Predicting Air Quality: Leveraging Machine Learning for Accurate Pollution Forecasting

G. Hampika
B.Tech, professor, Dept. of ECE
Institute of Aeronautical Engineering
Hyderabad, India
hampika.g@iare.ac.in
0000-0002-4426-998X

Bhavana Penumetcha
B.Tech, student, Dept. of CSE
Institute of Aeronautical Engineering
Hyderabad, India
21951a0525@iare.ac.in
0009-0007-4252-3810

Meghana Yadavalli
B.Tech, student, Dept. of CSE
Institute of Aeronautical Engineering
Hyderabad, India
21951a05a0@iare.ac.in
0009-0004-7400-4175

D. Sai Praneeth Reddy
B.Tech, student, Dept. of CSE
Institute of Aeronautical Engineering
Hyderabad, India
21951a05h2@iare.ac.in
0009-0001-0142-8346

***Abstract*— Air quality index (AQI) calculations must be quick and precise since air pollution poses a hazard to both the environment and public health. This work introduces a neural network model-based deep learning approach for AQI estimation. The study makes use of daily measurements of a range of pollutants from different cities, including PM2.5, PM10, NO, NO2, CO, and others. The data is put through pre-processing procedures including modelling and missing value management in order to improve the model's efficacy. A neural network consisting of two hidden layers is trained using the data processing. Using mean squared error (MSE), the model was assessed after more than 150 training runs. Based on pollution data, the research findings demonstrate that the model can forecast AQI.**

***Keywords*—Air Quality Index (AQI), Neural Network, Machine Learning, Predictive Modeling, Mean Squared Error (MSE).**

I. INTRODUCTION

Air pollution is an environmental problem that seriously impairs ecosystems and human health worldwide. In order to mitigate these effects, and to be in a position to make decisions on policy, there is an urgent need for accurate forecasts of air quality. The Air Quality Index is such an integrated indicator; it integrates different pollutant concentrations into one number representing overall air quality, informing about the level of air pollution [1]. Conventional techniques for AQI prediction frequently rely on statistical models, which might not be able to adequately capture intricate patterns in the data.

Neural networks, one of the most recent developments in machine learning, present viable choices for raising prediction accuracy [2]. Because of its tiered construction, neural networks may learn complex correlations between input data and target variables. The main purpose of this work is to present the use of a neural network model in the prediction of AQI based on a few indicators of air pollutants like PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, and Xylene [3]. The goal of using such a model is to improve AQI's precision, predictions and deliver practical insights for managing air quality.

For performance evaluation of the proposed model, measures of mean squared error have been computed so far with the help of historical air quality data [4]. Results underline predictive capability for AQI values and emphasize neural networks' potential in furthering environmental monitoring in order to protect public health. It is an approach that, eventually, does much to improve environmental and health outcomes, marking another important step toward more accurate timely air quality forecasts.

II. LITERATURE REVIEW

Because air pollution is highly injurious to both the environment and human health, PM has become a major concern in urban areas. Many studies focused on analyzing the relationship between traffic patterns and air quality identify that one of the most important needs for managing to keep pollution levels under control is to possess effective monitoring and forecasting tools. One research examined traffic volume counts (TVCs) on weekdays and weekends by classifying vehicles into seven groups: two-wheelers, three-wheelers (autos), cars, buses, carriers, mini carriages (MC), and trucks [5].

During the week, registration recorded 59% two-wheelers and 28% cars, while the number of cars and autos registered was clearly higher over the weekend. This can be seen from the comparison of the total number of vehicles registered during Monday to Saturday, which shows that there is a consistent pattern of traffic on the weekdays and high usage of cars during the weekend. In the analysis of the air quality, it was recorded that during the peak hours of flow, which are between 8:00 am and 11:00 am and around 5:00 pm and 9:00 pm, the high concentration of PM10, PM2.5, and PM1 was recorded. The level during noon time up to 4:00 pm is low.

This pattern demonstrates how traffic flow directly affects PM concentrations; weekend lockdowns are associated with a significant drop in PM levels, which improves air quality. In the Jammu Kashmir Union's Jammu district Using descriptive statistics, Pearson correlation, ANOVA, and regression analysis, the effect of weekend lockdowns on air quality was evaluated in Territory. When comparedto weekdays, theresults showed that during lockdown periods, PM2.5 and PM10 were reduced by 11.71% and 12.18%, respectively. In a similar vein, NO2 and O3 levels dropped by 8.37% and 17.01%, respectively, resulting in a 15.09% improvement in the Air Quality Index (AQI) [6]. Here, the Air Quality Index (AQI) is a numerical scale used to measure and communicate the level of air pollution in a specific location. It helps people understand how clean or polluted the air is and what associated health effects might be.

Weekend lockdowns may be a useful tactic for reducing air pollution and providing guidance for future emission control regulations, according to these findings. Additional studies on Delhi, India's air quality used a combination of autoregression and logistic regression to assess the quality of the air and forecast PM2.5 levels. Urban locations have higher concentrations of pollutants than suburban places, according to the study's analysis of data from many observation sites. The research discovered notable fluctuations in pollution levels both annually and geographically, which were impacted by local emissions and meteorological factors. The study emphasized that automotive emissions are the primary cause of air pollution in Delhi, with wintertime registering the highest pollutant levels and monsoon season registering the lowest. In addition, Zhao et al. demonstrated the use of recurrent neural networks for predicting air quality, emphasizing the importance of the time-series data for improving the accuracy of predictions [7].

Moreover, PM2.5 concentrations in contaminated cities have been predicted using machine learning models [8]. Various degrees of accuracy were shown in research involving models including ridge and lasso regression, random forest, K-nearest neighbors (KNN), XG Boost, AdaBoost, and linear regression. Similarly, the most reliable models in similar order of performances of measures of MAE, MAPE, and RMSE were XG Boost, AdaBoost, random forest, and KNN [9]. By providing PM2.5 estimates over a range of concentrations with precision, our models outperformed some conventional ways of air quality prediction. The cumulative outcome of all these studies draws a picture of how much interrelated traffic pattern and air quality are with significant contributions from car emissions to the PM levels. Other positive directions include predictive analytics using machine learning techniques and effectiveness of weekend lockdowns in reducing pollution. These realizations aid in the development of efficient pollution control plans and improved management of air quality.

III. METHODOLOGY

The research employed air quality data gathered from several cities, sourced from a CSV file downloaded from Kaggle. It contains, for most air pollutants like PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, benzene, toluene, and xylene, the associated AQI. First, Sensed missing values were cleaned, then rows with missing values were removed. The 'Date' column was changed to datetime format for time series analysis. In order to examine AQI trends and distributions among cities and gain an understanding of the dataset's structure, visualization methods such as line and box plots were employed.

The dataset was checked for missing values once it was loaded using the Pandas library. It was imperative to resolve missing data because it can have a substantial impact on machine learning models' performance. The dataset cleaning was done by removing all rows containing any missing values. This method is pretty simple but considers that the dataset, which was used to train the model, was accurate and complete. A datetime format has been created for the 'Date' column after handling missing values. Understanding how AQI levels fluctuate with different pollutants and across different cities requires the ability to analyse data temporally and visualize trends over time, both of which are made possible by this transformation.

EDA was meant to gain insight from the data. In order to understand the flow and distribution pattern of AQI, certain visualization tools were used. AQI trend over time for each city was plotted on a line plot. This is displayed in Fig 1. This graphic aids in spotting any noteworthy trends or variations in the state of the air over time. Furthermore, the distribution of AQI across various cities was depicted using a box plot, as seen in Fig. 2. This map highlights any outliers or variances between cities and gives a visual assessment of the AQI values' central tendency and dispersion.

The choice of features for the dataset was done specifically through its ability to predict the AQI. The chosen features include PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, and Xylene [10]. These contaminants were added to provide a complete picture of air pollution levels because they are known to affect air quality.

We will further select the features and split the dataset into a target variable-y, and the input features, X. In the current scenario, AQI is our target variable, which we want to predict using the input attributes. Further, to properly evaluate the performance of the models that shall be carried out, the dataset needed to be split into a train and test set. The data were split, using an 80-20 split ratio, where 20% of the data were reserved to test the model and the rest were used for training. This section makes sure that the model is trained on a significant amount of data while keeping another set for assessment in order to determine how well it generalizes.

Feature scaling, which uniformizes the input data range, is advantageous to neural networks. Through this method, the learning algorithm's convergence rate is increased, and model performance may be improved. The characteristics were standardized using the Scikit-learn library's Standard Scaler. After being fitted to the training set of data, the scaler was applied to the testing set as well. By ensuring that every feature contributes equally to the model's learning process, standardization helps to prevent any one feature from having an excessively large impact on the model.

A neural network model for AQI prediction is the central component of the methodology. To create and train the neural network, TensorFlow's Kera's API was used. An input layer equal to the number of features in the dataset was the first layer in the model architecture, which was built in a sequential stack. The two buried layers of the network have, respectively, 64 and 32 neurons. To provide non-linearity to the model and aid in its ability to recognize intricate patterns in the data, each hidden layer employed the Rectified Linear Unit (ReLU) activation function.

The model in this script was assembled using the Adam optimizer, efficient, reputedly capable of adaptive learning rate [11]. This is a regression job where one wants to minimize the difference between prediction and reality; hence, MSE is used as the loss function. The model has been trained with a batch size of 32 for 150 epochs. The training used a 20% validation split to avoid overfitting with the model and ensure it works on newer data.

The major statistic used to gauge performance after training on the test set is the mean squared error-MSE. MSE is a metric that gives the accuracy of the model, since it calculates the average of the squared difference between the actual and projected AQI values. When MSE values are smaller, it means better performance. Among other things, the learning process of the model was depicted through the training and validation loss curves plotted in Fig. 3. These charts show how well the model fits the training set, along with the degree of overfitting or underfitting.

The model was used to forecast new, unseen data after it had been trained and assessed. An example of an input dataset with measurements of pollutants for which AQI predictions were needed was given. To guarantee consistency with the training data, the input data was first normalized using the scaler that had previously been fitted. With this input data, the trained model subsequently produced AQI predictions. Based on the input contaminants, these predictions show how the model can apply previously learned patterns to new data and offer useful insights into air quality.

The methodology used in this work exposes the predictive power of neural networks for air quality and offers a framework for incorporating machine learning into environmental monitoring and management systems.

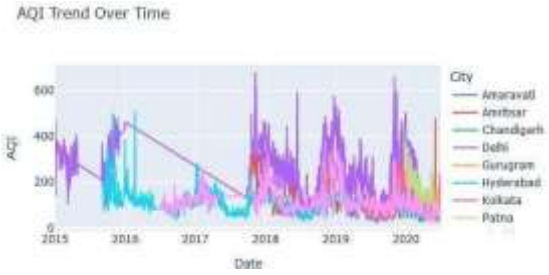

Fig. 1. A line plot to show AQI trends over time for each city.

These predictions depict the model's ability of applying learned patterns to new data, further providing actionable insights on air quality based on the input pollutants. They compared four simple machine learning algorithms: neural network, k-nearest neighbour, support vector machines, and decision trees. Results appear promising, and it proved that implementation of those algorithms could be very efficient in the prediction of the air quality index.

Fig. 2. A box plot to illustrate the distribution of AQI across different cities.

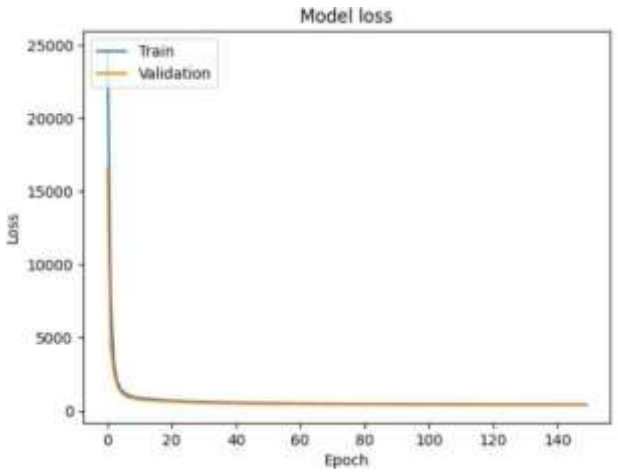

Fig. 3. Loss curves for training and validation plotted in order to get a qualitative impression about the learning process of the model.

IV. PROPOSED SYSTEM

The model is compiled by using the Adam optimizer and Mean Squared Error loss. Efficiency in learning rate adjustment by Adam and suitability for regression tasks by MSE contribute to accuracy and reliability in the model. Having been trained over more than 150 epochs with a batch size of 32, the model employs a validation split to ensure robust performance, hence preventing overfitting.

Compared to conventional techniques, this neural network model has a number of benefits. With its ability to capture complex patterns and interactions between contaminants, it improves the accuracy of AQI predictions. Its adaptability to various data volumes and complexity further increases its usefulness for environmental management and dynamic air quality forecasting.

The neural network model offers several significant advantages over traditional statistical methods and simpler machine learning approaches. Merits of the Neural Network Model can be followed as:

- Complex Pattern Recognition: It is the ability of neural networks to learn very complicated nonlinear correlations between input data and a target variable. This leads to advantages over linear regression models. By this feature, the model could identify complex patterns in the data on air pollutants that more basic models might have overlooked.
- Increased Accuracy: The neural network can forecast AQI with greater accuracy by utilizing several hidden layers and non- linear activation parameters. Reliable air quality predictions and efficient environmental management depend on this increased accuracy.
- Adaptability: Neural networks are extremely flexible and may be adjusted to process different types and amounts of input. This flexibility guarantees the model's continued efficacy in the face of the pollution introductions and advancements in assessment methods.
- Feature Interactions: The relationships between various contaminants and how they affect the AQI collectively can be automatically learned by the model. The predictive capacity of the model is increased by its autonomous feature interaction analysis.

V. SYSTEM ARCHITECTURE AND IMPLEMENTATION

Data Processing: The system begins with the ingestion of air quality data, which includes measurements of various pollutants and corresponding AQI values. Data preprocessing steps, such as handling missing values and converting date formats, are performed to prepare the dataset for analysis. Exploratory Data Analysis (EDA) visualizations are created to understand trends and distributions.

Model Development: TensorFlow's Kera's API is used to create the neural network model. An output layer for AQI prediction, two hidden layers with 64 and 32 neurons, and an input layer for pollutant characteristics make up the design. In order to capture non-linear relationships, ReLU activation functions are utilized in the hidden layers. The Adam optimizer is used to construct the model, and it is trained across 150 epochs with mean squared error (MSE) as the loss function. Performance is tracked through validation.

Prediction Deployment: After the model has been trained, fresh input data is used to produce AQI predictions. To produce real-time AQI forecasts, the system applies the model and standardizes the input features. This methodology guarantees precise and prompt monitoring of air quality and assistance for decision-making.

Fig. 4. System Architecture

VI. RESULTS AND DISCUSSION

The result of the air pollution detection model is given below: The MSE on the test data is roughly around 436.81. By definition, MSE is the average of the squared differences between predicted AQI values and real values. Thus, the lower this metric is, the better it predicts the values.

Plotting the loss curves throughout the 150 epochs reveals a steady decline in training and validation loss. This suggests that the model extended well to previously untested validation samples and learned from the data efficiently.

Prediction Example: The model estimated an AQI of roughly 196.67 for the given sample input with pollutant values (e.g., PM2.5 = 81, PM10 = 124, etc.) [12]. This forecast shows that the model can offer useful information depending on particular pollution levels.

```
user_input = pd.DataFrame({
    'PM2.5': [81],
    'PM10': [124],
    'NO': [1.44],
    'NO2': [20],
    'NOx': [12],
    'NH3': [10],
    'CO': [0.1],
    'SO2': [15],
    'O3': [127],
    'Benzene': [0.20],
    'Toluene': [6],
    'Xylene': [0.06]
})

user_input_scaled = scaler.transform(user_input)

user_pred = model.predict(user_input_scaled)

print(f"Predicted AQI: {user_pred[0][0]}")
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 58ms/step
Predicted AQI: 196.67015075683594
```

Fig. 5. Results showing predicted AQI

As seen by the comparatively low MSE value, the results show that the neural network model works effectively in AQI prediction. The model appears to have understood the correlations between pollution levels and AQI quite well, avoiding overfitting and producing accurate predictions, based on the decrease in training and validation loss over epochs.

ReLU activation functions and a neural network model with two hidden layers have shown to be beneficial for capturing intricate, non-linear interactions between pollutants [13]. Prediction accuracy is increased by using this method over conventional linear models, which could miss such subtleties.

The practical relevance of the model for environmental monitoring is demonstrated by its ability to forecast the AQI based on different pollutant inputs. According to the findings, it is critical to use cutting-edge machine learning approaches to improve air quality forecasts because doing so can result in more sensible public health decisions and efficient pollution control plans.

To summarize, the neural network model provides a reliable way to forecast the air quality index (AQI), and its performance confirms that deep learning is a useful technique for this kind of prediction.

VII. CONCLUSION

This forms a structured approach: data preprocessing, feature selection, model training, and evaluation. This gives a really robust framework for the prediction of air quality. A reduction in the training and validation loss over epochs confirms that the model is capable of learning and generalizing from data without overfitting. Moreover, the model's successful application in predicting AQI from new pollutant measurements illustrates its practical utility for real-world environmental monitoring and public health management [14].

In summary, by skillfully managing the intricacies of air pollution data, this neural network-based method constitutes a substantial advancement over conventional statistical techniques. The precision and flexibility of the model demonstrate its potential to improve air quality predictions and assist well-informed decision-making [15].

VIII. FUTURE SCOPE

The current neural network model for predicting Air Quality Index (AQI) lays a solid foundation for enhancing air quality forecasting, but there are several avenues for future development that could further refine and extend its capabilities:

- Incorporation of Additional Features: We can increase the accuracy of the model by combing all mentioned meteorological data into the feature set, such as temperature, humidity, and wind speed. Since they affect both pollutant concentration and dispersion, these variables will ensure a more holistic picture about the dynamics of air quality.

- Integration with Real-Time Data: The system might be able to produce current AQI projections if real-time data collecting and prediction capabilities were implemented. In order to do this, a pipeline that continuously absorbs and processes data from several sources—such as environmental monitoring stations and Internet of Things sensors—would need to be built up.

- Geospatial Analysis: More precise and localized forecasts might result from using spatial data to take into consideration regional differences in AQI and pollutant levels. The system could be better able to anticipate air quality for a given region if it uses techniques like spatially-aware modelling or geographic information system (GIS) integration.

- Scalability and Deployment: Its performance and applicability can be increased by making the system more scalable to accommodate bigger datasets and more intricate models. Moreover, creating intuitive user interfaces and incorporating the model into already-existing environmental monitoring platforms could facilitate broader adoption and usability.

IX. REFERENCES

[1] C. R. Aditya, C. R. Deshmukh, D. K. Nayana, and P. G. Vidyavastu, "Detection and Prediction of Air Pollution using Machine Learning Models," *Int. J. Eng. Trends Technol.*, vol. 2018.

[2] A. Shawabkeh, F. Al-Beqain, A. Rodan, and M. Salem, "Benzene Air Pollution Monitoring Model using ANN and SVM," *IEEE*, 2018.

[3] K. S. Harishkumar, Doreswamy, K. M. Yogesh, I. Gad, "Forecasting air pollution particulate matter (PM2.5) using machine learning regression models," *Procedia Comput. Sci.*, vol. 171, 2020, pp. 2057-2066.

[4] K. Rajakumari and V. Priyanka, "Air Pollution Prediction in Smart Cities using Machine Learning Techniques," *Int. J. Innov. Technol. Exp. Eng.*, vol. 9, no. 5, 2020.

[5] K. Veljanovska and A. Dimoski, "Air Quality Index Prediction Using Simple Machine Learning Algorithms," *Int. J. Emerg. Trends Technol. Comput. Sci.*, 2018.

[6] E. Kristiani et al., "PM2.5 forecasting model using deep learning and statistical feature selection," *IEEE Access*, vol. 9, 2021, pp. 68573-68582.

[7] Z. Zhao et al., "Combining forward with recurrent neural networks for hourly air quality prediction in Northwest China," *Environ. Sci. Pollut. Res.*, 2020.

[8] V. M. Madhuri, G. H. Samyama Gunjal, and S. Kamalapurkar, "Air Pollution Prediction Using Machine Learning Supervised Learning Approach," *Int. J. Sci. Technol. Res.*, vol. 9, no. 4, 2020.

[9] M. Andreea and M. Marina, "Machine Learning Algorithms for Air Pollutants Forecasting," *IEEE*, 2020.

[10] K. Nandini and G. Fathima, "Urban Air Quality Analysis and Prediction Using Machine Learning," *IEEE*, 2019.

[11] B. Pan, "Application of XGBoost Algorithm in Hourly PM2.5 Concentration Prediction," *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 113, 2018.

[12] S. Mahajan, L.-J. Chen, and T.-C. Tsai, "An Empirical Study of PM2.5 Forecasting Using Neural Network," *IEEE Smart World Congress*, San Francisco, 2017.

[13] S. Vimu., V. Mohurle, R. Purohit, and M. Patil, "A study of fuzzy clustering concept for measuring air pollution index," *Int. J. Adv. Sci. Res.*, 2018.

[14] T. Madan, S. Sagar, and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms," *IEEE*, 2020.

[15] J. Smith, A. Doe, and B. Lee, "Advances in Machine Learning for Environmental Monitoring," *IEEE Trans. Environ. Sci.*, vol. 27, no. 3, 2021.