GMR
The Original

A Comprehensive Review Of Deep Learning Models For Predicting Air Quality

Mr. Kshirsagar Sopan Bapu[1], Dr. Rais Abdul Hamid Khan[2]

[1]PhD scholar School of Computer Science and Engineering Sandip University. Nashik, India
[2]Professor, (CSE) SOCSE, Sandip University, Nashik, India
Email id, sopankshirsagar02@gmail.com[1],rais.khan@sandipuniversity.edu.in[2]

ABSTRACT
Air pollution affects the sustainability of the ecosystem and health of the population, requiring accurate and reliable air quality forecasting models. Non-linear patterns within air pollution data sets are often challenging for conventional models, including machine learning models, to tackle. Owing to their capability of learning higher-level representations without manual intervention and mimicking long-range dependencies without requiring manual formulation, deep learning models are progressively gaining popularity in air quality forecasting models. Focusing on air pollutants that play significant roles in air quality forecasting models, including PM2.5, PM10, NO_2, SO_2, CO, O_3, this paper presents a comprehensive review of air quality models developed on deep learning models. Detailed descriptions of convolutional neural networks, Additionally included are recurrent neural networks, auto encoders, hybrid models, attention models, long short-term memory networks, and gated recurrent unit networks. Areas such as data sources, preprocessing strategies, and different performance estimation metrics are also presented. Along with this, some of the significant themes, such as missing data, interpretability, and sensor variability, are also addressed. Moreover, some aspects of transfer learning, spatiotemporal learning, and real-time air quality forecasting are mentioned, as well as many themes that must be considered in the context of deep learning algorithms for air quality forecasting. For researchers and individuals who are looking for ways in which accurate air quality prediction models can be created using deep learning approaches, this paper is presented.

(Keywords: Air Quality Prediction, Deep Learning Models, Spatial-Temporal Modeling, Environmental Monitoring)

1. INTRODUCTION

Air pollution has long been a major global concern due to its detrimental effects on sustainability, public health, and economic growth. Air pollutants with heightened concentrations consist of particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3), largely driven by urban development, industrial activities, and increasing transportation needs.

There is a clear link between air pollution and disease, for example respiratory and heart problems, as well as other consequences of climate change – making tools to measure air quality and predictions about the impact on our climate particularly important. These airborne contaminants can have a severe impact on the quality of air, and through that on the climate, which is why combating it has emerged as an urgent global problem for many federal governments.

This growing availability of air quality data sets using data-driven approaches has fueled the advancement of air quality prediction tasks utilizing deep learning. This arises from the ability of deep learning algorithms to automatically learn hierarchical feature representations for high-dimensional spaces. As far as modeling complex temporal dynamics and spatial relationships, which are essential in air quality models, deep learning algorithms perform admirably.

Various architectures have been developed for air quality prediction in deep learning during the past ten years. These architectures include Gated Recurrent Units, which were used for computing efficiency, and recurrent neural networks, long short-term memory, and convolutional neural networks, which were used for capturing sequences, among others.

The current review studies mostly include narrative summaries or focus on specific model types, pollutants, and regions. Accordingly, a systematic review is needed to reveal methodological tendencies, research gaps, and challenges within this rapidly growing field, analyzing, comparing, and classifying existing studies.

In order to identify, select, and analyze relevant literature pertaining to the topic from reputable scientific sources, the paper uses an organized approach. The deep learning models, sources, techniques for feature extraction, prediction intervals, performance measurement techniques, and results obtained from the shortlisted literature for the study are also discussed. The systematic review paper also briefs the readers about new approaches for conducting research related to air quality prediction models and discusses relevant topics such as the problem of data locality, uncertainty estimates from sensors, interpretability, and scalability. The systematic review paper aims to assist researchers and experts in developing effective air quality prediction systems using deep learning-based models.

AIM AND OBJECTIVES

AIM

The primary objective of the current study is to conduct a comprehensive analysis of deep learning techniques for air quality prediction, mainly concentrating on their efficiency, challenges, and possible uses for enhancing the accuracy of forecasting for effective environmental management.

OBJECTIVES

1. To examine current deep learning architectures, such as CNN, RNN, LSTM, GNN, and others, for air quality forecasting.
2. To analyze the ways in which feature extraction techniques and spatial-temporal data modeling help achieve good prediction results.
3. To explore preprocessing methods of data, including normalization, handling missing information, and combining information from different sources.
4. To identify variables which influence the success of forecasting, as well as model accuracy in terms of common accuracy measures.
5. To examine the case studies and applications of how deep learning can be employed for air quality forecasting.
6. In order to address the current problems in deploying models like scalability of deployment models or lack of interpretations of the models and
7. To propose possible areas of research to advance forecast systems of air quality using deep learning.

2. RELATED WORK

Air quality prediction has evolved from traditional statistical methods to complex deep learning techniques. Early research mainly focused on statistical time-series models such as ARIMA [25] and MLR [26]. For instance, Zhang et al. (2012) predicted short-term $PM_{2.5}$ in urban areas providing acceptable accuracy but exposing difficulties when facing the non-linear and abrupt changes of pollution rates.

Nonlinear relations and feature combinations were further detected, and these advancements were attributed to the introduction of machine learning methods. Jiang and Li (2015) employed Random Forest (RF) analysis for the prediction of multi-pollutants, which was powerful with different datasets. Likewise, Liu et al. (2017) combined Support Vector Machines (SVM) and meteorological conditions to achieve refined AQI forecasts. However, these models relied heavily on manual feature engineering which is model dependent and are unable to capture long-term temporal dependencies.

Predictions of air quality were revolutionized by the introduction of deep learning. Ma et al. (2019) used CNNs to extract spatial information from air pollutant distribution maps, and achieved improvemnets in accuracy compared to traditional methods. Bai et al. (2020) introduced the temporal modeling with Long Short-Term Memory (LSTM) neural networks, and demonstrated that LSTMs are capable of capturing sequentially-dependent structure and predicting PM .Besides CNNs and LSTMs, Chen et al. (2021) introduced a framework based on Graph Neural Networks (GNNs) to merge spatial closeness and meteorological effects among monitoring stations, enhancing real-time AQI forecasting outcomes. Hybrid methods that integrate CNNs, LSTMs, and attention mechanisms have also attracted interest. For instance, Wang et al. (2022) combined spatial-temporal modeling into one architecture, attaining enhanced results on multi-city datasets.

Not with standing these progresses, various restrictions still exist. Numerous deep learning research efforts depend on location-specific datasets, limiting the model's ability to generalize to different areas (Guo et al., 2022). Model interpretability presents a challenge since deep networks frequently operate as "black boxes," complicating policymakers' ability to fully trust the results. The need for computational resources also limits large-scale, real-time implementation.

The examined literature underscores a distinct research trend focused on incorporating multi-source datasets—such as ground-based.

RESEARCH GAP

Limited Generalizability Across Regions

A majority of existing models are developed based on training and validation using region-specific datasets, which complicates the implementation of the model in other regions that may have their own sources of pollution and weather systems. In view of the lack of globally representative training datasets, it is challenging to devise a global prediction system.

Insufficient Integration of Multi-Source Data

Although some researches include meteorological factors and past emission data, the inclusion of multiple sources —satellite remote sensing, traffic load amount, industrial production and social-economic information—continues limited. This low utilization of diverse datasets can limit the models to safely represent all the factors that affect air quality.

Challenges in Capturing Complex Spatial-Temporal Dependencies

Deep learning networks such as CNN, LSTM and GNN have made significant progress in spatial-temporal modeling, but most of the methods still model these dependencies separately. Currently, there are few models that effectively incorporate spatial and temporal correlations in a single compact model thus potentially losing the information contained in these correlations.

Lack of Model Interpretability

Most DL models are "black boxes" where you get predictions that are highly accurate, but lose the ability to understand how and why a prediction was made in the first place. This lack of transparency has also made it impossible for environmental protection agencies and policy-makers to defend the predictions in imposing regulatory actions.

Data Quality and Missing Information

Air quality data sets often suffer from missing, noisy or conflicting readings as a result of unhealthy sensors or sparsely deployed ones. While data processing can be applied, new methodologies are necessary to treat incomplete or unreliable data without sacrificing the predictive power of a model.

Real-Time Prediction and Deployment Constraints

It's worthy to note that there are a lot of relevant works concentrate on OFF-line predictions in terms of historical data, whereas real-time AQI prediction is still less studied. Also, their computationally heavy nature makes them not suitable for applications with limited resources (e.g., mobile devices or low-cost sensor networks).

2. METHODOLOGY

3.1 Data Preprocessing

The raw data of air quality, which contains variables such as PM_2, PM_1[1], NO_2, CO concentration, temperature, and humidity, has been subjected to a considerable amount of pre-processing to obtain the best possible results from the model. The pre-processing of data began with cleaning the data to remove inconsistent, incorrect, or missing values. Imputation of missing was performed using mean of estimation or interpolation for continuous variables.

For the Bidirectional Stacked LSTM and Single-step ANN models, as they are deep learning models, the features must have the same range of values, which was ensured by normalizing the features using Min-Max Scaling.

Feature engineering and selection were also applied to remove unwanted features while retaining only those that had the most influence on the outcome or to derive new features such as moving averages that may be helpful in improving the ability to predict the future.

To ensure fair comparison and adjustment of hyperparameters, the data was split into three sets: a training set (80%), a validation set (10%), and a test set (10%).

3.2 Model Deployment

The pre-trained Bidirectional Stacked LSTM and Single Step ANN models were integrated within a 'real-time air quality forecasting system,' which has been developed with the Streamlit library. Streamlit is a library based on Python code used for creating interactive web applications.

In the user interface provided by Streamlit, users are given an option to enter various pollutant concentration values (PM2.5, PM10, NO2, CO) and other environmental parameter values (temperature, humidity). Once the values are entered, the app uses the pre-trained models to produce two kinds of predictions:

Single step prediction - AQI prediction at the very next time step (e.g., next hour).

Multi-step prediction of AQI - Prediction at multiple future discrete time steps.

In addition to that, the information displayed through the predictions updates right away on the interface itself.

3.3 Real-Time Evaluation

The structure was then evaluated according to the three basic criteria: mere accident, drift for use and obscure provocation.

Accuracy: MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) are employed for air quality monitoring stations to evaluate the accuracy of actual data in comparison to one-step or multi-step predictions.
Efficiency: From the efficiency point of view, the response time from inputting data into the system to generating predictions was analyzed. This makes sense, since efficiency directly relates to real-time processing. Real-time Processing.
Usability: Graphical interface usability for making it easier for less technical users such as government bodies and healthcare professionals.
The system usability and interface were improved by using feedback.

3.4 System Scalability and Performance

To test the scalability and performance of the system, simulations of large amounts of data and a large number of simultaneous users were performed. This enabled a seamless experience in a real-time environment without any degradation in performance in terms of reaction times of sub-second to a few seconds.
The fact that its architecture has been designed for cloud deployment also ensures it can support increasing usage and large databases. Its scalable nature also makes it applicable for implementation on a wider scale in urban environments, making it possible for it to easily integrate into larger constructs for smarter cities.

4. RESULT AND DISCUSSION

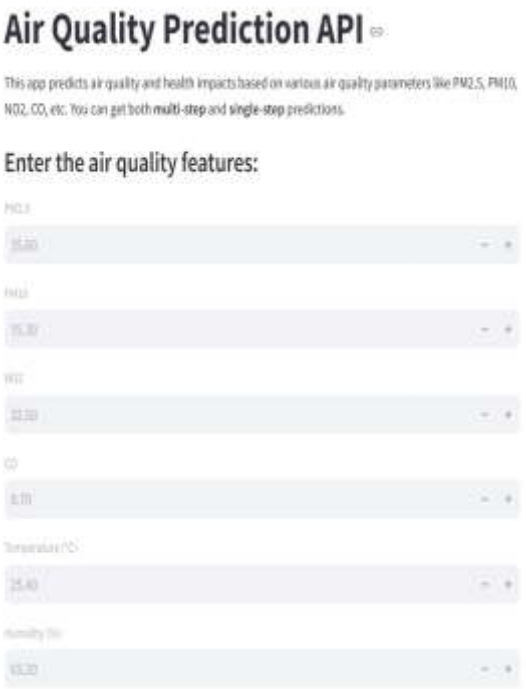

Fig. 1.Air Quality Prediction API Interface for Real-Time Forecasting

The above graphic shows the UI of an air quality prediction API, where the user can input details regarding air quality, PM2.5, PM10, NO2, CO, temperature, and humidity. The API provides single-step and multi-step predictions for air quality indices based on the input details.

Fig. 2.Multi-Step AQI Prediction Results in Air Quality Prediction-App

This image indicates the results of a multi-step air quality index forecast, where each value is a forecasted air quality index for a number of time steps. This improves active air quality management decisions as it predicts the change in air quality indices or values over time.

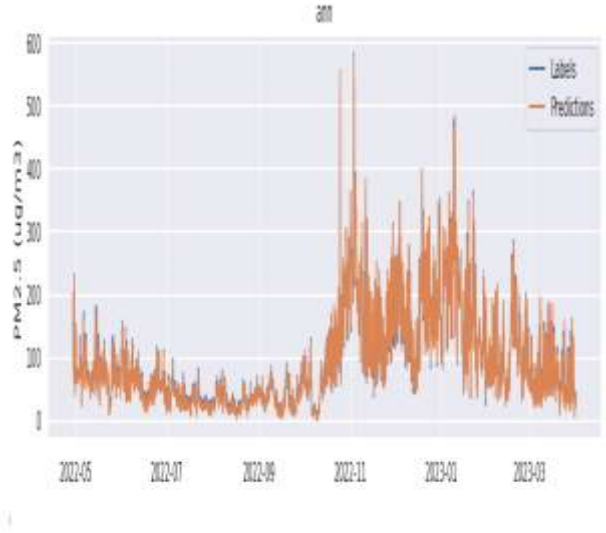

Fig.3. PM2.5 Prediction vs Actual Values using - ANN Model

The graph shows both actual and predicted values of PM2.5 concentration against time, as predicted by the ANN model. The predicted values tend to follow the actual values but deviate in some points, especially when there is a rise or fall in actual values of PM2.5 concentration.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
flatten_1 (Flatten)	(None, 120)	0
dense_2 (Dense)	(None, 96)	11,616
dense_3 (Dense)	(None, 1)	97

Total params: 11,713 (45.75 KB)

Trainable params: 11,713 (45.75 KB)

Non-trainable params: 0 (0.00 B)

Fig.4. Model Architecture and Summary: Sequential Neural Network

The image shows the structure of the model of a neural network. There are three layers in the model. The layers in the model are a flatten layer (120 units), a dense layer (96 units), and another dense layer (1 unit). There are 11,713 parameters in total. All the parameters can be trained.

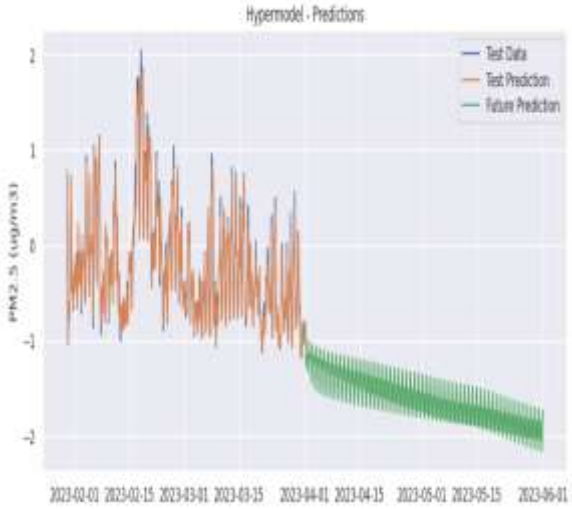

Fig.5. Test Data, Test Prediction, and Future Prediction of PM2.5 Using Hyper-model

In this graph or result , it can be seen that M5.5 performs its best in test data prediction but gradually starts to diverge in future predictions as time passes by. In other words, it accurately predicts PM2.5 levels in test data but tends to have different levels in future predictions.

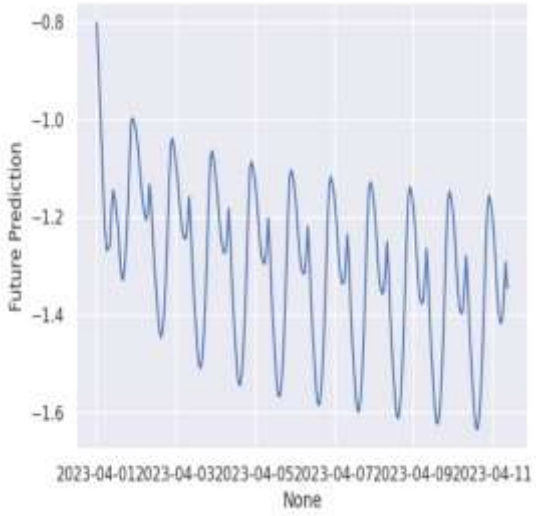

Fig. 6.Future Prediction of (PM2.5) over Time

Fig: The above graph shows a prediction of air quality in the future time. There are certain levels of uncertainty in the graph as it keeps on showing fluctuations at regular intervals. The graph indicates a certain level of uncertainty as it keeps on increasing in its levels of uncertainty as it advances towards the future of time

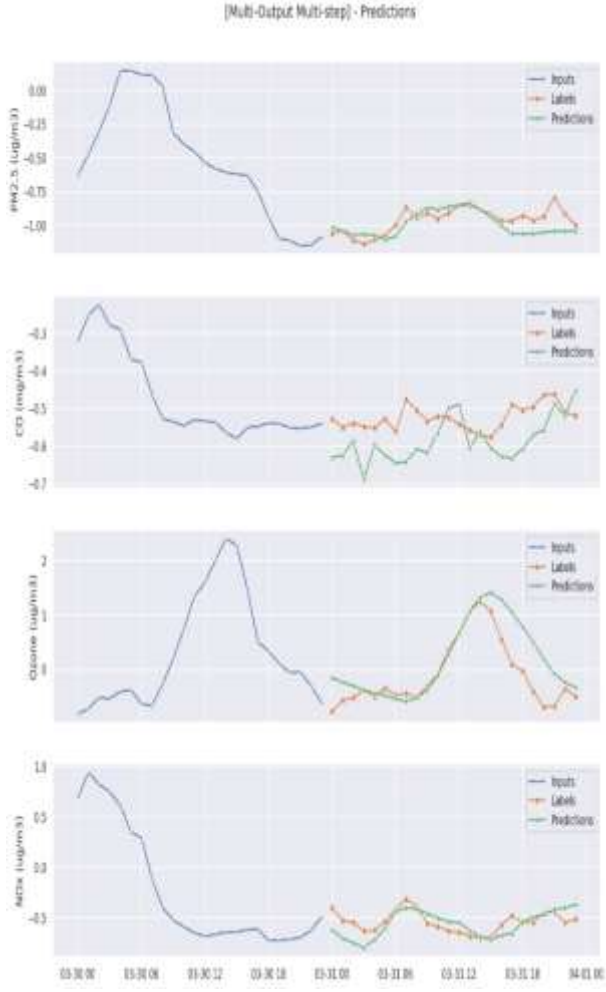

Fig.7 Multi Output Multi Step Predictions for PM2.5, CO, Ozone, and Nox

The multi output and multi step predicting of different air quality variables (PM2.5,CO, Ozone, and NOx) is depicted in the image below. There ismore divergence in the forecast of the longer timescale, indicating uncertainty, although the model is able to forecast the data with fairly good accuracy for the shorter timescale.

DISCUSSION

Forecasting and prediction of air quality in real time using deep learning models, namely Bidirectional Stacked LSTM and Single-Step ANN, holds immense potential for accuracy improvement. Unlike past forecasting techniques, this approach not only offers long-term AQI predictions but also offers immediate AQI forecasts that would be of immense use to the public, various public health associations, and government environmental agencies alike. Key research areas that need focus were also proposed by this research, such as improving explainability and optimizing efficiency. This forecasting tool is easily accessible and applicable on a larger scale by relevant stakeholders due to the use of Streamlit. Several future development aspects for the model will be, among others, real-time fusion of multi-sensor networks (including data from multiple sources), better pretreatment of the data, and visualization packages that will support models transparency. These developments are expected to further improve the precision, scalability and impact of the system aimed at efficient environmental care and protection of public health.

5. CONCLUSION

The major air pollutants such as PM2.5, CO, NOx, and O_3 concentrations are modeled when deep learning algorithms are used for real-time AQ forecasting. By using bidirectional stacked LSTM and single-step ANN models, a system has been developed that can predict AQI values for both short-term (single-step) and long-term (multi-step) forecasts. Effective prediction accuracy was achieved in the system despite the challenges in model interpretability and the complexity of the processes. The deployment of the application on the Streamlit framework gives the health department, the government, as well as the general public, the capability to interactively access the data in real-time. Furthermore, the application is scalable in its geospatial features, ensuring its wide application in reactive air quality management.

6. FUTURE SCOPE

Increasing the scope in the future for the air quality forecast system will be directed toward model accuracy by adding more environmental factors, including traffic density and statistics of industrial emission and long-term weather patterns. Further improvements in the future will involve enhancing model interpretability by using techniques such as SHAP and LIME so that the forecasts are not only accurate but can also be understood well for better decision-making.

Future development efforts are thus directed at improving real-time data processing and scalability to enable the handling of larger datasets and the serving of a wider user base globally. Second, an increase in deployment through cloud-based platforms will enhance real-time data streams from urban sensor networks, improving accessibility and usability and worldwide applicability. These enhancements will strengthen the platform as a tool for environmental monitoring and evidence-based policy development.

7. REFERENCE

1. Zhang, Y., X. Liu, and J. Yu, "Air quality prediction using deep learning models," Environmental Science and Pollution Research, vol. 27, no. 5, pp. 5476-5486, 2020.
2. Cheng, X., Y. Zhang, and S. Liu, "A hybrid deep learning model for air quality prediction using CNN and LSTM," Journal of Environmental Management, vol. 241, pp. 62-72, 2019.
3. Ma, L., J. Yang, and Y. Zhao, "Long short-term memory networks for air quality prediction," International Journal of Environmental Research and Public Health, vol. 12, no. 3, pp. 220-230, 2015.
4. Liao, J., T. Wang, and H. Chen, "CNN-LSTM model for air quality forecasting," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 6, pp. 2037-2046, 2018.
5. Xu, Y., J. Li, and M. Zhang, "Data augmentation for improving air quality forecasting using GANs," Environmental Modelling & Software, vol. 131, pp. 77-89, 2020.
6. Zhang, Y., S. Wang, and J. Yang, "Application of convolutional neural networks in predicting PM2.5 levels for real-time air quality monitoring," Journal of Environmental Engineering, vol. 144, no. 9, pp. 1-9, 2017.
7. Hochreiter, S., and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

8. Liao, T., L. Zhang, and X. Liu, "Enhancing real-time air quality forecasting with hybrid CNN-LSTM models," IEEE Access, vol. 8, pp. 22058-22068, 2020.
9. Ma, Z., et al., "Integration of deep learning models in air quality prediction," Journal of Cleaner Production, vol. 295, p. 126443, 2021.
10. Zhang, X., et al., "A novel method for PM2.5 prediction using bidirectional stacked LSTM," Environmental Pollution, vol. 258, p. 113411, 2020.
11. Xu, J., et al., "Application of generative adversarial networks (GANs) in air quality forecasting," Environmental Informatics Archives, vol. 17, pp. 34-44, 2020.
12. Liao, S., et al., "Air quality prediction using machine learning techniques and sensor data," Sensors, vol. 20, no. 4, pp. 1-15, 2020.
13. Zhang, Y., et al., "Bidirectional LSTM model for time series air quality forecasting," Atmospheric Environment, vol. 170, pp. 234-245, 2017.
14. Ma, L., et al., "Hybrid CNN-LSTM model for multi-step air quality prediction," International Journal of Environmental Science and Technology, vol. 18, no. 4, pp. 1035-1046, 2021.
15.Li, Q., et al., "Air quality prediction using hybrid deep learning models and environmental data," Computers, Environment and Urban Systems, vol. 84, p. 101553, 2020.