Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 11-2023
DOI: 10.5281/zenodo.10153164

A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR PREDICTING PM2.5 CONCENTRATION IN BEIJING: DATASET CHARACTERISTICS, HYPERPARAMETERS, AND TEMPORAL VARIABILITY

ALKISHRI MOOSA

PhD Student, Faculty of Communication, Visual Art and Computing, Universiti Selangor Malaysia.
Email: moosa.alkashari@utas.edu.om

NUR SYUFIZA AHMED SHUKOR

Associate Professor, Faculty of Communication, Visual Art and Computing, Universiti Selangor Malaysia.
Email: nur_syufiza@unisel.edu.my

JABAR H. YOUSIF

Associate Professor, Faculty of Computing and IT, Sohar University Oman. Email: jyousif@su.edu.om

Abstract

Researchers are making great efforts to develop novel, superior, and accurate machine learning (ML) models for air pollution prediction using area characteristics. However, a performance comparison is limited by several factors, as it is almost impossible to compare efficiency with all different models as the number of models proposed by researchers increases and different conditions and datasets are implemented. In addition, the results cannot be generalized to all future time periods because the characteristics of the area and the sources of pollution may vary from time to time. In this paper, we provide a periodic review of the state of the art in the application of ML techniques in the context of PM2.5 concentration prediction, focusing on the analysis of dataset size, hyper parameters, and preprocessing techniques applied in Beijing. Seven articles from 2015 to 2023 with 42 prediction models were collected and reviewed according to the same geographical area and dependent variable, PM2.5. In particular, we examined the hyper parameters of the models to describe the differences in model architecture. We also examine how using the same predictive model in a geographic area for the same pollutant at different times can result in different performance indices. The results show that it is not possible to prefer one predictive model over the other based on its performance at different times, even when applied at exactly the same location and with the same output.

Keywords: Air Pollution Prediction, Beijing, Machine Learning, PM2.5, Time Series Patterns.

1. INTRODUCTION

The environmental effects of total and partial pollutants are serious problems that directly or indirectly harm human, animal, and plant health as shown in Figure 1. The two most widely cited and regularly updated estimates for the death toll from air pollution come from the World Health Organization (WHO) and the IHME's Global Burden of Disease study. Their latest estimates are very close to each other – they estimate 7 million and 6.7 million deaths yearly, respectively [1]. Figure 2 shows the share of deaths attributed to air pollution. These deaths are attributed to indoor and outdoor pollution and – as explained below – stem from manufactured and natural sources of air pollution. PM2.5 causes many diseases, including respiratory, cardiovascular, and laryngeal cancers [2].

A study by [3] provides new longitudinal evidence that associates long-term exposure to PM2.5 components with increased mortality in Chinese adults. However, reducing the pollution risk can lower economic costs and preserve human lives by preventing stroke, lung cancer, and chronic and acute respiratory diseases.

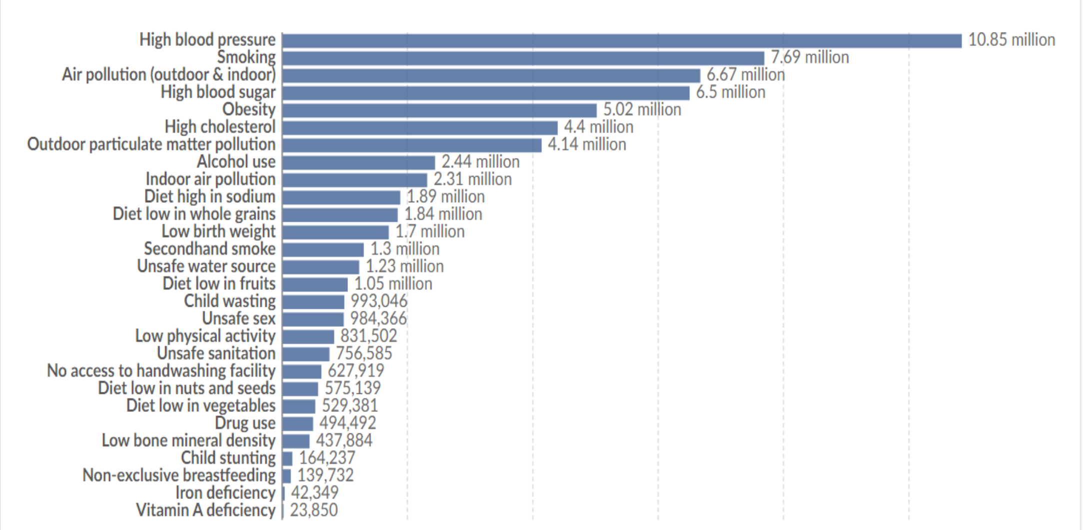

Figure 1:The estimated annual number of deaths attributed to each risk factor.[4]

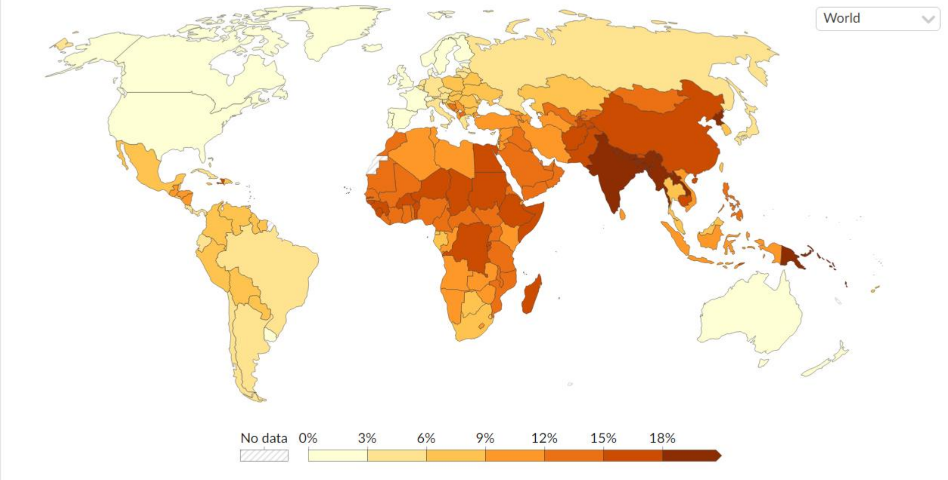

Figure 2: Share of deaths attributed to air pollution, 2019[4]

For instance, Song *et al.* found that implementing clean heating policies resulted in a 1.9 μgm^3 reduction in annual PM2.5 levels in mainland China between 2015 and 2021, potentially preventing 23,556 premature deaths in 2021 [5]. Researchers in whole the world developed several predictive models to support environmental management and prevent accidents to achieve this goal.

In Oman, for instance, a hybrid artificial neural network and mathematical models were suggested for evaluating particulate matter (PM2.5, PM10, and CO2) and assessing their impact on human health [6][7]. Reference [8] also proposed neural and mathematical predictive models to investigate the effects of particulate matter on human health in Oman. Given the severity of this issue, researchers have explored and recommended various models for examining gas levels and predicting future levels in Oman [9] and [10]. Furthermore, Reference [11] utilized a deep learning feedforward neural network model for predicting environmental risk factors. The area of air pollution prediction models is a crucial research area which using appropriate tools and mathematical models to analyze environmental data and accurately predict pollutant concentrations. Two main types of models are used in air pollution prediction: statistical and machine learning models (ML) [12]. Examples of statistical models include ARIMA [13] regression [14], grey models [15], and SVM [16]. The three models (APRIMA, SVM, regression-grey) provide excellent results. For instance, the SVM has been widely used in forecasting because of its outstanding performance in solving nonlinear problems [16]. On the other hand, ML is a branch of artificial intelligence (AI) that provides machines with the ability to learn to use a set of algorithms to process data and make predictions. Deep learning is a subfield of machine learning and the most advanced field in artificial intelligence. It designs machines capable of learning and thinking like humans, simulating the work of human brain neurons and neural networks (NNs) as shown in Figure 3.

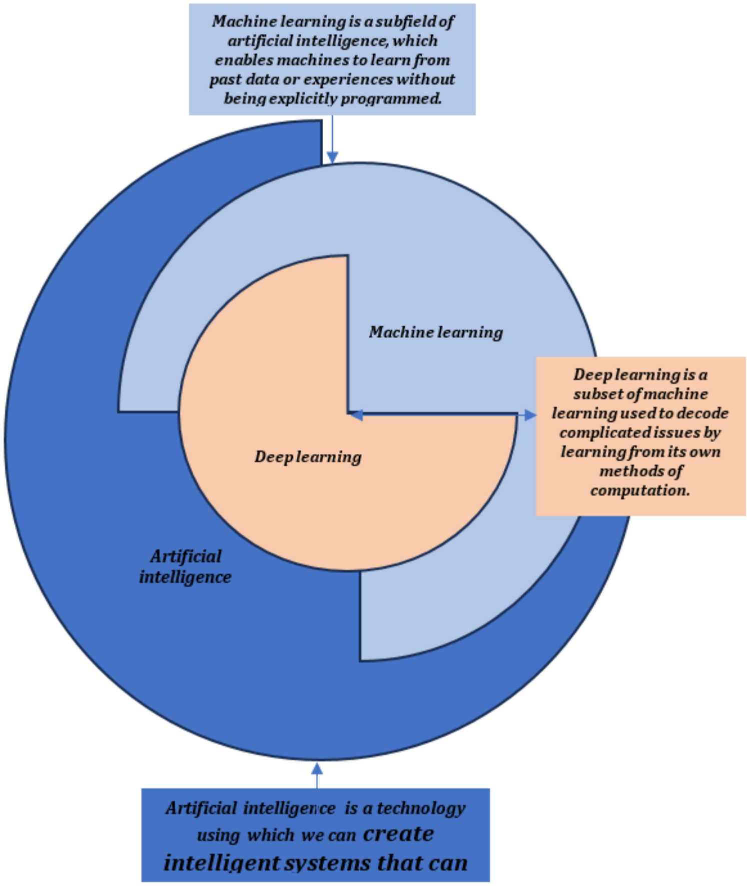
Machine learning is a subfield of artificial intelligence, which enables machines to learn from past data or experiences without being explicitly programmed.
Machine learning
Deep learning is a subset of machine learning used to decode complicated issues by learning from its own methods of computation.
Deep learning
Artificial intelligence
Artificial intelligence is a technology using which we can create intelligent systems that can

Figure 3: Relationship between artificial intelligence, machine learning, and deep learning.[17]

Throughout the literature review, it becomes evident that specific singular models and alternative methodologies have been amalgamated to create hybrid or ensemble models, thereby enhancing their performance. This combination aims to overcome specific issues the single models face. Several hybrid models can be considered: simple hybrid models, hybrid models based on data preprocessing methods, and intelligent hybrid models combining data processing and optimization algorithms [18]. Furthermore, there is another type of AI model called the meta-learning model. Meta-learning is a type of machine learning that deals with designing and applying algorithms that can learn from previous learning experiences [19]. Meta-learning models can learn to adapt to new tasks and data sets based on their previous experience. This model is helpful in settings with limited data available or where new data sets are constantly being introduced. Let us discuss each of these in more detail with an example for each:

Simple hybrid models involve combining multiple machine learning algorithms or models to improve performance. This improvement can be achieved through ensemble methods, where multiple models are trained independently, and their predictions are combined to make a final prediction. Popular ensemble methods include bagging (e.g., random forests) and boosting (e.g., AdaBoost, Gradient Boosting). These methods can help reduce bias variance and improve overall predictive accuracy. Random Forest is a popular example of a simple hybrid model that combines multiple decision trees to improve prediction accuracy. Each decision tree is trained on a different subset of the data, and their predictions are combined through majority voting or averaging. This ensemble approach helps to reduce overfitting and improve generalization [20].

Hybrid models based on data preprocessing methods combine different techniques before applying machine learning algorithms. Data preprocessing plays a crucial role in preparing the data for modeling, and different methods can be used to handle missing values, outliers, feature scaling, feature selection, and dimensionality reduction. By combining multiple preprocessing techniques, the hybrid models can leverage the strengths of each method and potentially improve model performance. For example, one might combine feature scaling techniques like standardization or min-max scaling with feature selection methods such as correlation-based feature selection or recursive feature elimination. The hybrid model can effectively preprocess the data and feed it into the subsequent machine-learning algorithm by applying a sequence of preprocessing steps. An example of a hybrid model based on data preprocessing methods is combining feature scaling and dimensionality reduction techniques. Feature scaling, such as standardization or min-max scaling, can be applied to normalize the numerical features. At the same time, dimensionality reduction techniques like Principal Component Analysis (PCA) can extract the most informative features and reduce the dimensionality of the dataset [21].

Intelligent hybrid models combine data processing techniques and optimization algorithms to improve model performance. These models aim to optimize not only the model parameters but also the data preprocessing steps to achieve the best possible results.

For instance, a hybrid model may use a metaheuristic optimization algorithm, such as genetic algorithms or particle swarm optimization, to search for the optimal combination of data preprocessing steps and hyperparameters of the machine learning model. The optimization algorithm can explore different preprocessing techniques, feature subsets, or parameter settings to find the best configuration that maximizes the model's performance on a given task. By combining data processing and optimization algorithms, intelligent hybrid models can adaptively learn and optimize both the data representation and the model parameters, leading to improved predictive accuracy and better overall performance. Genetic Algorithm-Based Feature Selection and Support Vector Machines (SVM) is an example of an intelligent hybrid model which combines a genetic algorithm-based feature selection technique with SVM as the classification model. Genetic algorithms can search for an optimal subset of features that maximizes the classification performance, while SVM is employed to build a predictive model based on the selected features [22].

Meta-learning algorithms are designed to quickly adapt to new tasks or data sets without retraining the model from scratch [23]. Researchers often use meta-learning algorithms when there is a need to adapt to changing conditions or when data are scarce rapidly. Meta-learning models have the advantage of being able to learn from multiple datasets and adapt to new datasets quickly, which can make them more efficient than traditional models. However, they can be more complex and require more training resources [24]. Incorporating hybrid, ensemble and meta-learning models, which combine multiple models to leverage their strengths, has also improved performance in many cases [25]. When comparing the performance of prediction models mentioned above within different geographic locations, various factors may arise, such as unique characteristics of the geographic location, meteorological, forecast horizon, and temporal features. Such external factors play a significant role in determining the accuracy of the predictions [14]. Each of them needs to be considered separately. As an example of the effect of temporal features, [14] observed that the characteristics of PM2.5 can vary hourly, daily, and monthly, with lower concentrations in the early morning and evening than at noon and late at night due to the sun effect. The daily activities might vary depending on the day of the week, with higher concentrations observed on Wednesdays, Fridays, and Sundays than the concentration on the other days. Furthermore, in terms of seasons, winter tends to have a greater PM2.5 concentration than spring and summer.

In summary, hybrid models can be simple combinations of different algorithms, leverage the strengths of various data preprocessing methods, or be intelligent models that combine data processing and optimization algorithms to improve performance in a more advanced and adaptive manner. The choice of the hybrid model depends on the specific problem at hand and the available resources and expertise. The results of the reviewed articles in this study support the statement in the literature that hybrid and ensemble models generally perform better than single models. According to [18], ensemble models outperformed single models in predicting stock prices. Similarly, Wang et al. found that a hybrid model combining deep learning and traditional machine learning techniques outperformed single models in predicting air quality [26].

2. BACKGROUND

2.1 Beijing

Beijing, the capital city of the People's Republic of China and home to over 21 million residents, has been grappling with significant air pollution issues, mainly related to PM2.5, a harmful particulate matter. The adverse effects of PM2.5 on human health have prompted the Chinese government to undertake various measures to reduce air pollution. These initiatives include implementing stringent air quality standards, investing in clean energy technologies, and promoting public transportation.

2.2 PM2.5

PM2.5 refers to tiny particles suspended in the air with a diameter of less than 2.5 microns. These particles comprise various solid and liquid materials, including ash, dust, and soot [27]. The sources of PM2.5 pollution can be diverse, including combustion processes in power generation, domestic heating, and vehicle emissions. Vehicles and industrial activities are the primary contributors to PM2.5 pollution. However, PM2.5 can also form through secondary processes in which sulfur emissions from industries react with oxygen and atmospheric water droplets, forming sulfuric acid, a secondary source of particulate matter [28].

PM2.5 is known to have significant adverse health effects. Exposure to these particles has been linked to respiratory issues, cardiovascular problems, and even laryngeal cancer [29]. In China, where PM2.5 pollution is a severe problem, high concentrations of PM2.5 contribute to approximately 1.6 million deaths annually [30]. The detrimental impact of PM2.5 on human health underscores the importance of monitoring and addressing this form of air pollution to protect public health.

3. PROBLEM STATEMENT

A study by Kokkinos et al. [31] concluded that the model is tied to characteristics of the geographic location, such as urban traffic and non-traffic, to predict pollutant concentrations accurately. The current review highlights some examples that indicate the model's correlation with the temporal characteristics of the area, which should not be overlooked. Although the researchers made great efforts to obtain the best results, the model performed differently when differences in the characteristics appeared. We came to the problem that the performance and efficiency of the model are limited to that place and time of the study, and the research must be continuous at each new study to obtain models with better results.

This challenge highlights the need to test multiple machine learning (ML) prediction models to identify the most accurate and suitable one for each area. Identifying the primary sources of pollutant emissions to enhance the forecasting capabilities is crucial. All input factors, including pollutant concentration and meteorological data, are considered inputs for the prediction model, and are divided into training and testing sets. When evaluating the models' performance, variations in evaluation methods, such as the selection of training and testing sets or the choice of performance metrics, can lead to differences in reported performance. [32].

To assess the performance of ML models, several commonly used performance indices are employed, including Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Percentage Absolute Error (MPAE). These indices help evaluate the accuracy of the models.

4. RESEARCH QUESTIONS

In this reviewed paper, several questions arise:

1) Is it feasible to develop a single model for a geographic area suitable for use at different times?
2) To what extent can an air pollution prediction ML model maintain accuracy within different periods in the same geographic area?
3) How well can an air pollution prediction ML model preserve accuracy within different distances in the same geographic area?
4) What factors influence the accuracy of the model?

In pursuit of our objectives, this paper undertakes a comprehensive review and exploration of pertinent literature, focusing on studies that have employed machine learning techniques to predict PM2.5 concentration. Our investigation is centered on analyzing key factors, including dataset size, hyperparameters, and preprocessing methodologies, specifically within the context of Beijing. Seven articles spanning the years 2015 to 2023 were meticulously gathered and evaluated, encompassing a total of 42 distinct prediction models. All the selected studies share a common geographic area and a singular dependent variable: PM2.5. Notably, we delve into a detailed examination of the hyperparameters utilized in these models to elucidate their architectural distinctions. Furthermore, our inquiry investigates how employing the same prediction model within a consistent geographic region for the same pollutant at different time intervals may yield divergent performance metrics.

This paper is structured in different sections. Section 1 presents an introduction to the problem statement and proposed method, and section 2 explores and reviews the recent literature studies and compares the performance of the suggested models. Section 3 presents the research methodology that will be implemented in this paper. Section 4 presents background about Beijing City and the levels of PM2.5 impact. Finally, it presents the discussion, results, conclusions, and future work.

5. LITERATURE REVIEW

This section explores and reviews the current literature of studies proposed for simulating and predicting mathematical models for classifying and examining the environmental issues and MP2.5 impact levels.

In a recent investigation conducted by Al-qaness [33], an updated Informer deep learning model known as the ResInformerStack was introduced. This model was systematically benchmarked against two other models, InformerStack and ResInformer, using multiyear datasets between 2014 to 2022. The outcomes of this comparative analysis highlighted the superior performance of the proposed ResInformerStack model, as evidenced by its lower RMSE and higher R^2 values, with the results (0.2852, 0.8285) for Informer, (0.2692, 0.8472) for InformerStack, and (0.2822, 0.8320) for ResInformer, respectively. The ResInformerStack model, in particular, demonstrated the most promising results with an RMSE of 0.2623 and an R^2 of 0.8549.

In [34], an attention mechanism was introduced for the prediction model ADST-ML (CNN + LSTM). This method was compared with several other models, including HA, Regression, ARIMA, Random Forest, MLP, LSTM, and LSTM-FC. The individual results for RMS (Root Mean Square) for a 1-hour ahead prediction are as follows: 56.866, 38.815, 36.235, 48.663, 27.534, 28.732, and 19.454, respectively. These results were obtained using PM2.5 and meteorological data from May 2014 to April 2015. Significantly, with a window size of 12, the RMSE measured 15.374. Meanwhile, for the ADST-ML model (CNN + LSTM) configured with a learning rate of 0.0009, the RMSE dropped to 10.974. These results underscore the model's effectiveness, particularly when it's combined with the attention mechanism. This study's limitation is that the missing rate in the data set for PM2.5 concentration is 13.26%, and the missing rate for weather is 14.52%. This limitation will make the proposed model unable to forecast or handle unforeseen events and limit its generality.

In the literature review, Du et al. [12] proposed an optimized extreme learning machine (ELM) model called the TVF-EMD-HHO-ELM model. They compared this model with several other models, including the Persistence model, TVF-EMD-ELM, VMD-HHO-ELM, ICEEMDAN-HHO-ELM, TVF-EMD-SCA-ELM, and TVF-EMD-HHO-ELM. The evaluation was performed using data collected from April 1, 2018, to August 10, 2019, specifically for the pollutant PM2.5.

The evaluation results were reported regarding two performance metrics: root mean square error (RMSE) and coefficient of determination (R^2). Here are the results for each model. The persistence Model is often used as a baseline model. The root mean square error (RMSE) for this model was 18.2972, indicating that, on average, the model's predictions were off by about 18.3 units. The R-squared (R^2) value, which measures how well the model explains the variance in the data, was 0.4654, suggesting that this model had limited predictive power. TVF-EMD-ELM model incorporates time-varying filter-based empirical mode decomposition (TVF-EMD) and extreme learning machine (ELM). It performed better than the Persistence model with an RMSE of 14.5847, making its predictions more accurate. The R^2 value also improved significantly to 0.8883, indicating a better fit to the data. VMD-HHO-ELM is a model that uses Variational Mode Decomposition (VMD), Harris Hawks Optimization (HHO), and ELM. It further improved the RMSE to 5.7581, signifying even more accurate predictions. The R^2 value reached 0.9721, indicating a solid explanatory capability. ICEEMDAN-HHO-ELM model employs Intrinsic Mode Functions (IMFs) Complementary Ensemble Empirical Mode

Decomposition with Adaptive Noise (ICEEMDAN) and HHO optimization with ELM. It outperformed the previous models with an RMSE of 5.4582 and an R^2 of 0.9524, demonstrating accurate predictions.

TVF-EMD-SCA-ELM is a model that combines TVF-EMD, Single Component Analysis (SCA), and ELM. It achieved an even lower RMSE of 1.8169, indicating highly accurate predictions. The R^2 value was exceptionally high at 0.9965, suggesting an excellent fit to the data. TVF-EMD-HHO-ELM is the proposed model in the study, incorporating TVF-EMD, HHO optimization, and ELM. It performed the best among all the models with an RMSE of 0.9442, which implies the most accurate predictions. The R^2 value was extremely high at 0.9986, indicating an exceptional ability to explain the variance in the data.

In summary, with the move from the Persistence model to more complex models, the RMSE decreases, indicating improved predictive accuracy. The R^2 value also increases, demonstrating a better fit of the model to the data. The TVF-EMD-HHO-ELM model stands out as the best-performing model in terms of RMSE and R^2, suggesting it is highly effective for PM2.5 concentration prediction. In a study [14] used a dataset of several meteorological factors, including wind speed, wind direction, temperature, precipitation, pressure, relative humidity, and pollutant concentrations (PM2.5, PM10, NO2, SO2, O3, CO). The performances of several models were compared, including linear regression (LR), autoregressive integrated moving average (ARIMA), support vector regression (SVR), backpropagation neural network (BPNN), long short-term memory (LSTM), gated recurrent unit (GRU), deep temporal convolutional neural network (DeepTCN), and other models that use decomposition methods. The study found that CEEMDAN + SVR had the lowest RMSE (1.2813) for one h, followed closely by CEEMDAN + LSTM (1.2892) and CEEMDAN + DeepTCN (1.1064). Some ML models can robustly handle missing data, but some might need extra supporting methods and techniques [35]. To overcome this issue in [14], fill in the missing values using the linear interpolation method to obtain a continuous time series. In a study by Sun and Li, various meteorological inputs, including humidity, were used along with air pollutants (humidity, PM10, SO2, NO2, O3, and CO) to evaluate the performance of several models, including BPNN, IBPNN, ELM, LSSVR, and stacking (PACF+SCC+BPNN+IBPNN+ELM) [36]. It found that the stacking model had the lowest RMSE (3.15) and the highest R^2 (0.999). The limitations of this work arise from adding two hidden layers to BPNN, the risk of overfitting and increased computational cost. On the other hand, two data processing methods were utilized: the partial autocorrelation function (PACF) and the Spearman correlation coefficient SCC.

In [37], they proposed a model for examining the air pollutants concentrations (PM2.5, SO2, NO2, CO, and O3), which achieved a better RMSE using ICEEMDAN as a decomposition tool and the prediction model ICA-BPNN. Feng et al. proposed an ICEEMDAN-ICA-BPNN model that utilized a one-year dataset covering November 2016 to July 2017. The proposed model was compared with several other models to prove its superiority, including the autoregressive integrated moving average (ARIMA), generalized regression neural network (GRNN), and neural networks (BPNN, SBO-BPNN, ICA-BPNN, and ICEEMDAN-ICA-BPNN). The results of the ICEEMDAN-ICA-

BPNN model outperform the other models, with RMSE of 1.8902 and R^2 of 0.9955. It is found that adding more layers can improve the generalization ability and avoid local minima, but this will increase the complexity of the model. Instead, advanced heuristic algorithms such as ICA have been proposed to overcome these challenges and to optimize the weight and thresholds of the BP network. The limitation of this study is the small size of the data set and the exclusion of the meteorological data. Feng et al. [32] used a one-year dataset from September 2013 to October 2014. This dataset includes meteorological data and pollutant concentrations (PM2.5, PM10, NO2, SO2, O3, CO). A backpropagation neural network (BP) compared to models that incorporate trajectory and wavelet transformation. The study proves that adding a trajectory and wavelet transformation improved the model performance regarding RMSE. For instance, the RMSE for one day using BP alone was 28.63, but the RMSE decreased to 15.65 with the BP + trajectory model + wavelet. The proposed model faced different challenges, such as BP suffering from local minima with complex mapping and missing data with a large standard deviation of PM2.5 concentration.

According to the literature of studies, the following challenges were examined:

- The size and quality of the data set
- The continuity of the accuracy of the prediction model
- The computational cost of the proposed model

The contribution of the reviewed papers and the research gaps are summarized in Table 1.

	Model	RMSE	R^2	Data Set	Contribution	Limitations
[38]	BP	28.63		September 1, 2013, to October 31, 2014 Meteorological, PM2.5, PM10, NO2, SO2, O3, CO)	a novel hybrid model that combines air mass trajectory analysis and wavelet transformation to enhance the forecast accuracy of daily average concentrations of PM2.5 two days in advance using an artificial neural network (ANN) BP alone suffered from local minima	Some data were missing due to instrument malfunctions. In contrast, the authors considered the days with consecutive hourly gaps of more than four hours or the cumulative amount of missing data exceeding eight hours to be discarded.
	BP + trajectory model	24.84				
	BP+ trajectory model +wavelet	15.65				
[37]	ARIMA	23.4705	0.4143	November 1, 2016, to July 31, 2017 Air pollutants (PM2.5, SO2, NO2, CO, O3)	1. Achieving good performance in accuracy and effectiveness in the designed early warning system. 2. Develop a real-time air quality forecasting system to support haze management.	The data set is small, and the meteorological data is excluded.
	GRNN	10.2774	0.8783			
	BPNN	6.9356	0.9402			
	SBO-BPNN	6.8130	0.9434			
	ICA-BPNN	6.7684	0.9440			
	ICEEMDAN-ICA-BPNN	1.8902	0.9955			
[36]	BPNN	31.31	0.969	Data set during Jan 2017 Humidity PM10 , SO2 , NO2 , O3, CO	Use of PACF and SCC for data processing.	The limitations of this work arise from adding two hidden layers to BPNN, the risk of overfitting and increased computational cost.
	IBPNN	23.88	0.984			
	ELM	28.39	0.974			
	LSSVR	35.77	0.947			
	Stacking (PACF+SCC + BPNN+IBPNN+ELM+LSSVR)	3.15	0.999			
[14]	LR	5.259	CEEMDANLR 1.4950	w s, w d, temp, precipitation, pressure, relative humidity, M2.5, PM10, NO2, SO2, O3, CO Two years data Without exogenous variables, 1.9275	Linear interpolation fills missing values in the dataset, enabling continuous time series. Investigating extracted multi-scale components of PM2.5 concentrations and multi-factor information of	This method may not work well for data with irregular or complex patterns, whereas outliers in the data can affect interpolation accuracy.
	ARIMA	5.2491				
	SVR	5.5224	CEEMDAN+SVR 1.2813			
	BPNN	5.6460	CEEMDAN + BPNN 1.4971			
	LSTM	4.7490	CEEMDANLSTM 1.2892			

	GRU	4.7750	CEEMDANGRU 1.3521	With pollutant concentrations, time variables and meteorological factors 1.1064	exogenous variables improves prediction performance.	
	DeepTCN	4.5710				
	EEMD + LSTM	2.9203				
	EMD + GRU	2.623				
	CEEMDAN + DeepTCN	1.1064				
[12]	Persistence model	18.2972	0.4654	PM2.5 only April 1, 2018, to August 10, 2019	obtaining higher prediction accuracy by using an optimized extreme learning machine (ELM) with the proposed model	The meteorological parameter and other pollutant concentrations are excluded from being input in this model to evaluate the computational cost of the proposed model.
	TVF-EMD-ELM	14.5847	0.8883			
	VMD-HHO-ELM	5.7581	0.9721			
	ICEEMDAN-HHO-ELM	5.4582	0.9524			
	TVF-EMD-SCA-ELM	1.8169	0.9965			
	TVF-EMD-HHO-ELM	0.9442	**0.9986**			
[34]	HA	56.866		PM2.5 and meteorological data from May 1, 2014 to April 30, 2015. Learning rate (**0.0009**) (RMSE **10.974**) Window size 12 (RMSE **15.374**)	When conducting experiments on a dataset containing Beijing air quality inspection records, ADST-ML outperforms baselines regarding prediction accuracy.	The dataset has a high missing rate (13.26% for PM2.5, 14.52% for weather), limiting forecasting capabilities.
	Regression	38.815				
	ARIMA	36.235				
	Random forest	48.663				
	MLP	27.534				
	LSTM	28.732				
	LSTM-FC	19.454				
	ADST-ML(CNN + LSTM)+Attention mechanism	**15.812**				
[33]	Informer (deep learning)	0.2852	0.8285	data set PM2.5 only, from January 1, 2014 to February 17, 2022,	1. Supporting the collection of daily symptoms and high-quality weather data, allowing examination of the relationship between weather and pain. 2. Demonstrating significant relationships between relative humidity, pressure, wind speed and pain, with correlations remaining even when accounting for mood and physical activity.	The effect of weather on pain was not fully explained by its day-to-day effect on mood or physical activity.
	InformerStack	0.2692	0.8472			
	ResInformer	0.2822	0.8320			
	ResInformerStack	0.2623	0.8549			

Table 1: Summary of the model's contribution and limitations

6. RESEARCH METHODOLOGY

This paper comprehensively reviews and explores the recent studies related to simulating and predicting mathematical models for classifying and examining the rate of PM2.5 impact. The research determines the method for selecting the needed papers in this review. Also, Beijing city in China is used as a case study due to its high levels of PM2.5. The selected models are then compared using various performance factors, which include Root Mean Square Error (RMSE) and coefficient of Determination (R^2), to determine the most effective method for predicting the impact of PM2.5 pollution as presented in equations 1, 2, and 3. The paper also discusses the limitations of current models and suggests areas for future research in this field.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_t - y_t^{\wedge}\right)^2} \qquad \text{.....(1)}$$

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N}\left|y_t - y_t^{\wedge}\right| \qquad \text{.....(2)}$$

$$\text{MPAE} = \frac{1}{N}\sum_{t=1}^{N}\left|1 - (y_t^{\wedge}/y_t)\right| \qquad \text{.....(3)}$$

7. RESULTS AND DISCUSSION

A spectrum of models has been explored in the realm of PM2.5 prediction in Beijing. Feng et al. offer a moderate RMSE of 15.65 [32]. In contrast, P. Jiang et al. shine with an impressive RMSE of 1.8902, catering to real-time forecasting needs [37].

Sun & Li employs a diverse stacking model with an RMSE of 3.15, hinting at potential optimizations [36]. F. Jiang et al. integrate advanced techniques, yielding an RMSE of 1.1064 [14]. Du et al. stand out with a notably low RMSE of 0.9442, emphasizing model simplicity [40]. On the other hand, Yang and Zhang exhibit an RMSE of 15.812, suggesting room for architectural improvements [34]. Al-qaness et al.'s model notably excels with an exceptionally low RMSE of 0.2623.

The choice of model hinges on specific requirements, with Du et al.'s and Al-qaness et al.'s models standing out for their precision. Other models may benefit from further fine-tuning, highlighting the evolving landscape of PM2.5 prediction in Beijing. Figure 4 demonstrates the enhanced accuracy of prediction models in Beijing between 2015 and 2023, except for the 2023(1) model proposed in reference [34]. This improvement can be attributed to the effective tuning of hyperparameters in deep learning models and advancements in dataset quality and size.

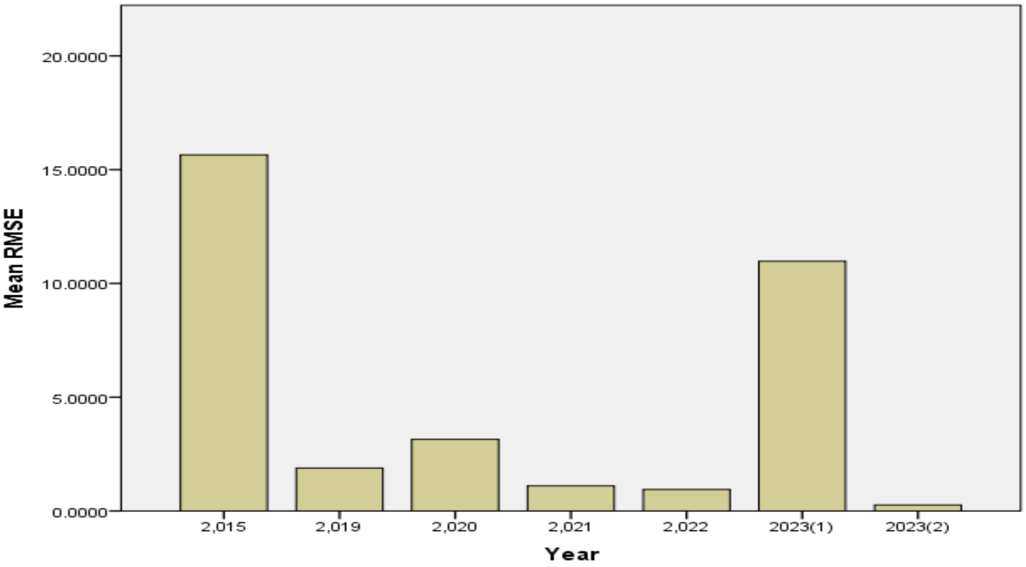

Figure 4: The RMSE of the reviewed papers

The choice of the source and the size of dataset is pivotal in machine learning. Some studies, like [39], [37], [12], [40], and [34], opt for one-year data, while others, such as [14] [33], extend to three and eight years, respectively as shown in Table 2. There is also variation in the inclusion of inputs. Feng et al. [39], F. Jiang et al. [14], and Yang & Zhang [34] integrate meteorological and air pollutant data. In contrast, P. Jiang et al. [37] focus on PM2.5 and air pollutants, and Du et al. [40] and Alqaness et al. [33] exclusively use PM2.5. This diversity underscores the challenges researchers face in collecting data and understanding the significance of certain input factors. For instance, Sun & Li removed temperature as an input due to its limited impact on PM2.5 hourly. The final choice of input features is often a complex, context-specific decision [36].

Following data collection and preprocessing, hyperparameter tuning becomes imperative. Hyperparameters are preset configurations that significantly influence how a model learns and performs. Each machine learning model demands its unique set of hyperparameters tailored to the dataset for optimal results. The complexity of a model, measured by the number of layers, impacts accuracy and the risk of overfitting. While the studies reviewed employ diverse hyperparameter tuning approaches (as detailed in Table 3), it's worth noting that Al-qaness et al. [33] achieved the best RMSE with low computational costs, highlighting the crucial role of hyperparameter optimization in enhancing predictive accuracy. This achievement underscores the critical role of hyperparameter optimization in enhancing predictive accuracy.

Table 2: Data set sources

Ref	Data set Source	Data set size	Data set elements
[39]	Jing-Jin-Ji Station	One year	Meteorological, PM2.5, PM10, NO2, SO2, O3, CO
[37]	Ministry of Environment Protection	One year	PM2.5, SO2, NO2, CO, O3
[36]	http://www.pm25.com/city.	One month	Humidity PM10 , SO2 , NO2 , O3, CO
[14]	https://www.heweather.com	Three years	Meteorological, PM2.5, PM10, NO2, SO2, O3, CO
[12]	https://www.aqistudy.cn/.	One year	PM2.5 only
[34]	Air Project	One year	PM2.5 and meteorological
[33]	(https://aqicn.org/data-platform/covid19/,.	Eight years	PM2.5 only

Table 3: Hyperparameters

Ref	Hyperparameter	The model	Horizon of Prediction	RMSE
[39]	First layer 10 Second layer 8	BP+ trajectory model +wavelet	Daly	15.65
[36]	First layers = 6 nods Second layers = 3 nods, learning rate = 0.1; iteration = 100; goal = 0.00004	Stacking (PACF+SCC + BPNN+IBPNN+ELM +LSSVR)	Hourly	3.15
[14]	The training epochs were 100, the batch size was 128, and the learning rate was 0.01.	CEEMDAN + DeepTCN	Hourly	1.1064
[34]	Three-layer CNN: First layer: - input channels = 8, - output channels = 32, kernel size = 3, stride = 1, padding = 1). Second layer: - input channels = 32, - output channels = 64, kernel size = 3, stride = 1, padding = 1). Third layer: - input channels = 64, - output channels = 128, kernel size = 3, stride = 1, padding = 1). inference network is that the number of input channels of the first layer CNN is 7, input in LSTM layer = 128. Learning rate. = (0.0009)	ADST-ML(CNN + LSTM)+ Attention mechanism	Hourly	10.974
[33]	Two encoder layers and one decoder layer with eight attention heads. The ADAM optimizer with an initial learning rate of 1E−4, batch size = 16 for 50 epochs. The early stop is within ten epochs.	ResInformerStack	Hourly	0.8549

Based on these findings, it can be concluded that utilizing hybrid, ensemble, and meta-learning approaches can enhance the performance, and we can conclude these points in the proposed problem and research questions (1, 2,3 and 4):

The answer of question 1 “Is it feasible to develop a single model for a geographic area suitable for use at different times?”

Feasibility of developing a single model for a geographic area across different times:

- Developing a single model for air pollution prediction in a specific geographic area poses challenges due to variations in pollutant emissions, meteorological conditions, and other factors over time.
- Temporal variations, such as seasonal changes, weather patterns, and human activities, can impact the effectiveness of a model trained on data from a specific time.
- Continuous research and model updates are necessary to account for temporal variations and improve the accuracy of air pollution prediction models over time.
- The answer of 2 “To what extent can an air pollution prediction ML model maintain accuracy within different periods in the same geographic area?”

Maintaining the accuracy of the same ML model across various periods within the same geographic area:

- The accuracy of an ML model for air pollution prediction can vary across different periods within the same geographic area.
- Changes in pollutant sources, emission patterns, and meteorological conditions can affect the model's performance.
- Regular model updates with new data and incorporating real-time data and adaptive modeling techniques can enhance accuracy across different periods.
- The answer of question 3 “How well can an air pollution prediction ML model preserve accuracy within different distances in the same geographic area?”

Maintaining the accuracy of the same ML model across different distances within the same geographic area:

- The distance between monitoring locations can influence the accuracy of an ML model for air pollution prediction within the same geographic area.
- Spatial variations in pollution sources, terrain, and local meteorological conditions can lead to variations in air pollution levels.
- Developing localized models or incorporating spatial factors as input features can improve accuracy across different distances within the same geographic area.

The answer of question 4 “What factors influence the accuracy of the model?”

Factors influencing the accuracy of air pollution prediction models:

- The quality and representativeness of input data, including pollutant concentrations, meteorological conditions, emission sources, and other relevant factors, are crucial for model accuracy.

- The variation in the Spatial and temporal of the study area impacts pollutant levels and model performance.
- The choice of ML algorithms, model complexity, and feature selection can affect prediction accuracy.
- Sufficient and diverse training data, representative of the target area and time are essential for building accurate models.
- Appropriate model evaluation and validation techniques, along with the selection of suitable evaluation metrics, are essential for assessing model performance.

8. CONCLUSION

Regarding the challenges associated with PM2.5 prediction, ordinary single models perform worse than hybrid, ensemble, and meta-learning models. Moreover, it appears from several studies that differences in RMSE values of the same model are to be expected even when performed at the same location. As mentioned earlier, there are several reasons for these differences:

1. Differences in data sets: Different studies may use different data sets with different sizes, temporal resolutions, and input parameters. These differences can affect the performance of the model.
2. Model configuration: even if the same model is used, different studies may use different formats, such as the number of hidden layers, number of neurons, learning rate, activation functions, and regularization techniques. These differences may also affect the performance of the model.
3. Evaluation metrics: although RMSE is a common evaluation metric, other metrics such as mean absolute error (MAE), coefficient of determination (R-squared), and mean absolute error percentage (MAPE) can also be used to evaluate model performance. The use of different metrics can lead to different results.
4. External Factors: There may be external factors that affect air pollution, such as weather patterns, traffic congestion, and industrial activities. These factors may change over time and affect the performance of the model.

It is important to carefully evaluate the results of different studies and consider the differences in data sets, model configurations, and evaluation scales when comparing the performance of other models.

Therefore, there is no definitive answer to the question of which model was the best among those used in these studies. Overall, the choice of the best model depends on the specific context and problem, as well as the amount and quality of available data and resources. Top of Form When the performance of the same model ML varies according to changes in the characteristics of the same area, the preference of one model over another is a relative issue and not a clear preference. The paper concludes that the comparison between air pollution prediction models must be made under the same conditions and with the same future characteristics. The solution proposed in this paper is the need for automatic tuning ML air pollution model is necessary. The current literature is limited to only seven studies. However, more studies need to be added for the same area to find more critical features that can be discussed.

References

1) M. Roser, “Data Review: How many people die from air pollution?,” *Our World Data*, 2021.

2) D. R. Liu, S. J. Lee, Y. Huang, and C. J. Chiu, “Air pollution forecasting based on attention-based LSTM neural network and ensemble learning,” *Expert Syst.*, vol. 37, no. 3, 2020, doi: 10.1111/exsy.12511.

3) L. Liu *et al.*, “Longitudinal Impacts of PM2.5Constituents on Adult Mortality in China,” *Environ. Sci. Technol.*, vol. 56, no. 11, pp. 7224–7233, 2022, doi: 10.1021/acs.est.1c04152.

4) H. Ritchie and M. Roser, “Air Pollution,” *Our World Data*, 2017.

5) C. Song *et al.*, “Attribution of Air Quality Benefits to Clean Winter Heating Polices in China: Combining Machine Learning with Causal Inference,” *Environ. Sci. Technol.*, 2022, doi: 10.1021/acs.est.2c06800.

6) N. Alattar and J. Yousif, “Evaluating Particulate Matter (PM2.5 and PM10) Impact on Human Health in Oman Based on a Hybrid Artificial Neural Network and Mathematical Models,” in *Proceedings - 2019 3rd International Conference on Control, Artificial Intelligence, Robotics and Optimization, ICCAIRO 2019*, 2019, pp. 129–135. doi: 10.1109/ICCAIRO47923.2019.00028.

7) J. H. Yousif, N. N. Alattar, and M. A. Fekihal, “Forecasting models based CO2 emission for sultanate of Oman,” *Int. J. Appl. Eng. Res.*, vol. 12, no. 1, pp. 95–100, 2017.

8) N. Alattar, J. Yousif, M. Jaffer, and S. A. Aljunid, “Neural and mathematical predicting models for particulate impact on human health in oman,” *WSEAS Trans. Environ. Dev.*, vol. 15, no. February, pp. 578–585, 2019.

9) Jabar yousif, “Implementation of Big Data Analytics for Simulating, Predicting and Optimizing the Solar Energy Production,” *Appl. Comput. J.*, vol. 3, no. Issue 3, pp. 133–140, 2021, doi: 10.52098/acj.202140.

10) D. K. Saini and J. H. Yousif, “Environmental Scrutinizing System based on Soft Computing Technique,” *International Journal of Computer Applications*, vol. 62, no. 13. pp. 975–8887, 2013.

11) Y. KHAMIS and J. H. Yousif, “Deep learning Feedforward Neural Network in predicting model of Environmental risk factors in the Sohar region,” *Artif. Intell. &amp; Robot. Dev. J.*, vol. 2, no. 3, pp. 201–2013, 2023, doi: 10.52098/airdj.202257.

12) P. Du, J. Wang, W. Yang, and T. Niu, “A novel hybrid fine particulate matter (PM2.5) forecasting and its further application system: Case studies in China,” *J. Forecast.*, vol. 41, no. 1, pp. 64–85, 2022, doi: 10.1002/for.2785.

13) D. J. Liu and L. Li, “Application study of comprehensive forecasting model based on entropy weighting method on trend of PM2.5 concentration in Guangzhou, China,” *Int. J. Environ. Res. Public Health*, vol. 12, no. 6, pp. 7085–7099, 2015, doi: 10.3390/ijerph120607085.

14) F. Jiang, C. Zhang, S. Sun, and J. Sun, “Forecasting hourly PM2.5 based on deep temporal convolutional neural network and decomposition method,” *Appl. Soft Comput.*, vol. 113, 2021, doi: 10.1016/j.asoc.2021.107988.

15) L. Wu, X. Gao, Y. Xiao, and S. Liu, “Using grey Holt – Winters model to predict the air quality index for cities in China,” *Nat. Hazards*, 2017, doi: 10.1007/s11069-017-2901-8.

16) Y. Xu, W. Yang, and J. Wang, “Air quality early-warning system for cities in China,” *Atmos. Environ.*, vol. 148, pp. 239–257, 2017, doi: 10.1016/j.atmosenv.2016.10.046.

17) M. Lotfian, J. Ingensand, and M. A. Brovelli, “The partnership of citizen science and machine learning: Benefits, risks and future challenges for engagement, data collection and data quality,” *Sustain.*, vol. 13, no. 14, 2021, doi: 10.3390/su13148087.

18) S. Zhang, X. Li, Y. Li, and J. Mei, “Prediction of Urban PM2.5 Concentration Based on Wavelet Neural Network,” *Proc. 30th Chinese Control Decis. Conf. CCDC 2018*, pp. 5514–5519, 2018, doi: 10.1109/CCDC.2018.8408092.

19) M. Noguer i Alonso, G. Batres-Estrada, and G. Yahiaoui, “A Meta-Learning approach to Model Uncertainty in Financial Time Series,” *SSRN Electron. J.*, no. April, pp. 1–27, 2021, doi: 10.2139/ssrn.3814938.

20) L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

21) T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

22) M. Dash and H. Liu, “Feature selection for classification,” *Intell. data Anal.*, vol. 1, no. 1–4, pp. 131–156, 1997.

23) C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” *34th Int. Conf. Mach. Learn. ICML 2017*, vol. 3, pp. 1856–1868, 2017, [Online]. Available: https://proceedings.mlr.press/v70/finn17a.html

24) K. Rakelly, A. Zhou, D. Quiilen, C. Finn, and S. Levine, “Efficient off-policy meta-reinforcement learning via probabilistic context variables,” *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 9291–9301, 2019.

25) M. Neshat *et al.*, “A deep learning-based evolutionary model for short-term wind speed forecasting: A case study of the Lillgrund offshore wind farm,” *Energy Convers. Manag.*, vol. 236, no. November 2020, p. 114002, 2021, doi: 10.1016/j.enconman.2021.114002.

26) J. Wang, P. Du, Y. Hao, X. Ma, T. Niu, and W. Yang, “An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting,” *J. Environ. Manage.*, vol. 255, 2020, doi: 10.1016/j.jenvman.2019.109855.

27) M. A. Kioumourtzoglou *et al.*, “Long-term PM2.5 exposure and neurological hospital admissions in the northeastern United States,” *Environ. Health Perspect.*, vol. 124, no. 1, pp. 23–29, 2016, doi: 10.1289/ehp.1408973.

28) World Health Organization, *Air Quality Guidelines: Global Update 2005*. [Online]. Available: https://books.google.com.om/books?id=7VbxUdlJE8wC&lpg=PR9&ots=w3a5qOPcvg&dq=World Health Organization%2C and UNAIDS Air Quality Guidelines%3A Global Update 2005%2C World Health Org.%2C Geneva%2C Switzerland%2C 2006.&lr&pg=PP1#v=onepage&q&f=false

29) D. R. Liu, S. J. Lee, Y. Huang, and C. J. Chiu, “Air pollution forecasting based on attention-based LSTM neural network and ensemble learning,” *Expert Syst.*, vol. 37, no. 3, pp. 1–16, 2020, doi: 10.1111/exsy.12511.

30) J. He and G. Christakos, “Space-time PM 2 . 5 mapping in the severe haze region of Jing-Jin-Ji (China) using a synthetic approach *,” *Environ. Pollut.*, vol. 240, pp. 319–329, 2018, doi: 10.1016/j.envpol.2018.04.092.

31) K. Kokkinos, V. Karayannis, E. Nathanail, and K. Moustakas, “A comparative analysis of Statistical and Computational Intelligence methodologies for the prediction of traffic-induced fine particulate matter and NO2,” *J. Clean. Prod.*, vol. 328, 2021, doi: 10.1016/j.jclepro.2021.129500.

32) S. S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, and A. Stolcke, “Sri ’ S 2004 Nist Speaker Recognition Evaluation System Sri International , Menlo Park , CA , USA,” *Word J. Int. Linguist. Assoc.*, pp. 173–176, 2005.

33) M. A. A. Al-qaness *et al.*, “ResInformer: Residual Transformer-Based Artificial Time-Series Forecasting Model for PM2.5 Concentration in Three Major Chinese Cities,” *Mathematics*, vol. 11, no. 2, pp. 1–17, 2023, doi: 10.3390/math11020476.

34) X. Yang and Z. Zhang, “An attention-based domain spatial-temporal meta-learning (ADST-ML) approach for PM2.5 concentration dynamics prediction,” *Urban Clim.*, vol. 47, no. October 2022, p. 101363, 2023, doi: 10.1016/j.uclim.2022.101363.

35) S. M. Miraftabzadeh, M. Longo, F. Foiadelli, M. Pasetti, and R. Igual, “Advances in the application of machine learning techniques for power system analytics: A survey†,” *Energies*, vol. 14, no. 16, 2021, doi: 10.3390/en14164776.

36) W. Sun and Z. Li, “Hourly PM2.5 concentration forecasting based on feature extraction and stacking-driven ensemble model for the winter of the Beijing-Tianjin-Hebei area,” *Atmos. Pollut. Res.*, vol. 11, no. 6, pp. 110–121, 2020, doi: 10.1016/j.apr.2020.02.022.

37) P. Jiang, C. Li, R. Li, and H. Yang, “An innovative hybrid air pollution early-warning system based on pollutants forecasting and Extenics evaluation,” *Knowledge-Based Syst.*, vol. 164, pp. 174–192, 2019, doi: 10.1016/j.knosys.2018.10.036.

38) X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, “Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation,” *Atmos. Environ.*, vol. 107, pp. 118–128, 2015, doi: 10.1016/j.atmosenv.2015.02.030.

39) X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, and J. Wang, “Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation,” *Atmos. Environ.*, vol. 107, pp. 118–128, 2015, doi: 10.1016/j.atmosenv.2015.02.030.

40) P. Du, J. Wang, W. Yang, and T. Niu, “A novel hybrid fine particulate matter (PM2.5) forecasting and its further application system: Case studies in China,” *J. Forecast.*, vol. 41, no. 1, pp. 64–85, 2022, doi: 10.1002/for.2785.