Air Quality Prediction using Deep Learning -A Review

Ayush Kumar Gupta
Computer Science and Engineering
Lovely Professional University
India
ayush.acclko123@gmail.com

Naval Chauhan
Computer Science and Engineering
Lovely Professional University
India
navallpu16@gmail.com

Tanima Thakur
Computer Science and Engineering
Lovely Professional University
India
tanima.23532@lpu.co.in

Abstract: Air pollution is a much greater threat known till date and the necessary measures must be taken by the government to anticipate the rising effects of air-quality on human health. The primary contributors to air pollution, which is measured by the air-quality-index, include pollutants including carbon dioxide, nitrogen dioxide, and carbon monoxide. Air pollution may cause fatal conditions including lung cancer, neurological disorders, and the greenhouse effect in addition to contributing to global warming. Forecasting air quality is essential for reducing air pollution. Although there is much research being done on this topic, ML and Deep Learning approaches are used to predict the air quality index. We have examined and contrasted the study findings of several writers who have worked on the topic in this paper. Comparing several models to Deep Learning models, it was shown that the Deep Learning models perform better when it comes to predicting the PM2.5 pollutant.

Keywords: AQI (Air Quality Index), AI, Air quality prediction, Deep learning, Machine learning.

INTRODUCTION

Air pollution is becoming a bigger problem, making it difficult to assess and forecast effectively. Natural pollution is divided into two categories: NO2, CO,CO2, SO2 and sulphate released by volcanic eruptions and forest fires, and man-made pollution caused by human activities, industrial production processes, burning of fossil fuels, and transportation emissions.[1] Environmental pollution has grown significantly due to industrialisation and the accompanying increase in haze, with one of the harmful pollutants being PM2.5[2].Air Quality Index is a technique for assessing air quality[3]. Technology improvements have made it easier to collect data using sensors, and the accuracy of the measurement depends on machine learning methods such as neural networks. The AQI is a critical first step for mitigating air pollution, and the accuracy of the measurement depends on machine learning methods such as ML and DL algorithms. [4] In this review paper we reviewed the various research done across the globe in the field of ML and DL to predict the AQI to overcome the issues related to lung cancer, asthma etc. Our aim is to review the different methodologies used by the authors, to find the better technology that can be implemented to find the better outcome. Deep Learning is a subset of ML, developed by the Google Brain project in 2012. It has three benefits: nonlinearity, generativity, and cross modularity. It has been used in recognition applications such as picture categorization and video understanding and can be improved by adding more layers. [5]. Machine learning is the foundation of artificial intelligence, as it takes data and algorithms as input and identifies the method used in the conventional system before allowing the computer to remember the algorithm. As time goes on, the computer gets smarter and can handle more issues. [5] Since AirQuality analysis and forecasting is a time-based process, so we need to implement to time-series analysis that is used in ML. Time series analysis forecasts models, which analyse the characteristics of the data points over some time intervals to forecast them in the future.

The rest of the paper is organised in the following way: Section 2 covers the literature review of the articles that were used to write this paper, Section 3 consists the research findings and Section 4 is the conclusion of the paper.

Literature Review

In [8], Patai et al. examined a variety of soft computing methodologies, with an emphasis on ANN approaches. Their suggested the soft computing approached model predicts the classification of various environmental air contaminants, producing the most optimal outcomes. In [9], Mokhtari et. al., the outcomes showed that the deep learning model they proposed outperformed cutting-edge methods using both machine and deep learning techniques. In [10], the study conducted by the authors Le et al., a model called ("Convolutional Long Short-Term Memory") Conv-LSTM, can automatically change data's temporal and spatial features. Their work showed that the recommended model outperform CNN and LSTM as well as comparable state of the art models. In [11], the study by Du et al. developed DAQFF an end-to-end model that uses a hybrid deep learning technique. The proposed model's enhanced prediction abilities were shown to outperform both standard shallow learning and baseline deep learning models. In [12], the study the authors Yi et al., have concluded the limitations of traditional air-quality prediction models, including the use of limited data sources and insufficient consideration of complex environmental factors. They reviewed models like recurrent neural networks, convolutional neural networks and hybrid models. In [13], Garg et al., the authors used four to calculate the PM2.5 levels at twelve locations in Beijing by using historical , meteorological and forecasting data, resulted that LSTM beat all other models. In second research, LSTM prediction model outperformed the DAE model, according to their comparison of the two models. In [2], the authors Bekkar et al., predicted Beijing, China's hourly PM2.5 concentration using a CNN-LSTM model. This model included a spatial-temporal characteristic by fusing historical pollutant and meteorological data with PM2.5 concentration from neighboring sites. The experimental results showed that the suggested "hybrid CNN-LSTM multivariate" model outperformed the traditional models in terms of predicting accuracy. In [14], the study conducted by Li et al., the spatial influencing component of data collected by remote sensing was examined by employing a deep CNN. Their research shows that spatially or remotely sensed data's spatial effect characteristic can be fully used by deep CNN technology. In [15], the study by Rijal et al., the three well-known Inception-v3, Resnet50 and CNNs—VGG-16 were suggested by the study's authors to be used as base learners in ensemble of DNN to predict PM2.5 concentrations from images. The results showed that the ensemble performed better than the individual DL networks.

In [16], the study by Pasupuleti et al., after contrasting LR, RF and decision tree models, the author drawn a conclusion that overfitting reduces errors, therefore Random Forest produces more precise results. However, Random Forest is costly and requires more RAM. In [17], the authors Jebamalar et al., used a combined light gradient and light tree boosting model. The hybrid model is found to be most successful in identifying PM 2.5, since boosting is sequential ensemble and corrects the errors. They also discovered that while it takes lesser space and can process enormous quantities of data, takes longer time. In [18], the authors of the research Jing H et al., employed XG Boost to forecast the AQI. This method combines weak classifiers, addressing the shortcomings of prior ones, to construct a powerful classifier, which reduces the discrepancy between actual and predicted values. In [19], the authors Wang S. et al. employed GRNN(Gas Recurrent Neural Networks) in comparison to SVR and MLP to anticipate the degree of air quality. Since sensor drift is less steady, GRNN perform better, but they are more sensitive to ambient humidity and fluctuation. In [1], the authors used 6 different ML models to predict the PM 2.5 pollutant by analysing historical datasets from the meteorological and PM2.5 pollution domains. They found that the suggested models KNN, Xgb, RF and Adab, are trustworthy models when compared to the LR model and RL model for forecasting PM2.5 pollution. In [20], the authors, Murukonda et al., concluded that algorithms including LASSO Regression, Logistic Regression, SVR and Ridge Regression are chosen. Ridge Regression was shown to have the greatest performance in predicting the AQI. It has the highest R Square and the lowest MAE and RMSE. In [21], Zheng H et al., said according to tests, which show a conclusion that model outperforms individual models, especially when stacking probability distributions and developing novel features. They evaluated the individual and hybrid versions of these models using these models, and they came to the conclusion that Catboost outperforms other ensemble approaches and is the best model all around. The strategy with the best performance is "B+N-PD+O-Cat," whereas LSTM and Adaboost models have the lowest results.

In [22] the authors Drewil et al., solved the issue of choosing the proper hyperparameters for the LSTM model, the author introduced model based on Genetic Algorithm approach and the LSTM deep learning algorithm. In this work, the performance of the LSTM model was improved using the GeneticAlgorithm (GA). The experiment results prove that their ED-LSTM model, when compared to cutting-edge PM2.5 prediction models, can raise the MAE by up to 53.7%. Additionally, their suggested strategy, which incorporates the GA-based feature selection technique, increases prediction accuracy by at least 13.7%. In [23], the author Dairi et al., showed the efficacy of IMDA-VAE(Integrated Multiple Directed Attention Variational Auto Encoder) techniques to forecast multiple pollutants in different locations is finally illustrated through a discussion of the data obtained, according to a model described by the author. Results also demonstrate that IMDA-VAE the proposed model effectively improves air quality forecasting -

performance and it provides more accuracy in forecasting the concentrations of four main pollutants when compare to uni-directional and bi-directional RNN. In [24], the authors, Al-Janabi et al., proposed a method that showed its effectiveness in handling time-series and large data. Application of optimisation techniques—PSO algorithms are used in this research—is the best method for obtaining the appropriate hyperparameter. In [25], the authors Huang et al., the author of this study proposed a method for forecasting AQI. It shows how the "PSO" algorithm and the BP neural network function together. The simulation results demonstrate how the improved "PSO-BP" algorithm better optimises the learning capabilities of the BP neural network.

Table 1: Review of various techniques

S.No	Paper No.	Objective	Technique/Tool Used	Dataset Name	Research Findings
1.	[1]	Compare different ML technique to predict better air quality.	Adab, RL, RF, LR, KNN and XGB models	For two years, meteorological data were gathered every hour of the day.	Adab, KNN, XGB and RF are superior models compared to current models.
2.	[2]	Predict air quality using deep learning technique.	Bi-LSTM, Bi-GRU, LSTM,GRU,CNN, CNN-LSTM model	UCI Machine Learning Repository	The proposed "hybrid CNN-LSTM multivariate" model outperformed traditional models.
3.	[6]	To verify soft computing methodology for AQI prediction	Linear Regression, ANN	Data collected from a national institute	Grouping air toxins into distinct categories provides improved outcomes.
4.	[7]	To predict the air quality by multi-point deep learning model	Conv-LSTM-based spatiotemporal deep learning model	Fusion Field Trial 2007 (FFT07)	Quantile regression outperforms MC dropout at low coverage intervals.
5.	[8]	To explain how to convert air pollution data into picture sequences that use the ConvLSTM model to interpolate and forecast air quality	Deep Air Learning Model with CovoLSTM	Weather information(Meteorological) from 28 observation stations	The proposed model outperforms LSTM and CNN in the study of spatial and temporal characteristics, surpassing contemporary methods.
6.	[9]	To introduce a novel model named DAQFF (Deep Air Quality Forecasting Framework) to predict PM 2.5	1D-CNN, Bi-LSTM	Beijing air quality data from the UCI Dataset was gathered for the Microsoft Research Urban Air project.	Experiments demonstrate how the suggested strategy works better than both conventional deep learning and shallow learning techniques.
7.	[10]	To build a DNN based technique to forecast air quality	1D-CNN, Bi-LSTM	Every hour, the system gathered information on the quality of the air in 302 Chinese cities official air quality monitoring stations.	They created two fusion-networks for long-term and short-term forecasting.
8.	[11]	Using the ARIMA, FBProphet, LSTM, and 1D-CNN models,	ARIMA, FBProphet, and LSTM, 1D-CNN	Beijing Multi-Site AirQuality, taken from the	LSTM performed better than ARIMA, FBProphet, LSTM and CNN and LSTM

		investigate the PM2.5 levels at 12 sites.		UCI Machine Learning Repository	performs better with relu activation fn. than tanh
9.	[12]	In order to do research on remotely sensed data using deep CNN.	Deep CNN (Deep convolutional neural network model)	Terrain information, LULC, and remotely sensed PM2.5 concentration, as well as population and GDP geographic distribution densities.	Terrain, LULC, GDP spatial density, and population spatial density may nearly all be used to predict the geographic pattern of PM2.5.
10.	[13]	Using outside photos and a suggested ensemble DNN(deep neural network) based regression, for estimating the concentrations of PM2.5.	ResNet50, Inception-v3 and VGG-16	Randomly collected data from Beijing tourist website	The PM2.5 forecasts from base learners may be effectively combined by the meta trainer.
11.	[14]	To Predict air quality using ML	Decision-Tree, Linear-Regression and Random-Forest	Random meteorological data	Compared to Decision Tree and Linear Regression, Random Forest performs better.
12.	[15]	Using a hybrid regression model, forecast the air quality	LightGDM and DecisionTree	From February 2017 to January 2019, PM 2.5 dataset was recorded by SDS011 Sensor and a Raspberry-Pi.	Compared to the separate models utilised to build the hybrid, the hybrid model performs better.
13.	[16]	To Research on Air Quality using ensemble learning of XGBoost	XGBoost	China's nationwide real-time urban air quality information platform	The XGBoost ensemble method significantly improves the accuracy, rate of inaccuracy, and interpretability of predictions.
14.	[17]	To predict gas concentration using gated RNN	Support vector regression, Gated RNN, Multi-layer Perception	Gas sensor Dataset	GRU model performs better than SVR and MLP models
15.	[18]	To Build a model using machine learning to predict AQI	Ridge regression, LASSO regression and SVR	Collected data set from the CPCB's official website (https://cpcb.nic.in/)	The performance of LASSO and Ridge Regression in AQI prediction is superior.
16.	[19]	To look into how ensemble approaches perform	SVM, ARIMA, LR, LSTM, ERT, Adaboost, RF, GBM, Catboost, XGBoost and LightGBM	Five years' worth of hourly PM2.5 data and meteorological data from the UCI ML repository China and two years' worth of hourly meteorological data from the Sha Tin air quality monitoring station in Hong Kong.	The strategy with the best performance is "B+N-PD+O-Cat," whereas LSTM and Adaboost models have the lowest results.

17.	[20]	To identify the ideal LSTM hyperparameters	Genetic Algorithm, LSTM	Past data from Kaggle	Air pollution predictions based on the metaheuristic principle are superior to methods that manually estimate parameters.
18.	[21]	To build a DL architecture named as IMDA-VAE (Integrated Multiple Directed Attention Variational Autoencoder) to forecast air quality.	GRU-A,VAE, LSTM-A, Gated GRUs, ConvLSTM, LSTM, BiLSTM, and BiGRU	The US Environmental Protection Agency gathers data.	More accurate forecasting is offered by the suggested IMDA-VAE model than by VAE, BiLSTM, LSTM-A, ConvLSTM, Gated GRUs, LSTM, BiGRU, and GRU-A.
19.	[22]	To create a smart prediction for the levels of air pollution during the following two days.	RNN, LSTM, PSO	Data from the 2018 KDD Cup is utilized, and the 35 stations' names are included.	They developed a brand-new technique called the Smart Sir Quality Prediction Model. (SAQPM)
20.	[23]	Utilizing a PSO-BP neural network with increased prediction capabilities	Particle Swarm Optimisation, Back Propagation	China's platform for tracking and analysing air quality	BP-PSO achieve a superior search ability as compared to a conventional Back Propagation neural network.

Research Findings

The literature review suggests that machine learning and deep learning are more accurate than traditional methods for forecasting air quality. Combining models like CNN-LSTM, ANN, LR, or Bi-LSTM-CNN can be used to build more precise models based on historical data. CNNs are effective in identifying geographical connections and patterns, while LSTMs are good at identifying temporal connections and patterns. Deep Learning algorithms can learn complex relationships and patterns from large datasets, making them a promising approach for air-quality prediction.

Conclusion

The increasing problem of air pollution has made it difficult to accurately estimate and forecast air quality. Various methods have been employed, with artificial intelligence being the most viable option. This review paper has investigated multiple studies conducted on the subject of air quality prediction. Deep learning models possess great potential for air quality prediction due to their ability to handle complex and high-dimensional data. Machine learning models are the most often used technology, but deep learning and ensemble learning algorithms may yield better results depending on the type of data used.

References

[1]Kothandaraman, D., Praveena, N., Varadaraj kumar, K., Madhav Rao, B., Dhabliya, D., Satla, S., & Abera, W. (2022). Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorption Science & Technology, 2022.

[2]Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in smart city, deep learning approach. Journal of big Data, 8(1), 1-21.

[3]Madhuri, V. M., Samyama, G. G., & Kamalapurkar, S. (2020). Air pollution prediction using machine learning supervised learning approach. Int J Sci Technol Res, 9(4), 118-123.

[4]Madan, T., Sagar, S., & Virmani, D. (2020, December). Air quality prediction using machine learning algorithms–a review. In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 140-145). IEEE.

[5]Guo, J. (2022, February). Research on Artificial Intelligence: Deep Learning to Identify Plant Species. In *2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)* (pp. 59-66). IEEE.

[6]Zhang, B., Zhu, J., & Su, H. (2023). Toward the third generation artificial intelligence. *Science China Information Sciences*, *66*(2), 1-19.

[7]Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X., & Wang, Z. (2020). Deep learning for air quality forecasts: a review. *Current Pollution Reports*, *6*, 399-409.

[8]Patni, J. C., & Sharma, H. K. (2019, April). Air quality prediction using artificial neural networks. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 568-572). IEEE.

[9]Mokhtari, I., Bechkit, W., Rivano, H., & Yaici, M. R. (2021). Uncertainty-aware deep learning architectures for highly dynamic air quality prediction. *IEEE Access*, *9*, 14765-14778.

[10]Le, V. D., Bui, T. C., & Cha, S. K. (2020, February). Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In *2020 IEEE international conference on big data and smart computing (BigComp)* (pp. 55-62). IEEE.

[11]Du, S., Li, T., Yang, Y., & Horng, S. J. (2019). Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2412-2424.

[12]Yi, X., Duan, Z., Li, R., Zhang, J., Li, T., & Zheng, Y. (2020). Predicting fine-grained air quality based on deep neural networks. *IEEE Transactions on Big Data*, *8*(5), 1326-1339.

[13]Garg, S., & Jindal, H. (2021, April). Evaluation of time series forecasting models for estimation of pm2. 5 levels in air. In *2021 6th International Conference for Convergence in Technology (I2CT)* (pp. 1-8). IEEE.

[14]Li, J., Jin, M., & Li, H. (2019). Exploring spatial influence of remotely sensed PM2. 5 concentration using a developed deep convolutional neural network model. *International Journal of Environmental Research and Public Health*, *16*(3), 454.

[15]Rijal, N., Gutta, R. T., Cao, T., Lin, J., Bo, Q., & Zhang, J. (2018, June). Ensemble of deep neural networks for estimating particulate matter from images. In *2018 IEEE 3rd international conference on image, Vision and Computing (ICIVC)* (pp. 733-738). IEEE.

[16]Pasupuleti, V. R., Kalyan, P., & Reddy, H. K. (2020, March). Air quality prediction of data log by machine learning. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1395-1399). IEEE.

[17]Jebamalar, J. A., & Kumar, A. S. (2019). PM2. 5 prediction using machine learning hybrid model for smart health. *International Journal of Engineering and Advanced Technology (IJEAT)*.

[18]Jing, H., & Wang, Y. (2020). Research on Urban Air Quality Prediction Based on Ensemble Learning of XGBoost. In *E3S Web of Conferences* (Vol. 165, p. 02014). EDP Sciences.

[19]Wang, S., Hu, Y., Burgués, J., Marco, S., & Liu, S. C. (2020, August). Prediction of gas concentration using gated recurrent neural networks. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (pp. 178-182). IEEE.

[20]Murukonda, V. S. N. M., & Gogineni, A. C. (2022). Prediction of Air Quality Index Using Supervised Machine Learning.

[21]Zheng, H., Cheng, Y., & Li, H. (2020). Investigation of model ensemble for fine-grained air quality prediction. *China Communications*, *17*(7), 207-223.

[22]Drewil, G. I., & Al-Bahadili, R. J. (2022). Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement: Sensors*, *24*, 100546.

[23]Dairi, A., Harrou, F., Khadraoui, S., & Sun, Y. (2021). Integrated multiple directed attention-based deep learning for improved air pollution forecasting. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1-15.

[24]Al-Janabi, S., Mohammad, M., & Al-Sultan, A. (2020). A new method for prediction of air pollution based on intelligent computation. *Soft Computing*, *24*(1), 661-680.

[25]Huang, Y., Xiang, Y., Zhao, R., & Cheng, Z. (2020). Air quality prediction using improved PSO-BP neural network. *Ieee Access*, *8*, 99346-99353.