) as Correlation-Based Feature Selection

Maximum Relevance Minimum Redundancy (mRMR) is a filter-based feature selection method by choosing a subset of features that are minimally redundant among themselves and extremely significant to the target variable which aims to enhance the predictive performance of models. The algorithm typically employs mutual information as a measure of relevance and redundancy, choosing features iteratively that maximize the balance between these two criteria. Relevance ensures that the chosen features are highly correlated with the target variable, which can enhance the accuracy of the model. A higher correlation suggests that the model can handle problems more effectively (Wu et.al, 2023). Conversely, redundancy is kept to a minimum to prevent the inclusion of several features that convey similar information, as this can cause overfitting and inefficiencies in the model. Thus, this mRMR method allows the building of simpler, more comprehensible, and reliable models. Moreover, this method is especially useful for high-dimensional datasets, where irrelevant features and multicollinearity can severely degrade model performance. The formula to calculates the maximum relevance shown in (3.9):

$$maxD(S, c), D = \frac{1}{|s|} \sum_{x_i \in S} I(x_i, c) \quad (3.9)$$

Based on the equation (3.9) above, 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 (3.10).

$$minR(S), R = \frac{1}{|s|^2} \sum_{x_i, x_J \in S} I(x_i; x_J) \tag{3.10}$$

3.5.3 Least Absolute Shrinkage and Selection Operator (Lasso) as Correlation-Based Feature Selection

Least Absolute Shrinkage and Selection Operator (Lasso) is an embedded-based feature selection method that enhances model performance by simultaneously performing variable selection and regularization. By adding a penalty equivalent to the absolute value of the magnitude of coefficients to the loss function, Lasso forces some of the coefficients to be exactly zero, effectively excluding less important features from the model. Furthermore Ibrahim (2020) stated that Lasso method goals are to lower the variance in models with a high number of unnecessary variables. It can simplify and facilitates the interpretation of the final model (Ibrahim, 2020). Moreover, Lasso tends to select one variable and ignore the other variables in that correlated group (Ibrahim, 2020). Thus, this characteristic makes Lasso particularly effective for high-dimensional datasets where the number of predictors exceeds the number of observations, or where multicollinearity is a concern. The formula to calculates Lasso is shown in (3.11). Where, $\sum_{j=1}^{\rho} |\beta_j|$ is known as L_1 penalty term and the value are 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).

$$\hat{\beta} = min\beta \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.11}$$

3.5.4 Adjusted Correlation Sharing T-test (AdjcorT) as Correlation-Based Feature Selection

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. In contrast, the AdjcorT method allows both positive and negative correlations between the variables from -1 to 1. The equation of AdjcorT is displayed in (3.13). As mentioned before, the air quality category based on PM2.5 has divided into two binary class $j = 0$ and 1 which is polluted and not polluted. $\bar{x}_{ij}$ is the mean of the i-th variable ($i = 1,2,3, \dots, p$) in group 0 or 1 while s_i is the pooled within-group standard deviation of the i-th variable. This are the formula of standard (unpaired) t-statistics as shown in (3.12).

$$T_i = \frac{\bar{x}_{i1} - \bar{x}_{i0}}{s_i} \tag{3.12}$$

$$r_i = sign(T_i) \cdot \left[max_{(0 \le \rho \le 1)} \frac{1}{W} \sum_{j \in C_\rho(i)} |T_j| \right] \tag{3.13}$$

Based on the equation (3.13) above, r_i is a score, which equals to the average of all t- statistics for variables having correlation (absolute) at least ρ with variable i. $C_p(i) = \{k: |corr(x_i, x_k)| \geq \rho\}$ is the set of the indices of the variables with correlation (absolute value) equal or larger than ρ with variable x_i. While w is the cardinal of the set $C_p(i)$.Let say $C_p(1) = \{x_1, x_3\}$, then $w = 2$. This method is applicable to continuous variables, as it relies on t-statistics and correlation measures among independent variables. In this study, AdjcorT was employed in the first stage as a feature ranking method to identify the optimal feature combination size. Although other feature selection methods serve similar purposes, AdjcorT explicitly accounts for inter-feature correlations without eliminating correlated variables. In contrast, methods such as mRMR and Lasso may suppress or remove variables exhibiting strong correlations due to redundancy minimisation or coefficient penalisation. The AdjcorT approach instead evaluates correlated variables collectively through correlation sharing, allowing

informative correlated predictors to be retained and appropriately ranked. This characteristic is particularly important for air quality data, where correlated pollutants often jointly influence PM2.5 levels.

3.6 Classification Method

This study applied two classification methods which are Artificial Neural Network (ANN) and Radial Basis Function Neural Network (RBFNN). However, these two methods were applied for different objectives. ANN were applied in this study to find the optimal number of features to predict PM2.5 levels. ANN model was used because it is widely used machine learning model known for its ability to capture complex relationship between input features and output predictions. Additionally, RBFNN is a subset of ANN model, the difference is RBFNN used radial basis function as activation function while ANN normally used sigmoid activation function in hidden layer. Furthermore, RBFNN was used to compare its accuracy of predict PM2.5 levels with two-stages feature selection method which are ReliefF-RBFNN, mRMR-RBFNN, Lasso-RBFNN and AdjcorT-RBFNN. The methodology of ANN and RBFNN were explained in sub section 3.6.1 and 3.6.2, respectively. All classification models were implemented using RStudio software.

3.6.1 Artificial Neural Network (ANN)

Next, this study will determine the number of best features combination based on AdjcorT values rank. To find the best features combination, we conduct ANN model to evaluate the performance of features combination. ANN are computational models inspired by the human brain's neural structure, consisting of interconnected nodes that process information. This method has been widely used in many areas for classification such as rock classification in mountain tunnels (Hasegawa et.al, 2019), estimating air pollution levels in urban, rural, and industrial areas (Kumbhar et.al, 2022) and others.

This study used feed forward backpropagation neural network (FFBP) follows study on PM10 prediction by Ul Saufie et.al (2022). There are three layers of neurons in FFBP which are input layer (independent variables), hidden layer and output layer (dependent variable). The input layers in this study consists of PM10, PM2.5, CO, NO2, O3, SO2, wind speed, wind direction, ambient temperature and relative humidity while

the output layer in this study is PM2.5 of the next day. Meanwhile the process of transferring from the input layer to output layer happen in the hidden layer. Before conducting the analysis, the dataset was split into 80% for training and 20% for testing. Additionally, this study set up the hyperparameters, including the number of hidden nodes and the learning rate. The formula to set up number of hidden nodes = (number of attributes + number of classes) / 2 + 1, while the learning rate values are set to 0.01 (Ul Saufie et.al, 2022).

3.6.2 Radial Basis Function Neural Network (RBFNN)

Radial Basis Function Neural Network (RBFNN) is a type of ANN. Due to the ability to correctly predict the dependent variables, simplicity and flexibility of the RBFNN, this method is frequently used in many areas including air quality classification. RBFNN is a machine learning method uses to model any continuous input-output mapping by using a two-layer neural network topology (Hao et al., 2022). There are three important layers in RBFNN which are the input layer, hidden layer, and output layer that are generally connected by weights. Firstly, the input layer consists of a source node that connects the network to its surrounding or called the independent variable, 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. The input layer and hidden layer are fully and directly connected, meanwhile, the hidden layer and output layer are fully and directly connected. Additionally, the data splitting for the RBFNN model follows an 80:20 ratio, where 80% of dataset is used for training and 20% for testing. Commonly, the training algorithm during the training process will suggest the number of hidden units. Figure 3.4 shows the general framework of RBFNN (Wu, 1998).

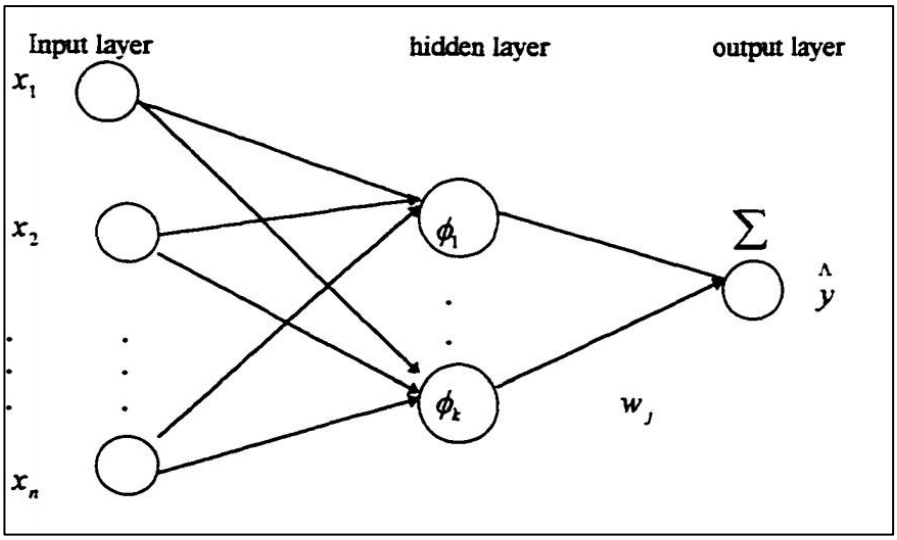

Figure 3.4 General Framework of a RBFNN

The hidden layer consists of each unit *j* that associated to transfer function ϕ_j, which is typically a Gaussian function. The radial basis function (RBF) has a radially symmetric shape that used as a transfer function. There are two categories of RBF which are localized and nonlocalized RBF. The number of hidden layer units is equal to the number of the RBF. The Gaussian radial basis function is formally defined in (3.14).

$$\phi(x) = exp\left(-\frac{||x-c||^2}{2\sigma^2}\right) \tag{3.14}$$

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 (3.15).

$$g(X) = \sum_{j=i}^{k} w_j \phi_j \left(r\left||X - C_j|\right|\right) \tag{3.15}$$

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. The RBFNN commonly employs a logistic (sigmoid) activation function in the output layer for applications involving binary classification. The logistic function shown in (3.16) is used to transform the weighted sum of the outputs from the hidden layer into a probability value ranging from 0 to 1.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.16}$$

Hence, the final output calculation of the RBFNN for binary classification is shown in (3.17), where w_0 is a bias term.

$$\hat{y} = \sigma\left(w_0 + \sum_{j=1}^{k} w_j \phi_j \left(r \left|\left| X - C_j \right|\right| \right) \right) \tag{3.17}$$

To determine the significant factors that contribute to air quality classification as mentioned in objective 1, AdjcorT may help RBFNN to identify the significant factors that contribute to air quality classification. It is because factor reduction is a key step in this study because air quality data consist of big data elements. Furthermore, the advantage of AdjcorT lies in its ability to rank factors using t-scores while accounting for correlations among variables, where multicollinearity remains a major limitation of RBFNN. Hence, by using AdjcorT, important uncorrelated factors are identified, and these factors are subsequently included in the RBFNN. Consequently, it overcomes the limitation of RBFNN where the top important factors are consisted of uncorrelated factors only.

3.7 Classification's Model Performance

The confusion matrix is a well-known tool to evaluate the performance of classification model. It is a table that compares the model's predicted classifications to the actual results, offering a complete overview of correct and incorrect classifications. The confusion matrix consists of four components which are true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are shown in Table 3.3 (Ibrahim et al., 2024).

Table 3.3
Confusion Matrix

Predicted Negative	Predicted Positive
True Negative (TN)	False Positive (FP)
False Negative (FN)	True Positive (TP)

The performance of the developed model is evaluated based on accuracy, sensitivity, specificity, precision, F1 score and AUROC. The accuracy is the number of correct classifications in a given number of classifications. The formula to calculates accuracy is shown in Equation (3.18):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.18}$$

Sensitivity and specificity are used to verify the model's reliability in the presence of uncertainty and to understand the relationship between the input and output variables of the system. The true positive (TP) rate is represented by sensitivity, and the true negative (TN) rate is represented by specificity. Equation (3.19) and (3.20) shows the formula to calculates sensitivity and specificity respectively.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.19}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.20}$$

Precision is used to measure the percentage of samples that are correctly classified among those that are classified as positives. While F1 score is a measure of harmonic mean of precision and sensitivity that demonstrate if the model performance is well-balanced. Equation (3.21) and (3.22) shows the formula to calculates precision and F1 score respectively.

$$Precision = \frac{TP}{TP + FP} \tag{3.21}$$

$$F1\ Score = \frac{2\ x\ Precision\ x\ Sensitivity}{Precision + Sensitivity} \tag{3.22}$$

Figure 3.5 is the example of ROC curve which is a plot of the true positive fraction (sensitivity) on the y-axis and false-positive fraction (1-specificity) on x-axis used to evaluate the classifier. While the AUROC stands for Area Under ROC curve shows which shows how well the model can differentiate between classes. The AUC has a value between 0 and 1. An AUC of 0.0 indicates a model whose classifications are 100% incorrect, and 1.0 indicates a model whose classifications are 100% correct. Hence, high values of accuracy, sensitivity, specificity, F1 score and AUROC indicates that the performance model is good.

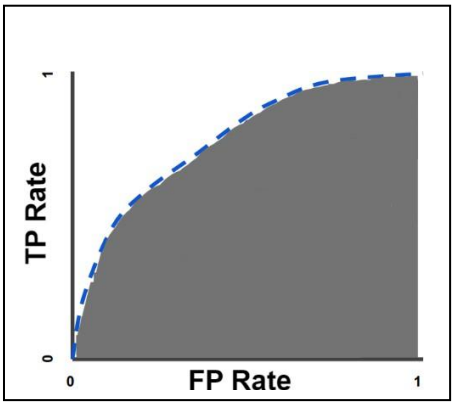

Figure 3.5 Example of Area Under ROC Graph

3.8 Connection Weight Approach

In an RBFNN, weights connect the input and hidden layers, as well as the hidden and output layers. The connections between the input and hidden layers are defined by the centers and spreads of the Gaussian radial basis functions. Each hidden layer neuron computes an activation based on a Gaussian function, which depends on the distance between the input vector and the center of the neuron, and the spread (sigma), which controls the width of the Gaussian. Typically, the centers of these Gaussian functions are determined using clustering techniques, such as k-means, while the spread (sigma) is treated as a hyperparameter, adjusted during training. However, this study are using the RSNNS package in R in this study. The rbf function, selects the centers using an internal method that approximates the input data distribution. These centers are chosen based on a distance criterion or by selecting representative data points from the training set.

In this study, the sigma is set to 0.01, as suggested by Ul-Saufie et al. (2022). The connections between the hidden and output layers are represented by output weights, initialized randomly or with small values. The training process focuses on adjusting the output weights to minimize the error between predicted and actual outputs (Cuevas et al., 2024). Unlike traditional neural networks, which use backpropagation to update weights in both the input-to-hidden and hidden-to-output layers, an RBFNN adjusts only the output weights. This is typically done using optimization methods such as least squares or gradient descent. The centers and spreads of the Gaussian functions remain fixed after initialization, while the output weights are iteratively updated to improve model performance. This training approach distinguishes RBFNNs from conventional neural networks, which iteratively adjust all weights. By focusing solely on optimizing the output layer weights, RBFNNs simplify the learning process, making them particularly useful for capturing highly nonlinear relationships between inputs and outputs.

Weight connection approach is a one of the methods to determine the relative impacts of input variables to the output created by Olden and Jackson (Santos et al., 2018). This method is widely recognized in recent research on neural network interpretability as a global feature-importance technique, because it directly quantifies the influence of each input from the trained network parameters while retaining both the magnitude and direction of effect. Recent studies have confirmed the relevance and

robustness of this method in various applications (Giam & Olden, 2021; Candotti et al., 2022). Unlike model-agnostic approaches, which only approximate importance externally, the connection weight approach leverages internal network weights to provide a direct measure of variable influence, making it particularly suitable for complex models such as RBFNN. Equation (3.23) is to measure the product of connection weight of input-hidden neuron (W_{ij}) and hidden-output neuron (W_{ij1}).

$$P_{ij} = W_{ij}\, X\, W_{ij1} \tag{3.23}$$

Then, the importance score of each input, Sj were computed by adding all the contribution of all each input to the output through all hidden nodes as shown in Equation (3.24). These scores are then ranked. Additionally, this study used a normalized rank following a study by Santos et al. (2018). Thus, their proportional difference to the variable with the highest relative importance can be determine (Santos et al., 2018).

$$S_j = P_{1j} + P_{2j} + P_{3j} + \cdots + P_{kj} \tag{3.24}$$

3.9 Simulation Setting

To achieve Objective 3, a Monte Carlo simulation study was conducted using RStudio software to verify the performance of the proposed AdjcorT-RBFNN model in comparison with the existing RBFNN model under controlled multicollinearity conditions (Ibrahim, 2020). The simulation design focuses on evaluating the ability of the proposed method to correctly identify important variables and to achieve reliable classification performance.

In each simulation run, a dataset consisting of 10 independent variables was generated. Among these variables, x1 and x2 were defined as true informative variables, while x3 to x10 were treated as non-informative noise variables. Variables x1 and x2 were generated from a bivariate normal distribution with specified correlation values (ρ), while x3 to x10 were generated independently from a standard normal distribution. To represent class separability, x1 and x2 were assigned different mean values for Class 0 and Class 1, whereas the remaining variables had identical distributions across both classes. The correlation between x1 and x2 was systematically varied to represent different multicollinearity scenarios, with correlation values set at ρ = −0.8, −0.5, −0.2,

0, 0.2, 0.5, and 0.8. Sample sizes of n = 50, 100, 150, and 200 were considered to examine the effect of sample size on model performance. The upper limit of 200 was selected because, for sample sizes of 150 and 200, the simulation results indicated that the proposed method was able to correctly select the true informative variables (x1 and x2) in all iterations for some scenarios, demonstrating stability and reliability of variable selection at higher sample sizes. These settings were adopted to evaluate model robustness under varying correlation strengths and data sizes.

For each simulated dataset, the observations were randomly split into 80% training data and 20% testing data (Krajnc et al., 2022). The AdjcorT method was first applied to the training data to rank the variables based on their adjusted t-statistics while accounting for correlation among variables. The top-ranked variables were then selected and used as inputs to the RBFNN classifier. In the simulation study, the feature combination optimisation stage used in the real data analysis was not applied, as the true informative variables were predefined. The simulation focuses on evaluating selection accuracy and resulting classification performance rather than searching for the optimal subset.

The RBFNN model was trained using the selected variables and subsequently evaluated on the testing data. This process was repeated for 100 Monte Carlo iterations for each combination of sample size and correlation level to ensure stability of results. During each iteration, the frequency with which x1 and x2 were correctly selected was recorded to assess the effectiveness of the feature selection stage. In addition, classification performance was evaluated using accuracy, sensitivity, specificity, precision, F1-score, and area under the receiver operating characteristic curve (AUROC). The simulation results were summarized by averaging performance metrics across iterations, allowing a comprehensive comparison between the proposed two-stage AdjcorT-RBFNN model and the existing RBFNN model under varying multicollinearity and sample size conditions. The R code used for the simulation is provided in the Appendix.

3.10 Conclusion

In conclusion, this study aims to build a two-stage feature selection method which is AdjcorT and RBFNN to predict PM 2.5 concentration of the next day with a better handling multicollinearity issue to enhance the classification accuracy. Moreover, this study also aims to compare the performance of existing noncorrelation-based, correlation-based feature selection method with proposed two-stage feature selection method by using air quality data in Shah Alam and Banting area. Based on the literature reviews, multicollinearity is a common issue that exist in big data. However, most of the previous studies did not consider the high correlation between factors in air quality data that will affect the accuracy of performance of model classifications. RBFNN is one of the neural networks has been used for classification modelling including in air quality classification. This study is expected to mitigate issue of the high correlation between factors that might exist in air quality data to produce more accurate classification. The study hopes this two- stages feature selection method will offer a good result that will give benefits to the collaborators of this study, DOE to monitor the PM2.5 concentration of the next day.

CHAPTER 4
RESULTS

4.1 Introduction

The result analysis of this study was present in this chapter that addressing three research objectives, which explore the role of feature selection methods in improving the accuracy of air quality classification models, specifically for next-day PM2.5 classification. In section 4.2, data pre-processing steps such as data imputation, data transformation, data scaling, and data augmentation are discussed. Moreover, this study examines the presence of multicollinearity in the air quality datasets from Shah Alam and Banting in section 4.2. Section 4.3 focuses on the optimal features to predict PM2.5 by feature ranking using AdjcorT and ANN model. Section 4.4 reviews the feature selection methods employed to determine the significant predictors, covering both non-correlation and correlation-based methods such as ReliefF, mRMR, Lasso, and AdjcorT. The chapter then discusses the comparative performance of these methods when integrated with radial basis function neural networks (RBFNN) for predicting PM2.5 levels in section 4.5. Furthermore, the connection weight analysis was discussed in section 4.6 to get better insight of feature importance of each model which are RBFNN, ReliefF-RBFNN, mRMR-RBFNN, Lasso-RBFNN, and AdjcorT-RBFNN model. Additionally, a monte carlo simulation study was conducted to assess the effectiveness of the proposed AdjcorT-RBFNN model under different correlation levels and sample sizes in section 4.7. Through this analysis, the results aim to provide insights into the effectiveness of correlation-based and non-correlation-based feature selection methods and their impact on the next-day PM2.5 classification.

4.2 Data Pre-processing

Table 4.1 shows the percentage of missing values for independent variables in Shah Alam and Banting. NO2 has the highest percentage of missing data in both Shah Alam and Banting, with 12.65% and 8.68%, respectively. In contrast, the variable with the least missing data differs between these two locations. The least missing data in Shah Alam is PM2.5 (1.29%), while Humidity has the fewest missing values in Banting with only 1.39% of the dataset. According to Zaman et al. (2024), missing air quality data in Malaysia are the result of insufficient information during model training and a lack of satellite data caused by cloud contamination, which may decrease PM2.5 prediction accuracy. Thus, ignoring missing values can lead to the loss of valuable information, which may compromise the ability to draw significant inferences from analysis of the dataset (Alsaber et al., 2021). To address this issue, this study employed linear interpolation to impute missing values following a study by Ul-Saufie et al. (2022).

Table 4.1
Percentage of Missing Values

Variable	N (Shah Alam)	Missing Value (Shah Alam)	N (Banting)	Missing Value (Banting)
PM2.5	43257	567 (1.29%)	43207	617 (1.41%)
PM10	43151	673 (1.54%)	43075	749 (1.71%)
SO2	41242	2582 (5.89%)	40938	2886 (6.59%)
NO2	38280	5544 (12.65%)	40021	3803 (8.68%)
O3	41198	2626 (5.99%)	41182	2642 (6.03%)
CO	40629	3195 (7.29%)	40735	3089 (7.05%)
WD	42925	899 (2.05%)	43135	689 (1.57%)
WS	42868	956 (2.18%)	42927	897 (2.05%)
Humidity	42907	917 (2.09%)	43217	607 (1.39%)
Temperature	42926	898 (2.05%)	42443	1381 (3.15%)

Table 4.2 shows descriptive statistics for independent variables in Shah Alam and Banting after data imputation and data transformation. The sample size (N) was reduced to 1825 for both locations as the original 43,824 hourly records were aggregated into daily data. This transformation was made because predicting air quality daily provides more practical and meaningful insights for public health monitoring and policy decision-making compared to hourly predictions (Adani et al., 2020). The scale and ranges of variable are differed significantly in both locations' dataset. For instance, in Shah Alam, the range of PM2.5 levels ranges are very wide which is from 5.466 to

144.917, while the range of SO2 values falls between 0 and 0.001. Similarly, in Banting, PM2.5 levels also have a wide range, from 4.485 to 143.872, while SO_2 values remain very low, ranging from 0 to 0.005. Furthermore, PM2.5 and PM10 have high skewness values in both Shah Alam and Banting, indicating asymmetric data distributions. Min-max normalization was used to address these issues to ensure that each variable contributes equally to the analysis. This technique rescales the data to a standardised range, usually between 0 and 1, to make the variables more comparable. According to Stajkowski et al. (2020) in their studies on river water monitoring, standardization can increase algorithm efficiency, allowing for faster and more reliable model training by removing trends and variations.

Table 4.2
Descriptive Statistics of Independent Variables

	Variable	N	Mean	Median	Std. Dev.	Skewness	Min	Max
Shah Alam	PM2.5	1825	23.321	21.187	11.712	4.142	5.466	144.917
	PM10	1825	32.554	30.244	13.739	3.086	7.184	156.553
	SO2	1825	0.001	0.001	0.000	1.33	0.000	0.001
	NO2	1825	0.015	0.015	0.005	0.308	0.002	0.037
	O3	1825	0.02	0.019	0.007	0.8	0.002	0.06
	CO	1825	0.77	0.754	0.266	0.447	0.148	1.967
	WD	1825	206.597	205.583	48.507	0.106	86.023	317.928
	WS	1825	0.82	0.783	0.24	1.468	0.386	2.504
	Humidity	1825	80.138	80.109	6.603	-0.151	56.347	99.155
	Temperature	1825	27.552	27.573	1.247	-0.155	22.31	31.939
Banting	PM2.5	1825	21.236	18.648	12.157	3.653	4.485	143.872
	PM10	1825	30.433	28.016	14.137	3.165	8.363	159.434
	SO2	1825	0.001	0.001	0.001	1.377	0.000	0.005
	NO2	1825	0.009	0.009	0.003	0.335	0.002	0.021
	O3	1825	0.019	0.018	0.006	0.881	0.002	0.05
	CO	1825	0.596	0.575	0.201	0.883	0.158	1.766
	WD	1825	163.226	159.162	38.15	0.787	37.413	316.281
	WS	1825	1.026	0.985	0.329	1.967	0.277	3.227
	Humidity	1825	83.366	83.827	6.023	-2.678	0.000	98.205
	Temperature	1825	27.123	27.138	1.14	-0.265	18.452	32.813

Moreover, the Mahalanobis Distance was employed to detect outliers in both locations' dataset. The Mahalanobis distance plot in Figures 4.1 and 4.2 illustrates the presence of outliers in the dataset for Shah Alam and Banting dataset, respectively. The plot uses colour to differentiate between outliers and non-outliers, where blue dots represent non- outliers, while orange dots indicate outliers. The red dashed line shows the critical threshold distance, set at 18.307, which is calculated using a chi-squared distribution with a 0.05 significance level. Observations with distances above this line are classified as outliers, while those below it are considered as non-outliers.

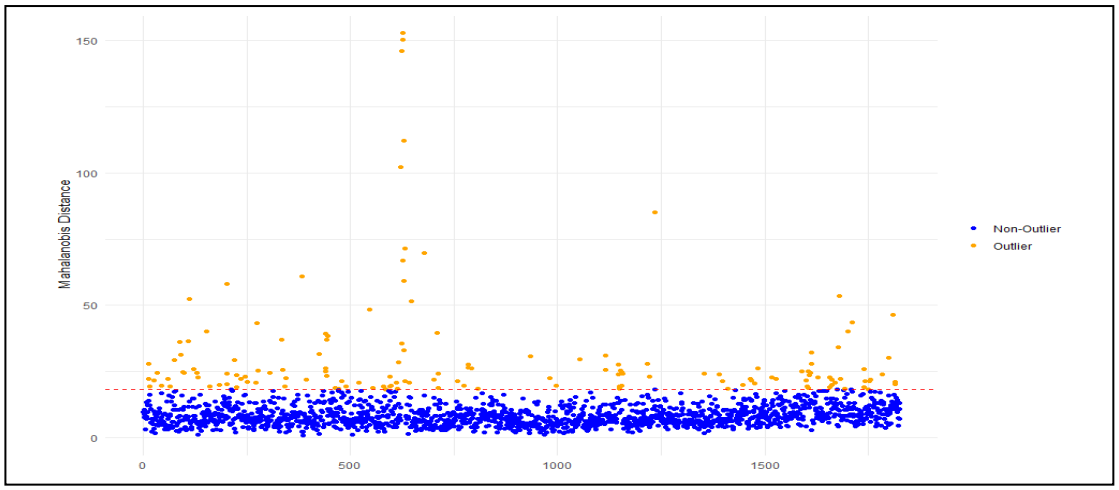

Figure 4.1 Scatter Plot of Mahalanobis Distance for Outlier Detection in Shah Alam

In Shah Alam, a total of 147 observations were identified as having Mahalanobis Distances exceeding the critical value of 18.307. Conversely, in Banting, the analysis yielded 135 observations with Mahalanobis Distances surpassing the same critical value. This represents approximately 8.06% and 7.39% of the total observations are identified as outliers in Shah Alam and Banting, respectively. This result emphasizes the variation in data characteristics between the two locations, indicating a different degree of outlier presence in Banting compared to Shah Alam. The higher occurrence of outliers in Shah Alam may be attributed to its status as an industrial and urban area, which experiences more pollution than Banting, a suburban area.

Although Mahalanobis Distance was applied to detect multivariate outliers, the identified extreme observations were not removed or transformed during data preprocessing. Recent air quality prediction frameworks have recommended careful consideration of seasonal and extreme data points and retaining them in model development after appropriate validation, rather than indiscriminate removal, to improve generalizability and real-world applicability (Dongre et al., 2025). Moreover, even though machine learning techniques can be employed for explicit outlier detection, these methods are conceptually distinct from ANN and RBFNN, which are used in this study as predictive and classification models rather than anomaly detection tools. Consequently, outlier identification and predictive modelling serve different methodological roles.

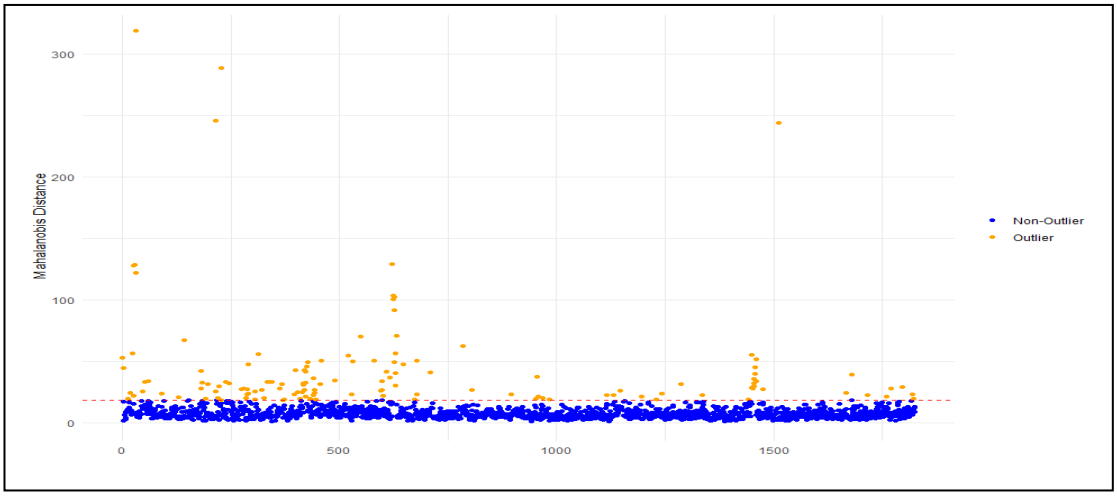

Figure 4.2 Scatter Plot of Mahalanobis Distance for Outlier Detection in Banting

Figures 4.3 and 4.4 are showing the distribution of polluted and not-polluted data in both locations. In Shah Alam, 1,676 observations (91.18%) are categorized as polluted, while 149 observations (8.2%) are categorized as not polluted. Meanwhile, in Banting, 1,692 observations (92.7%) are categorized as polluted, and 133 observations (7.3%) are categorized as not polluted. Based on the bar graphs, over 90% of the observations for both locations are non-polluted data, indicating a significant imbalance in the dataset. This imbalance in the data can have an impact on the effectiveness of the analysis. SMOTE oversampling technique can improve the classification accuracy of imbalanced datasets (Ghorbani & Ghousi, 2020). Hence, this study applied SMOTE

analysis to cater the imbalanced data issue due to extreme pollution events are less frequent in real data.

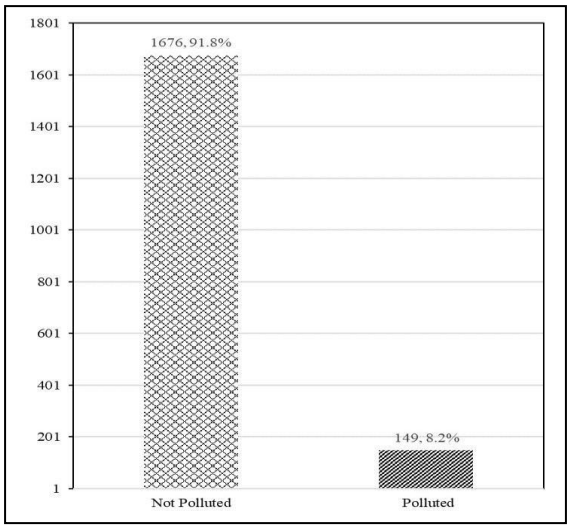

Figure 4.3 Air Quality Distribution of Shah Alam

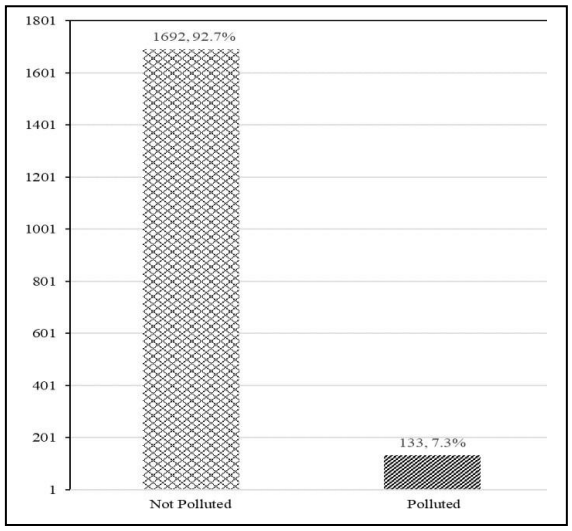

Figure 4.4 Air Quality Distribution of Banting

The descriptive statistics after data pre-processing for both locations are shown in Table 4.3. Now, the minimum and maximum values for all variables are scaled between 0 and 1. Furthermore, the mean and median values also within this range. For instance, in Shah Alam, the PM2.5 variable has a mean of 0.176 and a median of 0.144, indicating that the values are generally low within the standardized scale.

The skewness values, particularly for PM2.5 and PM10, are still positive and significant (greater than 1), indicating that despite the standardization, the data remains right skewed with a tail of higher values. This suggests that while the data has been scaled, the distribution's shape and the presence of outliers or extreme values have been preserved. Additionally, the number of observation (N), have increased due to SMOTE up-sampling which is 3315 and 3288 for Shah Alam and Banting dataset, respectively. The distribution of polluted and not-polluted data in each location seems more balance after employed SMOTE up- sampling as shown in Figures 4.5 and 4.6. The proportion of polluted cases increased to 49.4% and 48.5% of the data in Shah Alam and Banting, respectively.

Table 4.3
Descriptive Statistics After Data Pre-processing

	Variable	N	Mean	Median	Std. Dev	Skewness	Min	Max
Shah Alam	PM2.5	3315	0.176	0.144	0.142	3.239	0	1
	PM10	3315	0.219	0.19	0.146	2.82	0	1
	SO2	3315	0.227	0.215	0.1	1.332	0	1
	NO2	3315	0.426	0.423	0.165	0.279	0	1
	O3	3315	0.333	0.315	0.13	1.301	0	1
	CO	3315	0.395	0.388	0.165	0.512	0	1
	WD	3315	0.522	0.51	0.183	0.078	0	1
	WS	3315	0.214	0.201	0.1	1.174	0	1
	Humidity	3315	0.52	0.517	0.146	0.025	0	1
	Temperature	3315	0.559	0.565	0.118	-0.253	0	1
Banting	PM2.5	3288	0.172	0.136	0.143	2.778	0	1
	PM10	3288	0.201	0.167	0.152	2.597	0	1
	SO2	3288	0.278	0.255	0.137	1.314	0	1
	NO2	3288	0.433	0.435	0.17	0.105	0	1
	O3	3288	0.359	0.34	0.123	0.674	0	1
	CO	3288	0.319	0.307	0.139	0.764	0	1
	WD	3288	0.431	0.418	0.122	0.961	0	1
	WS	3288	0.24	0.229	0.094	2.104	0	1
	Humidity	3288	0.832	0.834	0.06	-1.616	0	1
	Temperature	3288	0.62	0.626	0.076	-0.476	0	1

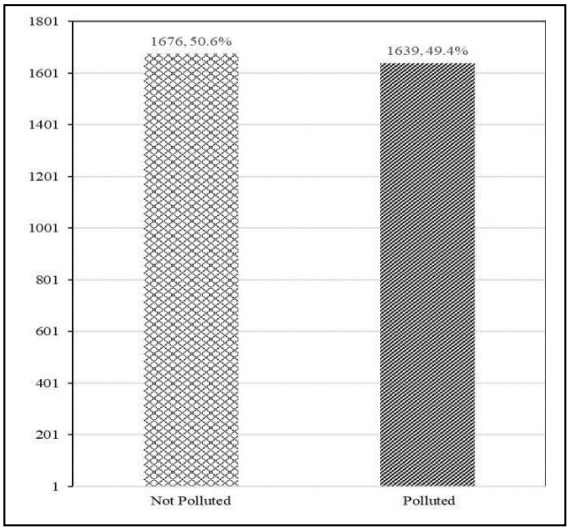

Figure 4.5 Air Quality Distribution of Shah Alam (After SMOTE)

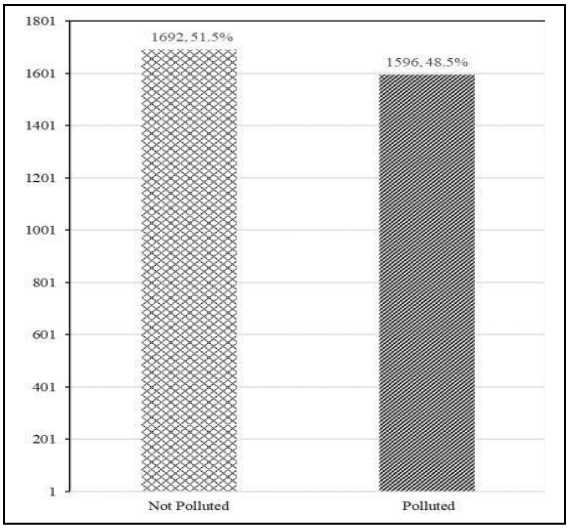

Figure 4.6 Air Quality Distribution of Banting (After SMOTE)

Before moving to the next analysis to achieve objective 1, this study examine the correlation between variables to identify any multicollinearity issues in the Shah Alam and Banting air quality data. Multicollinearity is a statistical phenomenon in which two or more independent variables in a regression model exhibit strong correlation, complicating the estimation of relationships between the variables and the dependent variable. This association has the potential to increase the variance of the coefficient estimates, making them unstable and challenging to understand. According to Shrestha (2020), this could jeopardize the validity of the statistical conclusions drawn from the model. The Spearman correlation coefficient is a nonparametric statistic that

evaluates the strength and direction of the monotonic association between two ranking variables. This correlation coefficient has an advantage over Pearson correlation coefficients because of its capacity to handle non-linear relationships and its robustness against violations of normality assumptions in the dataset. Furthermore, Spearman correlation coefficients also offer the advantage of being insensitive to outliers. Because it is based on ranks rather than raw data values, extreme values have less influence on the correlation result, making it more reliable for measuring correlations in datasets with anomalies. Hence, this study conduct Spearman correlation coefficient analysis to assess the correlation between independent variables in air quality data for both locations, Shah Alam and Banting.

Figures 4.7 and 4.8 show the Spearman correlation heatmap, which illustrates the relationships between independent variables for Shah Alam and Banting area, respectively. The heatmap's colour gradient ranges from red, indicating a strong negative correlation, to green, indicating a strong positive association. War and Barlis (2023) employed Dancey and Reidy's (2004) interpretation framework for Spearman's rho coefficient values, which categorizes relationships as very weak (0.01 to 0.2), weak (0.21 to 0.4), moderate (0.41 to 0.6), strong (0.61 to 0.8), and very strong (0.81 to 0.99). The figures reveal a strong positive link between PM2.5 and PM10, with correlation coefficients of 0.97 in Shah Alam and 0.98 in Banting. The correlation value for NO2 and CO in Shah Alam is 0.66, also suggesting a strong positive association. However, the correlation coefficient for NO2 and CO in Banting is only 0.4, suggesting a weak correlation between these variables. Besides, NO2 has moderate correlation with PM10 and PM2.5 for both locations. Furthermore, temperature and humidity in both Shah Alam and Banting have significant strong negative correlations, with values of -0.8 and -0.74, demonstrating an inverse association between these environmental factors. Moreover, humidity in Banting has moderate correlation with both particulate matter (PM10 and PM2.5) with correlation of -0.49 for both correlations. Similarly, temperature in Banting also has moderate correlation with both particulate matter (PM10 and PM2.5) with correlation of 0.44 for both correlations. The spearman correlation heatmap also shows a strong positive correlation between CO and particulate matter, PM2.5 and PM10 in Banting, but a moderate correlation in Shah Alam. In addition, ther variables exhibit weak correlations with particulate matter and other pollutants in both locations. The spearman correlation matrix shows that NO2 and

temperature has the lowest value in Shah Alam which is 0. This indicates that there is no correlation between NO2 and temperature in Shah Alam. However, the lowest Spearman correlation value in Banting is 0.01, observed between O3 and CO (0.01) as well as between O3 and SO2 (0.01).

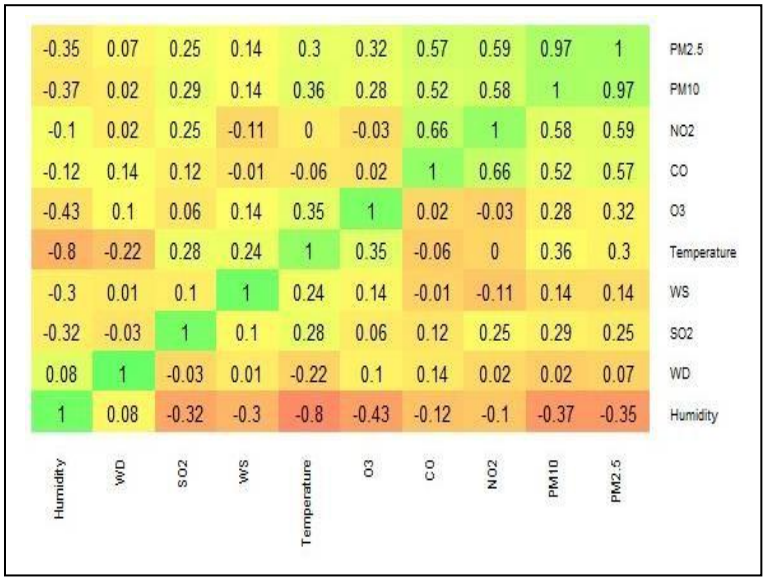

Figure 4.7 Spearman Correlation Heatmap of Shah Alam

Figure 4.8 Spearman Correlation Heatmap of Banting

In this study, the AdjcorT feature selection method is employed to carefully account for the high correlations observed between variables, particularly between PM2.5 and PM10, as well as other pollutants and meteorological factors, such as the correlations between NO2 and CO, and between temperature and humidity. Due to the significant positive correlations across these variables, AdjcorT is able to mitigate multicollinearity issues in the air quality data. AdjcorT enables the selection of features that are truly significant for predicting PM2.5 levels by evaluating the degree of correlation between variables, while avoiding redundant or highly correlated inputs. This ensures that the model focuses on variables that provide important information, ultimately improving the accuracy and reliability of the air quality classifications in Shah Alam and Banting. Thus, the dataset is now well prepared for the next analysis.

4.3 Feature Ranking Using AdjcorT

After the dataset has been go through data pre-processing, the AdjcorT feature selection were employed to rank the features. Figures 4.9 and 4.10 shows bar charts of the AdjcorT values for the independent variables for Shah Alam and Banting datasets, respectively. Based on the bar charts, NO2 is the most important variable to predict air quality in Shah Alam with AdjcorT value of 23.81, followed by PM2.5, PM10, and CO with AdjcorT values greater than 20. This indicates that these pollutants play a significant role in air quality forecasting in the region. Similarly, the same four pollutant were determined as the most influential for air quality classification in Banting. However, the order of importance differs, with PM10 being the most significant with AdjcorT value of 25.23, followed by PM2.5, CO, and NO2. Additionally, the least important features for predicting air quality differ between the two locations. In Shah Alam, temperature was chosen as the least significant variable with AdjcorT value of 14.99. However, in Banting, the AdjcorT method identifies O3 as the least important predictor with AdjcorT value of 16.49. These rankings reflect spatial differences in pollutant influence between the two monitoring stations.

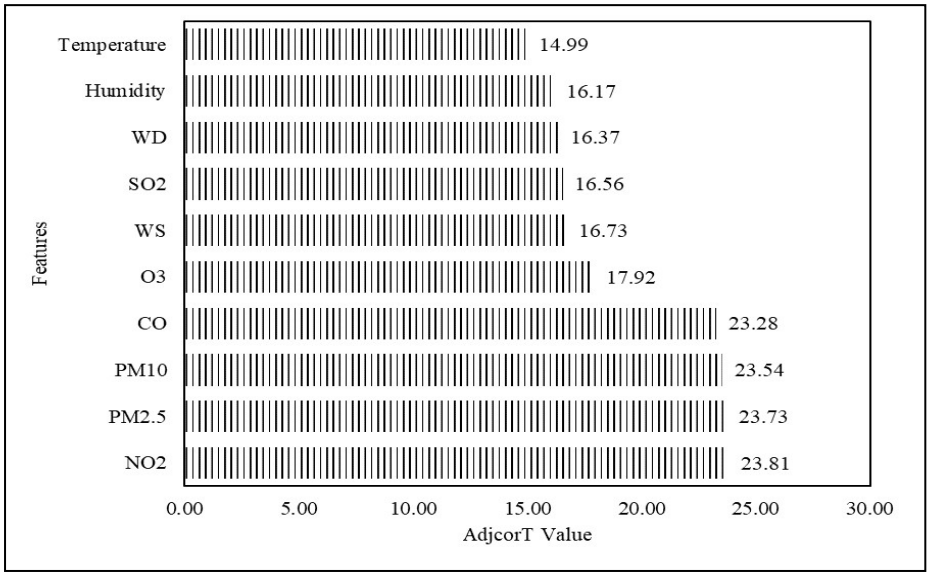

Figure 4.9 AdjcorT Values for Features in Shah Alam

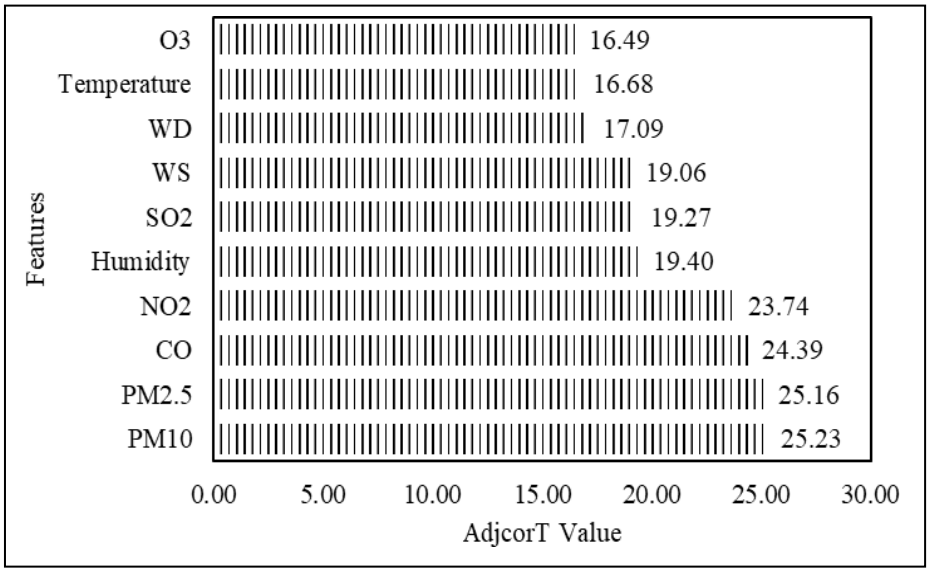

Figure 4.10 AdjcorT Values for Features in Banting

The AdjcorT ranking was subsequently used to evaluate model performance across progressively increasing feature combinations in order to determine the optimal number of features to be retained. The ANN model was run for 10 times for each location's dataset to find the best feature combination. Table 4.4 presents the performance metrics of models constructed using the top n features ranked by the AdjcorT method. Column 1 represents the model using only the highest-ranked feature (NO2 for Shah Alam), while Column 2 represents the model using the top two ranked features (NO2 and PM2.5 for Shah Alam), and so on, with each subsequent column incorporating an additional feature based on the AdjcorT ranking until all 10 variables

were included in the model. Six performance matrices, which are accuracy, sensitivity, specificity, precision, f1 score and AUROC was used to determine the best feature combinations. The features combination is based on AdjcorT's rank for each location. Based on Table 4.4, combinations with 8 features were chosen as the best model for Shah Alam dataset with highest accuracy (0.81), sensitivity (0.87), specificity (0.76), precision (0.76), F1 score (0.81), and AUROC (0.81). Similarly, in Banting, 8 features combination were chosen as the best model with four out of six performance metrics has highest values which are accuracy (0.85), sensitivity (0.91), F1 Score (0.85), and AUROC (0.85).

Table 4.4
ANN Model Performances for Different Numbers of Features Combination

	No of Features	1	2	3	4	5	6	7	8	9	10
Shah Alam	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
Banting	Accuracy	0.74	0.76	0.77	0.77	0.81	0.81	0.82	**0.85**	0.84	0.82
	Sensitivity	0.74	0.77	0.78	0.76	0.86	0.81	0.89	**0.91**	0.89	0.87
	Specificity	0.73	0.76	0.75	0.78	0.77	**0.81**	0.77	0.80	0.81	0.77
	Precision	0.73	0.76	0.75	0.78	0.77	**0.81**	0.77	0.80	0.81	0.77
	F1 Score	0.74	0.76	0.77	0.77	0.81	0.81	0.83	**0.85**	0.85	0.82
	AUROC	0.74	0.76	0.77	0.77	0.81	0.81	0.83	**0.85**	0.85	0.82

4.4 Feature Selection

In this section, this study presents the eight features that were selected and ranked by four feature selection method, ReliefF, mRMR, Lasso and AdjcorT. The selected eight features are subsequently used for integration with the RBFNN model in the next analysis. Table 4.5 shows eight features ranking for four feature selection method which is AdjcorT, Lasso, mRMR, and ReliefF. The table shows how feature importance varies significantly depending on the method used. PM10 was consistently ranked among the top three features to predict air quality in Shah Alam by all methods except for ReliefF. ReliefF method did not select PM10 among the top eight most important features in Shah Alam. Moreover, PM2.5 in Shah Alam was placed in top

two by both AdjcorT and Lasso method. However, ReliefF ranked PM2.5 as the sixth most important variable out of eight features. Conversely, mRMR did not include PM2.5 among the top eight most important features in Shah Alam. This may be because the mRMR method identified PM2.5 as having a high correlation with PM10 and excluded it, as the method aims to minimize redundancy among features. These findings in an industrial area align with previous studies that identify PM2.5 and PM10 as critical factors influencing air quality levels. According to a study by Cheng & Tsai (2022), they emphasized the important of PM2.5 and PM10 in predicting pollution levels in industrial area, consistent with its high ranking in Shah Alam by AdjcorT and Lasso method. However, exclusion of PM2.5 and PM10 from the top eight features by mRMR and ReliefF model, respectively, are contrasts with findings by (Cheng & Tsai, 2022) where PM2.5 and PM10 was identified a major pollutant that affect air quality. Although PM2.5 and PM10 are known to be highly correlated, their simultaneous selection by AdjcorT and Lasso does not indicate that multicollinearity was ignored. Instead, this reflects the different mechanisms by which feature selection methods handle correlated predictors. While mRMR explicitly removes redundant variables to minimise correlation, AdjcorT and Lasso manage multicollinearity by ranking and regularising correlated predictors rather than directly omitting them. The Lasso method mitigates multicollinearity through an L1 regularisation penalty that shrinks the coefficients of correlated variables and retains those with stronger predictive contributions. Similarly, AdjcorT explicitly incorporates both positive and negative correlation structures when ranking features, allowing correlated variables to be retained if they provide additional discriminatory information. Consequently, the inclusion of both PM2.5 and PM10 suggests that, despite their high correlation, each variable contributes complementary information for air quality classification in Shah Alam within the nonlinear modelling framework adopted in this study.

Furthermore, in Shah Alam, wind direction was identified as the least important feature by AdjcorT, Lasso and mRMR. In contrast, ReliefF's result shows that wind direction is the most important feature to predict air quality in Shah Alam. This contradicts result illustrates how ReliefF's algorithm doesn't account for correlation, prioritizes wind direction differently. In addition, AdjcorT were the only methods that did not include humidity and temperature among the top eight

features for predicting air quality in Shah Alam. This finding may be explained by the Spearman correlation analysis, which showed that humidity and temperature have a strong negative correlation. Among the feature selection methods used, only AdjcorT explicitly accounts for both strong positive and negative correlations between variables. As a result, AdjcorT may have excluded these two variables because they likely share similar information and retaining both may not add additional value to the model.

Table 4.5
Feature Ranking Across Feature Selection Methods

	Model	AdjcorT	Lasso	mRMR	ReliefF
Shah Alam	**PM2.5**	**2**	**1**		**6**
	PM10	3	2	1	
	SO2	7	6		
	NO2	1	4	6	2
	O3	5		3	5
	CO	4		7	3
	WD	8	8	8	1
	WS	6	7	2	7
	Humidity		3	5	4
	Temperature		5	4	8
Banting	PM2.5	2	2		7
	PM10	1	5		6
	SO2	6		1	8
	NO2	4	6	6	1
	O3			4	4
	CO	3	8	8	2
	WD	8	3	7	3
	WS	7	7	2	5
	Humidity	5	1	5	
	Temperature		4	3	

In Banting, the AdjcorT, Lasso and ReliefF methods include PM2.5 and PM10 among top eight important features. However, the mRMR method excludes PM2.5 and PM10 as the top eight features to predict air quality in Banting. This might happen because both PM2.5 and PM10 has high correlations between them but small contribution to predict air quality in suburban area, thus mMRM method drop both particulate matter as significant predictor. Furthermore, both AdjcorT and Lasso method exclude O3 in the model, while mRMR and ReliefF ranked it as fourth

important features. Moreover, Lasso and ReliefF method identify wind direction as top three important feature to classify air quality in Banting. This result is consistent with a study by Ling et al. (2020b), where wind direction is the major reason Banting were polluted due to its location in between the most polluted in Klang Valley, even though Banting has a high percentage of green land use and less industry or transportation land use. Additionally, ReliefF were the only methods that did not include humidity and temperature among the top eight features for air quality classification in Banting. In contrast, Lasso rank humidity and temperature with higher rank indicates that these two meteorological factors are important to classify air quality in Banting. Similarly, Bilal et al. (2019) highlighted that 72% of the changes in daily mean PM2.5 concentrations could be explained by changes in meteorological factors in suburban area. This indicates that Lasso has better selection on important features in suburban area compared to other method.

The feature rankings reflect real air quality conditions as they are derived from ground-based pollutant and meteorological measurements and align with known emission sources and dispersion characteristics of the study areas. The prominence of particulate matter and NO2 in industrial Shah Alam and the stronger influence of wind and meteorological factors in suburban Banting are consistent with established atmospheric behaviour. Despite some variation across feature selection methods, the repeated identification of key variables and their agreement with recent empirical studies indicate that the mathematical rankings capture meaningful real-world air pollution patterns (Bilal et al., 2019; Ling et al., 2020b; Cheng & Tsai, 2022).

4.5 Model Performances

Next, the performance of the original RBFNN model, which used all ten input variables, was compared with four two-stage models: ReliefF-RBFNN, mRMR-RBFNN, Lasso-RBFNN, and AdjcorT-RBFNN. These two-stage models were constructed using only eight input features, as the feature combination analysis indicated that eight variables were sufficient to classify PM2.5 levels accurately in both Shah Alam and Banting. Table 4.6 was included to address Objective 2, which is to compare the classification performance of models that incorporate existing feature selection methods with the proposed two-stage method, AdjcorT RBFNN, using real-world air quality data. The existing RBFNN model was used as a reference point to

evaluate the effect of applying feature selection techniques. By comparing the full RBFNN model, which uses all ten variables, with models that apply selected features, this study aims to determine whether applying feature selection, particularly those that consider the correlation between variables, can improve predictive accuracy.

Table 4.6
Comparison Model Performance

	Model	RBFNN	AdjcorT-RBFNN	Lasso-RBFNN	mRMR-RBFNN	ReliefF-RBFNN
Shah Alam	Accuracy	0.753	**0.759**	0.735	0.753	0.751
	Sensitivity	0.801	0.802	0.771	**0.818**	0.813
	Specificity	0.712	**0.721**	0.702	0.703	0.703
	Precision	0.712	**0.721**	0.702	0.703	0.703
	F1 Score	0.754	**0.759**	0.735	0.756	0.754
	AUROC	0.755	**0.761**	0.736	0.756	0.754
Banting	Accuracy	0.758	0.761	**0.771**	0.725	0.568
	Sensitivity	0.770	**0.770**	0.765	0.702	0.584
	Specificity	0.746	0.752	**0.778**	0.761	0.551
	Precision	0.746	0.752	**0.778**	0.761	0.551
	F1 Score	0.758	0.761	**0.771**	0.730	0.567
	AUROC	0.758	0.761	**0.769**	0.721	0.567

The computational time for each model varied significantly. ReliefF-RBFNN was the fastest, completing in 2.34 seconds, followed by RBFNN at 2.38 seconds, AdjcorT-RBFNN at 2.64 seconds, and mRMR- RBFNN at 2.91 seconds. Additionally, Lasso-RBFNN required the longest time at 6.75 seconds. These differences in computational time indicate that some models, like Lasso-RBFNN, required significantly more computational time to achieve its results. Despite the differences in computational efficiency, performance metrics reveal key insights. Based on Table 4.6, the AdjcorT-RBFNN method has the highest accuracy (0.759), specificity (0.721), precision (0.721), F1 Score (0.759), and AUROC (0.761). This indicates that this study proposed model, AdjcorT-RBFNN is the best model to predict the next day air quality in Shah Alam compared to other methods. This model correctly classifies approximately 75.9% of the instances in the dataset.

Moreover, the proposed model correctly identifies 72.1% of the true negative cases (not polluted). The proposed model also correctly identified 72.1% of the cases as polluted. The F1 score (0.759), indicates that the model maintains a good balance between precision and recall. Finally, the AUROC value of 0.761 shows that the model is capable of differentiate between polluted and non-polluted characteristics. Moreover, the Lasso-RBFNN model showed the lowest performance compared to the others, with the lowest accuracy, sensitivity, specificity, precision, F1 score, and AUROC which are 0.735, 0.771, 0.702, 0.702, 0.735, and 0.736, respectively. Additionally, in Banting, AdjcorT-RBFNN model correctly identified 77% of the actual polluted instances in the dataset, which is the highest sensitivity compared to other models. However, in contrast, the Lasso-RBFNN model was identified as the best model to predict the next day air quality in Banting, as it achieved the highest values across five out of six performance metrics which are accuracy (0.771), specificity (0.778), precision (0.778), F1 score (0.771), and AUROC (0.769).

In addition, ReliefF-RBFNN model is the worst model to classify the next day PM2.5 level in Banting with values below 0.6 for all performance metrics. This poor performance may be attributed to the nature of the ReliefF algorithm, which does not account for correlations between variables. As a result, it may have failed to capture important relationships or redundancies among highly correlated air quality variables, leading to less informative feature selection and weaker predictive accuracy. Similar observations were reported by Yadav et al. (2025), who found that feature subsets selected using Relief-based methods can reduce the generalisation capability of RBFNN models when informative correlated variables are removed.

4.6 Connection Weight Analysis

Based on the previous analysis, AdjcorT-RBFNN model and Lasso-RBFNN model have shown better performance than RBFNN model for air quality data in Shah Alam and Banting, respectively. This indicates that AdjcorT and Lasso method have significantly improved RBFNN model's accuracy by mitigating the high correlation between variables. Thus, the significant features to classify next-day PM2.5 levels in Shah Alam are NO2, PM2.5, PM10, CO, O3, wind speed, SO2, and wind direction. In Banting, the significant features to classify next-day PM2.5 levels are humidity, PM2.5, wind direction, temperature, PM10, NO2, wind speed and CO.

In this section, a connection weight analysis is conducted as a post-hoc interpretability approach to examine the relative contribution of each selected feature to the RBFNN-based classification models. It is important to clarify that individual connection weights in neural networks do not represent direct or causal importance. Instead, they reflect how input variables interact with hidden neurons to influence the model output in a nonlinear manner. Therefore, interpretation is not performed at the level of individual hidden-node weights, but through aggregated measures derived from these weights. The connection weight approach is performed to rank feature importance and to enable a comparative analysis between the baseline RBFNN and the proposed two-stage feature selection framework. This analysis is motivated by prior findings indicating that RBFNN performance is sensitive to the choice and number of input features (Yadav et al., 2025). By examining how feature importance rankings differ across models, this study provides insight into how correlation-based feature selection influences RBFNN behaviour and predictive performance.

Next, the feature importance of the RBFNN model is compared with the best-performing two-stage model for each location, namely AdjcorT-RBFNN for Shah Alam and Lasso-RBFNN for Banting, using the connection weight approach. Moreover, the features are ranked again based on the connection weight to identify the feature importance towards output of the model. Santos et al. (2018b) suggests that an iterative process for ranking variables could be advantageous, as it enables variables to be assessed independently of the score of the most significant variable. The primary purpose of applying the connection weight approach in this study is to provide a consistent and model-based method for ranking variables according to their contribution to classification performance, rather than to explain the real-world processes that cause PM2.5 pollution. Accordingly, for the Banting dataset, the connection weight analysis focuses on the Lasso-RBFNN model, as it achieved superior classification performance compared to AdjcorT-RBFNN, ensuring that feature importance interpretation is based on the most effective two-stage model for each location.

Tables 4.7 and 4.8 display the connection weight products of the RBFNN and AdjcorT- RBFNN models for the Shah Alam dataset. The table shows the influence of various input features on the hidden neurons in each model. RBFNN model has 7 hidden nodes, while AdjcorT-RBFNN model has 6 hidden nodes because RBFNN model consists of ten features, meanwhile AdjcorT-RBFNN model consists of eight features.

According to Ul Saufie et.al. (2022), the formula to set up number of hidden nodes = (number of attributes + number of classes) / 2 + 1. The connection weight products represent the strength of interaction between each input variable and the hidden neurons. Larger absolute values indicate stronger interactions; however, these values are not interpreted individually. Instead, they are aggregated across all hidden nodes to compute the overall contribution of each feature to the model output.

Table 4.7
Connection Weight Product of RBFNN in Shah Alam

Input	Hidden 1	Hidden 2	Hidden 3	Hidden 4	Hidden 5	Hidden 6	Hidden 7
PM2.5	-3.63	-33.59	-3.44	83.42	12.82	4.82	27.64
PM10	-8.86	-45.95	-12.42	73.39	11.96	2.22	29.91
SO2	-33.78	-30.46	-14.92	56.91	26.86	12.49	8.31
NO2	-40.09	-42.26	-30.33	102.82	14.29	30.90	21.08
O3	-5.97	-78.66	-13.74	88.40	30.95	16.20	17.63
CO	-43.79	-56.92	-21.32	98.97	22.31	28.35	26.03
WD	-64.44	-77.30	-19.77	94.72	84.99	26.69	7.09
WS	-7.88	-16.97	-30.53	44.21	42.46	4.15	13.17
Humidity	-60.19	-39.81	-80.48	115.08	49.11	53.55	9.84
Temperature	-30.42	-86.47	-28.49	80.91	40.51	25.19	29.01

Table 4.8
Connection Weight Product of AdjcorT-RBFNN in Shah Alam

Input	Hidden 1	Hidden 2	Hidden 3	Hidden 4	Hidden 5	Hidden 6
NO2	41.60	-4.45	-46.78	-53.83	48.08	53.18
PM2.5	6.01	-4.23	-19.95	-25.23	15.37	69.24
PM10	7.26	-7.30	-28.98	-37.97	6.85	71.15
CO	39.65	-4.59	-43.71	-64.00	31.02	64.64
O3	9.05	-14.73	-36.46	-53.10	78.40	46.40
WS	10.24	-11.80	-15.21	-52.54	59.71	34.93
SO2	36.96	-21.71	-40.16	-42.71	48.91	22.49
WD	63.06	-11.97	-41.63	-81.92	74.61	17.35

In the RBFNN model for Shah Alam, variables such as NO2, humidity, wind direction (WD), and CO exhibit larger cumulative connection weight magnitudes across hidden nodes, particularly in Hidden Nodes 4 and 5. This indicates that these variables contribute more strongly to the model output relative to other predictors. In contrast, PM2.5, wind speed (WS), and SO2 show smaller cumulative contributions across most nodes, suggesting lower relative importance within the model structure.

For the AdjcorT-RBFNN model, NO2, CO, and wind direction (WD) emerge as the most influential variables after aggregation, with consistently stronger connection weight contributions across multiple hidden nodes. Variables such as PM2.5 and PM10 exhibit lower aggregated weights, indicating that their relative contribution to classification performance is reduced once multicollinearity is addressed through AdjcorT. This shift reflects a refinement in feature ranking rather than a loss of predictive relevance.

Then, this study computed the relative and normalized importance to rank the features based on connection weight, as shown in Table 4.9 for the Shah Alam dataset. Relative importance values were obtained by summing the absolute connection weight products across all hidden nodes for each input variable, while normalized importance values were calculated to enable comparison on a common scale. These normalized values are used exclusively for ranking purposes.

Based on Table 4.9, both the RBFNN and AdjcorT-RBFNN models ranked PM2.5, NO2, and O3 as the most important features for classifying PM2.5 levels in Shah Alam. However, the AdjcorT- RBFNN model shows that NO2 and O3 have a greater influence on PM2.5 levels, with normalized importance values of 0.91 and 0.69, respectively, compared to the RBFNN model, where these features have normalized importance values of 0.49 and 0.47. This difference indicates that, although feature rankings may appear similar across models, the magnitude of contribution differs substantially once multicollinearity is mitigated.

Table 4.9
Relative Importance Based on Connection Weight Approach for Shah Alam Dataset

Model	Input	Relative Importance	Normalized Importance	Rank
RBFNN	PM2.5	88.04	1.00	1
	PM10	50.24	0.40	6
	SO2	25.42	0.00	10
	NO2	56.41	0.49	2
	O3	54.81	0.47	3
	CO	53.64	0.45	4
	WD	51.97	0.42	5
	WS	48.61	0.37	7
	Humidity	47.10	0.35	8
	Temperature	30.25	0.08	9
AdjcorT-RBFNN	NO2	37.81	0.91	2
	PM2.5	41.20	1.00	1
	PM10	11.01	0.19	7
	CO	23.02	0.51	5
	O3	29.56	0.69	3
	WS	25.34	0.58	4
	SO2	3.78	0.00	8
	WD	19.49	0.42	6

Moreover, PM10 was ranked sixth in the RBFNN model, with a normalized importance value of 0.40, while the AdjcorT-RBFNN model ranked PM10 in seventh place out of eight features, with a normalized importance value of 0.19, despite its high correlation with PM2.5. This suggests that multicollinearity inflated the apparent importance of PM10 in the RBFNN model, whereas AdjcorT-RBFNN provides a more conservative and stable ranking. Furthermore, SO2 was consistently ranked as the least important feature by both models. Although PM10 was ranked highly during the initial AdjcorT feature selection stage, its lower ranking after integration with the RBFNN model indicates that statistical relevance does not necessarily translate into strong nonlinear contribution when interactions among predictors are considered.

Tables 4.10 and 4.11 present the connection weight products of the RBFNN and Lasso-RBFNN models for the Banting dataset. In the RBFNN model, PM2.5, PM10, and NO2 exhibit relatively high connection weight magnitudes across several hidden nodes, particularly Nodes 4, 5, and 7, indicating their strong contribution to the classification output. In contrast, variables such as SO2, CO, and wind speed display

smaller connection weight values across most hidden nodes, suggesting lower relative importance.

Table 4.10
Connection Weight Product of RBFNN in Banting

Input	Hidden 1	Hidden 2	Hidden 3	Hidden 4	Hidden 5	Hidden 6	Hidden 7
PM2.5	-0.17	-2.79	2.35	8.12	0.90	-0.46	20.59
PM10	-0.19	-4.23	2.32	10.08	0.98	-0.40	20.65
SO2	-0.18	-4.10	3.62	8.86	3.28	-2.41	6.89
NO2	-0.31	-12.36	7.35	17.62	4.83	-1.83	14.69
O3	-0.19	-8.80	7.45	11.80	2.58	-2.08	10.85
CO	-0.19	-6.15	3.55	8.88	1.71	-0.72	16.20
WD	-0.22	-21.05	4.98	15.28	3.38	-2.97	11.45
WS	-0.11	-9.89	2.88	8.51	1.25	-3.13	2.72
Humidity	-0.45	-23.57	10.83	31.48	4.50	-9.54	17.76
Temperature	-0.39	-17.82	9.44	26.25	4.28	-6.67	13.94

Table 4.11
Connection Weight Product of Lasso-RBFNN in Banting

Input	Hidden 1	Hidden 2	Hidden 3	Hidden 4	Hidden 5	Hidden 6
Humidity	45.31	-9.24	16.03	8.89	-52.21	9.16
PM2.5	17.50	-1.70	3.65	0.45	-5.01	10.02
WD	24.73	-3.01	6.34	5.97	-37.80	5.83
Temperature	40.26	-6.10	10.44	5.76	-27.47	6.72
PM10	19.60	-1.95	3.90	0.75	-6.13	10.20
NO2	32.02	-6.52	8.86	3.70	-18.44	7.06
WS	11.73	-2.11	2.85	3.77	-8.88	1.11
CO	19.80	-1.81	6.87	0.86	-13.67	8.07

After applying Lasso feature selection, the Lasso-RBFNN model shows a redistribution of feature importance. Variables such as humidity, temperature, NO2, and wind direction exhibit stronger aggregated connection weight contributions, while PM2.5 and PM10 show reduced weights compared to the RBFNN model. This shift indicates that Lasso effectively suppresses redundant predictors by penalising correlated variables, leading to a refined ranking of feature importance.

Next, this study computed the relative and normalized importance values to rank the features for the Banting dataset, as shown in Table 4.12. Both the RBFNN and Lasso-RBFNN models identified meteorological parameters as the most significant

features for classifying PM2.5 levels in Banting. While the RBFNN model ranked humidity as the most important factor, the Lasso-RBFNN model identified temperature as the top feature. Additionally, both models ranked NO2 and PM10 as the second and third most important features.

Table 4.12
Relative Importance Based on Connection Weight Approach for Banting Dataset

Model	Input	Relative Importance	Normalized Importance	Rank
RBFNN	PM2.5	28.54	0.91	5
	PM10	29.22	0.94	3
	SO2	15.95	0.48	8
	NO2	29.98	0.96	2
	O3	21.62	0.67	7
	CO	23.27	0.73	6
	WD	10.85	0.30	9
	WS	2.24	0.00	10
	Humidity	31.01	1.00	1
	Temperature	29.02	0.93	4
Lasso-RBFNN	Humidity	17.94	0.58	6
	PM2.5	24.91	0.83	4
	WD	2.07	0.00	8
	Temperature	29.60	1.00	1
	PM10	26.37	0.88	3
	NO2	26.67	0.89	2
	WS	8.46	0.23	7
	CO	20.12	0.66	5

Although PM2.5 was ranked fifth in the RBFNN model and fourth in the Lasso-RBFNN model, the normalized importance values (0.91 and 0.83, respectively) indicate that its contribution to classification performance remained relatively similar. Overall, despite Lasso initially prioritising meteorological variables during feature selection, the Lasso-RBFNN results demonstrate that pollutant-related variables such as PM10, NO2, PM2.5, and CO still play a substantial role in PM2.5 classification once nonlinear interactions are considered. This finding differs from some previous studies that suggest meteorological factors dominate PM2.5 variability in suburban areas.

4.7 Simulation of AdjcorT-RBFNN and RBFNN Model

In this section, the Monte Carlo simulation is designed to evaluate the performance of the final AdjcorT-RBFNN model under controlled multicollinearity conditions, rather than to replicate the full feature selection optimization process applied to the real datasets. In the simulation, variables x1 and x2 were predefined as true informative predictors, while x3 to x10 were treated as non-informative noise variables. The correlation between x1 and x2 was systematically varied from strong negative to strong positive ($\rho = -0.8$ to 0.8) to assess the model's ability to correctly identify important features under varying multicollinearity scenarios. Sample sizes of n = 50, 100, 150, and 200 were also considered to examine the impact of dataset size on model performance.

For each iteration, the AdjcorT method ranked the variables based on adjusted t-statistics while accounting for correlations, and the top-ranked variables were used as inputs to the RBFNN classifier. By fixing the true informative variables, the simulation focuses on assessing selection accuracy and resulting classification performance rather than searching for the optimal feature subset. Results were averaged over 100 iterations per scenario and are reported in Table 4.13 and Figures 4.11–4.16. Table 4.13 shows the frequency with which x1 and x2 were correctly identified as top predictors, while the figures illustrate the corresponding classification metrics (accuracy, sensitivity, specificity, precision, F1-score, and AUROC) for both the proposed AdjcorT-RBFNN model and the existing RBFNN model. This approach allows a clear evaluation of how multicollinearity and sample size affect feature selection and classification performance.

Based on Table 4.13, the correlation between x1 and x2 varies from strong negative correlation ($\rho = -0.8$) to strong positive correlation ($\rho = 0.8$), and the sample sizes are 50, 100, 150 and 150. Thus, a total of 100 iterations for all 28 scenarios were considered in this study. When the correlation between variables is strong ($\rho = \pm 0.8$) and moderate ($\rho = \pm 0.5$), either positive or negative, x1 and x2 are almost always selected together across all sample sizes with more than 86% of probabilities. This is indicated that the model consistently selected x1 and x2 variables across all sample sizes. For example, at $\rho = 0.8$, both x1 and x2 were selected 88 times when N=50, 95 times when N=100, 100 times when N=150, and 100 times when N=200. This suggest

that the model correctly identifies important predictors when the correlation between variables is strong and moderate, regardless positive or negative correlation.

Additionally, the frequency of x1 and x2 was correctly selected by AdjcorT-RBFNN model becomes less frequent when the correlation between variables is weak ($\rho = \pm 0.2$). However, as the sample size increases, the frequency of correctly identifying x1 and x2 as important predictors also increases, though it decreases slightly when the sample size reaches 200. For instance, at $\rho = -0.2$, both x1 and x2 were selected 67 times when N=50, 84 times when N=100, 94 times when N=150, and 93 times when N=200. Furthermore, when there is no correlation ($\rho = 0$) between variables, the frequency of AdjcorT-RBFNN method successfully select x1 and x2 as important variable are very low. This suggested that this method would not perform well if there were no correlation between variables.

Table 4.13
Simulation of AdjcorT-RBFNN Correctly Select x1 and x2 as Top 2 Features

Correlation Between x1 & x2	**Selected Variable**	**N=50**	**N=100**	**N = 150**	**N = 200**
$\rho = -0.8$	x1 & x2	88	95	100	100
	x1 & others				
	x2 & others	12	5		
	Neither x1 nor x2				
	No selected variable				
$\rho = -0.5$	x1 & x2	86	95	100	100
	x1 & others				
	x2 & others	14	5		
	Neither x1 nor x2				
	No selected variable				
$\rho = -0.2$	x1 & x2	67	84	94	93
	x1 & others				
	x2 & others	33	16	6	7
	Neither x1 nor x2				
	No selected variable				
$\rho = 0$	x1 & x2	30	13	36	21
	x1 & others				
	x2 & others	70	87	64	79
	Neither x1 nor x2				
	No selected variable				
$\rho = 0.2$	x1 & x2	62	59	94	93

	x1 & others				
	x2 & others	38	41	6	7
	Neither x1 nor x2				
	No selected variable				
ρ = 0.5	x1 & x2	89	94	100	100
	x1 & others				
	x2 & others	11	6		
	Neither x1 nor x2				
	No selected variable				
ρ = 0.8	x1 & x2	88	95	100	100
	x1 & others				
	x2 & others	12	5		
	Neither x1 nor x2				
	No selected variable				

Figures 4.11(a)–4.16(a) and Figures 4.11(b)–4.16(b) show the performance metrics of the simulation, where the proposed AdjcorT-RBFNN model is compared with the existing RBFNN model. The input layers in RBFNN model has 10 nodes which is all the 10 independent variables. However, the AdjcorT-RBFNN model has 8 nodes in the input layers, which is the top 8 variables were selected by AdjcorT method.

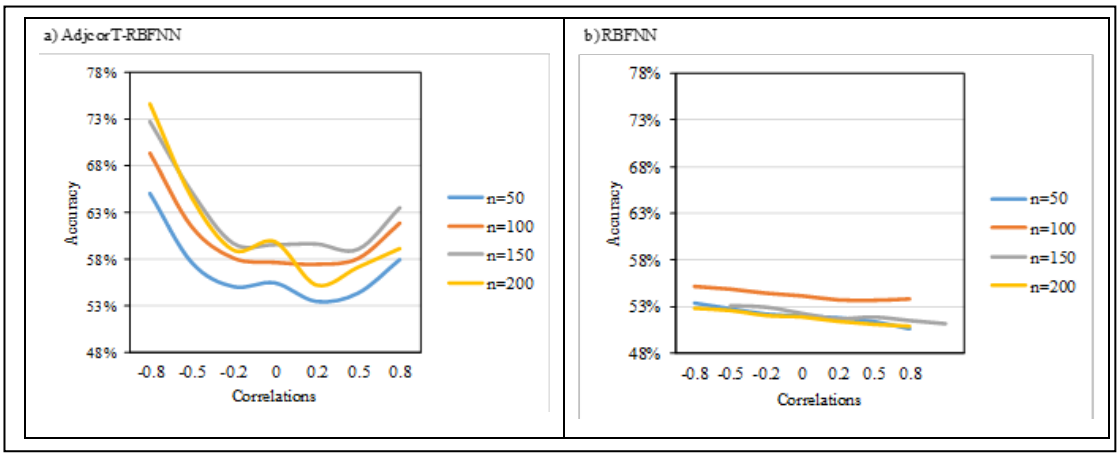

Figure 4.11 (a) Accuracy of AdjcorT-RBFNN, (b) Accuracy of RBFNN.

Based on Figure 4.11(a), the accuracy of AdjcorT-RBFNN model is highest when the variables have strong negative correlation (ρ = -0.8) across all sample size. The accuracy of proposed model decreases significantly as the negative correlation between the variables decreases and moves closer to zero. As the correlation becomes positive, the accuracy gradually improves. However, the accuracy does not return to the

high levels compared to negative correlation. Furthermore, the line graph of AdjcorT-RBFNN model also suggest that larger sample sizes generally lead to better accuracy regardless the correlation. Hence, the model's accuracy improves with more sample, as seen by the curves for N = 150 and N = 200, which are consistently higher than those for N = 50 and N = 100.

In contrast, Figure 4.11(b) shows that the accuracy of the RBFNN model exhibits less variation across different correlation levels. This flatness suggests that the model's performance is unaffected by the correlation between independent variables. Moreover, the accuracy of RBFNN model is lower than AdjcorT-RBFNN model across all correlation and sample size. For instance, the accuracy of RBFNN model is 50.88% while AdjcorT-RBFNN model is 60.3% when the variables has strong positve correlation (ρ = 0.8) and sample size is 200. This indicates that the AdjcorT-RBFNN model is better at selecting relevant features and have more accurate classifications compared to RBFNN model even when the predictors are highly correlated.

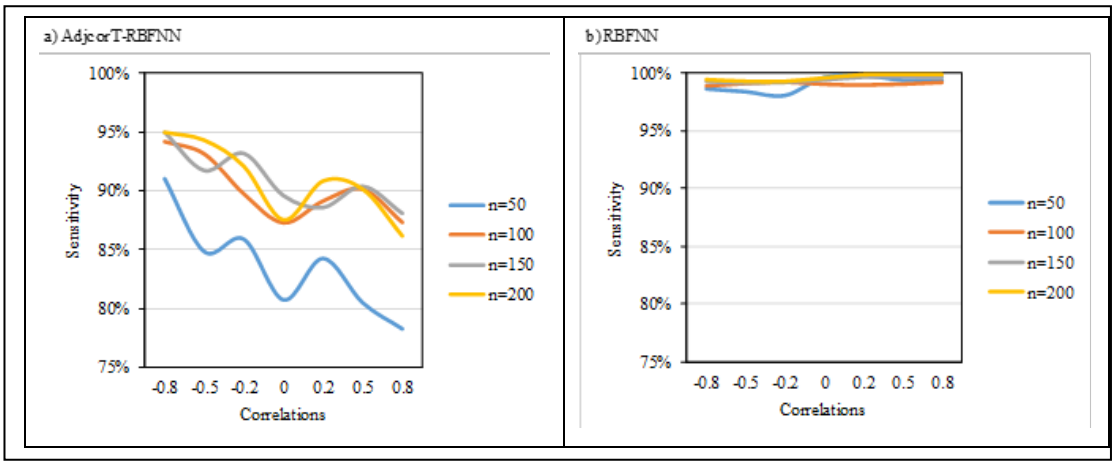

Figure 4.12 (a) Sensitivity of AdjcorT-RBFNN, (b) Sensitivity of RBFNN.

Figure 4.12(a) illustrates the sensitivity of the AdjcorT-RBFNN model, while Figure 4.12(b) presents the sensitivity of the RBFNN model. When the correlation is strongly negative (ρ = -0.8), the sensitivity of the model is relatively high across all sample sizes, though it decreases as the correlation becomes weaker or closer to zero. As the correlation turns positive, sensitivity improves again, but it does not reach the same levels seen with negative correlations. Smaller sample sizes, particularly n = 50, show more variability, with fluctuations in sensitivity observed throughout the range of correlation levels. On the other hand, larger sample sizes, such as n = 150 and n = 200,

result in higher and more stable sensitivity, indicating that the AdjcorT-RBFNN model performs better with more data.

In contrast, the RBFNN model's sensitivity remains essentially flat across all correlation levels, demonstrating that the strength and direction of the correlation between variables have no effect on its performance. Furthermore, RBFNN's sensitivity is consistently higher than that of the AdjcorT-RBFNN model across all sample sizes and correlation levels, with values near to 100%. This indicates that while AdjcorT-RBFNN is more sensitive to changes in correlation and benefits from larger sample sizes, RBFNN maintains a high level of sensitivity regardless of these factors.

The specificity of the simulation for both models is shown in Figures 4.13(a) and 4.13(b). Although the RBFNN model shows relatively high sensitivity values, the model's specificity is very low which is below 11% across all conditions. The AdjcorT-RBFNN model also shows low specificity across all conditions, ranging between 21% and 52%.

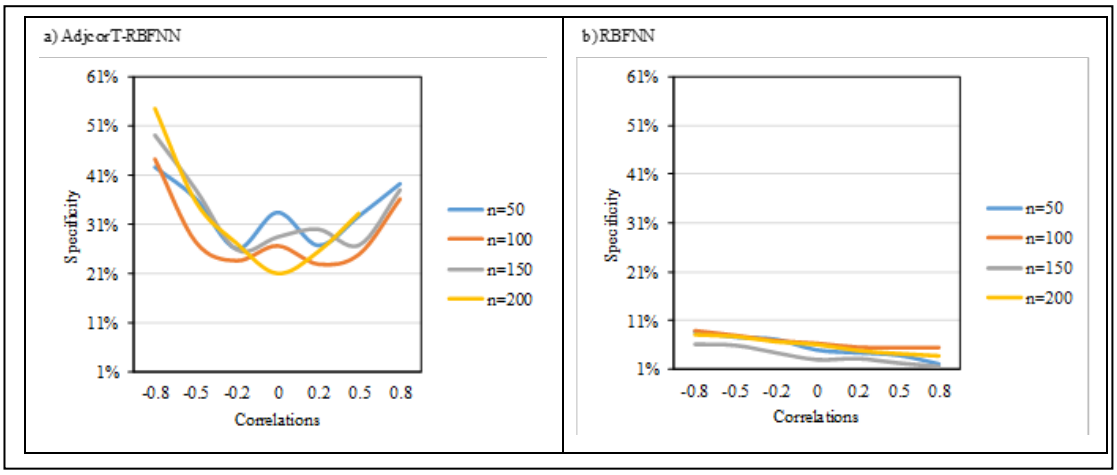

Figure 4.13 (a) Specificity of AdjcorT-RBFNN, (b) Specificity of RBFNN.

Based on the line graphs of AdjcorT-RBFNN model illustrates in Figure 4.13(a), the curves suggest a decline in specificity as the correlation between the variables closet to zero correlation. The proposed model performs higher specificity when the correlation is strong negative, specifically for the larger sample sizes. This line graphs suggest that while the AdjcorT-RBFNN model is slightly better at minimizing false positives across all sample sizes and correlations, both models struggle with correctly identifying true negatives especially RBFNN model.

Figures 4.14(a) and 4.14(b) show the precision of the simulation with different sample sizes and correlations between features for the AdjcorT-RBFNN and RBFNN models, respectively. Based on the line chart of AdjcorT-RBFNN model, the precision of the model is higher when the correlation between variables is strong, especially negative correlation for all sample sizes. However, when the correlation of the variables getting weaken, the AdjcorT-RBFNN's precision also becomes lower. Figure 4.14(b) representing the RBFNN model's precision where it is consistently lower across all levels of correlation, with minimal difference as sample size increases. Thus, the AdjcorT-RBFNN model perform better precision compared to RBFNN model, particularly in conditions with strong correlations and larger sample sizes. This suggests that the AdjcorT-RBFNN model is more effective at reducing false positives compared to the RBFNN model, especially when dealing with variables that have strong correlation between them.

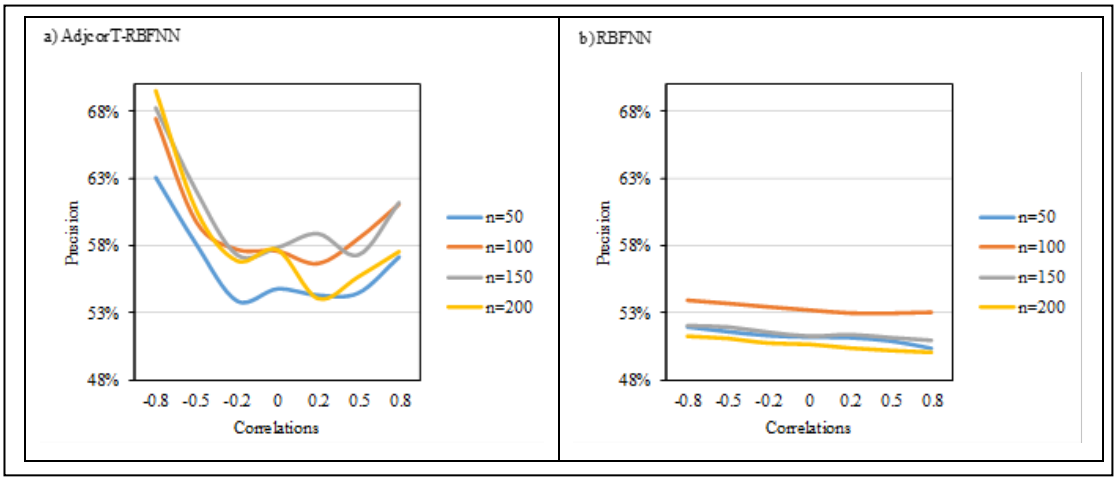

Figure 4.14 (a) Precision of AdjcorT-RBFNN, (b) Precision of RBFNN.

Figures 4.15(a) and 4.15(b) compare the F1-score of the simulation for the AdjcorT-RBFNN and RBFNN models, respectively.Based on the graphs, the lowest F1 score of the AdjcorT-RBFNN model are when the sample size is very small (N=50), across all difference correlations. Moreover, the chart suggests that for sample has more data, as the correlation between variables getting strengthen, the F1 score value also increasing. This indicates that the AdjcorT-RBFNN model perform better when there are strong correlations between the independent variables, especially strong negative correlations. Conversely, the RBFNN model displays a relatively stable F1 score

between 64% and 70% across all correlations and sample sizes, with only small differences.

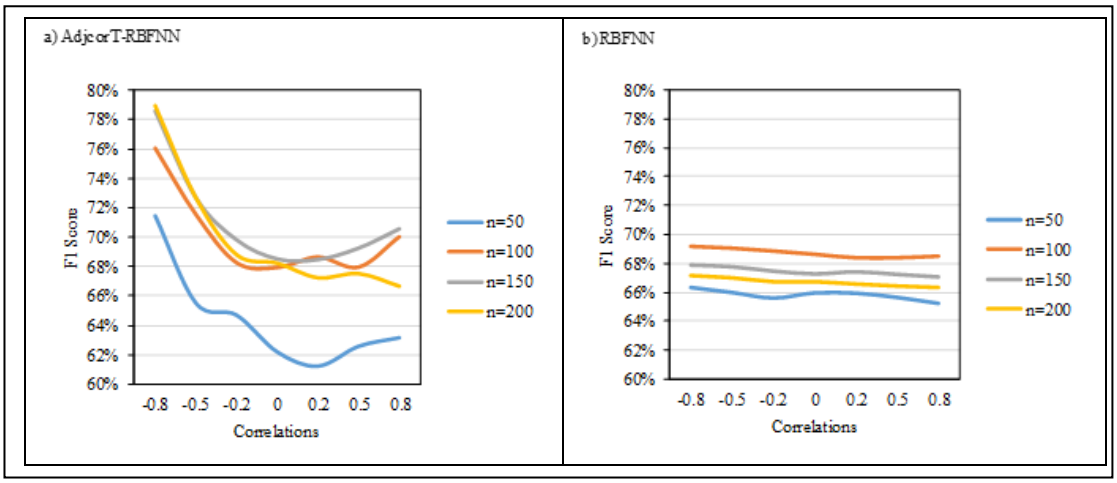

Figure 4.15 (a) F1 Score of AdjcorT-RBFNN, (b) F1 Score of RBFNN.

AUROC were computed to know how well the model can differentiate between classes. Figures 4.16(a) and 4.16(b) show the AUROC of the simulation for the AdjcorT-RBFNN and RBFNN models, respectively. Based on the graph, the AdjcorT-RBFNN model shows various AUROC values across different correlations between variables. However, AUROC of RBFNN model shows small differences for various sample sizes even though there is correlation between variables. The AUROC of RBFNN model remains consistently in between 50% to 55%, indicating that the model has limited discriminatory power regardless of the correlation strength or sample size.

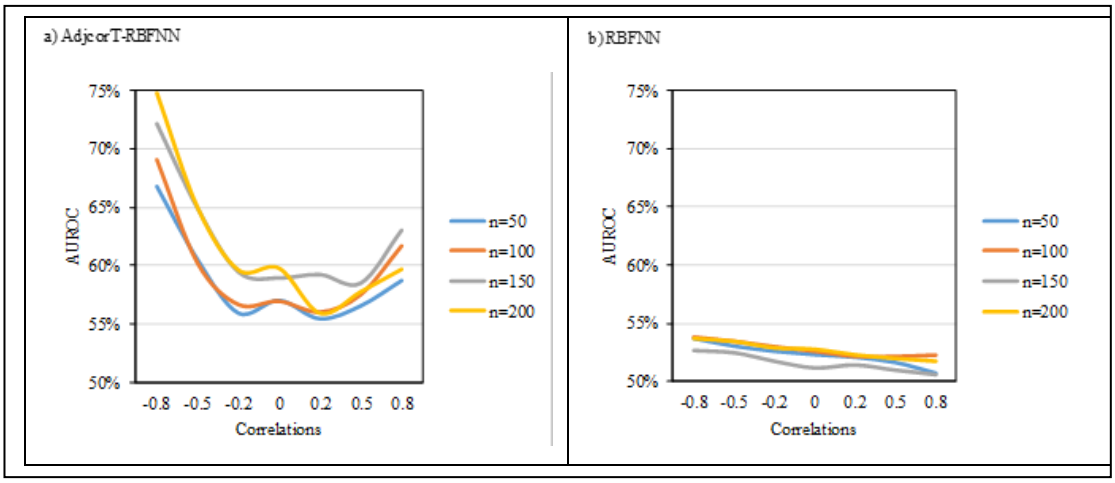

Figure 4.16 (a) AUROC of AdjcorT-RBFNN, (b) AUROC of RBFNN.

Additionally, the AdjcorT-RBFNN model shows better AUROC when there is strong correlation between variables regardless of the sample size, especially for variables that have strong negative correlation. For example, highest AdjcorT-RBFNN's AUROC is 73.9%, when the correlation is strong negative ($\rho = -0.8$) with a sample size of 200. Then, the AUROC slowly decrease when the correlation moves closer to 0 (no correlation). This suggests that the AdjcorT-RBFNN model's ability to discriminate between classes decreasing as the correlation between variables weakens. In summary, the AdjcorT- RBFNN model perform better to discriminate between classes when there is strong (positive or negative) correlation between the independent features, compared to the RBFNN model.

4.8 Conclusion

In conclusion, this study shows that AdjcorT-RBFNN and Lasso-RBFNN models performed the best in predicting next-day PM2.5 levels for Shah Alam and Banting, respectively. The significant combination features to classify next-day PM2.5 levels in Shah Alam are NO2, PM2.5, PM10, CO, O3, wind speed, SO2, and wind direction. While in Banting, the significant features to classify next-day PM2.5 levels are humidity, PM2.5, wind direction, temperature, PM10, NO2, wind speed and CO.

It emphasizes the importance of dealing with multicollinearity in air quality datasets, as AdjcorT and Lasso both consistently identified key predictors like PM2.5, PM10, and NO2 in Shah Alam. The simulations also highlight how well the AdjcorT-RBFNN model handles highly correlated variables, particularly when strong positive or negative correlations are present. Although the Lasso-RBFNN model achieved higher accuracy in Banting, AdjcorT-RBFNN was more computationally efficient and delivered similarly strong results. These findings highlight how important it is to choose the right feature selection method based on the characteristics of the dataset, especially when multicollinearity is a concern. This study contributes to developing more accurate and efficient models for air quality classification in mitigating the multicollinearity issues.

CHAPTER 5
DISCUSSION

5.1 Summary

The purpose of this study was to determine whether the AdjcorT feature selection method can help RBFNN in identifying significant features to classify the next-day PM2.5 levels, while taking into account the high correlation present in air quality data. Accordingly, this study addresses three research questions and objectives outlined in Chapter 1, focusing on multicollinearity in air quality data and the identification of significant predictors of PM2.5 concentration in Shah Alam and Banting, Malaysia. The findings highlight the performance of different feature selection methods which is AdjcorT, Lasso, mRMR, and ReliefF when integrated with radial basis function neural networks (RBFNN). AdjcorT, Lasso and mRMR are correlation-based feature selection method, while ReliefF is a noncorrelation-based feature selection method. The results shows that AdjcorT-RBFNN and Lasso-RBFNN models outperform other model in predicting next-day PM2.5 concentrations, with AdjcorT-RBFNN performing better in Shah Alam and Lasso-RBFNN performing better in Banting.

This study revealed there are strong positive correlations between several variables in both datasets, particularly between PM2.5 and PM10 with spearman correlation value of of -0.8 and -0.74 in Shah Alam and Banting, respectively. These findings are consistent with the results of previous studies, which also found strong positive correlations between PM2.5 and PM10. For instance, Su et al. (2021) found that there is a significant positive relationship between PM2.5 and PM10, with correlation coefficients of 0.99 in Henan Province, China. Although PM2.5 and PM10 are highly correlated, both variables were retained in this study because the modelling framework does not rely on linear regression assumptions, where multicollinearity typically necessitates variable exclusion. Instead, correlation-based feature selection methods and nonlinear classifiers were employed, allowing correlated predictors to be retained when each contributes additional discriminatory information to the target variable. Moreover, both PM2.5 and PM10 concentrations are influenced by meteorological parameters such as temperature, humidity, and wind speed, which affect

their dispersion and accumulation in the atmosphere (Su et al., 2021). In addition, this study found that there is very strong negative correlation between humidity and temperature in both locations. Similarly, a study on air pollution in Shah Alam by Ul Saufie et.al. (2022) revealed that temperature and relative humidity strong negative correlation with Pearson correlation value of -0.684. Despite their high correlation, humidity and temperature were retained because the two-stage feature selection framework evaluates correlated predictors based on their contribution to classification performance rather than excluding them solely on correlation strength.

This thesis aims to study three objectives outlined in Chapter 1. The first research question focused on identifying the most significant predictors of the next-day PM2.5 concentrations. This study applied the AdjcorT method to rank all the 10 features while considering the high positive and negative correlation between the variables in the first stage of feature selection method. Then, this feature ranking was used to determine the optimal subset of features that would give the high accuracy of the Artificial Neural Network (ANN) model. Based on the results, ANN model with eight features is sufficient to achieve high accuracy for both locations.

Hence, the AdjcorT, Lasso, Mrmr, and ReliefF feature selection method were employed to determine the eight most significant features to classify the next-day PM2.5 levels. The results reveal important differences between correlation-based and non-correlation-based feature selection methods. Where both correlation-based, AdjcorT and Lasso, consistently ranking PM2.5 and PM10 as the most significant predictors for both locations. This is in line with expectations, given the known multicollinearity between these variables and their strong association with air quality. However, mRMR also a correlation-based feature selection method but it consistently did not choose PM2.5 as significant features for both locations, and it also did not choose PM10 as significant features for Banting dataset. In addition, ReliefF, a noncorrelation-based method, also showed contrasting results by not selecting PM10 or PM2.5 as the top features in Shah Alam. This suggests that mRMR and ReliefF may underestimate the contribution of variables that are highly correlated with each other. Hence, this finding highlighting the advantage of using correlation-based methods like AdjcorT and Lasso in datasets where multicollinearity is the main concern. Moreover, AdjcorT, Lasso, and mRMR all identified wind direction as the least important feature in Shah Alam. On the other hand, ReliefF identified it as the most important feature in Shah Alam, suggesting the difference between correlation- based and non-correlation-

based feature selection methods. These findings suggest that mRMR and ReliefF may undervalue highly correlated predictors, whereas correlation-based methods such as AdjcorT and Lasso are better suited for datasets where multicollinearity is a dominant characteristic. This discrepancy indicates the importance of selecting the appropriate feature selection approach based on the data's multicollinearity characteristics.

The second research objective aimed to determine the most effective feature selection method for enhancing RBFNN's classification of air quality in Shah Alam and Banting. The results reveal that the AdjcorT-RBFNN and Lasso-RBFNN models outperformed in Shah Alam and Banting, respectively. Both model achieve higher accuracy, specificity, precision, F1 scores, and AUROC than RBFNN and other two-stages models. This shows that correlation-based feature selection methods such as AdjcorT and Lasso considerably enhanced model performance by selecting the most relevant and independent features while considering the high correlation between variables. These improvements confirm that correlation-based feature selection enhances RBFNN performance by mitigating multicollinearity while preserving relevant correlated predictors. Although the Lasso-RBFNN model outperformed the proposed AdjcorT-RBFNN model in terms of classification accuracy on the Banting dataset, it required much more processing time. This demonstrates the benefit of AdjcorT-RBFNN, which offers a more computationally efficient alternative while maintaining comparable performance. Besides, the Lasso- RBFNN model showed the lowest performance on the Shah Alam air quality dataset, indicating that it does not consistently perform well across different datasets.

In addition, the ReliefF-RBFNN model demonstrated the lowest performance for the Banting air quality dataset, suggesting that noncorrelation-based feature selection methods, like ReliefF, are not suitable for datasets with highly correlated variables. Similarly, Delgado-García et al. (2023) observed that the ReliefF algorithm did not select certain quantitative features due to their high correlation with other clinical factors. The authors suggested that having factors that are highly correlated could make other features less important, which could lead to poor results for the ReliefF feature selection. This result highlighting the importance of considering the relationships between variables when selecting features for air quality classification models.

Moreover, the results show that AdjcorT-RBFNN identified PM2.5, NO2, O3, wind speed, CO, wind direction, PM10, and SO2 as the best feature combinations to classify PM2.5 levels in Shah Alam. However, Lasso-RBFNN identified ambient temperature, NO2, PM10, PM2.5, CO, relative humidity, wind speed, and wind direction are the best feature combinations to classify PM2.5 levels in Banting. These findings are consistent with existing literature, which suggests that both pollutant concentrations and meteorological parameters play significant roles in determining air quality. According to Ling et al. (2020), the most polluted areas in the Klang Valley were the areas of Banting and Shah Alam, which are situated in the middle of the valley between the beach and the highlands. This is because, the air that was present in the environment was able to be purified by the wind that blew from the mountains to the ocean (Ling et al., 2020).

The differences in the significant factors between the two locations may be attributed to regional variations in air quality. For instance, the higher influence of pollutants such as PM2.5, NO2 and O3 in Shah Alam suggests that industrial emissions and traffic-related factors are more critical in urban area. For instance, Wu et al. discussed the effects of rapid urbanization on air quality, noting that cities experience high levels of particulate matter and surface ozone due to increased vehicular traffic and industrial emissions. They emphasized that urban areas are particularly affected by haze and photochemical smog, which are characterized by elevated levels of PM2.5 and O3 (Wu et al., 2020). In addition, Zhang et al. reported that vehicular emissions, particularly NOx, are a significant contributor to high NO2 concentrations in urban areas. They noted that volatile organic compounds (VOCs) and NOx emissions from vehicles are important precursors of O3, contributing significantly to summertime O3 pollution (Zhang et al., 2022).

In addition, a Monte Carlo simulation study was conducted to further validate the effectiveness of the proposed AdjcorT-RBFNN model under controlled multicollinearity conditions, thereby strengthening the findings related to third objective. The simulation results demonstrated thatthe AdjcorT-RBFNN model consistently identified x1 and x2 as important predictors when there was a strong or moderate correlation, whether positive or negative, between variables. This pattern was consistent across all sample sizes. For instance, with strong correlations ($\rho = \pm 0.8$), x1 and x2 were correctly selected over 86% of the time, demonstrating the model's robustness in identifying key features under these conditions.

However, when the correlation was weak ($\rho = \pm 0.2$), the frequency of correctly selecting these important predictors decreased. In cases where there was no correlation ($\rho = 0$), the model struggled, with a significantly lower frequency of success. Performance metrics further highlight AdjcorT-RBFNN's advantage in scenarios with strong correlations, particularly strong negative correlations. Additionally, the AdjcorT-RBFNN model demonstrated higher accuracy, sensitivity, precision, and F1 scores when dealing with strongly correlated variables, especially with larger sample sizes. In contrast, the RBFNN model maintained consistent performance across different correlations and sample sizes. This suggests that while RBFNN is stable, AdjcorT-RBFNN excels in classification where the relationships between variables are strong, making it a better option for highly correlated data. This finding similar to a study by Ibrahim (2020), where AdjcorT outperformed corT in terms of negative correlations and was comparable in terms of positive correlations across all sample sizes studied.

In summary, the study demonstrates that both AdjcorT and Lasso methods significantly improve the predictive performance of RBFNN by mitigating multicollinearity and selecting the most relevant features. AdjcorT-RBFNN was more effective in the Shah Alam dataset, while Lasso-RBFNN excelled in Banting, indicating that the choice of feature selection method may depend on the specific characteristics of the dataset. Nevertheless, the AdjcorT-RBFNN model demonstrates low computational time and competitive results, with strong performance that shows only a slight difference compared to the Lasso-RBFNN model for the Banting area. This technique emphasizes the necessity of tackling multicollinearity issues to avoid biased estimates and increase model accuracy. These findings contribute to a better understanding of how multicollinearity affects air quality classification and provide insights into the development of more accurate models for environmental monitoring.

CHAPTER 6
CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

In conclusion, this study successfully achieved all three objectives and contributes to the growing body of research on next-day PM2.5 classification by enhancing the RBFNN model with an integrated correlation-based feature selection method to address multicollinearity in air quality data. Rather than focusing solely on maximising predictive accuracy, this study emphasises the importance of addressing multicollinearity as a fundamental methodological issue that affects model robustness, feature stability, and interpretability in air quality classification.

The proposed two-stage feature selection method, AdjcorT-RBFNN, showed a strong performance compared to other models, including RBFNN, Lasso-RBFNN, mRMR-RBFNN, and ReliefF-RBFNN in the Shah Alam dataset. Moreover, the Lasso-RBFNN outperformed other models, including AdjcorT- RBFNN, in the Banting dataset. However, the AdjcorT-RBFNN showed competitive results, with only a small difference in performance metrics, and had a shorter computational runtime than Lasso-RBFNN. This indicates that AdjcorT-RBFNN performs well in both urban and suburban areas, specifically Shah Alam and Banting.

The findings suggest that AdjcorT successfully enhances the performance of the RBFNN model by effectively identifying significant predictors while considering high correlations between variables in air quality data. Although the overall numerical improvements over the standard RBFNN are sometimes modest, the results consistently demonstrate more stable and reliable feature selection when correlation-based methods are applied. Additionally, this study highlights the importance of considering for both positive and negative correlations in feature selection for accurate air quality classification. The AdjcorT-RBFNN model proved to be particularly effective in improving predictive accuracy when strong correlations were present, outperforming existing models in such cases.

Importantly, the contribution of the proposed AdjcorT-RBFNN lies in improving the robustness and reliability of the modelling framework, rather than producing marginal gains in accuracy alone. Standard RBFNN models do not explicitly

account for correlation structures, which may lead to unstable feature rankings and reduced interpretability. By incorporating correlation sharing through AdjcorT, the proposed two-stage approach provides a more principled mechanism for handling multicollinearity, which exists in air quality data. The novel application of AdjcorT to air quality classification therefore represents a methodological advancement in addressing the limitations of existing feature selection approaches.

6.2 Limitations

However, there were a few limitations in the study. First, the availability of data was a major challenge since Malaysia started monitor PM2.5 levels since 2018. This limited dataset might have restricted the ability to completely capture long-term trends and correlations in the data. Secondly, the dataset lacked a significant number of cases with high negative correlations, which limited AdjcorT's full potential. Simulations show that the method works best when there are strong positive and negative correlations. However, it is not effective as well when there are mostly weak or no correlations in the dataset. Furthermore, only used ten independent variables in this study, which might not cover all the factors that influence PM2.5 levels. This could potentially impact the accuracy of the overall model.

6.3 Recommendations

In this study, RBFNN was integrated with correlation-based feature selection methods such as AdjcorT and Lasso to enhance classification performance while addressing the multicollinearity issues in the air quality dataset. Therefore, this research recommend that future researchers explore integrating other types of machine learning models such as Support Vector Machine (SVM) with AdjcorT and Lasso feature selection methods to improve classification accuracy and address multicollinearity challenges more effectively. Additionally, future researcher could apply other feature selection methods to compare thier performance with AdjcorT-RBFNN and Lasso-RBFNN models. Moreover, future researcher should try to include other independent variable incorporating additional meteorological and pollutant factors that may influence PM2.5 levels. Adding more detailed data from various geographical areas such as rural area and extending the time periods will improve the model's robustness

and accuracy. Finally, future research is encouraged to apply the AdjcorT-RBFNN framework to datasets with stronger and more diverse correlation structures to fully evaluate its advantages under pronounced multicollinearity conditions.

6.4 Implication of The Study

This study offers meaningful insights into how air quality classification can be improved, especially in the context of urban and suburban areas in Malaysia. By applying the AdjcorT feature selection method together with the RBFNN model, the study shows the importance of selecting relevant variables while considering how strongly they relate to one another, both positively and negatively correlation. This is especially relevant for real-world environmental datasets, where multicollinearity is unavoidable rather than exceptional.

For policymakers and environmental authorities, the proposed model can serve as a practical tool. It helps pinpoint which factors are most responsible for changes in PM2.5 levels, which could make monitoring efforts more focused and responsive. Since the model works well even when fewer variables are used, it becomes easier to apply it regularly, such as for daily predictions. From a research perspective, the results highlight the value of combining feature selection techniques with machine learning, particularly in situations where many variables are strongly related. While AdjcorT may not be as widely known as other methods like Lasso or mRMR, this study suggests that it can perform just as well, or even better, in certain cases. It also opens up opportunities to test this method in other fields that deal with highly correlated data, such as healthcare or finance.

Overall, this study contributes to the ongoing discussion about how to improve classification models by making better choices about which features to include. The success of the AdjcorT-RBFNN model in handling real-world air quality data shows that there is still room for innovation in this area. Future research could explore its use in other locations or with more diverse datasets to see how well it performs under different conditions.

REFERENCES

Aamir, M., Li, Z., Bazai, S., Wagan, R. A., Bhatti, U. A., Nizamani, M. M., & Akram, S. (2021). Spatiotemporal change of air-quality patterns in Hubei province—a pre-to post- COVID-19 analysis using path analysis and regression. Atmosphere, 12(10), 1338.

Aarthi, C., Ramya, V. J., Falkowski-Gilski, P., & Divakarachari, P. B. (2023). Balanced Spider Monkey Optimization with Bi-LSTM for Sustainable Air Quality Prediction. Sustainability, 15(2), 1637. https://doi.org/10.3390/su15021637

Adani, M., Guarnieri, G., Vitali, L., Ciancarella, L., D'Elia, I., Mircea, M., … Zanini, G. (2020). Evaluation of air quality forecasting system FORAIR_IT over Europe and Italy at high resolution for year 2017. Geoscientific Model Development Discussions, 2020, 1–49.

Adesunkanmi, R., Al-Hamdan, A., & Nlenanya, I. (2023). Prediction of Pavement Overall Condition Index based on Wrapper Feature-Selection techniques using municipal pavement data. Transportation Research Record Journal of the Transportation Research Board, 2678(6), 208–221. https://doi.org/10.1177/03611981231195054

Ahmed, A., & Hussein, S. E. (2020). Leaf identification using radial basis function neural networks and SSA based support vector machine. PLoS ONE, 15(8), e0237645. https://doi.org/10.1371/journal.pone.0237645

Albraikan, A. A., Alzahrani, J. S., Negm, N., Hussain, L., Duhayyim, M. A., Hamza, M. A., ... & Yaseen, I. (2022). Future challenges of particulate matters (PMs) monitoring by computing associations among extracted multimodal features applying Bayesian network approach. Applied Artificial Intelligence, 36(1), 2112545.

Alsaber, A. R., Pan, J., & Al-Hurban, A. (2021). Handling complex missing data using random forest approach for an air quality monitoring dataset: A case study of Kuwait environmental data (2012 to 2018). International Journal of Environmental Research and Public Health, 18(3), 1333. https://doi.org/10.3390/ijerph18031333

Amann, M., Kiesewetter, G., Schöpp, W., Klimont, Z., Winiwarter, W., Cofala, J., Rafaj, P., Höglund-Isaksson, L., Gomez-Sabriana, A., Heyes, C., Purohit, P., Borken-Kleefeld, J., Wagner, F., Sander, R., Fagerli, H., Nyiri, A., Cozzi, L., &

Pavarini, C. (2020). Reducing global air pollution: the scope for further policy interventions. Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences, 378(2183), 20190331. https://doi.org/10.1098/rsta.2019.0331

Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access, 7, 128325–128338. https://doi.org/10.1109/ACCESS.2019.2925082

Barid, A. J., Hadiyanto, H., & Wibowo, A. (2024). Optimization of the algorithms use ensemble and synthetic minority oversampling technique for air quality classification. Indonesian Journal of Electrical Engineering and Computer Science, 33(3), 1632. https://doi.org/10.11591/ijeecs.v33.i3.pp1632-1640

Balakrishnan, K., Steenland, K., Clasen, T., Chang, H., Johnson, M., Pillarisetti, A., … Peel, J. L. (2023). Exposure–response relationships for personal exposure to fine particulate matter (PM2.5), carbon monoxide, and black carbon and birthweight. The Lancet Planetary Health, 7(5), e387–e396.

Banga, A., Ahuja, R., & Sharma, S. C. (2021). Performance analysis of regression algorithms and feature selection techniques to predict PM2.5 in smart cities. International Journal of System Assurance Engineering and Management, 12, 1–14.

Bărbulescu, A., Dumitriu, C. S., Ilie, I., & Barbeş, S. B. (2022). Influence of anomalies on the models for nitrogen oxides and ozone series. Atmosphere, 13(4), 558.

Bhimavarapu, U., & Sreedevi, M. (2022). Kurtosis-Based Feature Selection Method using Symmetric Uncertainty to Predict the Air Quality Index. Computer Science Journal of Moldova, 90(3), 360-375.

Bilal, M., Nichol, J. E., Nazeer, M., Shi, Y., Wang, L., Kumar, K. R., Ho, H. C., Mazhar, U., Bleiweiss, M. P., Qiu, Z., Khedher, K. M., & Lolli, S. (2019). Characteristics of Fine Particulate Matter (PM2.5) over Urban, Suburban, and Rural Areas of Hong Kong. Atmosphere, 10(9), 496. https://doi.org/10.3390/atmos10090496

Cahyani, N., & Irsyada, R. (2025). Performance comparison of SelectKBest and permutation importance in feature selection for diabetes prediction. Brilliance Research of Artificial Intelligence, 5(1), 529–541. https://doi.org/10.47709/brilliance.v5i1.6507

Cano, A., Arévalo, P., Benavides, D., & Jurado, F. (2024). Integrating discrete wavelet transform with neural networks and machine learning for fault detection in microgrids. International Journal of Electrical Power & Energy Systems, 155, 10961

Çekik, R., & Kaya, M. (2024). A new performance metric to evaluate filter feature selection methods in text classification. JUCS - Journal of Universal Computer Science, 30(7), 978–1005. https://doi.org/10.3897/jucs.111675

Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the multicollinearity problem and its machine learning approach: a review. Mathematics, 10(8), 1283.

Chen, B., Zhu, G., Ji, M., Yu, Y., Zhao, J., & Liu, W. (2020). Air quality prediction based on Kohonen clustering and ReliefF feature selection. Computational Materials & Continua, 64, 1000–1012.

Chen, Z., Yu, H., Geng, Y. A., Li, Q., & Zhang, Y. (2020, December). EvaNet: An extreme value attention network for long-term air quality prediction. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 4545-4552). IEEE.

Cheng, C. H., & Tsai, M. C. (2022). An Intelligent Time Series Model Based on Hybrid Methodology for Forecasting Concentrations of Significant Air Pollutants. Atmosphere, 13(7), 1055.

Cherrington, M., Lu, J., Airehrour, D., Thabtah, F., Xu, Q., & Madanian, S. (2019, November). Feature selection: Multi-source and multi-view data limitations, capabilities and potentials. In 2019 29th International Telecommunication Networks and Applications Conference (ITNAC) (pp. 1-6). IEEE.

Chinatamby, P., & Jewaratnam, J. (2023). A performance comparison study on PM2. 5 prediction at industrial areas using different training algorithms of feedforward-backpropagation neural network (FBNN). Chemosphere, 317, 137788.

Candotti, A., De Giglio, M., Dubbini, M., & Tomelleri, E. (2022). A Sentinel-2 based Multi-Temporal Monitoring framework for wind and bark beetle detection and damage mapping. Remote Sensing, 14(23), 6105. https://doi.org/10.3390/rs14236105

Cuevas, E., Ascencio-Piña, C. R., Pérez, M., & Morales-Castañeda, B. (2024). Considering radial basis function neural network for effective solution generation in metaheuristic algorithms. Scientific Reports, 14(1), 16806. https://doi.org/10.1038/s41598-024-67778-0

da Costa, N. L., de Lima, M. D., & Barbosa, R. (2021). Evaluation of feature selection methods based on artificial neural network weights. Expert Systems with Applications, 168, 114312.

De Sa, C. R. (2019). Variance-based feature importance in neural networks. In P. Kralj Novak, T. Šmuc, & S. Džeroski (Eds.), Discovery Science (Lecture Notes in Computer Science, Vol. 11828, pp. 1–15). Springer.

Delgado-García, G., Engbers, J. D. T., Wiebe, S., Mouches, P., Amador, K., Forkert, N. D., White, J., Sajobi, T., Klein, K. M., & Josephson, C. B. (2023). Machine learning using multimodal clinical, electroencephalographic, and magnetic resonance imaging data can predict incident depression in adults with epilepsy: A pilot study. Epilepsia, 64(10), 2781–2791. https://doi.org/10.1111/epi.17710

Ding, J., Du, J., Wang, H., & Xiao, S. (2025). A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning. Scientific Reports, 15(1), 16828.

Dongre, P. K., Patel, V., Bhoi, U., & Maltare, N. N. (2025). An outlier detection framework for Air Quality Index prediction using linear and ensemble models. Decision Analytics Journal, 14, 100546. https://doi.org/10.1016/j.dajour.2025.100546

Duan, J., & Ren, Q. (2023). Air Quality Prediction Based on Wavelet Analysis and Machine Learning. Strategic Planning for Energy and the Environment, 119-136.

Dutta, D., & Pal, S. K. (2023). Prediction and assessment of the impact of COVID-19 lockdown on air quality over Kolkata: a deep transfer learning approach. Environmental Monitoring and Assessment, 195(1), 223.

Eşsiz, E. S., Kilic, V. N., & Oturakçi, M. (2023). Firefly-Based feature selection algorithm method for air pollution analysis for Zonguldak region in Turkey. Turkish Journal of Engineering, 7(1), 17-24.

Fan, W., Xu, L., & Zheng, H. (2022). Using Multisource Data to Assess PM2. 5 Exposure and Spatial Analysis of Lung Cancer in Guangzhou, China. International Journal of Environmental Research and Public Health, 19(5), 2629.

Gao, M., Yang, H., Xiao, Q., & Goh, M. (2022). COVID-19 lockdowns and air quality: Evidence from grey spatiotemporal forecasts. Socio – economic planning sciences, 83, 101228. https://doi.org/10.1016/j.seps.2022.101228

Gao, Z., Do, K., Li, Z., Jiang, X., Maji, K. J., Ivey, C. E., & Russell, A. G. (2024). Predicting PM2. 5 levels and exceedance days using machine learning methods. Atmospheric Environment, 120396.

Ghanizadeh, A. R., Heidarabadizadeh, N., & Jalali, F. (2020). Artificial neural network back-calculation of flexible pavements with sensitivity analysis using Garson's and connection weights algorithms. Innovative Infrastructure Solutions, 5(2). https://doi.org/10.1007/s41062-020-00312-z

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE Access, 8, 67899–67911. https://doi.org/10.1109/access.2020.2986809

Giam, X., & Olden, J. D. (2021). Connection weight methods for assessing variable importance in artificial neural networks. Neural Computation, 33(10), 2543–2560.

Gibergans Bàguena, J., Hervada Sala, C., & Jarauta Bragulat, E. (2020). The quality of urban air in Barcelona: A new approach applying compositional data analysis methods. Emerging Science Journal, 4(2), 113–121.

Goudarzi, G., Hopke, P. K., & Yazdani, M. (2021). Forecasting PM2. 5 concentration using artificial neural network and its health effects in Ahvaz, Iran. Chemosphere, 283, 131285.

Guan, Y., & Fu, G. H. (2022). A double-penalized estimator to combat separation and multicollinearity in logistic regression. Mathematics, 10(20), 3824.

Guo, Z., Wang, X., & Ge, L. (2023). Classification prediction model of indoor PM2. 5 concentration using CatBoost algorithm. Frontiers in Built Environment, 9, 1207193.

Hao, G., Guo, J., Zhang, W., Chen, Y., & Yuen, D. A. (2022). High-precision chaotic radial basis function neural network model: Data forecasting for the Earth electromagnetic signal before a strong earthquake. Geoscience Frontiers, 13(1), 101315. https://doi.org/10.1016/j.gsf.2021.101315

Hasegawa, N., Hasegawa, S., Kitaoka, T., & Ohtsu, H. (2019). Applicability of neural network in rock classification of mountain tunnel. Materials Transactions, 60(5), 758- 764.

Hassan, M. H., Mostafa, S. A., Mustapha, A., Saringat, M. Z., Al-Rimy, B. a. S., Saeed, F., Eljialy, A., & Jubair, M. A. (2022). A new collaborative Multi-Agent Monte

Carlo Simulation model for Spatial Correlation of Air Pollution Global Risk Assessment. Sustainability, 14(1), 510. https://doi.org/10.3390/su14010510

Hosni, M., Idri, A., & Abran, A. (2021). On the value of filter feature selection techniques in homogeneous ensembles effort estimation. Journal of Software: Evolution and Process, 33(6), e2343.

Houndekindo, F., & Ouarda, T. B. (2023). Comparative study of feature selection methods for wind speed estimation at ungauged locations. Energy Conversion and Management, 291, 117324. https://doi.org/10.1016/j.enconman.2023.117324

Huang, W., & Yang, Y. (2019). Water quality sensor model based on an optimization method of RBF neural network. Computational Water, Energy, and Environmental Engineering, 9(1), 1–11.

Hussain, L., Aziz, W., Saeed, S., Rafique, M., Nadeem, M. S. A., Shim, S. O., ... & Pirzada, J. U. R. (2020). Extracting mass concentration time series features for classification of indoor and outdoor atmospheric particulates. Acta Geophysica, 68, 945- 963.

Ibrahim, N. (2020). Variable selection methods for classification: Application to metabolomics data (Doctoral dissertation, University of Liverpool). ProQuest Dissertations & Theses Global.

Ibrahim, N., Ishak, U. M., Ali, N. N. A., & Shaadan, N. (2024). MACHINE LEARNING-BASED APPROACHES FOR CREDIT CARD DEBT PREDICTION. Malaysian Journal of Computing,9(1), 1722–1733. https://doi.org/10.24191/mjoc.v9i1.25656

Imam, M., Adam, S., Dev, S., & Nesa, N. (2024). Air Quality Monitoring Using Statistical Learning Models for Sustainable Environment. Intelligent Systems with Applications, 200333.

Iskandaryan, D., Ramos, F., & Trilles, S. (2023). Graph Neural Network for Air Quality Prediction: A Case Study in Madrid. IEEE Access, 11, 2729-2742.

Jalali, S., Karbakhsh, M., Momeni, M., Taheri, M., Amini, S., Mansourian, M., & Sarrafzadegan, N. (2021). Long-term exposure to PM2.5 and cardiovascular disease incidence and mortality in an Eastern Mediterranean country: findings based on a 15 -year cohort study. Environmental Health, 20(1). https://doi.org/10.1186/s12940-021-00797-w

Jebamalar, J. A., & Kumar, A. S. (2019). PM2. 5 prediction using machine learning hybrid model for smart health. Int. J. Eng. Adv. Technol., 9(1), 6500-6504.

Jia, Z., Shi, C., Wang, Y., Yang, X., Zhang, J., & Ji, Z. (2019). Nondestructive determination of salmon fillet freshness during storage at different temperatures by electronic nose system combined with radial basis function neural networks. International Journal of Food Science & Technology, 55(5), 2080–2091. https://doi.org/10.1111/ijfs.14451

Kalajdjieski, J., Zdravevski, E., Corizzo, R., Lameski, P., Kalajdziski, S., Pires, I. M., ... & Trajkovik, V. (2020). Air pollution prediction with multi-modal data and deep neural networks. Remote Sensing, 12(24), 4142.

Kao, C. M., Chen, Y. M., Huang, W. N., Chen, Y. H., & Chen, H. H. (2023). Association between air pollutants and initiation of biological therapy in patients with ankylosing spondylitis: a nationwide, population-based, nested case–control study. Arthritis Research & Therapy, 25(1), 75.

Kayanan, M., & Wijekoon, P. (2020). Variable selection via biased estimators in the linear regression model. Open Journal of Statistics, 10(01), 113–126. https://doi.org/10.4236/ojs.2020.101009

Ko, U. W., & Kyung, S. Y. (2022). Adverse effects of air pollution on pulmonary diseases. Tuberculosis and Respiratory Diseases, 85(4), 313.

Krajnc, D., Spielvogel, C. P., Grahovac, M., Ecsedi, B., Rasul, S., Poetsch, N., Traub-Weidinger, T., Haug, A. R., Ritter, Z., Alizadeh, H., Hacker, M., Beyer, T., & Papp, L. (2022). Automated data preparation for in vivo tumor characterization with machine learning. Frontiers in Oncology, 12, 1017911. https://doi.org/10.3389/fonc.2022.1017911

Kristiyanti, D. A., Purwaningsih, E., Nurelasari, E., Al Kaafi, A., & Umam, A. H. (2020, November). Implementation of neural network method for air quality forecasting in Jakarta region. In Journal of Physics: Conference Series (Vol. 1641, No. 1, Article 012037). IOP Publishing.

Kumar, P. R., Ali, M. K. M., & Ibidoja, O. J. (2024). Identifying heterogeneity for increasing the prediction accuracy of machine learning models. Journal of the Nigerian Society of Physical Sciences, 2058. https://doi.org/10.46481/jnsps.2024.2058

Kumbhar, V. S., Sohi, S. S., Jayaram, V., Sreelekshmy, P. G., Shukla, S. K., & Abhilash, K. S. (2022). Hybrid artificial neural network algorithm for air pollution estimation.International Journal of Health Sciences,6(S5), 2094–2106.

Liang, Y., Sengupta, D., Campmier, M. J., Lunderberg, D. M., Apte, J. S., & Goldstein, A. H. (2021). Wildfire smoke impacts on indoor air quality assessed using crowdsourced data in California. Proceedings of the National Academy of Sciences, 118(36). https://doi.org/10.1073/pnas.2106478118

Li, Y., Wang, B., Li, B., Zhang, X., Liu, X., Wang, Q., ... & Xiu, R. (2021). Common Microcirculatory Framework for Monitoring Integrated Microcirculation.

Ling, O. H. L., Marzukhi, M. A., Qi, J. K., & Mabahwi, N. A. (2020). Impact of urban land uses and activities on the ambient air quality in Klang Valley, Malaysia from 2014 to 2020. Planning Malaysia, 18.

Liu, B., & Zhang, Y. (2022). Calibration of miniature air quality detector monitoring data with PCA–RVM–NAR combination model. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-13531-4

Liu, J., Christensen, J. H., Ye, Z., Dong, S., Geels, C., Brandt, J., Nenes, A., Yuan, Y., & Im, U. (2024). Impact of meteorology and aerosol sources on PM 2.5 and oxidative potential variability and levels in China. Atmospheric Chemistry and Physics, 24(18), 10849–10867. https://doi.org/10.5194/acp-24-10849-2024

Liu, Y., Li, Z., & Ning, Y. (2023). Integrated navigation model based on TDCP constrained algorithm. Measurement Science and Technology, 34(12), 125137. https://doi.org/10.1088/1361-6501/acf77c

Mahajan, M., Kumar, S., Pant, B., & Khan, R. (2021). Improving accuracy of air pollution prediction by two step outlier detection. 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 5069 lncs, 1–7. https://doi.org/10.1109/icaect49130.2021.9392404

Mamouei, M., Zhu, Y., Nazarzadeh, M., Hassaine, A., Salimi-Khorshidi, G., Cai, Y., & Rahimi, K. (2022). Investigating the association of environmental exposures and all - cause mortality in the UK Biobank using sparse principal component analysis. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-13362-3

Mansor, M. A., Jamaludin, S. Z. M., Kasihmuddin, M. S. M., Alzaeemi, S. A., Basir, M. F. M., & Sathasivam, S. (2020). Systematic boolean satisfiability

programming in radial basis function neural network. Processes, 8(2), 214. https://doi.org/10.3390/pr8020214

Mcculloch, W. S., & Pitts, W. (1990). A logical calculus nervous activity. Bulletin of Mathematical Biology, 52(l), 99–115.

McDuffie, E. E., Martin, R. V., Spadaro, J. V., Burnett, R., Smith, S. J., O'Rourke, P., … Brauer, M. (2021). Source sector and fuel contributions to ambient PM2.5 and attributable mortality across multiple spatial scales. Nature Communications, 12(1), 1–12.

Meena, K. K., Bairwa, D., & Agarwal, A. (2024). A machine learning approach for unraveling the influence of air quality awareness on travel behavior. Decision Analytics Journal, 100459.

Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: a survey. Artificial Intelligence Review, 1-36.

Menéndez García, L. A., Menéndez Fernández, M., Sokoła-Szewioła, V., Álvarez de Prado, L., Ortiz Marqués, A., Fernández López, D., & Bernardo Sánchez, A. (2022). A method of pruning and random replacing of known values for comparing missing data imputation models for incomplete air quality time series. Applied Sciences, 12(13), 6465.

Mohammadi, F., Teiri, H., Hajizadeh, Y., Abdolahnejad, A., & Ebrahimi, A. (2024). Prediction of atmospheric PM2. 5 level by machine learning techniques in Isfahan, Iran. Scientific Reports, 14(1), 2109.

Mokhtari, I., Bechkit, W., Rivano, H., & Yaici, M. R. (2021). Uncertainty-Aware deep learning architectures for highly dynamic air quality prediction. IEEE Access, 9, 14765– 14778. https://doi.org/10.1109/access.2021.3052429

Murugan, R., & Palanichamy, N. (2021, May). Smart city air quality prediction using machine learning. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1048-1054). IEEE.

Nemade, S. (2019). A survey on different machine learning techniques for air quality forecasting for urban air pollution. International Journal for Research in Applied Science & Engineering Technology, 7(4), 2185–2194. https://doi.org/10.22214/ijraset.2019.4395

Nouri, A., Ghanbarzadeh Lak, M., & Valizadeh, M. (2021). Prediction of PM2. 5 concentrations using principal component analysis and artificial neural network

techniques: a case study: Urmia, Iran. Environmental Engineering Science, 38(2), 89-98.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in bioinformatics, 2, 927312.

Radočaj, D., Rapčan, I., & Jurišić, M. (2023). Indoor Plant Soil-Plant Analysis Development (SPAD) prediction based on multispectral indices and soil electroconductivity: a deep learning approach. Horticulturae, 9(12), 1290. https://doi.org/10.3390/horticulturae9121290

Raffee, A. F. (2021). Multivariate time series analysis for short-term forecasting of ground level ozone (O3) in Malaysia (Doctoral dissertation, Universiti Tun Hussein Onn Malaysia).

Rahardja, U., Aini, Q., Sunarya, P. A., Manongga, D., & Julianingsih, D. (2022). The use of TensorFlow in analyzing air quality artificial intelligence predictions (PM2.5). Aptisi Transactions on Technopreneurship, 4(3), 313–324. https://doi.org/10.34306/att.v4i3.282

Ramli, N., Abdul Hamid, H., Yahaya, A. S., Ul-Saufie, A. Z., Mohamed Noor, N., Abu Seman, N. A., ... & Deák, G. (2023). Performance of Bayesian Model Averaging (BMA) for Short-Term Prediction of PM10 Concentration in the Peninsular Malaysia. Atmosphere, 14(2), 311.

Razali, C. M. C., Hussein, S. F. M., Harudin, N., & Abdullah, S. S. (2018). Estimation of building energy efficiency performance using radial basis function neural network. International Journal of Engineering & Technology, 7(4.35), 755–759. https://doi.org/10.14419/ijet.v7i4.35.23102

Saha, P. K., Presto, A. A., & Robinson, A. L. (2023). Hyper-local to regional exposure contrast of source-resolved PM2.5 components across the contiguous United States: implications for health assessment. Journal of Exposure Science & Environmental Epidemiology. https://doi.org/10.1038/s41370-023-00623-0

Saminathan, S., & Malathy, C. (2023). Ensemble-based classification approach for PM2. 5 concentration forecasting using meteorological data. Frontiers in big Data, 6, 1175259.

Santos, R. P. D., Dean, D. L., Weaver, J. M., & Hovanski, Y. (2018). Identifying the relative importance of predictive variables in artificial neural networks based on data produced through a discrete event simulation of a manufacturing

environment. International Journal of Modelling and Simulation, 39(4), 234–245. https://doi.org/10.1080/02286203.2018.1558736

Sethi, J. K., & Mittal, M. (2021). An efficient correlation based adaptive LASSO regression method for air quality index prediction. Earth Science Informatics, 14(4), 1777-1786.

Shaadan, N., & Din, W. N. W. M. (2022). Application of Functional Time Series Model in Forecasting Monthly Diurnal API Curves: A Comparison between Multi-Step Ahead and Iterative One-Step Ahead Approach. Malaysian Journal of Fundamental and Applied Sciences, 18(1), 124-137

Shaheen, N., Shah, I., Almohaimeed, A., Ali, S., & Alqifari, H. N. (2023). Some Modified Ridge Estimators for Handling the Multicollinearity Problem. Mathematics, 11(11), 2522.

Shaziayani, W. N., Ul-Saufie, A. Z., Ahmat, H., & Al-Jumeily, D. (2021). Coupling of quantile regression into boosted regression trees (BRT) technique in forecasting emission model of PM10 concentration. Air Quality, Atmosphere & Health, 14(10), 1647-1663.

Shin, Y., Kim, Y. J., Jin, J., Lee, S., Kim, H., & Kim, Y. (2023). Machine learning model for predicting immediate postoperative desaturation using spirometry signal data. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-49062-9

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. American Journal of Applied Mathematics and Statistics, 8(2), 39–42. https://doi.org/10.12691/ajams-8-2-1

Sokhi, R. S., Moussiopoulos, N., Baklanov, A., Bartzis, J., Coll, I., Finardi, S., ... & Kukkonen, J. (2021). Advances in air quality research–current and emerging challenges. Atmospheric Chemistry and Physics Discussions, 2021, 1-133.

Song, D., & Zhou, X. (2023, August). A practical three-stage hybrid feature selection method using discrete state transition algorithm. In Sixth International Conference on Advanced Electronic Materials, Computers, and Software Engineering (AEMCSE 2023) (Vol. 12787, pp. 80-86). SPIE.

Spyrou, E. D., Stylios, C., & Tsoulos, I. (2023). Classification of CO Environmental Parameter for Air Pollution Monitoring with Grammatical Evolution. Algorithms, 16(6), 300.

Stajkowski, S., Zeynoddin, M., Farghaly, H., Gharabaghi, B., & Bonakdari, H. (2020). A methodology for forecasting dissolved oxygen in urban streams. Water, 12(9), 2568. https://doi.org/10.3390/w12092568

Su, B., Wu, D., Zhang, M., Bilal, M., Li, Y., Li, B. L., ... & Howari, F. M. (2021). Spatio- temporal characteristics of PM2. 5, PM10, and AOD over the Central Line Project of China's South-North water diversion in Henan Province (China). Atmosphere, 12(2), 225.

Suleiman, A., Tight, M. R., & Quinn, A. D. (2020). APPLYING HYBRID FEATURE SELECTION METHODS FOR STATISTICAL MODELLING OF ROADSIDE PARTICLE CONCENTRATIONS (PM. Air Pollution Studies, 1.

Tateo, A., Campanaro, V., Amoroso, N., Bellantuono, L., Monaco, A., Pantaleo, E., ... & Maggipinto, T. (2023). Predicting Air Quality from Measured and Forecast Meteorological Data: A Case Study in Southern Italy. Atmosphere, 14(3), 475.

Tella, A., Balogun, A. L., Adebisi, N., & Abdullah, S. (2021). Spatial assessment of PM10 hotspots using random forest, K-nearest neighbour and Naïve Bayes. Atmospheric Pollution Research, 12(10), 101202.

Tibshirani, R., & Wasserman, L. (2006). Correlation-sharing for detection of differential gene expression. Carnegie Mellon University.

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.

Usmani, R. S. A., Saeed, A., Abdullahi, A. M., Pillai, T. R., Jhanjhi, N. Z., & Hashem, I. A. T. (2020). Air pollution and its health impacts in Malaysia: a review. Air Quality, Atmosphere & Health, 13, 1093-1118.

Wang, H., Jin, X., Du, Y., & Zhang, N. (2023). An adaptive node network based on multitask deep learning.

Wang, J., Jin, L., Li, X., He, S., Huang, M., & Wang, H. (2022). A hybrid air quality index prediction model based on CNN and attention gate unit. IEEE Access, 10, 113343- 113354.

Wang, J., Shi, L., Chen, J., Wang, B., Qi, J., Chen, G., Kang, M., Zhang, H., Jin, X., Huang, Y., Zhao, Z., Chen, J., Song, B., & Chen, J. (2021). A novel risk score system for prognostic evaluation in adenocarcinoma of the oesophagogastric junction: a large population study from the SEER database and our center. BMC Cancer, 21(1). https://doi.org/10.1186/s12885-021-08558-1

Wang, R., Bian, J., Nie, F., & Li, X. (2022). Nonlinear feature selection neural network via structured sparse regularization. IEEE Transactions on Neural Networks and Learning Systems, 34(11), 9493-9505.

Warneke, C., Schwarz, J. P., Dibb, J., Kalashnikova, O., Frost, G., Al-Saad, J., ... & FIREX-AQ Science Team. (2023). Fire influence on regional to global Environments and air quality (FIREX-AQ). Journal of Geophysical Research: Atmospheres, 128(2), e2022JD037758.

War, W., & Barlis, M. J. P. (2023). The Correlation of Financial Management Practices and Controlling Mechanisms for Financial Management Challenges and Issues The Case of Franciscan Missionaries of Mary Congregation. Business Economic, Communication, and Social Sciences Journal (BECOSS), 5(2), 101-114.

Wattimena, E. M. C., Annisa, A., & Sitanggang, I. S. (2022). CO and PM10 Prediction Model based on Air Quality Index Considering Meteorological Factors in DKI Jakarta using LSTM. Scientific Journal of Informatics, 9(2), 123–132. https://doi.org/10.15294/sji.v9i2.33791

Wu, C. L. R., Wagterveld, R. M., & Van Breukelen, B. M. (2024). Reactive transport modeling for exploring the potential of water quality sensors to estimate hydrocarbon levels in groundwater. Water Resources Research, 60(4). https://doi.org/10.1029/2023wr036644

Wu, F., Min, P., Jin, Y., Zhang, K., Liu, H., Zhao, J., & Li, D. (2023). A novel hybrid model for hourly PM2. 5 prediction considering air pollution factors, meteorological parameters and GNSS-ZTD. Environmental Modelling & Software, 167, 105780.

Wu, H., Yang, T., Li, H., & Zhou, Z. (2023). Air quality prediction model based on mRMR–RF feature selection and ISSA–LSTM. Scientific Reports, 13(1), 12825.

Wu, L., Hang, J., Wang, X., Shao, M., & Gong, C. (2020). APFoam-1.0: Integrated CFD simulation of O_3–NO_x–VOCs chemistry and pollutant dispersion in a typical street canyon. Geoscientific Model Development, 13, 4367–4383.

Wu, Y.-J. (1998). Exchange rate forecasting: an application of radial basis function neural networks. Iowa State University.

Yadav, V., & Nath, S. (2018, February). Daily prediction of PM 10 using radial basis function and generalized regression neural network. In 2018 Recent Advances

on Engineering, Technology and Computational Sciences (RAETCS) (pp. 1-5). IEEE.

Xu, B., You, X., Zhou, Y., Dai, C., Liu, Z., Huang, S., Luo, D., & Peng, H. (2020). The study of emission inventory on anthropogenic air pollutants and source apportionment of PM2.5 in the Changzhutan Urban Agglomeration, China. Atmosphere, 11(7), 739. https://doi.org/10.3390/atmos11070739

Yadav, V., Yadav, A. K., & Khargotra, R. (2025). Relief-Based feature selection for ANN-Driven air pollution prediction and school safety policy in India. Water Air & Soil Pollution, 237(1). https://doi.org/10.1007/s11270-025-08759-5

Yan, R., Liao, J., Yang, J., Sun, W., Nong, M., & Li, F. (2021). Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Systems with Applications, 169, 114513.

Yang, W., Jiang, J., Schnellinger, E. M., Kimmel, S. E., & Guo, W. (2022). Modified Brier score for evaluating prediction accuracy for binary outcomes. Statistical methods in medical research, 31(12), 2287-2296.

Yang, Y., Mei, G., & Izzo, S. (2022). Revealing influence of meteorological conditions on air quality prediction using explainable deep learning. IEEE Access, 10, 50755-50773.

Yoo, C., & Cho, E. (2019). Effect of multicollinearity on the bivariate frequency analysis of annual maximum rainfall events. Water, 11(5), 905.

Zaini, N. A., Ean, L. W., Ahmed, A. N., Abdul Malek, M., & Chow, M. F. (2022). PM2. 5 forecasting for an urban area based on deep learning and decomposition method. Scientific Reports, 12(1), 17565.

Zaman, N. A. F. K., Kanniah, K. D., Kaskaoutis, D. G., & Latif, M. T. (2021). Evaluation of machine learning models for estimating PM2. 5 concentrations across Malaysia. Applied Sciences, 11(16), 7326.

Zaman, N. a. F. K., Kanniah, K. D., Kaskaoutis, D. G., & Latif, M. T. (2024). Improving the quantification of fine particulates (PM2.5) concentrations in Malaysia using simplified and computationally efficient models. Journal of Cleaner Production, 448, 141559. https://doi.org/10.1016/j.jclepro.2024.141559

Zhang, B., Li, Y., & Chai, Z. (2022). A novel random multi-subspace based ReliefF for feature selection. Knowledge-Based Systems, 252, 109400.

Zhang, W., Wu, Y., & Calautit, J. K. (2022). A review on occupancy prediction through machine learning for enhancing energy efficiency, air quality and thermal

comfort in the built environment. Renewable and Sustainable Energy Reviews, 167, 112704.

Zhang, Y., Yang, D., Wu, R., Li, Y., Yang, X., Wang, R., & Xu, H. (2022). Characterization of roadside air pollutants: An artery road of Lanzhou city in northwest China. In E3S Web of Conferences (Vol. 360, p. 01039). EDP Sciences.

Zhao, Z., Wu, J., Cai, F., Zhang, S., & Wang, Y. G. (2022). A statistical learning framework for spatial-temporal feature selection and application to air quality index forecasting. Ecological Indicators, 144, 109416.

Zhao, Z., Wu, J., Cai, F., Zhang, S., & Wang, Y. G. (2023). A hybrid deep learning framework for air quality prediction with spatial autocorrelation during the COVID-19 pandemic. Scientific Reports, 13(1), 1015.

Zhou, Z. H., Xie, S. F., Li, G. H., Zhao, Y., & Zhang, W. (2020). Analysis and prediction of PM2.5 concentration in Guilin City. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 1081–1089.

Zouhri, W., Rostami, H., Homri, L., & Dantan, J. Y. (2021). A genetic-based SVM approach for quality data classification. In Artificial Intelligence and Industrial Applications: Smart Operation Management (pp. 15-31). Springer International Publishing.

APPENDICES

APPENDIX 1

R code for Simulation

```
# Load necessary libraries
library(MASS)
library(caret)
library(pROC)
library(RSNNS)

# Set seed for reproducibility
set.seed(110)

# Set mean and standard deviation for each group
meanx1group0 <- 0
meanx2group0 <- 0
meanx1group1 <- 0.5
meanx2group1 <- 1
stddev <- c(1, 1)

# Set variance correlation and covariance
corMat <- matrix(c(1, 0.8, 0.8, 1), ncol = 2)
corVar <- stddev %*% t(stddev) * corMat

# Set sample sizes and generate data for x1 and x2
x1x2_group0 <- mvrnorm(100, c(meanx1group0, meanx2group0), Sigma = corVar,
empirical = TRUE)
x1x2_group1 <- mvrnorm(100, c(meanx1group1, meanx2group1), Sigma = corVar,
empirical = TRUE)
x1x2 <- rbind(x1x2_group0, x1x2_group1)

# Generate data for x3 to x10
x3to10 <- matrix(rnorm(200 * 8, 0, 1), 200, 8)

# Column bind all variables
```

```
x <- cbind(x1x2, x3to10)

# Set column names
colnames(x) <- paste0("x", 1:ncol(x))

# Set dependent variable into category 0 and 1
y <- data.frame(rep(0:1, each = 100))
colnames(y) <- "y"

# Combine y and x into a data frame
data <- data.frame(y, x)

# Initialize performance metrics
accuracy_vals <- numeric(100)
sensitivity_vals <- numeric(100)
specificity_vals <- numeric(100)
precision_vals <- numeric(100)
f1_score_vals <- numeric(100)
auroc_vals <- numeric(100)

# Initialize feature counts
feature1_and_other <- 0
feature2_and_other <- 0
both_features <- 0
other_features <- 0

# Perform 100 iterations
for (i in 1:100) {
  set.seed(110 + i)
  smp_size <- floor(0.80 * nrow(data))
  train_ind <- sample(seq_len(nrow(data)), size = smp_size)
  trainset <- data[train_ind, ]
  testset <- data[-train_ind, ]
```

```
  # adjcorT method
  adjcorTfunction <- function(trainset, data) {
    tmp <- centroids(as.matrix(trainset[, 2:ncol(data)]), as.matrix(trainset[, 1]),
var.groups = FALSE, centered.data = TRUE,
                     lambda.var = 0, lambda.freqs = 0, verbose = TRUE )
    diff <- tmp$means[, 1] - tmp$means[, 2]
    n1 <- tmp$samples[1]
    n2 <- tmp$samples[2]
    v <- tmp$variances[, 1]
    sd <- sqrt((1/n1 + 1/n2) * v)
    R <- cor(tmp$centered.data)
    t <- diff / sd
    p <- length(t)
    cst.vec <- numeric(p)
    for (i in 1:p) {
      idx <- order(abs(R[i, ]), decreasing = TRUE)
      nonneg <- sum(abs(R[i, ]) >= 0)
      z <- cumsum(abs(t[idx[1:nonneg]])) / 1:nonneg
      cst.vec[i] <- max(z) * sign(t[i])
    }
    idx <- order(abs(cst.vec), decreasing = TRUE)
    return(idx)
  }

  adjcorT_results <- adjcorTfunction(trainset, data)
  selected_features <- adjcorT_results[1:2]
  print(paste("Iteration:", i, "Selected Features:", paste(selected_features, collapse = ",
")))

  # Combination selected variables
  if (length(selected_features) == 1) {
    # No change for feature 1 or 2 only cases
  } else if (length(selected_features) == 2) {
    # Check for both features or one feature with other features
```

```
    if (1 %in% selected_features & 2 %in% selected_features) {
      both_features <- both_features + 1
    } else {
      # Check if only one of feature 1 or 2 is selected with other features
      if ((1 %in% selected_features & !2 %in% selected_features) | (!1 %in%
selected_features & 2 %in% selected_features)) {
        if (1 %in% selected_features) {
          feature1_and_other <- feature1_and_other + 1
        } else {
          feature2_and_other <- feature2_and_other + 1
        }
      }
    }
  } else {
    # Check if any features other than 1 and 2 are selected
    if (any(selected_features > 2 & selected_features < ncol(data))) {
      other_features <- other_features + 1
    }
  }

  # Train RBFNN model
  nn_model <- rbf(trainset[, selected_features + 1], as.numeric(trainset$y), size = 6,
linOut = FALSE)

  # Prediction with RBFNN
  predicted_probs <- predict(nn_model, testset[, selected_features + 1])
  binary_predictions <- ifelse(predicted_probs > 0.5, 1, 0)

  y_test <- testset$y

  # Convert to factors
  y_test_factor <- factor(y_test, levels = c(0, 1))
  binary_predictions_factor <- factor(binary_predictions, levels = c(0, 1))
```

```
  # Calculate performance metrics
  accuracy_vals[i] <- mean(binary_predictions == y_test)
  sensitivity_vals[i] <- sensitivity(binary_predictions_factor, y_test_factor)
  specificity_vals[i] <- specificity(binary_predictions_factor, y_test_factor)
  precision_vals[i] <- posPredValue(binary_predictions_factor, y_test_factor)
  f1_score_vals[i] <- 2 * (precision_vals[i] * sensitivity_vals[i]) / (precision_vals[i] +
sensitivity_vals[i])
  auroc_vals[i] <- auc(y_test, predicted_probs)
}

# Calculate mean metrics
mean_accuracy <- mean(accuracy_vals, na.rm = TRUE)
mean_sensitivity <- mean(sensitivity_vals, na.rm = TRUE)
mean_specificity <- mean(specificity_vals, na.rm = TRUE)
mean_precision <- mean(precision_vals, na.rm = TRUE)
mean_f1_score <- mean(f1_score_vals, na.rm = TRUE)
mean_auroc <- mean(auroc_vals, na.rm = TRUE)

cat("Mean Accuracy:", mean_accuracy, "\n")
cat("Mean Sensitivity:", mean_sensitivity, "\n")
cat("Mean Specificity:", mean_specificity, "\n")
cat("Mean Precision:", mean_precision, "\n")
cat("Mean F1 Score:", mean_f1_score, "\n")
cat("Mean AUROC:", mean_auroc, "\n")
```

AUTHOR'S PROFILE

Siti Khadijah Arafin obtained a Bachelor of Science in Statistics (Hons.) in 2022 from Universiti Teknologi MARA, Shah Alam and is pursuing a Master of Science in Statistics at Universiti Teknologi MARA, Shah Alam. Her research focuses on feature selection, machine learning, and statistical techniques for air quality prediction, particularly in addressing multicollinearity issues in environmental datasets. Her study emphasizes the development and application of the adjusted correlation t-score (AdjcorT) feature selection method integrated with Radial Basis Function Neural Networks (RBFNN) to improve PM2.5 classification and prediction performance. Her research interests include feature selection, multicollinearity analysis, machine learning applications, and environmental data analytics. In addition to her academic journey, she is currently working full-time as a Data Analytics Executive at Daythree Digital Berhad, specializing in data analytics, reporting, and dashboard development.

LIST OF PUBLICATIONS

Arafin, S. K., Ibrahim, N., Mansor, M. M., & Ghani, N. A. M. (2025, October). A Comparative Study of Filter-Based Feature Selections for Improved PM2. 5 Prediction in Urban Area. In 2025 IEEE International Conference on Computing (ICOCO) (pp. 18-23). IEEE.

Arafin, S. K., Mazumdar, S., & Ibrahim, N. (2025). Improving Air Quality Prediction Models for Banting: A Performance Evaluation of Lasso, mRMR, and ReliefF. International Journal of Advanced Computer Science & Applications, 16(2).

Arafin, S. K., Ghani, N. A. M., Rosli, M. M., & Ibrahim, N. (2025). AdjcorT-RBFNN for Air Quality Classification: Mitigating Multicollinearity with Real and Simulated Data: 10.32526/ennrj/23/20250006. Environment and Natural Resources Journal, 23(3), 242-255.

Arafin, S. K., Ul-Saufie, A. Z., Md Ghani, N. A., & 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).

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.

Arafin, S. K., Ibrahim, N., Rahim, N. A. A., Mansor, M. M., Ul-Saufie, A. Z., & Ghani, N. A. M. (2024, December). Two-Stage Feature Selection for PM2. 5 Prediction: adjcor-TRBFNN Approach to Address Multicollinearity in Air Quality Data in Urban Area. In 2024 IEEE International Conference on Computing (ICOCO) (pp. 463-468). IEEE.

Arafin, S. K., Ibrahim, N., Mansor, M. M., Zahid, Z., & Md Ghani, N. A. (2024). A two-stage feature selection method to enhance prediction of daily PM2.5 concentration air pollution in Cheras. In Proceedings of the 5th International Conference on Computational Science and Information Management (ICoCSIM 2024) (pp. 103–108).