Air Quality Monitoring System Based on Deep Learning: Hybrid CNN-BiLSTM Model.

Abstract

Background: Air quality monitoring is essential for protecting public health and the environment. Reliable short-term forecasting of the Air Quality Index (AQI) can help authorities anticipate pollution episodes and take appropriate decisions.

Methods: This study proposes a hybrid CNN-BiLSTM model to predict the AQI one hour ahead from the preceding 24 hours of observations. The CNN extracts relevant features from 14 input variables, while the BiLSTM learns temporal dependencies. Two datasets from India were used: a multi-station dataset from the Central Pollution Control Board (CPCB) and an hourly single-station dataset from the NSIT Dwarka monitoring station in Delhi (2024–2025). The model was compared with linear regression, support vector machine (SVM) and a simple LSTM using the R^2, MSE, RMSE and MAE metrics.

Results: On the first dataset, the CNN-BiLSTM achieved an R^2 of 0.9942, an MSE of 53.2003, an RMSE of 7.2938 and an MAE of 2.6339. On the second, shorter dataset collected from a single station, it obtained an R^2 of 0.9756, an MSE of 300.7757, an RMSE of 17.3428 and an MAE of 10.9171, outperforming the three baseline models on both datasets.

Conclusion: The model maintains good predictive performance on the shorter single-station series, although its results are lower than those obtained in the first experiment. These findings highlight the potential of combining convolutional and bidirectional recurrent networks for short-term AQI prediction; further evaluation across additional monitoring stations is needed.

Key words:-
Air Quality, Air Pollution, Deep Learning, CNN-BiLSTM, Time Series, Forecasting, Air Quality Index (AQI)

Introduction:-

Air quality monitoring is a major concern for public health and sustainable development. Its deterioration, mainly caused by industrialization, road traffic and urbanization, affects both human health and the environment. Therefore, intelligent systems capable of monitoring and forecasting pollution levels are needed to help authorities make appropriate decisions. Air pollution is one of the main environmental factors affecting life expectancy and contributes to approximately 5% of deaths worldwide [1].

Air pollution refers to the presence of chemical, physical or biological substances in the atmosphere at concentrations that may harm human health and the environment [2]. It can cause serious diseases, reduce life expectancy and disrupt ecosystems [3]. Industrial activities, transportation, agriculture and urbanization contribute to increasing concentrations of harmful particles and gases [4]. These pollutants can cause respiratory and cardiovascular diseases, reduce agricultural yields and affect biodiversity [5].

Air pollution therefore poses a major threat to populations and ecosystems [6]. A large proportion of the world's population is exposed to polluted air, which is responsible for millions of deaths each year [7]. Indoor air pollution also represents a significant health risk [8]. Ambient air pollution is associated with several million premature deaths, particularly in low- and middle-income countries [9]. In Africa, this issue also remains a concern, particularly in South Africa, where pollution is associated with cardiopulmonary diseases and respiratory cancers [10].

Fine particulate matter (PM2.5) is one of the main contributors to air pollution-related mortality. Because of its small size, it can penetrate deep into the lungs and enter the bloodstream. Ozone

UNDER PEER REVIEW IN IJWR

(O_3) also contributes to the health effects of air pollution [11]. According to the World Health Organization, 99% of the world's population is exposed to air pollution levels exceeding the recommended limits [12]. Advances in deep learning have encouraged the development of hybrid models to improve air quality forecasting. Dong et al. (2024) proposed a model combining EMD, Transformer and BiLSTM for short-term Air Quality Index (AQI) forecasting [13], highlighting the potential of hybrid architectures for modeling air pollution time series. Statistical and machine learning methods may nevertheless face limitations when modeling complex relationships and temporal dependencies. Wang et al. (2024) developed a VMD-GAT-BiLSTM architecture that exploits the spatial and temporal characteristics of the data [14], further demonstrating the relevance of deep learning for air quality forecasting.

To investigate model performance under different observation conditions, this study uses two datasets from India. The first combines measurements from several monitoring stations over a period of several years, while the second comes from the NSIT Dwarka station in Delhi and covers the period 2024–2025. This second dataset, consisting of measurements aggregated at an hourly frequency, makes it possible to assess model performance on a shorter time series collected from a single station. Therefore, using both datasets enables a comparison of the predictive performance of the models under different data collection conditions and data volumes. In this study, we propose a hybrid CNN-BiLSTM model in which the CNN extracts relevant features from the data, while the BiLSTM models temporal dependencies. The model's performance is evaluated using the R^2, MSE, RMSE and MAE metrics and compared with that of several prediction methods.

Materials and Methods:-

Data Description:

Air pollution forecasting requires a large and reliable dataset. In this study, we used a public dataset available on Kaggle entitled *"Time Series Air Quality Data of India (2010–2023)"* [15]. It contains 13 years of historical data, from January 1, 2010, to July 1, 2023, covering more than 450 cities in India. The data were collected by the Central Pollution Control Board (CPCB), the official agency responsible for pollution monitoring in India. The dataset includes 32 variables related to weather conditions, air pollutants and local context, with measurements from major cities such as Delhi, Mumbai, Kolkata, Chennai and Bangalore. These data provide an important basis for developing a reliable air quality prediction model. Further details are available in [15].

A second dataset was used to investigate model performance under different monitoring conditions. It was obtained from the NSIT Dwarka air quality monitoring station in Delhi, India, and covers the period from January 2024 to December 2025 [20]. The original data, collected at 15-minute intervals, were aggregated to an hourly frequency to obtain a time series suitable for the prediction protocol. After data preparation and quality checks, the final dataset contained approximately 17,447 hourly observations from a single station. It includes concentrations of various air pollutants, meteorological variables and the Air Quality Index (AQI), which was used as the target variable. Unlike the first dataset, which combines observations from numerous stations over several years, this dataset enables model performance to be assessed on a shorter time series collected at a single site.

UNDER PEER REVIEW IN IJAP

Data Preprocessing:

The first dataset, covering the period 2016–2022, was extracted, grouped by semester, merged and combined with the metadata. Columns containing more than 50% missing values were removed, and incomplete rows and duplicates were handled. Outliers were examined using the interquartile range (IQR) method (threshold of 2 × IQR), fixed thresholds and the Local Outlier Factor (LOF) algorithm (threshold of 0.05). The remaining missing values were imputed using K-nearest neighbors (KNN, k = 3), resulting in a residual outlier rate of 0.95%. Finally, dates, measurement units and variable names were standardized. The AQI and AQI_Bucket variables were also included in the analysis.

The second dataset, obtained from the NSIT Dwarka station for the period 2024–2025, underwent preprocessing adapted to its characteristics. Physically impossible values were removed, and outliers were examined using the IQR method, physical thresholds, LOF and temporal variations in the measurements. This process helped distinguish measurement errors from potential pollution peaks. Depending on the case, erroneous values were removed or corrected, while missing values were handled through linear interpolation or KNN imputation (k = 3). Final checks were performed to identify inconsistent negative values and duplicates.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) was conducted to examine the distribution of major pollutants and their variations over time. It also helped identify relationships between environmental variables and pollutants and better understand the characteristics of the dataset. A correlation matrix was then constructed to analyze relationships among the variables (Figure 1). Based on this analysis, 14 relevant variables were selected as input features for developing and training the CNN-BiLSTM model.

For the second dataset, obtained from the NSIT Dwarka station and covering the period 2024–2025, no separate exploratory data analysis was performed. The processing mainly focused on data quality checks, outlier detection and verification, and the treatment of impossible and missing values. The same 14 input variables were used for modeling to maintain a consistent basis for comparison between the two datasets.

UNDER PEER REVIEW IN IJAR

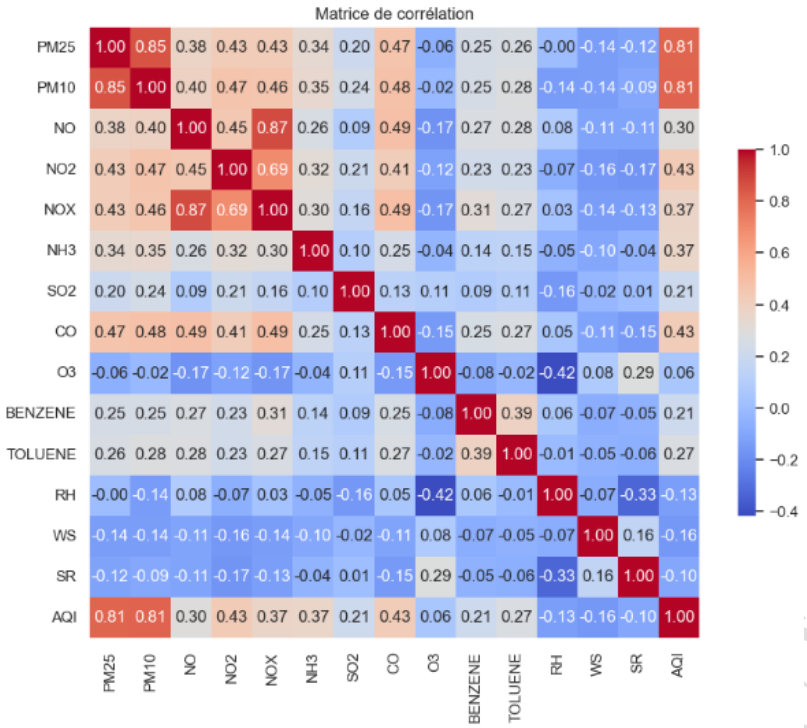

Figure 1:- Correlation matrix of the variables (Dataset 1)

Data Preparation for the Model:

The hourly data were standardized using *StandardScaler(a python function)*, which was fitted exclusively on the training set to prevent data leakage. The data were then organized into 24-hour windows to predict the AQI one hour ahead (t+1). The observations were converted into PyTorch tensors and processed in batches of 128.

The data were split chronologically, with 70% allocated to training, 15% to validation and 15% to testing. Random shuffling was applied only to the training data. The same protocol was used for both datasets to compare their performance.

CNN-BiLSTM Model Architecture and Training:

The CNN-BiLSTM model combines a convolutional neural network (CNN), which extracts features from the input data, with a bidirectional long short-term memory network (BiLSTM), which learns temporal dependencies. This approach builds on the studies presented in [16] and [17].

The model uses 14 input features over a 24-hour window to predict the AQI one hour ahead. It consists of three CNN layers with 32, 64 and 128 filters, respectively, followed by a BiLSTM layer with 128 hidden units per direction. The BiLSTM output is passed through a dropout layer with a rate of 25%, followed by two fully connected layers, the last of which produces the predicted AQI value (Figure 2).

Training uses batches of 128 observations, the AdamW optimizer and the mean squared error (MSE) loss function, with a maximum of 60 epochs. The learning rate is adjusted during training using the *OneCycleLR* scheduler. Early stopping, with a patience of 10 epochs, terminates training when the validation loss no longer improves sufficiently.

UNDER PEER REVIEW IN IJAR

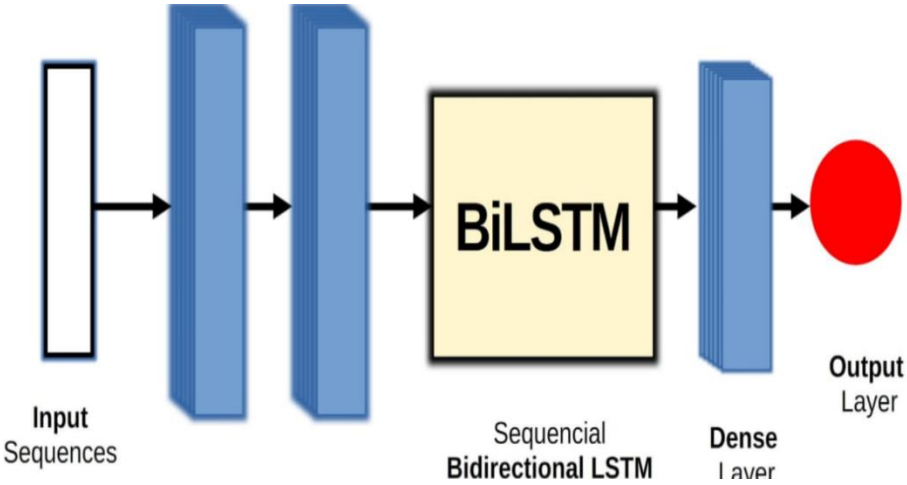

Figure 2:- Hybrid CNN-BiLSTM architecture

Input (14 features, 24-hour window) → CNN (32 → 64 → 128 filters) → BiLSTM (128 hidden units per direction; 256-dimensional output) → Dropout (25%) → Fully connected layers (256 → 64 → 1) → Predicted AQI

Results:-

Performance of the CNN-BiLSTM Model on the First Dataset (Dataset 1):

Table 1:- Performance of the CNN-BiLSTM model on Dataset 1 (test set)

Metric	Value
R^2	0.9942
MSE	53.2003
RMSE	7.2938
MAE	2.6339

As shown in Table 1, the CNN-BiLSTM model achieved a coefficient of determination (R^2) of 0.9942, an MSE of 53.2003, an RMSE of 7.2938 and an MAE of 2.6339. These results show that the predicted values are close to the actual AQI values. The R^2 coefficient indicates that the model captures a large proportion of the variations in the index, while the MAE and RMSE values provide information about the magnitude of the prediction errors.

UNDER PEER REVIEW IN IJAR

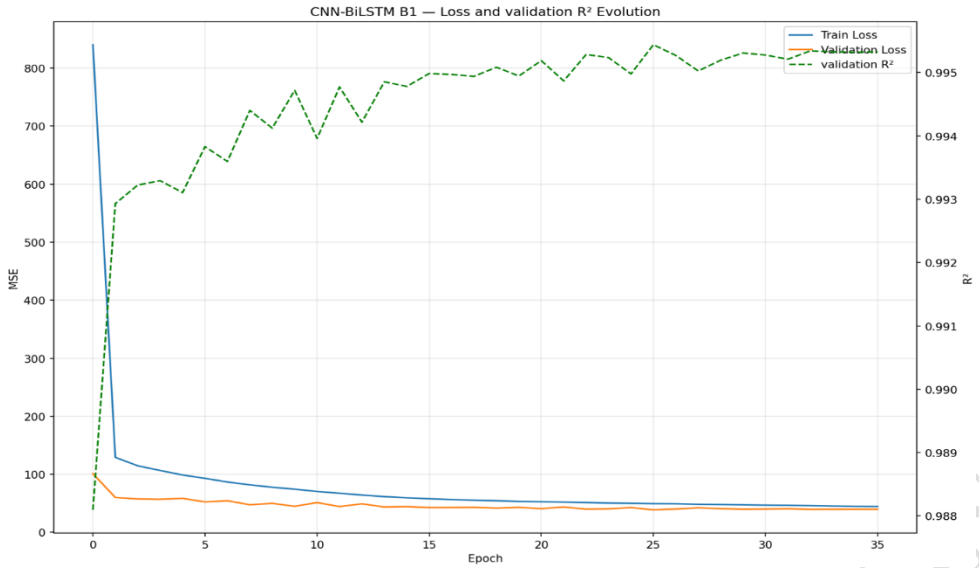

Figure 3:- Learning curve (training and validation loss) – Dataset 1

The learning curve (Figure 3) shows a rapid decrease in loss during the first few epochs, followed by stabilization at a low level. This trend indicates that the model learns effectively from the training data. The coefficient of determination (R^2) reaches approximately 0.996, indicating good agreement between the predicted and actual values. These results are encouraging; however, evaluation on the test set remains necessary to assess the model's ability to predict data that were not used during training.

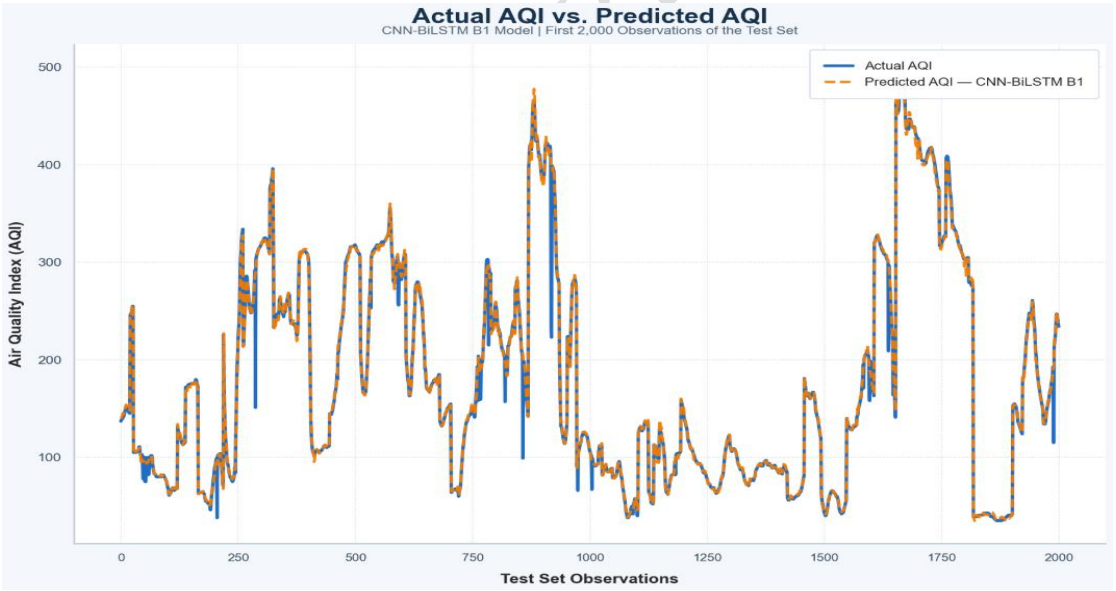

Figure 4:- Overlaid curves of actual and predicted AQI values – Dataset 1

Figure 4 compares the actual AQI values with those predicted by the CNN-BiLSTM model across 7,000 test observations. The two curves follow similar patterns, indicating that the model reproduces the variations in the index reasonably well. This agreement is also evident at several pollution peaks. The graph therefore illustrates the model's ability to estimate the AQI from the input variables.

PREVIEW IN IJAR

Performance of the CNN-BiLSTM Model on the Second Dataset (Dataset 2):

Table 2:- Performance of the CNN-BiLSTM model on Dataset 2 (test set)

Metric	Value
R^2	0.9756
MSE	300.7757
RMSE	17.3428
MAE	10.9171

On the second dataset (Table 2), the CNN-BiLSTM model achieved a coefficient of determination (R^2) of 0.9756, indicating that it explains approximately 97.56% of the variability in the AQI. The RMSE of 17.3428 and the MAE of 10.9171 show that the predictions deviate from the actual values. The MSE, at 300.7757, also reflects the magnitude of the squared errors. Overall, these results indicate that the model has a good ability to predict the AQI on the test data.

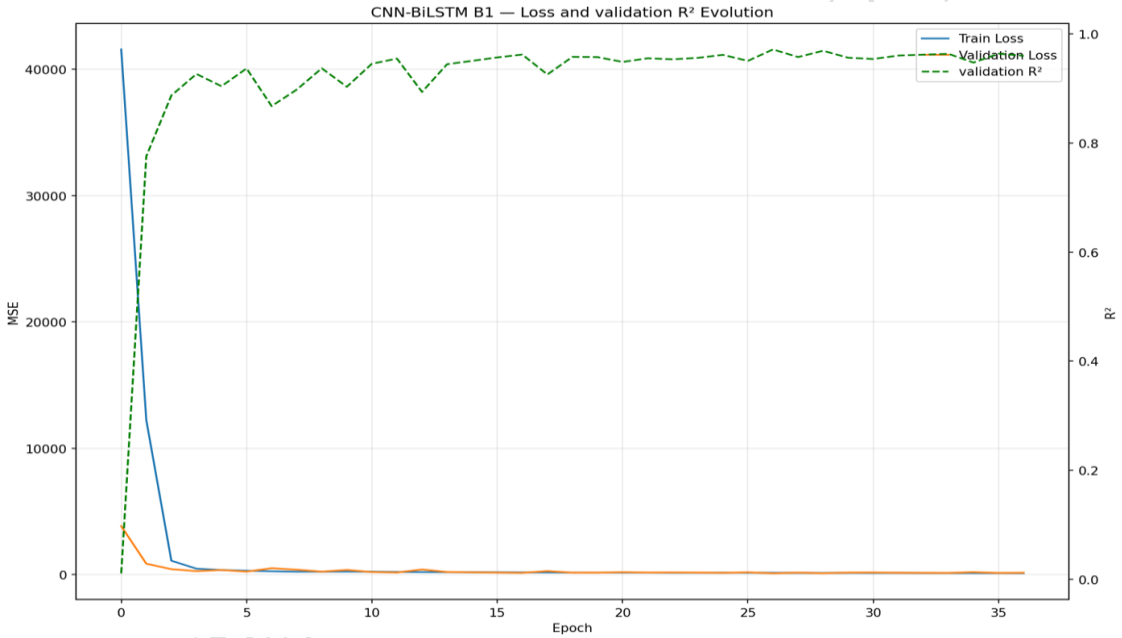

Figure 5:- Learning curve (training and validation loss) – Dataset 2

The curves in Figure 5 show a rapid decrease in training and validation losses during the first few epochs. The validation coefficient of determination (R^2) increases rapidly and remains high despite some fluctuations. These results indicate satisfactory convergence of the CNN-BiLSTM model.

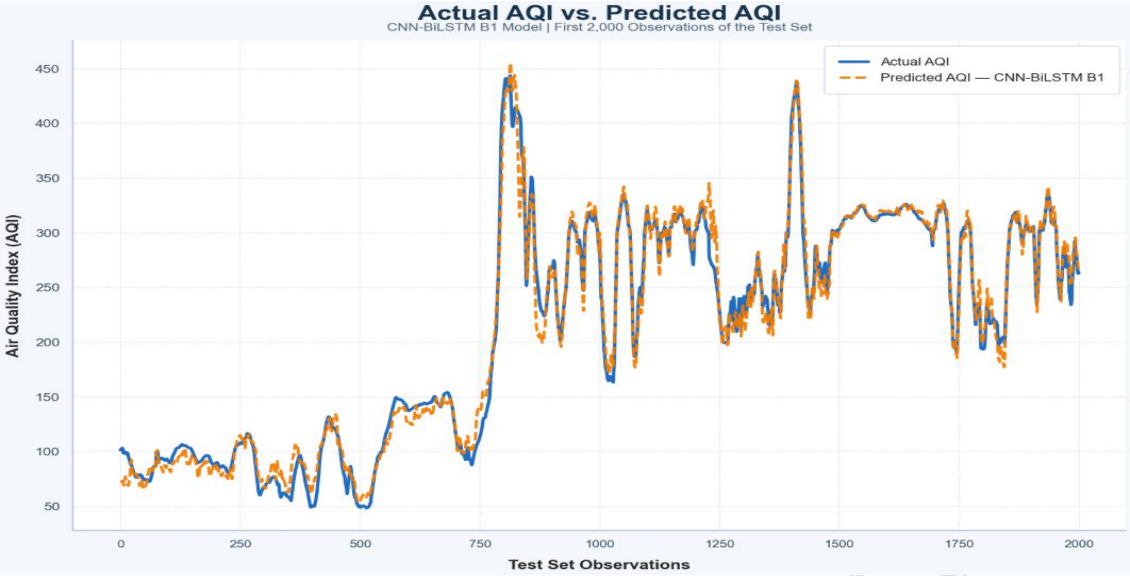

Figure 6:- Overlaid curves of actual and predicted AQI values – Dataset 2

As shown in Figure 6, the predicted values generally follow the actual AQI values, both at low levels and during pollution peaks. Some discrepancies occur during abrupt changes, but overall, the model reproduces the main variations in the index.

Comparison of Model Performance Across the Two Datasets:

Table 3:- Comparison of model performance on the two datasets

Model	Dataset	MSE	RMSE	MAE	R^2
Linear Regression	Dataset 1	619.4210	24.8881	16.8250	0.9329
	Dataset 2	839.1418	28.9679	22.1437	0.9319
SVM	Dataset 1	375.3024	19.3727	12.3213	0.9593
	Dataset 2	2180.1899	46.6925	33.1735	0.8232
Simple LSTM	Dataset 1	54.7693	7.4006	2.5810	0.9940
	Dataset 2	3131.0461	55.9557	34.9684	0.7462
CNN-BiLSTM	Dataset 1	53.2003	7.2938	2.6339	0.9942
	Dataset 2	300.7757	17.3428	10.9171	0.9756

Table 3 shows that the CNN-BiLSTM achieves the highest scores on both datasets, with an R^2 of 0.9942 on Dataset 1 and 0.9756 on Dataset 2. On the first dataset, the simple LSTM achieves very similar performance, with an R^2 of 0.9940. Linear regression and SVM obtain scores of 0.9329 and 0.9593, respectively.

On the second dataset, the performance gaps between the models are larger, although the same modeling protocol was applied to each of them. The CNN-BiLSTM achieves an R^2 of 0.9756, compared with 0.9319 for linear regression, 0.8232 for SVM and 0.7462 for the simple LSTM.

Discussion:-

The decrease in the simple LSTM's performance may be related to the characteristics of the second dataset. It covers a shorter period and comes from a single monitoring station, unlike the

UNDER PEER REVIEW

first dataset, which contains observations from multiple stations over several years. Consequently, the simple LSTM has less data available to learn the different variations in the AQI. However, these results do not establish that dataset size is the sole cause of this decline.

The SVM score also decreases substantially. Its performance may depend on the choice of hyperparameters and the way the data are partitioned. The relationships between pollutants and the AQI may also differ between datasets. These differences may partly explain the results, but further analysis would be needed to confirm this.

The CNN-BiLSTM achieves better results on the second dataset than the three baseline models. By combining CNN layers with a BiLSTM, the model can capture features in the input variables and their temporal relationships [18]. Despite a decrease in its score between the two datasets, it maintains a good level of predictive performance. These results show that model performance varies depending on the data used. The CNN-BiLSTM delivers the most consistent results across our two experiments, whereas the simple LSTM and SVM experience more pronounced performance declines on the second dataset.

Overall, the CNN-BiLSTM achieves the best performance among the compared models on both datasets. This performance may be explained by the complementary roles of the CNN, which extracts features from the input variables, and the BiLSTM, which captures temporal dependencies. These findings are consistent with previous studies on hybrid CNN-LSTM architectures and Explainable Artificial Intelligence (XAI) applied to AQI forecasting [19].

Limitations and Future Perspectives:

Nevertheless, the model has certain limitations, particularly its complexity, training time and the difficulty of interpreting its predictions. To address these limitations, temporal Transformers, such as Informer and Autoformer, could be investigated. Graph Neural Networks (GNNs) could also be used to account for spatial relationships between monitoring stations. In addition, SHAP and LIME methods could help examine the contribution of individual variables to predictions, while model compression could facilitate deployment on IoT devices.

Another future direction is to integrate real-time data from sensors, weather stations and satellites. This approach could enable faster detection of pollution peaks and the triggering of alerts. The ultimate goal would be to develop an air quality monitoring system suited to real-world operating conditions.

Conclusion:-

Air quality forecasting is an important issue in atmospheric pollution monitoring. This study proposes a hybrid CNN-BiLSTM model to predict the Air Quality Index (AQI) one hour ahead using observations from the preceding 24 hours. The model's performance is compared with that of three approaches: linear regression, SVM and a simple LSTM. The experiments are conducted on two datasets with different characteristics.

The results show that the CNN-BiLSTM achieves the best performance, with a coefficient of determination (R^2) of 0.9942 on the first dataset and 0.9756 on the second. Unlike the SVM and simple LSTM, whose performance declines substantially on the second dataset, the CNN-BiLSTM maintains a high level of predictive performance. These findings highlight the potential benefits of combining convolutional networks with bidirectional recurrent networks for short-

UNDER PEER REVIEW IN IJAR

term AQI prediction. However, further evaluations across additional monitoring stations will be necessary to better assess the model's ability to adapt to different environments.

References:-

1. Camfil. Understanding particulate matter: PM1 (booklet). Available: https://camfil.com/dfsmedia/baeefe6b74df44be8a2bccfb2c57af8e/410-source/bouklet-pm1-undertanding-particulate-mater-eng-us
2. Listyarini S., Warlina L., Sambas A. Air quality monitoring system in South Tangerang based on Arduino Uno: from analysis to implementation. Department of Environment, Universitas Terbuka, Indonesia. Available: https://doi:10.1088/1757-899X/1115/1/012046
3. World Health Organization. Air quality and health. Available: https://www.who.int/docs/default-source/documents/publications/air-quality-and-health.pdf
4. Ayturan Y.A., Öztürk A., Ayturan Z.C. (2017): Modelling of PM10 pollution in Karatay district of Konya with artificial neural networks. Journal of International Environmental Application and Science, 12(3): 256–263.
5. Kumar S.R., Arumugam T., Anandakumar C., Balakrishnan S., Rajavel D. (2013): Use of plant species in controlling environmental pollution – a review. Bulletin of Environment, Pharmacology and Life Sciences, 2(2): 52–63.
6. Debo-Saiye Y., Ajose-Ismail B.M. (2020): IoT based air quality detection and monitoring system. International Journal of Research and Innovation in Applied Science (IJRIAS), 5(12). Available: https://rsisinternational.org/virtual-library/papers/iot-based-air-quality-detection-and-monitoring-system/
7. World Health Organization (2018): Air pollution and child health: prescribing clean air. WHO, Geneva, Switzerland. Available: https://www.who.int/ceh/publications/AdvancecopyOct24_18150_Air-Pollution-and-Child-Healthmergedcompressed.pdf
8. United States Environmental Protection Agency (2018): Managing air quality – air pollutant types. Available: https://www.epa.gov/air-quality-management-process/managing-air-quality-air-pollutant-types
9. World Health Organization (2014): Burden of disease from ambient air pollution. WHO, Geneva, Switzerland.
10. Wright C.Y. (2012): Quick best practices guide to consider air-related human health in your area – for South African air quality officers and environmental health practitioners. CSIR Natural Resources and the Environment, Pretoria, South Africa.
11. UNICEF France. Available: https://www.unicef.fr/
12. Saminathan S., Malathy C. (2024): PM2.5 concentration estimation using Bi-LSTM with osprey optimization method. Nature Environment and Pollution Technology, 23(3): 1631–1638. https://doi.org/10.46488/NEPT.2024.v23i03.033
13. Dong J., Zhang Y., Hu J. (2024): Short-term air quality prediction based on EMD-transformer-BiLSTM. Scientific Reports, 14: 20513. https://doi.org/10.1038/s41598-024-67626-1
14. Wang X., Zhang S., Chen Y., He L., Ren Y., Zhang Z., Li J., Zhang S. (2024): Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD–GAT–BiLSTM. Scientific Reports, 14: 17841. https://doi.org/10.1038/s41598-024-68874-x

UNDER PEER REVIEW IN IJAR

15. Central Pollution Control Board (CPCB), India. Time Series Air Quality Data of India (2010–2023) [dataset, distributed on Kaggle]. Available: https://cpcb.nic.in

16. Agbor B.A., et al. (2025): Hybrid CNN–BiLSTM–DNN approach for detecting cybersecurity threats in IoT networks. Computers, 14(2): 58. https://doi.org/10.3390/computers14020058

17. Liu S., Hu Y. (2025): Air quality prediction based on factor analysis combined with Transformer and CNN-BILSTM-ATTENTION models. Scientific Reports, 15: 20014. https://doi.org/10.1038/s41598-025-03780-4

18. Liu X., Su K., Wang S., Ghazali K.H. (2025): Intelligent prediction of air quality index based on the transformer-BiLSTM model. Scientific Reports, 15: 41838. https://doi.org/10.1038/s41598-025-25865-w

19. Mohapatra E., Das M., Rath S. (2025): Deep learning-based AQI forecasting: a CNN-LSTM model with visual insights from SHAP-LIME and PDP. Discover Applied Sciences, 7: 1326. https://doi.org/10.1007/s42452-025-07845-x

20. OpenCity. Delhi hourly air quality reports [dataset]. Available: https://data.opencity.in/dataset/delhi-hourly-air-quality-reports/resource/f7d16626-68d5-4de3-b6e6-e6a1b8831db3

UNDER PEER REVIEW IN IJAI