Machine Learning Approach to Forecast PM 2.5 Levels in Gurugram City

Sanjeev Kumar[1], Yogesh Kumar[2], Dhiraj Khurana[3], Ameet[4], Neeraj Kumar[5]

Department of Computer Science and Engineering, UIET, MD University Rohtak Haryana[1 2 3 4 5]
Sanjeevbumbra07@gmail.com[1], dryogeshkumar.uiet@mdurohtak.ac.in[2], dhirajkhurana@mdurohtak.ac.in[3], ameety429@gmail.com[4] , Neerajchawaria@gmail.com[5]

KEYWORDS
PM2.5, Time series analysis, regression, RMSE, LSTM, prediction

ABSTRACT
Over the years, the air quality in the capital city of Delhi and its adjoining states has reached an alarming state, primarily due to the increasing concentration of air pollutants such as PM2.5. This heightened concentration has been linked to severe respiratory problems. In this study, we assess the forecasting of PM2.5 concentration specifically in the Gurugram region of Haryana. We employ time series analysis and various regression models on weather data obtained from the Indian Meteorological Department (IMD) to construct a prediction model on an hourly basis. The study involves a comprehensive comparison of different models, encompassing data processing, exploratory data analysis (EDA), model development, and the application of performance metrics such as Root Mean Squared Error (RMSE) and R2 Score. The objective is to evaluate the error, which represents the difference between actual and predicted values. Upon comparing various time series and regression models, it is observed that the Long Short-Term Memory (LSTM) based analysis outperformed Adaboost and other regression models, exhibiting lower performance metrics. This suggests that the LSTM model provides a more accurate prediction of PM2.5 concentration in the specified region and timeframe.
..

1. Introduction

The air around us is like a special mix of gases that keeps us alive. Most of it is nitrogen and oxygen, with a bit of other stuff like carbon dioxide and water vapor(Chawala & Sandhu, 2020). These gases work together in a delicate balance to create the air we breathe. It's kind of like a magical dance of molecules that happens all around us. This mix of gases is super important for life on Earth, making it possible for us and other living things to survive. So, every time we take a breath, we're inhaling this amazing blend that keeps our planet humming with life.

The contemporary challenges posed by air composition are notably accentuated by the presence of a new form of pollution: PM2.5 particles. These fine particulate matter, measuring 2.5 micrometers or smaller, originate from various sources such as vehicle emissions, industrial processes, and combustion activities(Davidson et al., 2005). Unlike larger particles, PM2.5 can penetrate deep into the respiratory system, posing significant risks to human health. As these minute pollutants become increasingly pervasive, the dynamics of air composition undergo distinct shifts, necessitating a heightened understanding of their sources, dispersion patterns,

and potential long-term effects. Addressing this evolving facet of air quality requires a multifaceted approach, encompassing stringent regulations, innovative technologies, and public awareness initiatives to mitigate the impact of PM2.5 particles and preserve the integrity of our atmospheric composition.

The danger lies in their small size, allowing them to bypass the body's natural defense mechanisms in the upper respiratory tract and penetrate into the lungs. Once inhaled, PM2.5 particles can cause or exacerbate respiratory and cardiovascular issues such as asthma, bronchitis, and cardiovascular diseases(Yadav & Rana, 2022). Additionally, long-term exposure to elevated levels of PM2.5 has been associated with an increased risk of chronic health conditions, including lung cancer and premature death. Vulnerable populations such as children, the elderly, and individuals with pre-existing health conditions are particularly at risk, emphasizing the importance of addressing and mitigating the impact of PM2.5 pollution on air quality to safeguard public health.

In this paper, the work has been done on the air pollution data of Gurugram district of Haryana. The air pollution data is collected from Central Pollution Control board. The objective of the proposed research is to enhance the efficiency in predicting PM2.5 pollutant particle concentrations in the air of gurugram. The primary goal is to develop a robust predictive model that can forecast PM2.5 levels well in advance. This proactive approach aims to provide sufficient time for implementing preventative measures, thereby safeguarding human lives from the hazardous impacts of elevated PM2.5 concentrations. By advancing our ability to anticipate and monitor air quality changes, the research seeks to contribute to timely interventions and public health strategies, ultimately mitigating the adverse effects associated with high levels of PM2.5 pollution in the atmosphere of gurugram.

2. Related Works

This area has been inadequately addressed by a limited number of models in terms of quantifying PM 2.5 pollutant particle concentrations(Janarthanan et al., 2021). Forecasting the Air Quality Index (AQI) for Chennai City has been conducted using regression and LSTM techniques. A gradient boosting regression and LSTM model were employed to predict PM 2.5 concentrations, with LSTM outperforming the regression model, achieving an R2 score of 0.9821(Jamei et al., 2022).

Research conducted by the Indian Meteorological Department (IMD) in Delhi involved the measurement of air quality using regression and machine learning models. An ET+Adaboost model was proposed and demonstrated superior performance compared to other models(Kumar et al., 2020). In the prediction of PM 2.5 concentration changes, ARIMA, Facebook Prophet, and LSTM techniques were utilized. The LSTM model yielded an RMSE value of 0.1002(Gladkova & Saychenko, 2022). The AQI prediction employing deep learning models, including GRU, LSTM, KNN, and SVM, showed that LSTM-GRU outperformed other models, with R2, MAE, and RMSE values of 0.84, 36.11, and 0.52, respectively(Sarkar et al., 2022).

Logistic regression and Random Forest models were employed for forecasting the Air Quality Index in India, achieving an accuracy of 93%(Mangayarkarasi et al., 2021). Additionally, the prediction of PM 2.5 concentrations using Random Forest, SVM, and LSTM revealed that

LSTM outperformed RF and SVM, demonstrating a correlation coefficient of 0.972(Ghose et al., 2022).

3. Data Collection and Study Area

An hourly basis Air Quality data has been collected from Central Pollution Control board. The acquired dataset was in .csv (comma-separated values) format, encompassing various features. After extracting relevant information, irrelevant features were eliminated. The dataset is comprised of hourly data collected from November 25, 2020, to January 24, 2023 for gurugram region of Haryana. This work is implemented on Google colaboratory environment. To study the regression and time series models, a special library called Keras which is open source library is used for the purpose.

A correlation heat map using different air pollutants has been generated (Fig.1). Missing values and outliers are taken care using python libraries. Impute missing values using mean imputation for numerical features. Outliers are removed by observing Boxplot method. Feature scaling has been executed on the raw data utilizing the Min-max Scaler from the scikit-learn library in Python. The hourly data is split into training and testing data in ratio of 80:20. A comparison between PM2.5 and other pollutants are depicted using scatterplot in fig.2.

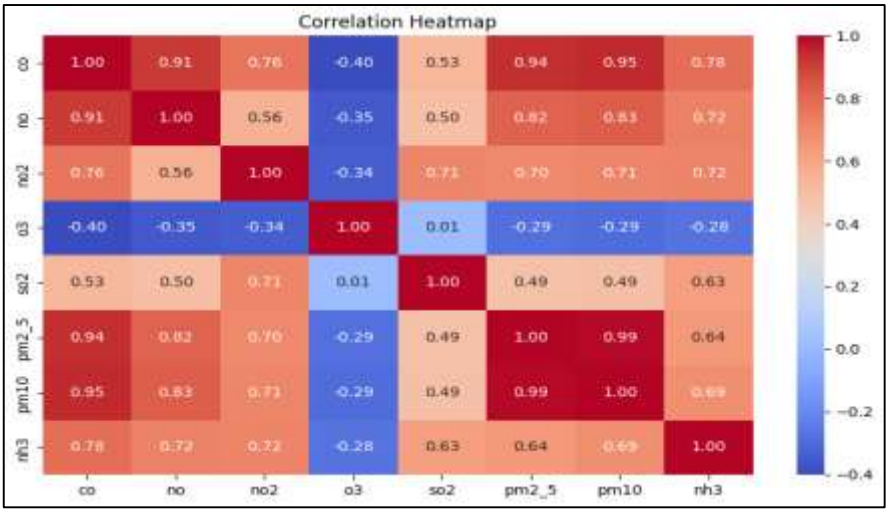

Fig.1. Correlation Heat map of pollutants

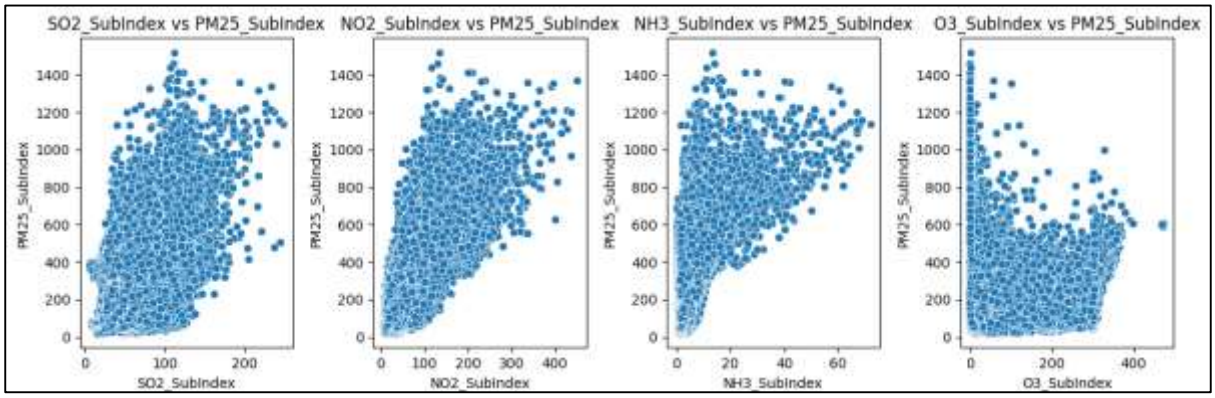

Fig.2. Scatterplot representation of PM2.5 Vs other pollutants

4. Methodology

In this paper, a comparative study is performed using regression and LSTM model, the description of these models are as follows:

1. Linear Regression: Linear regression stands out for its simplicity, making it easily understandable. The interpretability of coefficients allows for a clear grasp of relationships between variables. It's computationally efficient, handling large datasets with ease. As a baseline model, linear regression serves as a benchmark for more complex models, aiding in effective model comparison.

2. Decision Tree Regression: Decision tree regression models have ability to capture complex, non-linear relationships, handle both numeric and categorical data, and they are robust to outliers and ease in handling missing data.

3. Random Forest Regression: Random Forests are robust to outliers and demonstrate strong performance in capturing complex relationships, as well as handling missing data without significant impact on results. The ease of implementation, scalability, and high predictive accuracy, even for large datasets.

4. Gradient Boosting Regression: The Gradient Boosting regression operates by sequentially building a series of weak learners, typically decision trees, each correcting the errors of its predecessor. This iterative learning process allows Gradient Boosting to capture intricate relationships in the data and provide highly accurate predictions. The model is resilient to overfitting due to its sequential nature and the use of regularization techniques.

5. AdaBoost Regression: This regression take decisions using adaptive boosting.

6. LSTM: Long Short-Term Memory (LSTM) models is a type of recurrent neural network (RNN), effective in handling sequential data and addressing the vanishing gradient problem. LSTMs use a memory cell with gating mechanisms, allowing them to selectively retain or discard information at different time steps.

A comparative study is analysed using various performance metrics RMSE(Root Mean Squared Error), R-Squared.

1. RMSE(Root Mean Squared Error): It measures the average magnitude of the residuals, or the differences between predicted and actual values, providing a comprehensive assessment of the model's precision. A lower RMSE indicates better model performance.

2. R Squared: The coefficient of determination, commonly known as R-squared (R^2), is a performance metric used in regression analysis to assess the proportion of variance in the dependent variable explained by the independent variables. R-squared values range from 0 to 1, where a higher R^2 indicates a better fit of the model to the data.

5. Results and Discussions

An hourly basis Air Quality data is used for predicting PM 2.5 particulate matter of Gurugram district of Haryana over the period from November 25, 2020 to January 24, 2023. For the training data set, 80% of 18762 entries are reserved and rest 20% for testing data. In this work, a comparative study is performed using regression and LSTM model. For LSTM model, we use three hidden layers and dropout function uses 0.2 as percentage of deactivated neurons on training. The lose rate is below 0.0453.

The performance of the models used in the study is analysed through bar plots (Fig.3 , Fig.4). Root Mean Squared Error(RMSE) value of LSTM which is lowest among all other regression model. It is observed from RMSE bar plot(Fig.3), RMSE of LSTM(0.0916) is closed to Adaboost regression(0.124). and other models have much higher RMSE value.

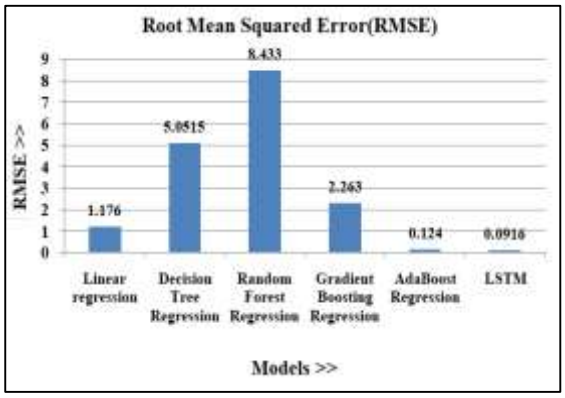

Fig.3. RMSE comparison

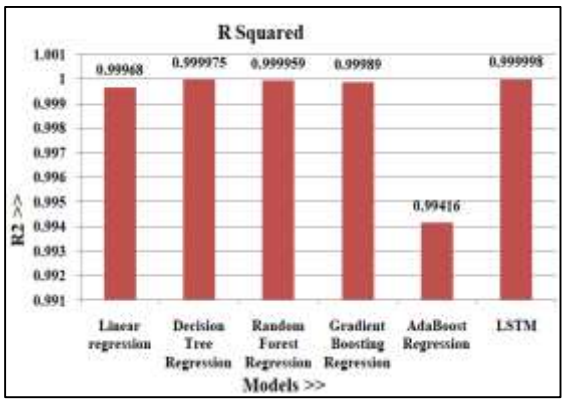

Fig.4. R2 comparison

The Comparative analysis is observed using R Squared performance metrics through bar plot(fig.). R2 value of LSTM is highest(0.999998) among the other models used. Therefore the comparative study of performance metrics (RMSE, R2), it is observed that a deep learning model LSTM is performing significantly better than other regression models with lower RMSE and higher R2 values. In summary, the LSTM model stands out as the most suitable choice for predicting PM2.5 levels.

6. Conclusion

A deep learning and machine learning methods are used to predict the PM2.5 concentrations in the Gurugram city. Time series analysis is performed using LSTM and regression model i.e. Linear regression, Decision Tree regression, Random Forest regression, Gradient Boosting regression, and Adaboost regressor. A deep learning model LSTM is performed significantly better than other regression methods by performing better in evaluated metrics. The memory cell enhances the learning rate. Thus these models can be applied to other geographical area and can consider other air composition to enhance the effectiveness of the prediction. We can apply to AQI prediction over different areas and then compare them on the basis of these composition to contribute to timely interventions and public health strategies, ultimately mitigating the adverse effects associated with high levels of pollution in the atmosphere.

7. References

Chawala, P., & Sandhu, H. A. S. (2020). Stubble burn area estimation and its impact on ambient air quality of Patiala & Ludhiana district, Punjab, India. *Heliyon*, *6*(1), e03095. https://doi.org/10.1016/j.heliyon.2019.e03095

Davidson, C. I., Phalen, R. F., & Solomon, P. A. (2005). Airborne particulate matter and human health: A review. *Aerosol Science and Technology*, *39*(8), 737–749. https://doi.org/10.1080/02786820500191348

Ghose, B., Rehena, Z., & Anthopoulos, L. (2022). A Deep Learning based Air Quality Prediction Technique Using Influencing Pollutants of Neighboring Locations in Smart City. *Journal of Universal Computer Science*, *28*(8), 799–826. https://doi.org/10.3897/jucs.78884

Gladkova, E., & Saychenko, L. (2022). Applying machine learning techniques in air quality prediction. *Transportation Research Procedia*, *63*, 1999–2006. https://doi.org/10.1016/j.trpro.2022.06.222

Jamei, M., Ali, M., Malik, A., Karbasi, M., Sharma, E., & Yaseen, Z. M. (2022). Air quality monitoring based on chemical and meteorological drivers: Application of a novel data filtering-based hybridized deep learning model. *Journal of Cleaner Production*, *374*. https://doi.org/10.1016/j.jclepro.2022.134011

Janarthanan, R., Partheeban, P., Somasundaram, K., & Navin Elamparithi, P. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. *Sustainable Cities and Society*, *67*. https://doi.org/10.1016/j.scs.2021.102720

Kumar, S., Mishra, S., & Singh, S. K. (2020). A machine learning-based model to estimate PM2.5 concentration levels in Delhi's atmosphere. *Heliyon*, *6*(11). https://doi.org/10.1016/j.heliyon.2020.e05618

Mangayarkarasi, R., Vanmathi, C., Khan, M. Z., Noorwali, A., Jain, R., & Agarwal, P. (2021). COVID19: Forecasting Air Quality Index and Particulate Matter (PM2.5). *Computers, Materials and Continua*, *67*(3), 3363–3380. https://doi.org/10.32604/cmc.2021.014991

Sarkar, N., Gupta, R., Keserwani, P. K., & Govil, M. C. (2022). Air Quality Index prediction using an effective hybrid deep learning model. *Environmental Pollution*, *315*. https://doi.org/10.1016/j.envpol.2022.120404

Yadav, A., & Rana, C. (2022). *A sytematic study of covid-19 prediction models of India*.