

Application of deep learning techniques for analysis and prediction of particulate matter at Kota city, India

Lovish Sharma, Hajari Singh*, Mahendra Pratap Choudhary

Department of Civil Engineering, Rajasthan Technical University, Kota (Raj.), India

*Corresponding author Email: hajari.phd21@rtu.ac.in

Article info

Received 12/11/2024; received in revised form 9/12/2024; accepted 16/12/2024

DOI: [10.6092/issn.2281-4485/20687](https://doi.org/10.6092/issn.2281-4485/20687)

© 2025 The Authors.

Abstract

Air pollution significantly threatens human health and the environment, making accurate prediction of pollutant concentrations crucial for effective mitigation. This study leverages deep learning models, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, to predict concentrations of PM_{10} and $PM_{2.5}$. The analysis utilizes hourly air quality data from July 1, 2017, to December 30, 2022, collected from the portals of the Central Pollution Control Board (CPCB) and Rajasthan State Pollution Control Board (RSPCB) for Kota city Rajasthan. Data preprocessing involves cleaning, normalization using a min-max scaler, and handling missing values with Multiple Imputation in XLSTAT. The methodology encompasses dataset loading, preprocessing, and data splitting, followed by model training and evaluation. Python libraries such as Pandas, Numpy, TensorFlow, and Matplotlib are employed for data analysis and visualization. Performance metrics, including Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 score, are calculated to assess the models' predictive accuracy. The results demonstrate that GRU model effectively capture temporal dependencies in air quality data, offering reliable predictions for PM_{10} and $PM_{2.5}$ concentrations with 41.85 and 17.73 RMSE values for PM_{10} and $PM_{2.5}$. These findings underscore the potential of deep learning models in air pollution forecasting, providing valuable insights for policymakers to implement timely interventions.

Keywords: *Air Pollution, Machine learning, PM_{10} , $PM_{2.5}$, LSTM, GRU*

Introduction

Air pollution, a significant environmental issue, has been linked to various health problems, environmental degradation, and climate change. Accurate prediction of air pollution levels is crucial for mitigating its impacts on public health and the environment. In recent years, machine learning techniques have emerged as powerful tools for predicting air quality, enabling the development of models that can accurately forecast the concentrations of pollutants such as Particulate Matter (PM_{10} and $PM_{2.5}$).

Numerous studies have explored the application of various machine learning models for air pollution prediction. One study analyzed $PM_{2.5}$ pollutants in polluted cities using a range of machine learning models, including linear regression, random forest, KNN, ridge and lasso regression, XGBoost, and AdaBoost. The results indicated that XGBoost, AdaBoost, random forest, and KNN models provided the most reliable predictions, with low error metrics across several performance indicators (Kothandaraman et al. 2022). Another investigation focused on the hourly concentration of $PM_{2.5}$ over a

station in Canada using the Group Method of Data Handling Neural Network (GMDHNN), Extreme Learning Machine (ELM), and Gradient Boosting Regression (GBR) tree models. This research highlighted the impact of data splitting on model performance, with the ELM model demonstrating superior accuracy in predicting $PM_{2.5}$ concentrations (Alomar et al. 2022). Further studies have delved into the use of deep learning approaches. Time series experiments on $PM_{2.5}$ concentrations were conducted using a combination of 1D convolutional neural networks and bidirectional gated recurrent units (CBGRU), with results showing that the CBGRU model outperformed traditional machine learning and conventional deep learning models in terms of prediction accuracy (Tao et al. 2019). Additional research explored machine learning models for air pollution prediction in various regions, such as Poland and Taiwan, demonstrating the effectiveness of models like the e-APFM (Enhanced Air Pollution Forecasting Model) and Gradient Boosting Regression in predicting pollutant concentrations with lower deviations between measured and predicted values (Domańska et al. 2014; Doreswamy et al. 2020). In another study, Long Short-Term Memory (LSTM) models were applied, combined with statistical methods, to forecast $PM_{2.5}$ concentrations in Taichung City, Taiwan. This research found that training models with different feature sets yielded varying results, with one model achieving the lowest Root Mean Square Error (RMSE) and the best accuracy (Kristiani et al. 2021). Another approach involved the development of an interpolated convolutional neural network (ICNN) model for predicting PM_{10} and $PM_{2.5}$ concentrations, which, by incorporating spatio-temporal data, achieved high prediction performance (Chae et al. 2021). Additionally, an artificial neural network (ANN) model was constructed to forecast air quality in Manila, Philippines, based on the Air Quality Index (AQI). The model demonstrated high accuracy, with a correlation coefficient (R^2) of 0.999, indicating strong applicability for air pollutant concentration forecasting (Viñas and Gerardo, 2022). Despite the progress made in air pollution prediction using machine learning, several research gaps remain. Many

studies have focused on PM_{10} and $PM_{2.5}$ prediction in urban settings, with limited research on the applicability of deep learning models like LSTM and Gated Recurrent Unit (GRU) in semi-urban or less polluted areas. Based on the existing research gap, the present study applies deep learning models, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), for predicting PM_{10} and $PM_{2.5}$ concentrations in Kota City, Rajasthan, India. Kota represents a semi-urban area with unique air quality challenges, making it an ideal case study for assessing the effectiveness of these models in a different geographical and pollution context. The aim of the study is to develop and evaluate deep learning models for the accurate prediction of air pollution levels in Kota City. Specifically, the study focuses on applying Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models to predict the concentrations of PM_{10} and $PM_{2.5}$. The performance of these models will be assessed using key metrics, including Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 score, to determine their effectiveness in forecasting air pollution levels in the city. The novelty of this paper lies in its focused approach to capturing localized trends and the impact of air pollution specifically in Kota city through deep learning. Utilizing Long Short-Term Memory (LSTM) networks, a widely recognized and reliable machine learning model, this study aims to accurately capture and predict air quality variations, providing a tailored analysis for the region. This research fills a significant gap, as no prior studies have employed deep learning models to analyze and forecast air pollution specifically for Kota, making it a pioneering effort in understanding and managing air quality in this area.

Materials and Methods

Study area

Kota city, an industrial and educational center in Rajasthan, faces significant challenges in managing air quality, particularly with particulate matter pollution. Recent studies have shed light on pollution patterns in the region. Kuldeep et al. (2022b) specifically analyze PM_{10} and $PM_{2.5}$ levels in Kota, noting an increase in the dust ratio ($PM_{10}/PM_{2.5}$) from 0.36 to

0.51 over recent years. This trend indicates a shift toward finer particles, with $PM_{2.5}$ now making up over 45% of total particulate matter, posing greater health risks due to the particles' ability to penetrate the respiratory system (Kuldeep et al., 2022). Several studies attribute temporary reductions in air pollution to the 2020 COVID-19 lockdown, which led to significant drops in PM_{10} and $PM_{2.5}$ levels across Rajasthan (Sharma et al., 2020; Yadav et al., 2022; Kuldeep et al., 2022). For instance, Yadav et al. (2022) observed reductions of 19–30% in Kota and nearby cities like Jaipur and Jodhpur. In addition, Sharma et al. found a 30% decrease in PM_{10} and $PM_{2.5}$ in Kota during the lockdown, along with a 45% rise in ozone levels, suggesting complex pollution dynamics influenced by restricted human activity. Singh et al. (2022) similarly emphasize the temporary effects of human activity restrictions, such as the ban on Diwali fireworks, which notably reduced $PM_{2.5}$, PM_{10} , CO , and SO_2 concentrations and improved air quality (Singh et al. 2022). However, Kuldeep et al. (2022b) stress that while lockdowns and activity bans can re-

duce pollution in the short term, they are unsustainable for long-term air quality management. Instead, Kota's growing industrialization and increasing fine particulate concentrations highlight the need for targeted pollution control policies tailored to the city's specific urban-industrial environment (Kuldeep et al., 2022). These studies collectively underscore the importance of developing effective, long-term strategies to address air quality in Kota, where PM_{10} remains a dominant contributor to the AQI year-round (Kamboj et al. 2022).

Research methodology

In the study, deep learning models LSTM & GRU are used to predict concentrations of particulate matter in Kota city of Rajasthan. The data has been obtained from the website of Central Pollution Control Board (CPCB) and Rajasthan State Pollution Control Board (RSPCB). The Figure 1 illustrates the flow chart of methodology wherein the data is preprocessed and the missing values are removed using XLSTAT software. Data has been splitted in the ratio of 80:20 (Training: Testing).

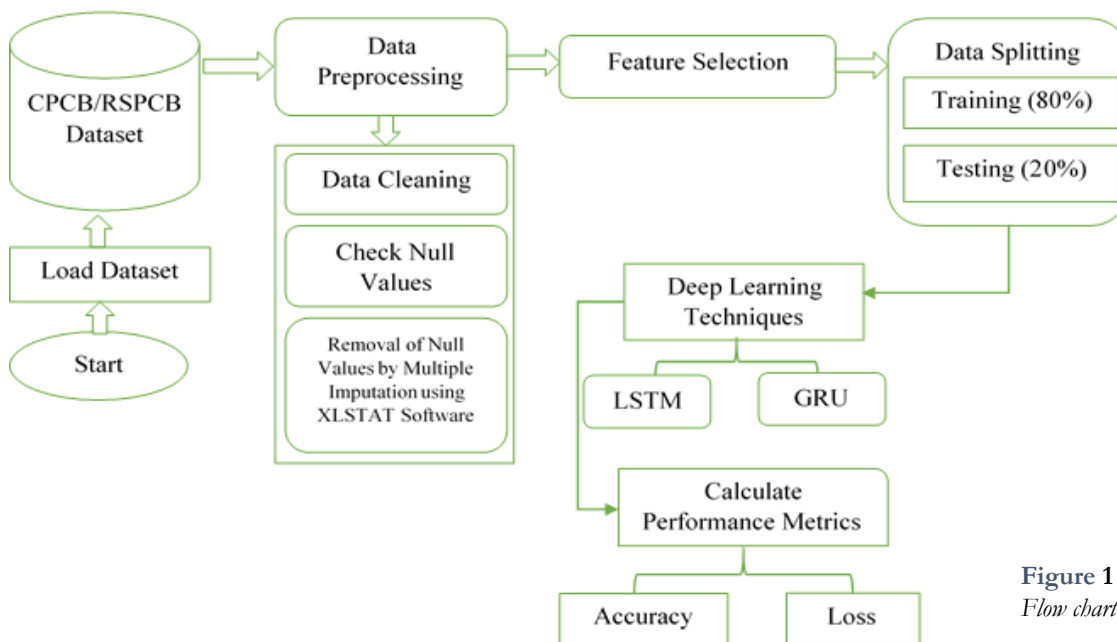


Figure 1
Flow chart of methodology

Data pre-processing

Once the dataset is collected, a series of data preparation steps are undertaken. Python is utilized for initial data processing, with the Pandas library employed for data cleaning. This includes tasks such as identifying and handling null values, filling in mis-

sing data, and labeling the dataset appropriately. Normalization of the data is performed using the min-max scaler, which scales the data to a range suitable for deep learning models. To address any missing values in the dataset, XLSTAT software is used, which provides various imputation methods

including mean, median, and more sophisticated techniques. Specifically, multiple imputation is applied, which creates multiple versions of the dataset with different estimated values for the missing data, thereby improving the accuracy of the subsequent analyses.

Deep learning classifiers

The final stage of the methodology involves utilizing a variety of deep learning classifiers to predict air pollution levels. By comparing the performance and limitations of different classifiers, the study aims to identify the most reliable and accurate model for predicting PM_{2.5} and PM₁₀ concentrations.

LSTM Classifier

The Long Short-Term Memory (LSTM) network, a type of recurrent neural network (RNN), is employed for sequence prediction tasks in this study. LSTMs are particularly well-suited for problems particularly well -

-suited for problems that require long-term dependencies, making them ideal for time-series data like air quality measurements. An LSTM network is composed of units that include an input gate, an output gate, a forget gate, and a cell. These gates control the flow of data into and out of the cell, allowing the network to retain or discard information as needed as shown in Figure 2. The forget gate decides what information from a previous state should be discarded, assigning values between zero and one, where zero indicates that the information should be forgotten, and one indicates that it should be retained. The input gate determines which new information should be added to the current state, while the output gate considers both the current and previous states to decide what data should be output. By effectively managing these long-term dependencies, LSTM networks can make accurate predictions for both current and future time-steps in air quality forecasting (Naresh et al. 2024; Xayasouk et al. 2020).

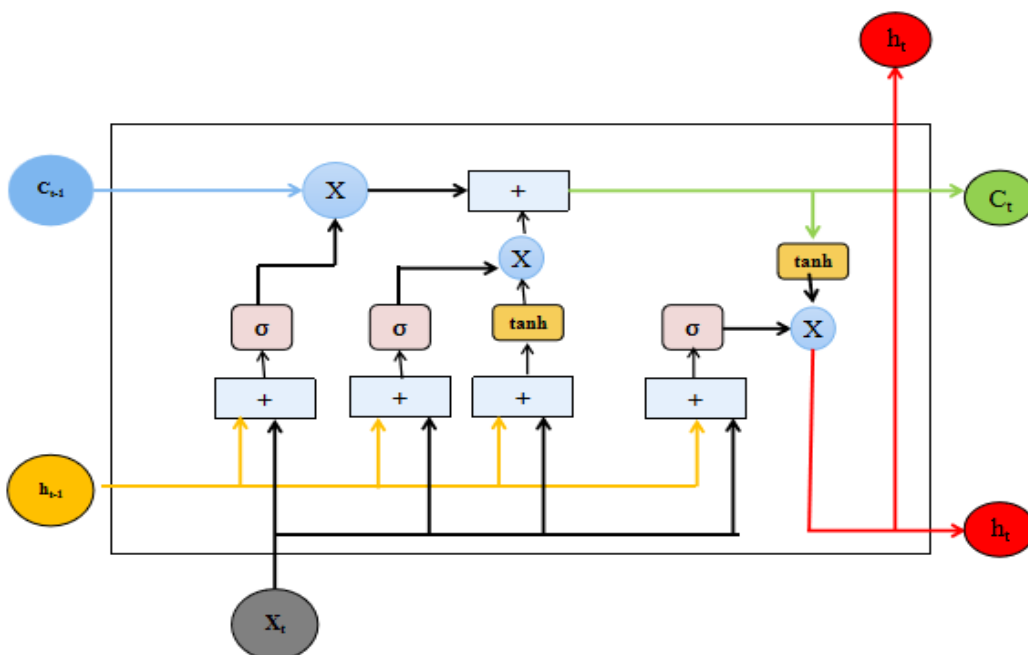


Figure 2
Long short-term memory model.

GRU classifier

The Gated Recurrent Unit (GRU) is another type of RNN that is explored in this study. GRUs are similar to LSTMs but have a simpler architecture, lacking an output gate and utilizing fewer parameters (Zhou et al. 2019). Despite these differences, GRUs perform similarly to LSTMs in many tasks, including modeling polyphonic music, analyzing speech sounds, and understanding spoken language as shown in Figure 3.

GRUs address the problem of vanishing gradients in recurrent neural networks by using gating mechanisms that regulate the flow of information, much like LSTMs. Due to their architectural similarities and comparable performance, GRUs are often viewed as a variant of LSTM networks. In this study, GRUs are evaluated alongside LSTMs to determine the most effective model for predicting air pollution levels in Kota city.

Results and discussion

Correlation analysis

A correlation analysis of the pollutants is conducted using Python. Data from July 1, 2017 to December 30, 2022 have been taken for analysis. The process involves loading the dataset, filtering it for the specified date range selecting relevant pollutant columns a

and computing the correlation matrix A heat map is generated using seaborn to visualize the correlations where values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation and values near 0 indicate a neutral correlation. Figure 4 represents the heat map of correlation.

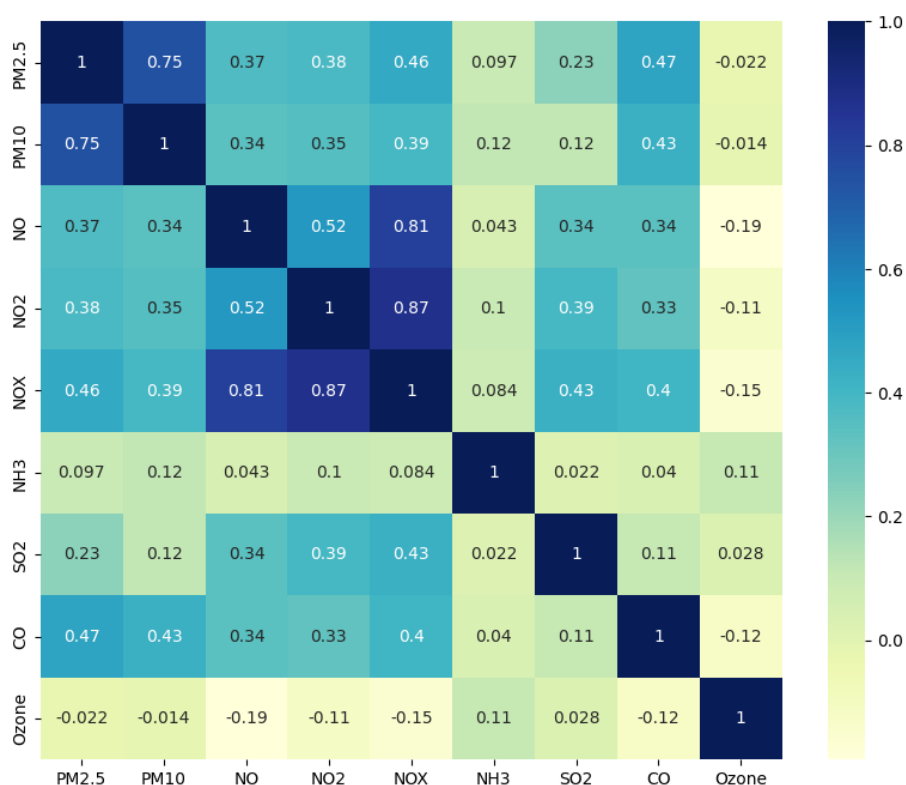


Figure 4
Heat map correlation

Figure 4 illustrates that NOx is strongly positively correlated with NO₂ (r= 0.87) which shows that NO₂ is a major component of NOx emissions. NOx primarily consists of nitrogen oxide (NO) and nitrogen dioxide (NO₂) and negatively correlated with Ozone. PM_{2.5} & PM₁₀ showing positive correlation with CO (r=0.47, 0.43), which shows that sources emitting CO, such as vehicle exhaust, also contribute to particulate matter pollution. Incomplete combustion processes release both CO and PM, leading to their concurrent rise in concentrations. CO shows negative correlation with Ozone. Particulate matter PM_{2.5} is positively correlated with PM₁₀ (r= 0.75).

Model performance

The performance comparison between the Long Short-Term Memory (LSTM) and Gated Recurrent

Unit (GRU) models for predicting PM₁₀ and PM_{2.5} concentrations in Kota city reveals that the GRU model generally outperforms the LSTM model across all evaluation metrics. For PM₁₀, the GRU model achieves a lower MSE of 1751.98 compared to 1877.5 for LSTM, along with a lower RMSE of 41.85 versus 43.33 for LSTM. Similarly, the GRU model shows a slightly better MAE of 23.49, compared to 23.96 for LSTM. The R² score, which indicates the proportion of variance explained by the model, is higher for GRU at 0.65, compared to 0.62 for LSTM. For PM_{2.5}, the GRU model again outperforms LSTM, with a significantly lower MSE (314.56 vs. 451.29), RMSE (17.73 vs. 21.24), and MAE (11.53 vs. 14), and a higher R² score of 0.74 compared to 0.62 for LSTM. These results suggest that the GRU model is more effective in predicting air pollution levels, particularly

Model	MSE		RMSE		MAE		R ²	
	PM ₁₀	PM _{2.5}	PM ₁₀	PM _{2.5}	PM ₁₀	PM _{2.5}	PM ₁₀	PM _{2.5}
LSTM	1877.50	451.29	43.33	21.24	23.96	14.00	0.62	0.62
GRU	1751.98	314.56	41.85	17.73	23.49	11.53	0.65	0.74

Table 1
Comparative performance of models

for PM_{2.5} concentrations in Kota city. Table 1 below shows the comparative performance of both the models. Figure 5 shows a Taylor diagram, with the black point marking the reference point that represents ideal model performance. The blue point corresponds to the LSTM model, and the green point corresponds to the GRU model.. The closer a point is

to the reference, the better the model's performance. In this diagram, the GRU model (green point) is positioned closer to the reference point than the LSTM model (blue point), indicating that the GRU model achieves superior performance in terms of correlation, standard deviation, and RMSE.

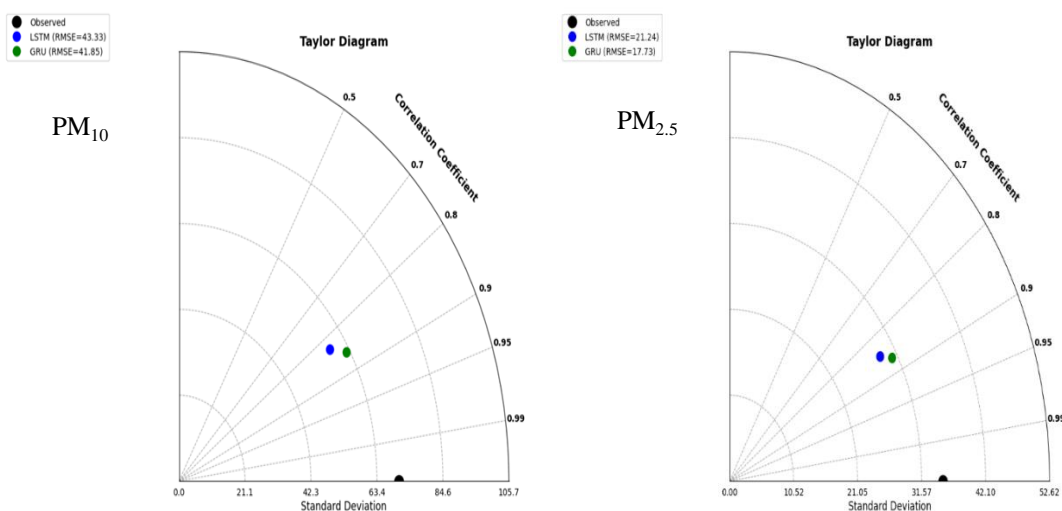


Figure 5
Taylor Diagram for PM₁₀ & PM_{2.5}

Training & validation graphs

The GRU training & validation MSE's for PM₁₀ are

shown in Figure 6. It shows the training and validation mean squared errors (MSE) of a machine

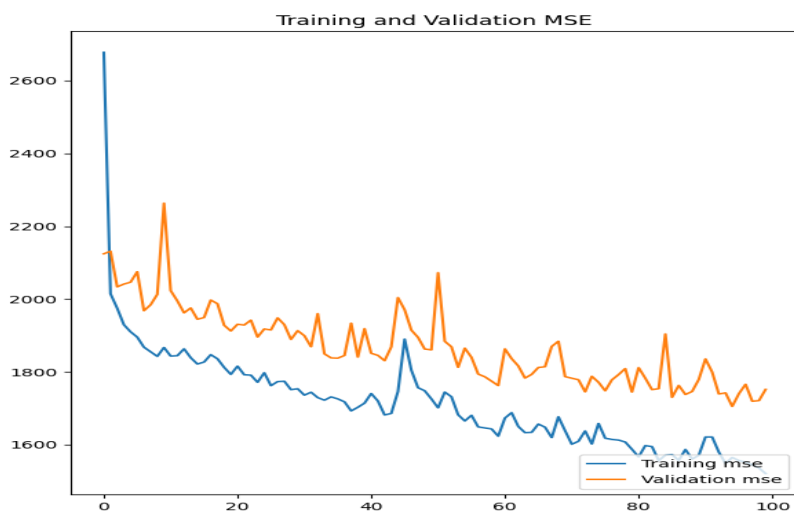


Figure 6
Training and validation of gated recurrent unit for PM₁₀.

learning model over 100 training epochs. Initially, the validation MSE of GRU is at 2124.47 and over the course of the training, it decreases and reaches 1751.98 by the 100th epoch. The GRU training & validation MSE's of PM_{2.5} are shown in Figure-7. It shows the training and validation mean squared errors (MSE) of a machine learning model over 100 training epochs. Initially, the validation MSE of GRU is at 436.48 and over the course of the training, it decreases and reaches 314.56 by the 100th epoch.

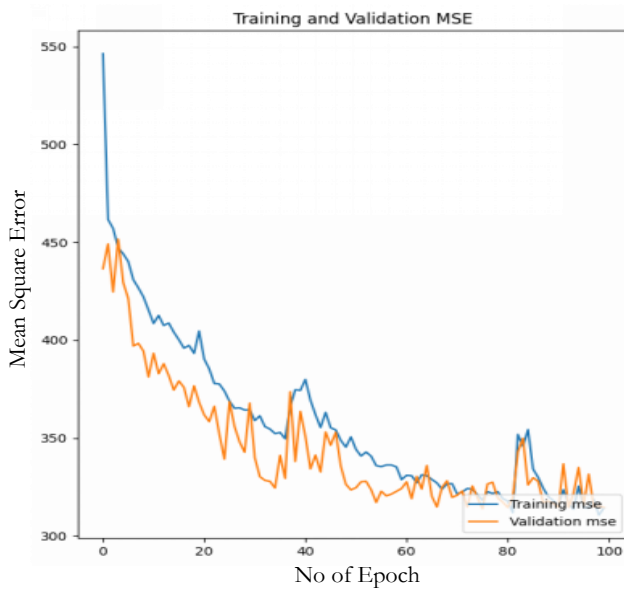


Figure 7. Training and validation of gated recurrent unit for PM_{2.5}.

From Figures 6 & 7, it is concluded that the results of the GRU model are better having the lowest scores of MSE being 1751.98 and 314.56 for PM₁₀ and PM_{2.5} respectively. In comparing the performance of the GRU model for predicting PM_{2.5} concentrations from the current study with that reported by previous study, several key differences are evident (Huang et al. 2021). The current study's GRU model exhibits a Root Mean Square Error (RMSE) of 17.73, which is lower than the RMSE of 20.309 ± 0.053 reported by previous study. This suggests that the GRU model in the current study achieves better average prediction accuracy, as it has a lower magnitude of error. However, the Mean Absolute Error (MAE) for the current study's GRU model is 23.49, significantly higher than the MAE of 11.039 ± 0.049 reported by previous study. This indicates that, despite the lower RMSE, the current model tends to produce larger deviations from the actual values in some cases. Additionally, the R² value for the current study's GRU model is 0.65, which is notably lower than the R² value of 0.9531 ± 0.0002 reported by previous study. This disparity highlights that previous study's GRU model explains a much larger proportion of the variance in PM_{2.5} concentrations, demonstrating a superior overall fit and explanatory power. Overall, while the current study's GRU model shows improved prediction accuracy in terms of RMSE, it does not perform as well in explaining variability or



Figure 8. Original v/s Predicted PM₁₀ & PM_{2.5} using GRU

minimizing absolute errors compared to previous study model. These comparisons underscore both the strengths and limitations of the current GRU model relative to established benchmarks in the literature.

Prediction of PM₁₀ and PM_{2.5}

Figure 8 illustrates the performance of a predictive model for PM₁₀ & PM_{2.5} levels, comparing actual values (green colour) with predicted values (orange colour) over data points. The x-axis represents the number of testing values and y-axis represents the levels of PM₁₀ and PM_{2.5}. The close alignment between the two colored lines indicates that the model accurately predicts PM₁₀ and PM_{2.5} concentrations, effectively capturing overall trends and significant peaks.

Conclusions

The study conducts a correlation analysis and model performance evaluation to predict air pollution levels in Kota City, focusing on PM₁₀ and PM_{2.5} concentrations. The correlation analysis reveals strong relationships between various pollutants, with NO_x and NO₂ showing a high positive correlation ($r=0.87$), indicating that NO₂ is a significant component of NO_x emissions. PM_{2.5} and PM₁₀ are positively correlated with CO ($r=0.47$ and $r=0.43$, respectively), suggesting that sources of CO, such as vehicle exhaust, also contribute to particulate matter pollution. The performance evaluation between Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models demonstrates that the GRU model outperforms LSTM in predicting both PM₁₀ and PM_{2.5} concentrations. The GRU model achieves lower MSE, RMSE, and MAE values and a higher R² score for both pollutants, indicating better accuracy and reliability. Specifically, the GRU model attains MSE values of 1751.98 for PM₁₀ and 314.56 for PM_{2.5}, with corresponding R² scores of 0.65 and 0.74, respectively. The training and validation graphs further confirm the GRU model's superior performance, with a steady decrease in MSE over 100 epochs, reaching the lowest scores for both PM₁₀ and PM_{2.5}. Additionally, the prediction results show a close alignment between the predicted and actual values, demonstrating the model's ability to accurately capture the trends and peaks in pollutant concentrations. In conclusion, the GRU model proves to be a more effective tool for predicting air pollution levels in Kota, making it a valuable asset in air quality management and mitigation strategies.

Acknowledgement

The authors would like to express their gratitude to the Central Pollution Control Board (CPCB) for providing access to the air quality data used in this study. The data collected from the CPCB portal have been invaluable to the analysis and have significantly contributed to the outcomes of this research.

References

- ALOMAR M.K., KHALEEL F., ALSAADI A. A., HAMEED M. M., ALSAADI M. A., AL-ANSARI N. (2022) The Influence of Data Length on the Performance of Artificial Intelligence Models in Predicting Air Pollution. *Advances in Meteorology*. <https://doi.org/10.1155/2022/5346647>.
- CHAE S., SHIN J., KWON S., LEE S., KANG S., LEE D. (2021) PM₁₀ and PM_{2.5} real-time prediction models using an interpolated convolutional neural network. *Scientific Reports* 11(1): 1–9. <https://doi.org/10.1038/s41598-021-91253-9>.
- DOMAŃSKA D., WOJTYLAK M. (2014) Explorative forecasting of air pollution. *Atmospheric Environment*, 92:19–30. <https://doi.org/10.1016/j.atmosenv.2014.03.041>.
- DORESWAM Y., HARISHKUMAR K.S., KM Y., GAD I. (2020) Forecasting air pollution particulate matter (PM_{2.5}) using machine learning regression models. *Procedia Computer Science*, 171(2019):2057–2066... <https://doi.org/10.1016/j.procs.2020.04.221>
- HUANG G., LI X., ZHANG B., REN J. (2021) PM_{2.5} concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. *Science of the Total Environment* 768:144–156. <https://doi.org/10.1016/j.scitotenv.2020.144516>.
- KAMBOJ K., SISODIYA S., MATHUR A.K., ZARE A., VERMA P. (2022) Assessment and spatial distribution mapping of criteria pollutants. *Water, Air, and Soil Pollution*, 233(3). <https://doi.org/10.1007/s11270-022-05522-y>.
- KOTHANDARAMAN D., PRAVEENA N., VARADARAJKUMAR K., MADHAV RAO B., DHABLIYA D., SATLA S., ABERA W. (2022) Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning. *Adsorption Science and Technology* 2022. <https://doi.org/10.1155/2022/5086622>.
- KRISTIANI E., KUO T. Y., YANG C. T., PAI K. C., HUANG C. Y., NGUYEN K. L. P. (2021) PM_{2.5} Forecasting Model Using a Combination of Deep Learning and Statistical Feature Selection. *IEEE Access* 9:68573–68582. <https://doi.org/10.1109/ACCESS.2021.3077574>.

- KULDEEP K., KUMAR P., KAMBOJ P., MATHUR A.K. (2022a) Air Quality Decrement After Lockdown in Major Cities of Rajasthan, India. *ECS Transactions*, 107(1):18479–18496. <https://doi.org/10.1149/10701.18479ecst>.
- KULDEEP K., SISODIYA S., MATHUR A. (2022b) Environmental Risk Assessment Ascribed to Particulate Matter for Kota City, Rajasthan (India). *ECS Transactions* 107(1):543–559. <https://doi.org/10.1149/10701.0543ecst>
- NARESH G., INDIRA B. (2024) Air Pollution Prediction using Multivariate LSTM Deep Learning Model. *International Journal of Intelligent Systems and Applications in Engineering IJISAE* 2024(8s). <https://ijisae.org/index.php/IJISAE/article/view/4111>
- SHARMA M., JAIN S., LAMBA B.Y. (2020) Epigrammatic study on the effect of lockdown amid Covid-19 pandemic on air quality of most polluted cities of Rajasthan (India). *Air Quality, Atmosphere and Health* 13(10):1157–1165. <https://doi.org/10.1007/s11869-020-00879-7>
- SINGH B., NAGDA C., KUMAR K., KAIN T., JHALA L. S., RATHORE D. S. (2022) COVID-19 Implicated ban on Diwali fireworks: a case study on the air quality of Rajasthan, India. *EQA*, 47:22–30. <https://doi.org/10.6092/issn.2281-4485/13698>.
- TAO Q., LIU F., LI Y., SIDOROV D. (2019) Air pollution forecasting using a deep learning model based on 1D Convnets and Bidirectional GRU. *IEEE Access* 7:76690–76698. <https://doi.org/10.1109/access.2019.2921578>.
- VINÑAS M.J.D., GERARDO B.D., MEDINA R.P. (2022) Forecasting PM2.5 and PM10 Air Quality Index using Artificial Neural Network. *Journal of Positive School Psychology*, 6(5):6863–6871. ISSN: 2717-7564 <https://journalppw.com/index.php/jpsp/index>
- XAYASOUK T., LEE H.M., LEE G. (2020) Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. *Sustainability (Switzerland)* 12(6). <https://doi.org/10.3390/su12062570>.
- YADAV R., VYAS P., KUMAR P., SAHU L. K., PANDYA U., TRIPATHI N., GUPTA M., SINGH V., DAVE P. N., RATHORE D. S., BEIG G., JAAFFREY S. N.A. (2022) Particulate Matter Pollution in Urban Cities of India During Unusually Restricted Anthropogenic Activities. *Frontiers in Sustainable Cities*, 4(3):1–14. <https://doi.org/10.3389/frsc.2022.792507>.
- ZHOU X., XU J., ZENG P., MENG X. (2019) Air Pollutant Concentration Prediction Based on GRU Method. *Journal of Physics: Conference Series* 1168(3). <https://doi.org/10.1088/1742-6596/1168/3/032058>.