7th Intercontinental Geoinformation Days

igd.mersin.edu.tr

PM10 air pollutant prediction using deep learning LSTM Model: A case study of Istanbul, Türkiye

Omar Wisam Alqaysi*[1], **Dursun Zafer Şeker**[2]

[1] *Istanbul Technical University, Graduate School, Applied Informatics Department, GIT Program, Istanbul, Türkiye*
[2] *Istanbul Technical University, Civil Engineering Faculty, Department of Geomatics Engineering, Istanbul, Türkiye*

Keywords
Deep learning
LSTM
Air pollution
PM10
GRU

Abstract

Accurate forecasting of PM10 concentrations is crucial for air quality management and public health protection. This study proposes a deep learning-based model for predicting PM10 concentrations in Istanbul, Türkiye, utilizing a combination of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. Historical air pollution data from the Ministry of Environment, Urbanization, and Climate Change of Turkey and meteorological data from NASA for the period of January 2018 to August 2023 were employed for model development. Ümraniye district was selected as the study area due to its comprehensive air quality data availability. An extensive model development process involved identifying the optimal input sliding window, input features, and model architecture through parameter tuning. The LSTM+GRU model resulted in the best metrics, achieving an RMSE of 6.71, R2 of 0.86, and MAPE of 15.9%. The model demonstrated strong generalization capabilities when tested on data from eight different stations in Istanbul. While the proposed model exhibited promising performance, certain limitations warrant further investigation. The effectiveness of the model for air pollutants other than PM10 remains unexplored. Additionally, an evaluation of feature importance ranking for the input parameters is necessary to identify the most influential factors contributing to PM10 concentrations. Future research endeavors will address these limitations and refine the model for broader applicability.

1. Introduction

Air pollution, especially PM10 and PM2.5, is a major health and environmental hazard. PM air pollution causes cardiovascular and respiratory diseases, according to numerous studies. (Anderson et al., 2011; Liang et al., 2014). In addition, there are a number of important environmental effects of air pollution, such as the greenhouse effect, acid rain, ozone layer alterations, decreased visibility, and lower-quality products. (Lehadus et al., 2019).It is widely regarded as a primary concern for the environment and public health on a global scale. It is considered one of the main problems for the environment and public health globally (Estuardo-Moreno et al., 2022). Particulate matter also impacts climate and precipitation, making it a crucial issue associated with air pollution (Kamarehie et al., 2017).

Understanding the effects of PM10 and PM2.5 on the environment and public health is essential for developing effective strategies to mitigate air pollution and protect human well-being.

1.1 Machine Learning & Deep Learning Prediction Models

Machine learning and deep learning methods have become effective tools for predicting air pollutants because they can comprehend intricate connections among many components and provide precise forecasts. Machine learning methods such as Extreme-Gradient Boosting (XGBoost), Random Forest (RF), and Deep Neural Networks (DNNs) have effectively been utilized for PM10 prediction RF has been used to estimate daily concentrations of pollutants in Sweden (Stafoggia et al., 2020), while XGBoost has been applied to assess the role of atmospheric circulation in PM10 in urban areas with complex topography (Sekula et al., 2022) DNNs have demonstrated their effectiveness in comprehensive air-quality index prediction, incorporating multiple variables and factors (Kim et al., 2022). These machine learning and deep learning models offer promising solutions for air quality monitoring and forecasting, enabling timely interventions and mitigating the adverse health and environmental impacts of air pollution.

* **Corresponding Author**

*(alqaysi21@itu.edu.tr) ORCID ID 0009-0009-1085-9621
(seker@itu.edu.tr) ORCID ID 0000-0001-7498-1540

Cite this study

Alqaysi O & Şeker D Z (2023). PM10 Air Pollutant Prediction Using Deep Learning LSTM Model: A Case Study of Istanbul, Türkiye. Intercontinental Geoinformation Days (IGD), 7, 254-257, Peshawar, Pakistan

1.2 Long-Short Term Memory (LSTM)

LSTM (Long Short-Term Memory) recurrent neural network (RNN) architecture overcomes typical RNNs' shortcomings in learning long-term dependencies (Kratzert et al., 2018). Since 1995, it has been employed in natural language processing, time series forecasting, and trajectory prediction (Dai & Li, 2019; Greff et al., 2017).

Due to its ability to capture and store data across lengthy sequences, LSTM is useful for processing and predicting. Gates and memory cells control network information flow (Gers et al., 2000). Each LSTM memory cell has input, forget, and output gates. These gates control information flow into, out of, and within each memory cell, allowing the network to recall or forget information at specific time steps. (Gers et al., 2000).

Studies show that LSTM is an effective and accurate PM10 prediction model. W. Li & Jiang (2023) suggested a TCN-BiLSTM-DMAttention model with strong prediction accuracy and generalization performance to help prevent air pollution. Istiana et al. (2022) evaluate deep learning applications, notably LSTM, for PM2.5 concentration prediction, emphasizing their efficiency and cost-effectiveness.

1.3 Gated Recurrent Unit (GRU)

Recurrent neural networks (RNNs) like the Gated Recurrent Unit (GRU) capture dependencies and patterns in sequential input. It has gating techniques to keep long-term dependencies and solve the vanishing gradient problem in traditional RNNs. Reset and update gates allow the GRU model to adaptively acquire and interpret sequential information, making it ideal for time series data and sequential modeling. The GRU model has been used to predict PM10 and PM2.5 air pollution in several studies. Dairi et al. (2021) used deep learning models using RNN, LSTM, and GRU architectures to forecast air quality using the AirNet dataset, which comprises meteorological time series and air quality data. Qing et al. (2019) suggested a deep learning-based short-term PM2.5 concentration forecasting model using convolutional-based bidirectional gated recurrent unit (CBGRU) neural networks and 1D convents. Yang et al. (2020) also evaluated CNN—LSTM and CNN—GRU with different stand-alone PM concentration prediction algorithms in Seoul.

2. Method

This study tests the performance of both LSTM & GRU deep learning models on PM10 concentrations forecasting concentrations. The LSTM model, inspired by the human brain, processes and learns sequential data well, making it suited for modeling air pollution data's temporal patterns and correlations. The study's methodology includes data collecting, preprocessing, model construction, and evaluation.

2.1 Study area & Data collection

Istanbul, a heavily populated megacity on the European and Asian continents, is a vital air pollution forecast research hub. For air quality assessments, Istanbul's comprehensive air pollution monitoring station network provides valuable data. Air pollution from heavy traffic, industrial activities, and high population density makes the city a good air quality research site. Istanbul map with metrological and air quality stations is in Figure 1.

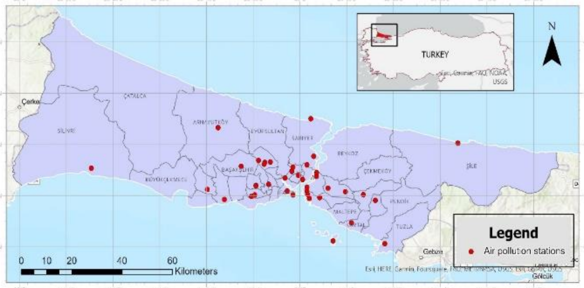


Figure 1. Air pollution stations across Istanbul.

Ümraniye district was chosen for model training because it had the most consistent PM10 data. Hourly time series data of air quality for 38 stations in Istanbul from January 2018 to August 2023 was obtained from the Ministry of Environment, Urbanization and Climate Change of Turkey (SIM, 2023). Eight air pollutants were monitored: PM10, PM2.5, SO2, CO, NO, NOX, NO2, and O3. However, subsequent inspection revealed that not all stations measured all contaminants.

Meteorological data was obtained from NASA's MERRA2 satellite data using its Earth Science research program. The following abbreviations were used to denote wind speed at 10 m, wind direction, surface pressure, specific humidity, and temperature at 2 m: WS10M, WD10M, PS, QV2M, and T2M. (NASA, 2023).

2.2 Data preprocessing

After exploring the data, missing values and outliers were detected thus data cleaning and missing data imputation were done in the following steps. Figure 2 shows the missing data for PM10 in white areas.

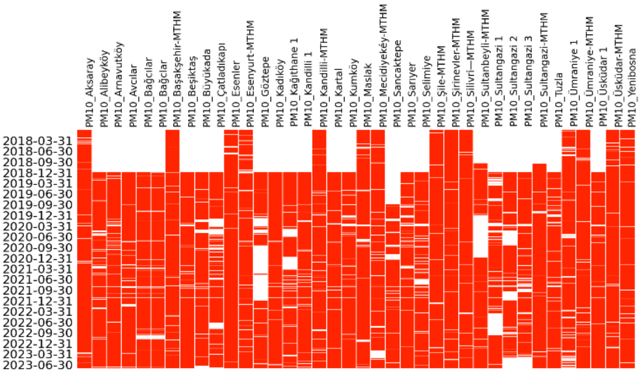


Figure 2. PM10 Data availability.

Imputation of missing values follows outlier removal. Outliers were identified using the Python ADTK library's InterQuartileRangeAD detector (Arundo, 2023). This detector utilizes historical data to compare time series values with the 1st and 3rd quartiles. Anomalous data points are identified when their differences exceed the product of the interquartile range (IQR) and a user-defined factor c.

The data was standardized using StandardScaler after outlier elimination. After that, "Fancyimpute" (Rubinsteyn & Feldman, 2016) was used to fill in the missing data. Before modeling, the dataset was split into 60% training, 20% validation, and 20% testing sets.

2.3 Model development

A variety of LSTM and GRU model architectures were trained and evaluated then the best-performing model was selected based on evaluation metrics. The study utilizes five criteria to evaluate the models' predictive performance and ascertain the effectiveness of the suggested strategy.

Sliding window inputs were made which is a common technique used in time series analysis and is often used in conjunction with LSTMs. The sliding window technique involves splitting a time series into smaller windows of fixed length and using these windows as input to a model. Many trials were done until the best window size was selected.

Prior to training, it is necessary to normalize the dataset. Applying data normalization can enhance the efficiency of deep learning models and enhance their resilience to fluctuations in the input data.

The model development was conducted in three steps, which involved determining the optimal input size by utilizing all input features. Then choose input features alternatively. Finally, adjusting the number of hidden layers and neurons in the model architecture until optimal results are achieved.

Table 1. Parameters used during model development.

Model Architecture	Input Features	Input window(hr)	Hidden units
LSTM	Metrological data	50	32
LSTM,LSTM	Metrological data + NO	30	32,32
LSTM,LSTM,LSTM,LSTM	Metrological data + NO2+NOX+PM2.5	20	64,64
GRU		10	64
GRU,GRU		5	16,16,16,16
GRU,GRU,GRU			
GRU,LSTM			
LSTM,GRU			

The assessment measures employed include Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R2). Figure 3 illustrates the sequence of steps followed in our study. The parameters that were examined to determine the optimal combination are illustrated in Table 1.

3. Results & Discussion

Experimental trials were done to get the lowest MSE & RMSE by changing the combination of parameters used for model training. Table 2 shows some of the trials and their respective evaluation metrics.

As mentioned in the model development section, the first step is finding the best input window size, and trials showed that 10 hours of input data performed best. Using metrological data and (NO) pollutants improved feature selection predictions.

Finally, choosing the model architecture, hidden layers, and neurons. After testing, one-layer LSTM and one-layer GRU with 32 neurons in each layer produced the most accurate prediction model with an RMSE of 6.71 and R2 of 0.86. Figure 4 shows PM10 predictions vs. actual PM10 values.

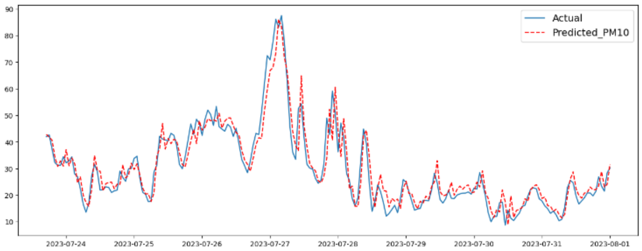


Figure 4. Predicted PM10 vs. Actual concentrations.

The best model generalization ability was tested on other stations and the results is shown in table 2. Although the model would predict more accurately if PM2.5 data were used. But using PM2.5 data would limit our model's usability on other stations. Since not all stations have data for PM2.5 it was not added for our model.

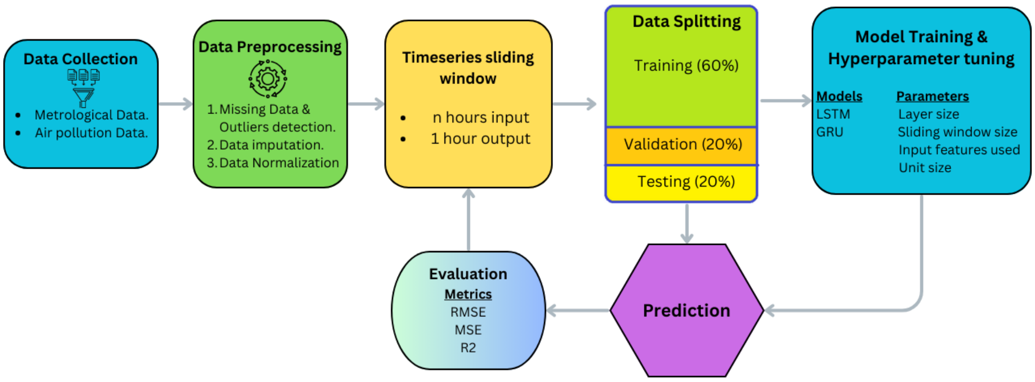
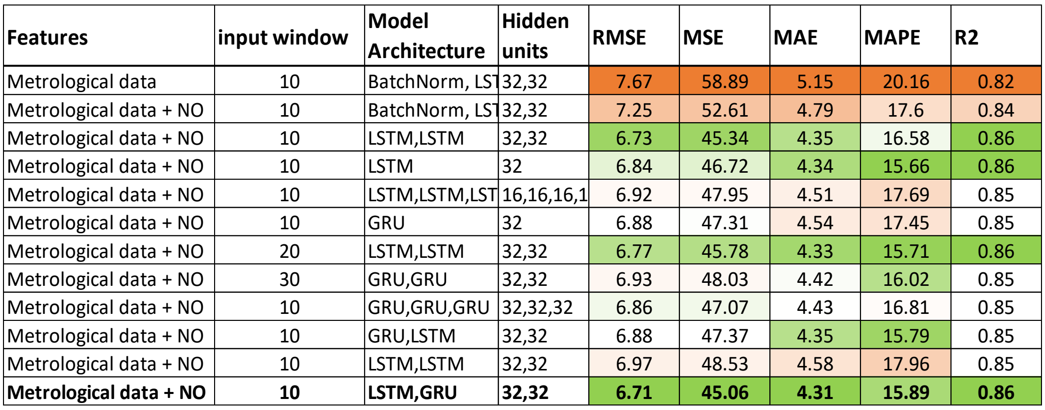


Figure 3. PM10 prediction workflow.

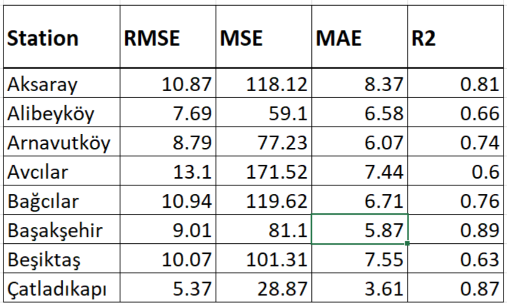
It's worth noting that usage of other outlier detectors or other station's data for modeling may yield better results and that can be performed on later research.

Table 2. PM10 prediction trials with different parameter combination.



Features	input window	Model Architecture	Hidden units	RMSE	MSE	MAE	MAPE	R2
Metrological data	10	BatchNorm, LST	32,32	7.67	58.89	5.15	20.16	0.82
Metrological data + NO	10	BatchNorm, LST	32,32	7.25	52.61	4.79	17.6	0.84
Metrological data + NO	10	LSTM,LSTM	32,32	6.73	45.34	4.35	16.58	0.86
Metrological data + NO	10	LSTM	32	6.84	46.72	4.34	15.66	0.86
Metrological data + NO	10	LSTM,LSTM,LST	16,16,16,1	6.92	47.95	4.51	17.69	0.85
Metrological data + NO	10	GRU	32	6.88	47.31	4.54	17.45	0.85
Metrological data + NO	20	LSTM,LSTM	32,32	6.77	45.78	4.33	15.71	0.86
Metrological data + NO	30	GRU,GRU	32,32	6.93	48.03	4.42	16.02	0.85
Metrological data + NO	10	GRU,GRU,GRU	32,32,32	6.86	47.07	4.43	16.81	0.85
Metrological data + NO	10	GRU,LSTM	32,32	6.88	47.37	4.35	15.79	0.85
Metrological data + NO	10	LSTM,LSTM	32,32	6.97	48.53	4.58	17.96	0.85
Metrological data + NO	**10**	**LSTM,GRU**	**32,32**	**6.71**	**45.06**	**4.31**	**15.89**	**0.86**

Table 3. Model's PM10 prediction performance on other stations.



Station	RMSE	MSE	MAE	R2
Aksaray	10.87	118.12	8.37	0.81
Alibeyköy	7.69	59.1	6.58	0.66
Arnavutköy	8.79	77.23	6.07	0.74
Avcılar	13.1	171.52	7.44	0.6
Bağcılar	10.94	119.62	6.71	0.76
Başakşehir	9.01	81.1	5.87	0.89
Beşiktaş	10.07	101.31	7.55	0.63
Çatladıkapı	5.37	28.87	3.61	0.87

4. Conclusion

This study constructed a deep-learning model utilizing LSTM and GRU deep-learning algorithms to estimate PM10 concentrations on an hourly basis.

Compared to earlier models, the LSTM+GRU model has demonstrated great predictive capabilities across all the models that have been built. In addition, the model that was created was tested on data from eight distinct areas and showed strong generalization abilities.

Nevertheless, this study has specific constraints. The efficacy of the suggested model has not been investigated for air pollutants other than PM10. Moreover, there has been no assessment of the ordering of the input parameters in terms of their relevance. Subsequent investigations will focus on overcoming these constraints.

References

Anderson, J. O., Thundiyil, J. G., & Stolbach, A. (2011). Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health. Journal of Medical Toxicology. https://doi.org/10.1007/s13181-011-0203-1

Dai, S., & Li, L. (2019). Modeling Vehicle Interactions via Modified LSTM Models for Trajectory Prediction. Ieee Access. https://doi.org/10.1109/access.2019.2907000

Estuardo-Moreno, H., Gomez-Alvarez, A., Lucero-Acuña, J. A., Almendariz-Tapia, F. J., Esparza-Ponce, H. E., & Ramirez-Leal, R. (2022). Physical and Chemical Morphology of Organic Compounds at PM10 by TEM-EDS and GC-SM. Microscopy and Microanalysis. https://doi.org/10.1017/s1431927622011977

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. Ieee Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/tnnls.2016.2582924

Istiana, T., Kurniawan, B., Soekirno, S., & Prakoso, B. (2022). Deep Learning Implementation Using Long Short Term Memory Architecture for PM$_{2.5}$ Concentration Prediction: A Review. Iop Conference Series Earth and Environmental Science. https://doi.org/10.1088/1755-1315/1105/1/012026

Kamarehie, B., Ghaderpoori, M., Jafari, A., Karami, M. A., Mohammadi, A., Azarshab, K., Ghaderpoury, A., & Noorizadeh, N. (2017). Estimation of Health Effects (Morbidity and Mortality) Attributed to PM10 and PM2.5 Exposure Using an Air Quality Model in Bukan City, From 2015-2016 Exposure Using Air Quality Model. Environmental Health Engineering and Management. https://doi.org/10.15171/ehem.2017.19

Kim, D., Han, H., Wang, W., Kang, Y., Lee, H.-Y., & Kim, H. S. (2022). Application of Deep Learning Models and Network Method for Comprehensive Air-Quality Index Prediction. Applied Sciences. https://doi.org/10.3390/app12136699

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff Modelling Using Long Short-Term Memory (LSTM) Networks. Hydrology and Earth System Sciences. https://doi.org/10.5194/hess-22-6005-2018

Lehadus, M. P., Nedeff, V., Barsan, N., Sandu, A. V., Mosnegutu, E., Tomozei, C., Irimia, O., Andrioai, G., & Sandu, I. (2019). Monitoring the Particulate Matter (PM10) Emissions from Bacau City Termo-Energetic Industry. Revista De Chimie. https://doi.org/10.37358/rc.19.8.7446

Liang, Y., Fang, L.-Q., Pan, H., Zhang, K., Kan, H., Brook, J. R., & Sun, Q. (2014). PM2.5 in Beijing – Temporal Pattern and Its Association with Influenza. Environmental Health. https://doi.org/10.1186/1476-069x-13-102

Rubinsteyn, A., & Feldman, S. (2016). fancyimpute: An Imputation Library for Python. https://github.com/iskandr/fancyimpute