Environmental Protection: Pollution Level Prediction

ALEXANDER ALBOUKHARI, Applied Data Science, Noroff University College, Norway

In order to reduce the negative health impacts of air pollution especially (PM2.5) which is prevalent in cities, reliable and up to date prediction models are required. Using past air quality and weather data, this research investigates the feasibility of using machine learning methods to forecast PM2.5 concentrations. The effectiveness of many models in predicting PM2.5 levels was assessed. These models included Random Forest, LSTM , and multiple regression techniques. With the best accuracy measured by the lowest (MAE and RMSE) the Random Forest model successfully captured intricate patterns in the dataset. In contrast, the LSTM model demonstrated excellent skills in capitalizing on temporal relationships, which rendered it appropriate to time-series analysis. The potential of combining classic regression with advanced deep learning approaches to improve prediction accuracy is demonstrated by our results. The incorporating the machine learning models into air quality monitoring systems in real-time is highlighted by this research, which provides useful information for lawmakers to take preventative actions for the environment and public health

1 INTRODUCTION

One of the biggest problems affecting people's health especially in cities is air pollution. The particulate matter (PM2.5) that enters the respiratory system and causes a number of health problems, such as respiratory infections and heart disease, makes it one of the most harmful pollutants. Early alerts give the ability to act quickly, and better public health outcomes all depend on accurate PM2.5 level forecasts. However, predicting air pollution concentrations is difficult, as the environmental factors that impact them are multiple and variable.

Recent years have seen an increase of machine learning approaches as powerful air quality prediction tools, thanks to their ability to process massive datasets and identify non-linear correlations between variables. In order to effectively estimate PM2.5 levels, this study will examine several machine learning models, such as Random Forest and Long Short Term Memory (LSTM) among other regression approaches. The need to improve air quality forecasts using advanced algorithms that can draw on past data and a variety of meteorological elements is what drives this study. Not only does this study help improve air quality forecasting technology, it also helps authorities protect public health by creating reliable models to determine the effective model for PM2.5 prediction.

2 RELATED WORK

The studies reviewed focus on the wider impacts of air pollution and the various methodologies for the prediction of PM2.5. Together, they make a substantial contribution to the understanding of advanced modeling techniques and underscore the impact of precise air quality forecasting in terms of public health and socioeconomic consequences. The following section will discuss the alignment of each study with the objectives of this research, which are to evaluate various machine learning techniques, develop a robust model for PM2.5 prediction, and comprehend the implications for urban environments.

By integrating a variety of data sources, including meteorological data such as dew point and wind speed, [4]introduced a novel method for long-term PM2.5 predictions that employs the Informer architecture. Compared to more conventional models, their methodology demonstrated substantial increases in accuracy, particularly for expanded prediction horizons. This objective is in line with the objective of the present investigation of assessing advanced deep learning models to improve the accuracy of PM2.5 forecasting. This study validates the methodology of using multisource data

Author's address: Alexander Alboukhari, Applied Data Science, Noroff University College, Tordenskjolds gate 9, Kristiansand, Agder, Norway, 4612.

to improve model performance by highlighting the importance of combining multiple data sources and addressing long-term prediction challenges, a fundamental component of our research objectives.

In a similar vein [5] investigated the utility of Random Forest model structures for time series forecasting with a particular focus on their ability to capture intricate relationships, both linear and non-linear. Random forest models are valuable instruments for air quality data due to their interpretability and efficacy in modeling interactions among multiple features. This fits with the fact that many machine learning techniques are still being studied to see which ones can best predict PM2.5. It also shows how useful the Random Forest model can be when time series complexity and readability are crucial .

In addition to the technical aspects [1] emphasized the larger social effects of air pollution. There were large variations in environmental consciousness between the urban and suburban Beijing residents surveyed, but all groups reported a marked decrease in frequency and contentment with eating out as pollution levels increased. This emphasizes the need for reliable air quality predictions to guide public actions and improve health in general. Our study aims to produce models that are technically accurate and also give practical insights that could help with urban planning and policymaking. This coincides with our research objectives, as improving the precision and timeliness of PM2.5 forecasting is relevant because it helps us understand the larger socioeconomic consequences of air pollution .

According to [3], who conducted an extensive analysis of machine learning algorithms used for air quality forecasting, deep learning models are now the gold standard because they can accurately represent complicated, non-linear connections. Among the best models for predicting PM2.5, they found Convolutional Neural Networks (CNNs) and Long-Short-Term Memory (LSTM) to be the most effective. In line with the present research objective, which is to improve the accuracy of PM2.5 forecasts, this study investigates advanced machine learning methods, particularly deep learning. Handling non-linearity and achieving generalizability across multiple datasets are particular challenges in PM2.5 prediction. The survey also allows review of different architectures and approaches that most effectively address these challenges .

The Jing-Jin-Ji region was the focus of comparative research by [2] that looked at four machine learning models for PM2.5 prediction, including Linear SVR, K-Nearest Neighbor, Lasso Regression, and Gradient Boosting. Compared to other models, they found that Gradient Boosting performed better, particularly in winter when air pollution was worse. This is quite congruent with the aims of the present study, which is to determine the best method for PM2.5 forecasting in diverse environmental situations by comparing several machine learning architectures. The method for thoroughly testing models in various temporal settings is inspired by the work of [2], which focuses on comparing and analyzing the performance of models throughout the year. In this way, we can be sure that the model we choose will operate well in different seasons with varied amounts of pollution.

All of these studies together give a solid foundation for current research on PM2.5 prediction. They show the importance of developing air quality forecasting models by integrating many data sources, using advanced machine learning approaches, and understanding the broader social consequences of air pollution. Models like Long-Short-Term Memory (LSTM),Random Forest, and Informer architecture have great promise in improving prediction accuracy. Reliable PM2.5 predictions can help reduce negative health effects, and they also provide crucial information for urban planning and policy decisions, as highlighted in these studies. The present study aims to help public health and urban policy objectives by advancing technology and providing practical information based on rigorous methodology and real-world implications.

3 METHODOLOGY

3.1 Dataset Preparation

An air quality dataset of 18 columns was evaluated using exploratory data analysis (EDA), as detailed in this study. Several pollution and weather measurements collected from various monitoring sites are included in the dataset. Data integrity and the extraction of useful insights from missing values (both numerical and categorical) are the main concerns of the current study, which also includes feature engineering and exploratory data analysis (EDA).

PM2.5, PM10, SO2, NO2, CO, O3, temperature, and wind direction are included in the dataset. A monthly mean imputation strategy was implemented to address missing values in numerical features such as PM2.5, PM10, SO2, NO2, CO, O3, temperature, pressure, dew point, rainfall, and wind speed. In particular, the mean value of each numerical column was determined for each distinct month and utilized to complete all gaps throughout that month. In order to maintain the seasonal variability that is inherent in environmental data, this method was implemented. The general mean for the column was used as a backup in the event that a specific month lacked data to ensure that no missing values remained. also it is important to understand that an excessive reliance on the overall mean can reduce data variance.

An individual imputation approach for each station was used to fill in missing values in categorical characteristics, especially for the WindDir attribute, which represents the direction of the wind. We used the most often observed value (mode) for each monitoring station to fill in the missing values in the wind direction column after grouping the dataset by each station. This imputation approach helps make sure that the imputed values are indicative of each station's regular data, which is important because wind direction is often affected by local topography and individual station features. This method is superior to a global mode in that it reduces bias while still preserving the distinct features of each monitoring site.

In order to extract new details from the existing dataset, feature engineering was an essential step. The Air Quality Index (AQI), daily high and low temperatures, classification feature encoding, wind direction transformation, and seasonal labels for each data point were among the characteristics that were created.

To calculate AQI, the highest levels of PM2.5, PM10, SO_2, NO_2, CO, and O_3 were used. The (`calculate_aqi`) function divided air quality into several categories, such as *"Excellent," "Good,"* and *"Lightly Polluted"* based on the highest pollutant concentration measured. To allow for a more detailed seasonal examination of air quality changes, the recorded month was used to give seasonal labels such as Winter, Spring, Summer and Autumn. Air quality may change widely depending on seasons; therefore, identifying pollutants according to their season helps shed light on how pollution levels shift throughout the year. The granularity of the data set, measured in terms of daily maximum and lowest temperatures, was further improved by organizing the data by year, month and day. This is especially helpful in understanding the effects of temperature on pollution levels.

To allow for a more detailed seasonal examination of air quality changes, the recorded month was used to give seasonal labels such as Winter, Spring, Summer, and Autumn. Air quality may change widely depending on the seasons, therefore identifying pollutants according to their season helps shed light on how pollution levels shift throughout the year. Dataset granularity, as measured in terms of daily maximum and lowest temperatures, was further improved by organizing the data by year, month, and day. This is especially helpful for understanding the effects of temperature on pollution levels.

In order to better understand cycle relationships, additional feature engineering work involved converting wind direction data and encoding category characteristics. To prepare Season, AQI, and Station, which are categorical variables,

for use in machine learning models, one-hot encoding was employed. For environmental modeling applications, where wind direction is important for pollution dispersion, the wind direction was converted to degrees and then modified using circular functions while keeping its cyclical attributes.

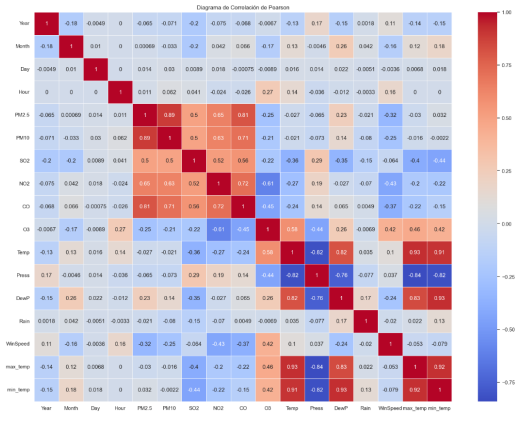

Fig. 1. Heatmap applying Spearman's rank correlation

The goal of the EDA phase was to evaluate data properties and discover correlations among attributes. The EDA relied heavily on the generation of a correlation heatmap applying Spearman's rank correlation to show feature to feature correlations. For the purpose of choosing good predictors and comprehending feature relationships, this heatmap **Fig 1** was useful in identifying strongly linked features. It is worth mentioning that there were major links found between certain pollutants, such PM2.5 and PM10, which often occur together. There were positive associations between CO and NO2, which might indicate that both gases are emitted by the same sources, such cars. Conversely, temperature exhibited negative correlations with CO and NO2, indicating that higher temperatures may facilitate the dispersion of these pollutants.

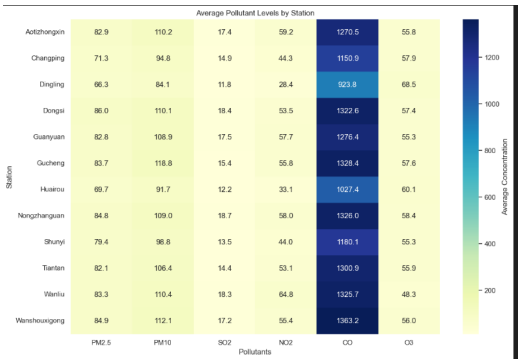

Fig. 2. Heatmap: Average pollutant concentrations across different monitoring stations

The average pollution concentrations across several monitoring sites were shown on a heatmap. Consequently, learned a lot about the distribution across regions of pollution levels. In particular, stations like Dongsi, Nongzhanguan, Tiantan, and Wanshouxigong showed unusually large amounts of CO, showing that pollution levels were higher there. Similarly, while some stations showed consistently high levels of PM10 and PM2.5, others, like Dingling, showed much better air quality with lower amounts of SO_2 and CO. Thanks to these findings, we can narrow down the most urgent issues and conduct extensive research into their causes.

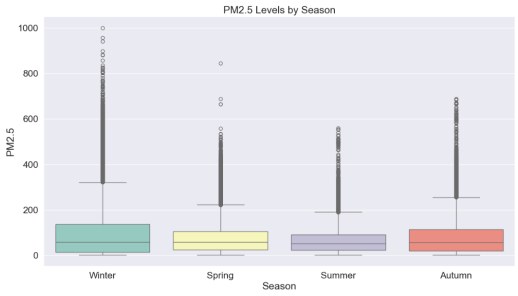

Fig. 3. Boxplot depicting PM2.5 levels across different season

A boxplot showing PM2.5 levels during the winter, spring, summer, and fall seasons was made in order to examine seasonal patterns. According to the data, PM2.5 concentrations were much greater in the winter and autumn than in the spring and summer. the Air quality was more changeable in the winter and decrease, which may explain why there were more outliers during those seasons. Air quality often increases in winter, possibly as a result of increased heating activities and decreased atmospheric dispersion, since median PM2.5 concentrations were noticeably greater during colder months for create effective methods to handle air quality, it is crucial to understand these seasonal changes.

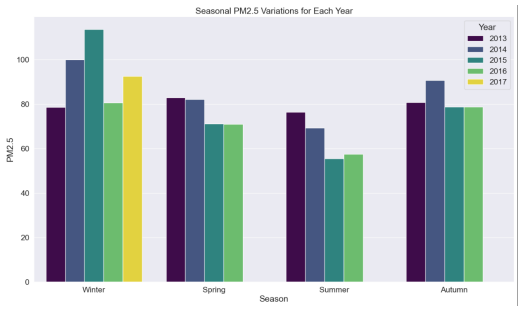

Fig. 4. bar plot that visualized PM2.5 levels for each year

Another way to look at seasonal patterns was using a bar chart that showed the PM2.5 values for every year from 2013 to 2017. There were clear seasonal trends, with PM2.5 levels being high in autumn and winter and lowest in the summer. The differences from one year to the next brought attention to the fact that air quality varies with the seasons and the years, which may be linked to regional or national policy shifts. When evaluating the long-term impact of air pollution management measures, a temporal perspective, as provided by such visualizations is important.

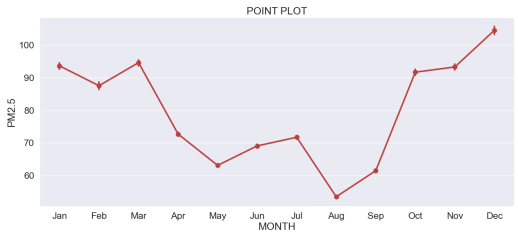

Fig. 5. the average monthly PM2.5 levels throughout the year

The average monthly PM2.5 values throughout the year were shown in a point plot so that insights could be gained at the monthly level. During the winter, when heating activities are at a peak and pollutant dispersion is at a minimum, the graph showed that PM2.5 levels were highest in the coldest months, especially January, November, and December. On the other hand, levels were lower in the summer in July and August , which may be because there are fewer emission sources and the air is better mixed.

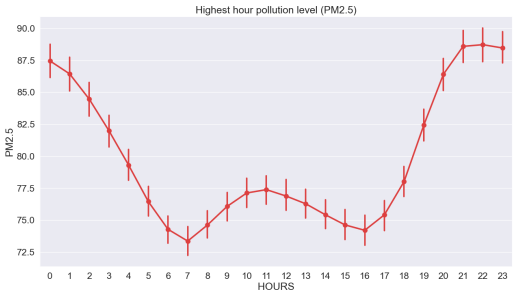

Fig. 6. Highest hour pollution level (PM2.5)

Intervals between hours of PM2.5 levels As a result of reduced dispersion and higher emissions from both vehicle and domestic heating activities, the investigation showed that PM2.5 concentrations peaked in the late evening (between 20:00 and 23:00). When temperatures were higher and air mixing was better in the afternoon (07:00 to 15:00), on the other hand, PM2.5 levels were lower.

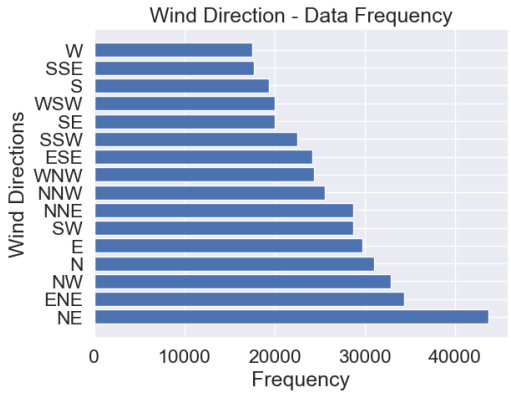

Fig. 7. frequency wind directions

It represented the occurrence of various wind directions. While directions pointing south-southeast (SSE) and west (W) were among the least common, the most common were NE and ENE. Important for understanding the mechanics of pollution spread, this variance sheds light on typical wind patterns in the areas under observation.

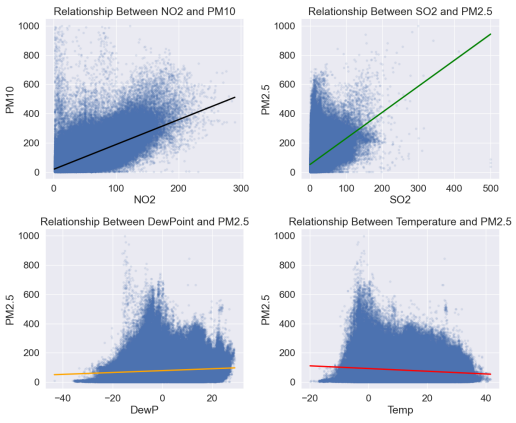

Fig. 8. PM2.5 and various meteorological

ln attempt to demonstrate the connections between PM2.5 and several other weather factors, including but not limited to temperature, dew point, SO2, and NO2. There was a negative correlation between temperature and (PM2.5) which means that warmer temperatures make particulate matter easier to disperse. It was also suggested that increasing moisture levels might somewhat reduce particle concentrations by a small negative association between dew point and PM2.5. As a result of their positive relationships, SO2 and PM2.5 and NO2 and PM10 likely originate from the same combustion processes and vehicle emissions.

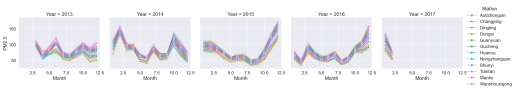

Fig. 9. Yearly trends in PM2.5 levels across different stations

Using a set of line plots., annual trends in PM2.5 concentrations at many monitoring locations from 2013 to 2017 were presented. This method underlined seasonal variations and station differences. Regular records of elevated PM2.5 concentrations for the most of the years and monitoring locations revealed localized emission sources and the impact of seasonal weather conditions. Wanliu and Nongzhanguan stations showed consistently high pollution levels, suggesting persistent local pollution problems.

In conclusion, the data preparation techniques utilized, including localized imputation using monthly means and mode imputation per station, preserved significant temporal and geographic variability essential for environmental modeling. The constructed attributes, including AQI, seasonal labels, and wind direction translation, substantially enhanced both analysis and model development. Exploratory data analysis provided important insights into the relationships among variables, underscored seasonal and regional pollution trends, and facilitated the finding of problem locations. Although technical factors were taken into account to scale the imputation methods to larger datasets, the entire strategy maintained data integrity and facilitated rigorous analysis. Future studies may concentrate on improving imputation and feature engineering methodologies to increase efficiency and forecast accuracy. The analysis identified critical factors, including meteorological conditions (temperature, wind direction) and concurrent pollutants (PM10, CO, NO2), that significantly influence pollution levels.

3.2 Modelling Process

Making predictions about PM2.5 concentrations required creating and testing several models. Random Forest, Long-Short-Term Memory (LSTM) and Multiple Regression models were among those developed.

(1) **Multiple Regression Models** Analyzing the prediction ability of different regression models using the mean absolute error (MAE) and the root mean square error (RMSE) showed significant disparities. Regressors such as linear regression, K-neighbors, decision tree, random forest, and gradient booster were among many that were integrated.

(2) **LSTM Model** Several preprocessing operations were performed on the data set to ensure data quality and model compliance before modeling could begin. Initially, MinMaxScaler was used to scale all the features to the range [0, 1] to address the scales of the different features. Models like LSTM and Random Forest, due to their sensitivity to large input values, relied heavily on this normalization to stabilize their training processes.
The development of the **LSTM** model to forecast PM2.5 levels required constructing a neural network capable of capturing temporal links within the data. The design consisted of two layers of Long Short-Term Memory units. The initial layer, which included 50 units, was set to return sequences to the subsequent **LSTM** layer, which also included 50 units but returned only the final output. Dropout layers with a rate of 20% were applied after each **LSTM** layer to reduce overfitting by randomly deactivating some neurons during training. The model concluded with two Dense layers: the first with 25 units and the last featuring a single output unit to forecast PM2.5 levels.
To minimize the error between predicted and actual PM2.5 levels, the model was constructed using the Adam optimizer with the loss function set to Mean Squared Error (MSE). To balance model complexity and the

risk of overfitting, key hyperparameters, including LSTM units, dropout rate, batch size, and epoch count, were carefully chosen. Specifically, 50 units per LSTM layer provided adequate learning capacity without overwhelming the model. A 20% dropout rate improved the robustness of the model, while a batch size of 64 allowed efficient use of resources during training.

To account for temporal data dependencies, the **LSTM** model used sequences. The model used the last 24 hours of data to predict the PM2.5 concentration for the next hour, with a 24-hour sequence length selected. Overlapping data segments were created with feature values from the previous 24 hours in each sequence; the target output was the PM2.5 value for the next hour. This setup allowed the **LSTM** model to identify and understand patterns throughout daily periods, potentially capturing periodic changes in air quality.

(3) **Random Forest Model** A random forest regression coefficient was developed to forecast the levels of PM2.5 pollution. To distinguish between low- and high-pollution scenarios, a threshold-based classification was implemented to balance computational efficiency with predictive accuracy. The model was trained with 100 decision trees (`n_estimators=100`). Given its strong capability to manage feature interactions and its robustness in capturing complex nonlinear correlations, the Random Forest model was chosen. As part of feature engineering, lag features were created for the **Random Forest** model to introduce a temporal component. Lagged versions of PM2.5 concentrations—`PM2.5_lag1`, `PM2.5_lag2`, and `PM2.5_lag3`—were developed to capture historical information and short-term dependencies. By considering past PM2.5 levels with these lag features, the Random Forest model was able to better represent temporal patterns.

Although all models considered air quality characteristics, meteorological variables, such as temperature, pressure, and wind speed, were also included. These factors were crucial for reliable forecasts due to their known impact on PM2.5 variability.

The data set was divided into training and testing subsets using an 80-20 ratio. The models were fitted on the training data and evaluated on unseen data using the test set, ensuring proper assessment of the models' generalizability.

After training, the was evaluated using regression metrics to determine its performance. Its predictions were also classified as "high" or "low" according to a PM2.5 threshold value of 75, according to environmental regulations. The capability of the **LSTM** model to capture sequential dependencies, along with the strong performance of the **Random Forest** model in regression tests, demonstrates the feasibility of predicting future pollution levels based on past data.

4 EXPERIMENTATION AND RESULTS

4.1 Experimental setup

Training and evaluation of all the models developed using relevant metrics and methodologies provided the experimental setup.

Multiple Regression Models

Multiple regression models have been evaluated using the root mean square error (RMSE) and the mean absolute error (MAE). These measures provide insight into the average amount of error and the importance of larger errors, respectively. Of all the regression models tested, the Random Forest Regressor was found to be the most accurate.

LSTM Model

The data was divided into two sets, one for training and one for validation, in an 80/20 ratio. The Adam optimizer and MSE loss function were used to train the model for 10 epochs with a batch size of 64. Because training LSTM networks is computationally intensive, the training method used GPU-based computing resources. After categorizing PM2.5 levels into "Low" and "High" using a threshold of 75 units, the model performance was evaluated using regression accuracy measures such as MAE and RMSE, as well as classification metrics including precision, recall and F1 score.

Random Forest Model

The dataset was also divided 80/20 for training and testing purposes, and the model was developed in Python with the help of the scikit-learn package. Precision, recall, F1-score, and accuracy were used to evaluate classification performance, whilst MAE and RMSE were used to evaluate regression performance. Using a threshold of 75 units, the actual and forecasted PM2.5 levels were transformed into binary classifications for the classification process.

4.2 Discussion of Results

(1) **Multiple Regression Models**

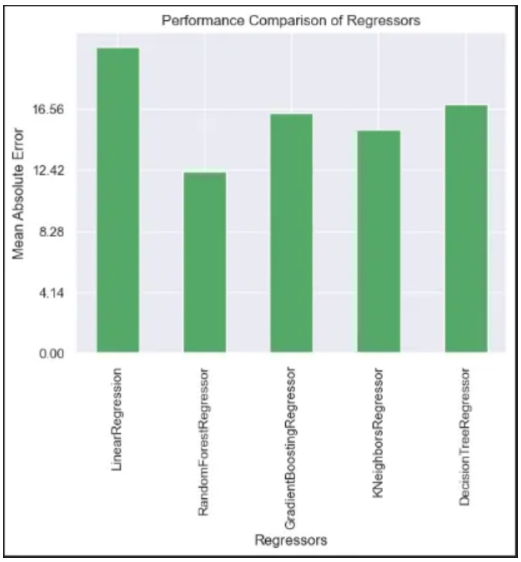

Fig. 10. Performance Comparison of Regressorsl

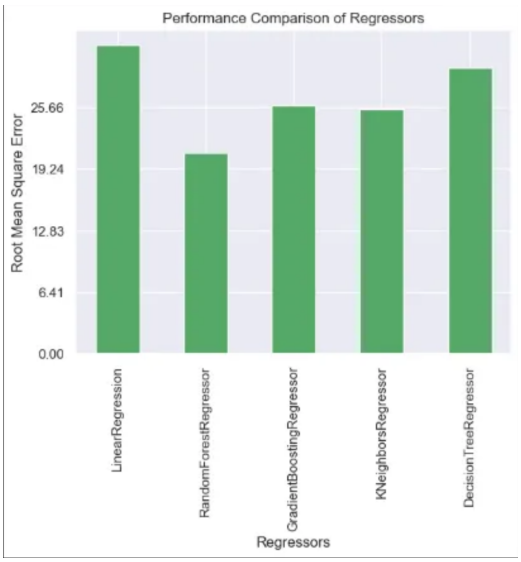

Fig. 11. Performance Comparison of Regressors

The comparative analysis of the regression models, as shown in Figure 10 and Figure 11indicates that Random Forest consistently yielded the lowest MAE and RMSE, demonstrating its capability in handling complex patterns. Conversely, Linear Regression exhibited the highest errors, revealing its limitations in capturing intricate relationships. Both Gradient Boosting Regressor and KNeighbors Regressor performed reasonably well but lagged behind Random Forest, indicating a need for further hyperparameter tuning. The Decision Tree Regressor showed potential but still underperformed compared to Random Forest.

(2) **Random Forest Model**

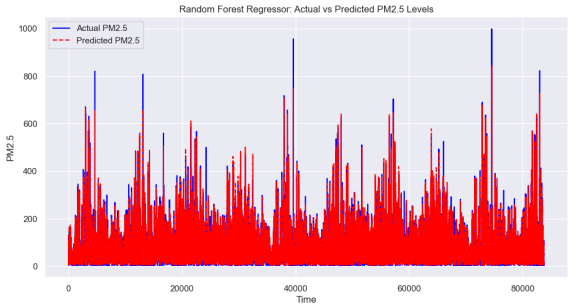

Fig. 12. Actual vs Predicted PM2.5 levels for RF

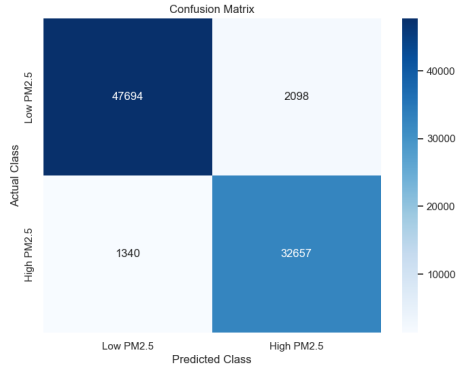

Fig. 13. Confusion Matrix

Table 1. Classification Report for PM2.5 Prediction using Random Forest Model

Class	**Precision**	**Recall**	**F1-score**	**Support**
Low PM2.5	0.97	0.96	0.97	49,792
High PM2.5	0.94	0.96	0.95	33,997
Accuracy	**0.96** (83,789 samples)			
Macro avg	0.96	0.96	0.96	83,789
Weighted avg	0.96	0.96	0.96	83,789

The Random Forest model showed good general performance in tracking PM2.5 levels, achieving an MAE of 7.69 and an RMSE of 14.69. Figure 12 provides visual confirmation of the model's ability to follow trends over time. However, the model struggled to accurately predict sudden shifts or severe peaks, possibly due to the unpredictable nature of pollution incidents and their sensitivity to sudden weather changes. In the binary classification, the model effectively distinguished PM2.5 levels as either "High" or "Low" using a threshold of 75, with the classification report shown in 1 indicating strong recall, accuracy, and F1 scores for both categories. Figure 13 presents the confusion matrix for the Random Forest model, highlighting mostly accurate predictions with minimal outliers. The overall classification accuracy of 95.08% indicates that the model reliably differentiates between high and low pollution levels. However, the model's limitations became evident during times of abrupt change, likely due to inadequate representation of these events in the training data or the inability of lag features to capture complex temporal relationships. Future improvements could involve incorporating advanced temporal features or employing deep learning models like LSTMs to enhance temporal pattern detection. But some restrictions showed up it is worth mentioning that the model struggled during times of abrupt change, which might be because these events weren't correctly captured in the training data or because lag features can't properly capture complicated temporal connections. Furthermore, the lagged characteristics tracked trends over the short term, but they failed to take into consideration the larger seasonal changes that might be a major factor in the variability of PM2.5. To improve future versions temporal pattern identification can

look at using deep learning models like Long Short-Term Memory (LSTM) networks or other more advanced temporal features.

(3) **LSTM model**

Fig. 14. Training and Validation Loss of LSTM Model

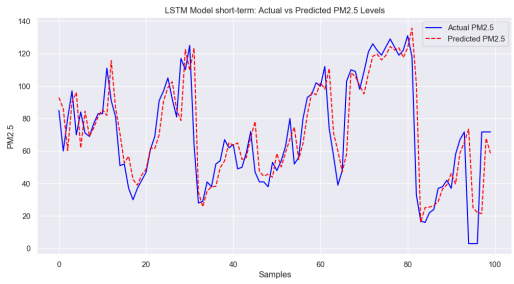

Fig. 15. Actual vs. Predicted PM2.5 Levels

The LSTM model experimental results indicated that it effectively used the temporal patterns present in the air quality data, resulting in reasonably accurate predictions of PM2.5 levels. Figure 14 shows the training and validation loss curves and showed a substantial increase in loss as the epochs progressed. The training deficit experienced a notable reduction at the beginning, suggesting that the process of learning was accelerated. Even so, the validation loss began to slightly increase around epoch 6 indicating the potential for overfitting with this behavior, suggesting the model was beginning to get patterns that are specific to the training set rather than generalizing well to unseen data. As shown in figure 15, the comparison from actual vs predicted PM2.5 levels for the first 100 samples in the test set showed that the model was able to follow the trend of the actual data rather reasonably well, catching peaks and troughs with good accuracy. Some differences were noted especially where PM2.5 levels suddenly spiked. These differences might be explained by the inherent difficulties in forecasting unexpected changes which usually arise from factors like human activity or sudden capture by the input features from the weather.

Table 2. Classification Metrics Summary for PM2.5 Prediction

Metric	Class	Precision	Recall	F1-Score	Support
Low PM2.5	Low PM2.5	0.95	0.96	0.95	49,970
High PM2.5	High PM2.5	0.94	0.92	0.93	34,160
Accuracy	-	-	-	0.94	84,130
Macro Avg	-	0.94	0.94	0.94	84,130
Weighted Avg	-	0.94	0.94	0.94	84,130

The classification based evaluation provided further insights into the ability of the model to forecast PM2.5 levels beyond regression analysis. The model classification accuracy, precision, recall, and (F1-score) were assessed by classifying PM2.5 concentrations as (Low or High) using a threshold of 75 units. The findings, as

shown in Table 2 indicated that the model effectively classified between low and high pollution levels and providing high precision and recall values for both classes.

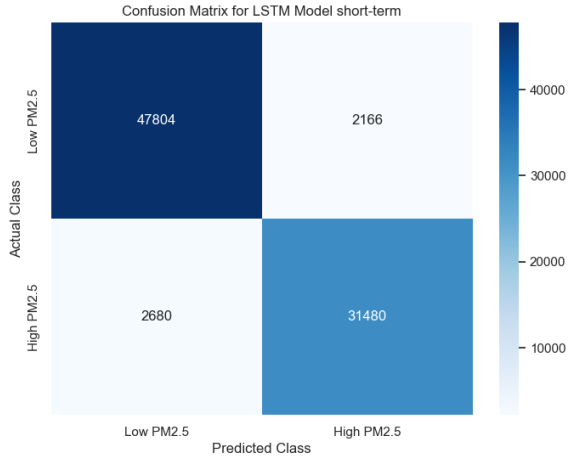

Fig. 16. Confusion Matrix for LSTM Model

The confusion matrix shown in Figure 16 reveals that the majority of predictions were accurate, with a small number of false positives and false negatives. The model's performance indicates its potential utility in early warning systems, where timely categorization of pollution levels is crucial. The MAE and RMSE values suggest that the model performs well under normal conditions, but struggles with extreme or rare events, which may be due to outliers in the dataset.

These results were positive but there were still potential areas to inaccuracy that would have impacted the model's predicted performance. The dataset's completeness and quality were a key cause of error and for the model accuracy predictions is highly dependent on the quality of the input data. Furthermore, not all situations require a 75 unit categorization threshold air quality requirements differ by area thus a more complex and area specific strategy might provide better results. In addition, the model failed to consider several external variables that may cause substantial changes in (PM2.5) levels such as unexpected emissions from industries or nearby fires.

5 CONCLUSION

In conclusion, several ML methods for the prediction of the PM2.5 level were studied, with a particular focus on the Random Forest and LSTM networks. Lower (MAE and RMSE) scores for the Random Forest model compared to other regression models showed that it performed better when dealing with complicated patterns and non-linear connections. At the same time, the (LSTM) model demonstrated encouraging performance in capturing temporal dependencies, which is highly beneficial when trying to predict future trends in air pollution.

research shows that public health initiatives and early warning systems can greatly benefit from more accurate PM2.5 forecasts made using a combination of weather and air quality data. One reason why the LSTM model might be useful for air quality monitoring systems in real time is its capacity to use past data. Overfitting in deep learning

models and unexpected spikes in pollution are two examples of the problems that call for additional hyperparameters tuning and data augmentation to help with generalization.

To improve the ability to detect changes in the seasons and unexpected increases in pollution levels, and future research may investigate ensemble learning methods that combine many models or use advanced deep learning structures, and the models may be stronger if the dataset includes a wider environmental variables and time periods. This work helps us better understand pollution dynamics, which in turn helps us build better health and urban planning strategies by making PM2.5 projections that are precise and easy to understand.

REFERENCES

[1] Rong Gao, Hua Ma, Hongmei Ma, and Jiahui Li. 2020. Impacts of Different Air Pollutants on Dining-Out Activities and Satisfaction of Urban and Suburban Residents. *Sustainability* 12 (03 2020), 2746. https://doi.org/10.3390/su12072746

[2] Xin Ma, Tengfei Chen, Rubing Ge, Caocao Cui, Fan Xu, and Qi Lv. 2022. Time series-based PM2.5 concentration prediction in Jing-Jin-Ji area using machine learning algorithm models. *Heliyon* 8, 9 (2022), e10691. https://doi.org/10.1016/j.heliyon.2022.e10691

[3] Manuel Méndez, Mercedes G. Merayo, and Manuel Núñez. 2023. Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review* (2023), 1 – 36. https://api.semanticscholar.org/CorpusID:256939252

[4] Meng Niu, Yuqing Zhang, and Zihe Ren. 2023. Deep Learning-Based PM2.5 Long Time-Series Prediction by Fusing Multisource Data—A Case Study of Beijing. *Atmosphere* 14, 2 (2023). https://doi.org/10.3390/atmos14020340

[5] Evangelos Spiliotis. 2022. Decision Trees for Time-Series Forecasting. (01 2022), 30–44.