International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Comparative Study of ML and DL Techniques for AQI Prediction with Explainable AI

Sandeep Kaur Goraya[1], Harjot Kaur[2]

[1]Research Scholar, Department of Computer Science, Guru Nanak Dev University, Amritsar
[2]Assistant Professor, Department of Computer Science, Guru Nanak Dev University, Amritsar

Abstract:
The growing impacts of air pollution on public health, ecosystems, and climate have made the Air Quality Index (AQI) prediction a more critical research concern. Traditional statistical methods often fail to capture the nonlinear and complex relationships in AQI data. Machine learning (ML) and Deep learning (DL) techniques are becoming more and more popular due to their ability to process large amounts of data and to identify complex patterns. This paper highlights some machine and deep learning techniques, including Random Forest (RF), Support Vector Machines (SVM), and XGBoost, along with deep learning architectures, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), transformers. These models were mostly used because they are scalable, flexible, and better at combining data from multiple sources about air quality. The present study emphasizes comparative analysis, their strengths, weaknesses, application domains, and the trade-offs between computational cost and prediction accuracy. This paper focuses on air prediction through emerging technologies, including explainable AI alongside the integration of machine learning and deep learning techniques. The outcomes of this review are intended to guide the development of effective AQI prediction systems for real-world applications.

Keywords: Air Quality Index, Prediction, Machine Learning models, Deep Learning, Internet of Things (IoT)

1. INTRODUCTION

Air quality poses a serious threat to environmental sustainability and directly impacts public health. Global air quality is declining due to rapid industrialization, urbanization, and vehicular emissions (World Health Organization 2021). Therefore, accurate predictions are required to monitor and forecast air quality. For the key pollutants like particulate matter (PM2.5 and PM10), ozone (O3), nitrogen dioxide (NO2), and sulfur dioxide (SO2) (Kumar et al. 2019), the Air Quality Index (AQI) provides perspicuity.

1.1 Need for AQI predictions:

It is essential to provide accurate and timely Air Quality Index predictions for the following reasons:

1.1.1 Reporting health advisories to sensitive clusters: Air pollution is associated with severe health problems, especially among high-risk masses, including children, the elderly, or people with persistent respiratory or cardiovascular illnesses. Machine learning-based forecasts and real-time aura quality monitoring provide these inhabitants with immediate health caution (Chen et al. 2018).

1.1.2 Guiding policymakers in implementing air quality control measures: Both historical and current pollution data are assessed by hybrid ML and DL algorithms to differentiate regions at stake, seasonal

patterns, and critical sources of pollution. Predictive insights can be utilized to propose future pollution scenarios based on myriad policy strategies (Zhang et al. 2022).

1.1.3 Supporting urban planning to minimize pollution hotspots: Urban planning is essential for enabling pollution hotspots (Sharma 2020), particularly in the areas in which the concentrated sources of pollution constantly result in inadequate air quality. Data collected by IoT sensors and ML models can assist in recognizing and visualizing pollution hotspots. According to predictive analytics, metropolitan structures may simulate airflow, human susceptibility, and the distribution of toxins.

1.2 The Shift Towards ML and DL Techniques: Traditional AQI prediction approaches such as linear Regression and ARIMA (Autoregressive Integrated Moving Average) were not fully able to capture non-linear and vigorous dependencies in existing air pollution data (Wang et al. 2018). Linear regression relies on simple relationships, whereas ARIMA is effective for linear and stationary time-series data. Managing the non-linear character of air pollution dynamics presents a great challenge. Conversely, by using their capacity to replicate intricate data relationships, new ML and DL methods solve these limitations. Precise air quality forecasts depend on the capacity to replicate complex and related interactions between contaminants and meteorological variables. ML and DL methods, including hybrid algorithms and neural networks, can handle these intricate data interactions. They can integrate several characteristics and find latent trends that conventional statistical approaches would not find (Li et al. 2020). Many sources, including meteorological stations, satellite images, and IoT sensors, can generate constantly fresh data on air quality and temperature. These datasets are enormous, have large dimensions, and demonstrate spatiotemporal variability (Chen et al. 2022). The efficient handling of this big data streamlines dynamic policymaking, proposes actionable insights, and improves public transparency regarding air quality. Real-time air quality monitoring and forecasting are crucial for pollution management and public health safety. Multi-step and real-time forecasting capabilities permit stakeholders to be well-prepared, which improves environmental management and minimizes health risks (Yang, L., et al. 2021).

Hybrid models integrate multiple ML and DL approaches to improve prediction accuracy, robustness, and generalization. Hybrid techniques that integrate different ML and DL models to maximize their benefits and minimize their drawbacks. Some new hybrid models that can provide strong solutions are Convolutional Neural Networks (CNNs)-LSTM, XGBoost-LSTM, Random Forest-CNN model, Genetic Algorithms and SVR, Recurrent Neural Networks (RNNs), and Bi-LSTM with attention mechanisms.

2. LITERATURE REVIEW: COMPARATIVE ANALYSIS OF PREVALENT ML AND DL TECHNIQUES

Researchers used a lot of different machine learning techniques to make predictions about air pollution. These included Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees (DTs), Adaptive Boosting (AdaBoost), Gaussian Naive Bayes (GNB), Gaussian Process Regression (GPR), Extreme Learning Machines (ELM), Extreme Gradient Boosting (XGBoost), and many more. In the literature review, it was found from the authors' perspective that RF, SVM, and XGBoost were extensively used due to their scalability, robustness, and ability to handle heterogeneous air quality datasets. Over the past decade, DL techniques have been extensively used to capture complex spatial-temporal relationships and to improve prediction accuracy. The various DL techniques, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), Gated Recurrent Units (GRUs), Recurrent Neural Networks (RNNs), Transformers, Autoencoders, and Generative

Adversarial Networks (GANs), etc., have all been extensively used by the authors. But CNN gained significant prominence due to its ability to identify spatial air pollution trends from satellite images and sensor networks. LSTM is found effective at predicting time series because it learns how changes in pollutants are associated over time. Transformers is now a powerful sequence-based model that employs attention mechanisms to make AQI predictions better. In this part, we go into more detail about the most common ML and DL methods.

2.1. Random Forest (RF):

Random Forest is one of the ensemble learning techniques demonstrated to be quite successful and efficient in the resolution of both classification and regression challenges. This approach enhances the fundamental idea of decision trees by using their strengths and lowering their shortcomings, especially their inclination to overfit. Random Forest builds a "forest" of decision trees and uses their outputs to make predictions that are more accurate and reliable (Breiman 2001). Random Forest has been employed to identify germane features in the framework of AQI prediction and to project pollution concentrations. People were interested in Random Forest models because they could accurately find important AQI characteristics, such as PM2.5 and PM10 concentrations that changed with temperature and humidity (Liu et al. 2019). (Gupta et al., 2020) They found that the RF method, which predicted AQI in cities by combining pollution and weather data, performed exceptionally well. Because it naturally chooses the best features, it is the best way to predict AQI in places that are changing and not staying the same, like big cities with many datasets that contain different dimensions.

2.2 Support Vector Machines (SVM):

Support Vector Machines have been utilized for forecasting the AQI in both classification and regression tasks. Using kernel methods (Vapnik, V. 1995), SVM may map data into higher dimensions for model non-linear connections. It can be used for multi-class classification of air quality as good, moderate, and unhealthy air quality (Zhou et al. 2020). This classification helps to evaluate air quality based on certain pollutant limits. Support Vector Regression (SVR) is a variant of SVM applied to regression issues. One can estimate pollution levels using data on emissions from certain sources, including factories, vehicles, and other human-made activities. This approach lets lawmakers make more targeted plans for reducing pollution. By approximating the levels of particular pollutants, including PM2.5, PM10, NO2, SO2, etc., SVR offers a basis for estimating regression-based pollution concentrations and therefore enabling real-time AQI forecasting (Zhao et al. 2021). On short datasets, SVM has shown strong performance for PM2.5 level prediction over other ML methods (Sun et al. 2022). It is the appropriate solution for AQI forecasts since it can properly generalize on sparse data and control non-linear relations.

2.3. XGBoost:

XGBoost's scalability, efficiency, and predictive powers (Chen et al. 2016) have made it the approved model for many regression and classification projects. Scalability, efficiency, and prediction accuracy follow as one increases the principles of gradient boosting. It is considered a possible candidate for AQI prediction and pollution analysis because it can handle complex interactions, missing values, and large datasets. One important XGBoost used in metropolitan settings is real-time AQI forecasts. Fast-changing air quality in urban areas is the outcome of various dynamic elements, including industrial pollutants, traffic in cars, and weather. Considered most appropriate for time-sensitive applications, XGBoost, due to its quick training and inference speeds, is According to (Zhang et al. 2022), XGBoost consists of important metrics that evaluate the impact of many components on AQI prediction, including pollution levels and temperature, thereby including temperature. By raising awareness of significant air pollutants like PM2.5

and NO_2, the feature ranking helps focus efforts on reducing pollution, such as lowering emissions from specific sources.

At estimating AQI values, XGBoost was faster and more accurate than RF and SVM models. This was largely because it's able to better manage missing data (Li et al. 2021). This study also revealed that the main determinant of AQI values was PM2.5 and meteorological parameters like wind speed. Such information is crucial for developing accurate AQI forecast models and finding practical means of lowering them.

2.4 Convolutional Neural Networks (CNNs):

Strong DL architectures for processing and analysis of spatial data are convolutional neural networks, or CNNs. Their capacity to extract significant features from structured data—such as pictures or grid-based information—allows them immense utility in forecasting the Air Quality Index (AQI) in metropolitan regions. CNNs can search for patterns in spatial data thanks to convolutional layers. Still another component of AQI prediction is determining spatial connections between pollution concentrations and environmental or geographical factors. Since CNNs are amazing at processing spatial data, they fit very well for estimating AQI in cities where air quality changes significantly due to elements including traffic, industrial pollution, terrain, and weather (LeCun et al. 2015). CNNs and other spatial modeling techniques were used to forecast AQI and extract features from maps. Pinpointing pollution hotspots and providing localized forecasts enables better decision-making for human health and improved policy development. A real-time AQI prediction system could use both satellite data and measurements from ground sensors to make detailed maps that let city planners make changes that are specific to each area (Zhu, L. et al. 2020). Geographic Information Systems (GIS) let geospatial data and environmental monitoring systems be merged since they provide strong instruments for spatial analysis. GIS techniques let researchers map, investigate, and show spatial changes in air quality. Therefore, ML models can improve the utility of GIS by facilitating the diagnostic and predictive analysis of air quality (Li et al. 2021). A CNN-based system was developed in a review (Lee et al. 2020) to forecast AQI for different districts and analyze spatial trends in air quality. This study illustrates that AI-powered systems offer scalable solutions for urban areas and have the potential to completely redefine air quality prediction and monitoring.

2.5 Long Short-Term Memory Networks (LSTMs):

LSTMs are a subclass of RNNs. They are specifically designed to capture long-range dependencies in sequential data. So, they can be used to predict AQI over time, where past trends and outside factors like weather can change pollutant levels (Hochreiter et al. 1997). LSTM was used to analyze seasonal fluctuations, weather conditions, and historical patterns to predict daily AQI levels (Yu et al. 2021). Moreover, they facilitate multi-step predictions that enable planners to forecast trends in air quality over a couple of days or weeks (Singh, R., & Sharma, A. 2021). LSTM networks were better at predicting PM2.5 levels over a 2-hour period than traditional methods like ARIMA and other common machine learning algorithms, according to a large study. (Agarwal, A., & Sharma, S. 2021) They highlighted the ability of LSTM to effectively capture complex temporal dependencies, which were often overlooked by traditional statistical ML models.

2.6 Transformers:

Transformers have become an important advancement in sequence modeling because of their self-attention mechanism. They can overcome the limitations of traditional RNN and LSTM models, particularly in tasks requiring sequential modeling (Vaswani, A., et al. 2017). Unlike their predecessors,

transformers do not rely on sequential data processing, which often leads to difficulty in retaining long-range dependencies.

Instead, transformers can more effectively and efficiently capture long-term dependencies than LSTMs because they are capable of analyzing entire data sequences simultaneously. This capability makes them ideal for regional AQI forecasting across multiple locations (Wong, K., & Zhang, S. 2022). Transformers can predict AQI (Chen, M., et al. 2023–2) by weighing multiple data types, including weather, pollution levels, and traffic movement, and how these elements change throughout time and location. Transformer-based models were used to project the AQI of several cities. Transformers can effectively investigate and forecast long-term trends in AQI prediction, particularly regarding the complex changes in pollution levels and weather conditions over time. Modern prediction operations are thus more accurate, scalable, and dependable (Li, Y., et al. 2022).

3. DISCUSSION

The adoption of various ML and DL algorithms for air prediction comes with several advantages and disadvantages. The above models are great at dealing with different types of data sources and capturing complex spatial-temporal relationships, but their performance changes depending on how much computing power is needed, how easy the data is to understand, and the quality of the data. This section elaborates on strengths, weaknesses, and trade-offs between computational cost and prediction accuracy to evaluate their applicability in real-world applications

3.1. Strengths and Weaknesses:

The literature review above says that these ML/DL techniques were most often used because they could handle large and complex datasets, show how things change over time and space, and model how air pollutants are affected by things like weather, traffic, and industrial emissions in a way that is not linear. Figure 1 highlights the strengths and weaknesses of ML and DL techniques mentioned above. The adoption of various ML and DL algorithms for air prediction comes with several advantages and disadvantages. The above models are great at dealing with different types of data sources and capturing complex spatial-temporal relationships, but their performance changes depending on how much computing power is needed, how easy the data is to understand, and the quality of the data. This section elaborates on strengths, weaknesses, and trade-offs between computational cost and prediction accuracy to evaluate their applicability in real-world applications

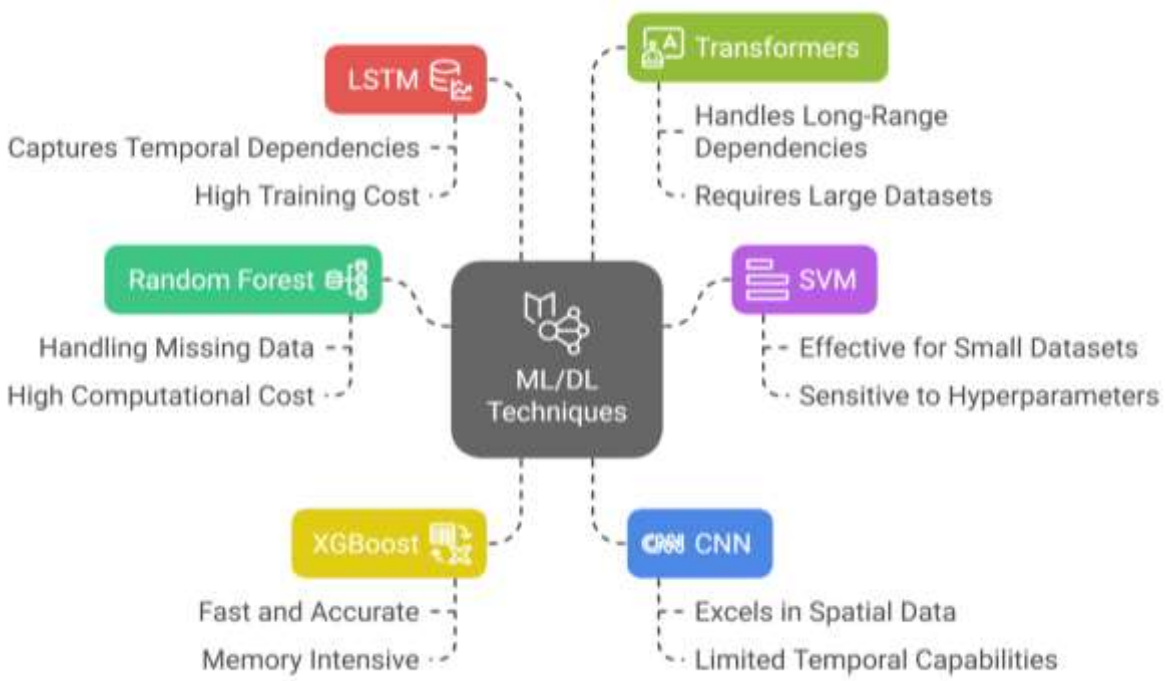

Figure 1—Strengths and Weaknesses of reviewed ML and DL models

The quality, completeness, and granularity of input data, which may vary significantly across regions, have a substantial impact on how well these techniques operate. Also, many advanced models are difficult to understand, which makes it hard for policymakers and environmental agencies to adopt and use them. These groups need clear and useful information to make decisions. In the research, their pros include being accurate and able to work with different types of data sources, like sensors, satellites, and the Internet of Things (IoT). On the other hand, their cons include needing a lot of processing power, being sensitive to data quality, and not being simple to use in real life. They often need substantial computational resources, particularly when deep learning models are being trained on large datasets.

3.2. Trade-offs between Computational Cost and Prediction Accuracy:

Depending on research goals and resource availability, researchers occasionally sacrifice prediction accuracy and computing efficiency while selecting ML and DL models. ML methods, including Random Forest and Extreme Gradient Boosting (XGBoost), were commonly used due to their computational efficiency and capacity to manage rather big structured datasets. Moreover, they give feature importance values that enable one to understand how input variables affect forecasts (Huang, A., et al. 2021).

On the other hand, DL models, including transformers and LSTM networks, were judged more effective for managing large volumes of unstructured, spatiotemporal data. At spatiotemporal modeling, they achieve better accuracy; they are perfect at identifying intricate links and trends in data. Higher processing capabilities and a lot of training data, however, will help to determine the deployment resources required (Singh, R., & Verma, M. 2022).

4. EMERGING TRENDS IN AQI PREDICTION

As air quality becomes a growing global concern, advancements in AQI prediction are evolving rapidly. This section covers the emerging trends essential for enhanced accuracy, transparency, scalability, and

energy efficiency for proactive environmental management. Figure 2 explains the different approaches, such as edge computing and explainable AI, that balance predictive power with efficiency, interpretability, and sustainability.

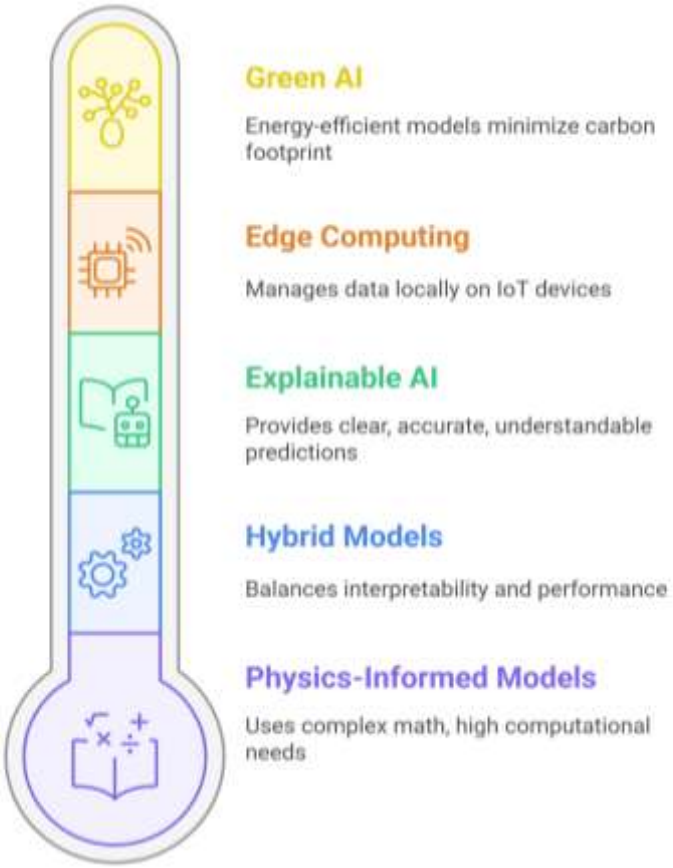

Figure 2—AQI Prediction model range from theoretical to practical application

4.2.1 Inclusion of Explainability: By showing how humidity, temperature, and air pollutants affect air pollution predictions, explainable AI (XAI) techniques make AQI predictions more clear and accurate. Traditional models, such as neural networks, often behave in a "black box" way that makes it hard to figure out how they make decisions. Several AI methods, including Shapley Additive Explanations (SHAP) (Doshi, T., & Joshi, H. 2022) and Local Interpretable Model Explanations (LIME) (Ribeiro et al. 2016), give numerical or visual explanations for certain forecasts. These help to figure out how important a feature is. These methods boost trust and openness by letting everyone understand and react to the model results correctly.

4.2.2 Integration of Hybrid Models: The hybrid models combine ML and DL techniques to minimize their limitations. Hybrid models can handle multi-source datasets and balance interpretability and performance. For instance, a hybrid model RF-LSTM combining Random Forest (RF) and Long Short-Term Memory (LSTM) can be employed to predict AQI. RF is applied in this system (Zhang, Y., Li, X., Wang, J., & Chen, L. 2022) to identify significant variables, including PM2.5, PM10, and other weather variables, including temperature and humidity. Known for its ability to identify temporal connections, LSTM then employs these selected features to learn sequential patterns and trends in data across time, hence predicting AQI.

RF-XGBoost—which consists of Random Forest (RF) and XGBoost—can be utilized to find characteristics and fine-tune predictions by boosting strategies to lower mistakes and increase accuracy (Alghamdi, A. S. 2020). CNNs can be used to extract spatial information, and Extensive Gradient Boosting (XGBoost) (Zhou, J., & Chen, W. 2023) can be incorporated for last predictions.

4.2.3 Bridging Physics-Informed Models with AI: Community Multiscale Air Quality (CMAQ) and Weather Research and Forecasting with Chemistry (WRF-Chem) are two conventional physics-based models. To show how physics and chemistry work in the atmosphere, they use challenging math problems. However, their real-time application is difficult due to their high computational requirements and need for precise data to predict air pollution. ML/DL models such as transformers and LSTMs use historical as well as real-time data to efficiently predict AQI. Hybrid approaches integrating physics-based models with ML/DL models are getting attention because they provide more intelligent and accurate air quality forecasts by combining the best aspects of both scientific knowledge and data-driven flexibility. For instance, a Physics-guided neural network, AirPhyNet (Hettige et al. 2024) exhibited considerable performance in capturing spatiotemporal relationships, showcasing a notable reduction in Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

4.2.4 Edge Computing: Models including RF, XGBoost, CNNs, LightGBM, and LSTM with pruning are applied on edge devices, including IoT/sensors in edge computing, to manage data locally instead of on central servers. Remote places allow one to monitor and forecast the AQI in real-time, and environmental changes can be swiftly responded to to enable rapid movement of people. Edge devices lack a lot of processing capability; hence, they may not be able to manage rather complex models (Wang, F., et al. 2022). Such an issue drives the need for improved DL models using quantization and cutting. More complex real-time apps could be created by linking 5G networks to new IoT devices.

4.2.5 Green AI: Sustainable AQI forecasting using pre-trained models: Green AI focuses on the development of energy-efficient DL models to minimize carbon footprints (Green, A., & Patel, M. 2023). Due to AI adoption in environmental sciences, ensuring sustainability in AI practices has become critical. Model optimization techniques such as sparse training, efficient architectures (e.g., TinyML, MobileNet), and transfer learning reduce resource usage. Using pre-trained models and fine-tuning them for environmental applications reduces the need for training from scratch. This method will make real-time AQI forecasting more accessible and environmentally friendly by adopting Green AI practices

5. CHALLENGES AND FUTURE DIRECTIONS

5.1. Challenges:

Some ethical and technical problems like interpretability, sensitivity to data quality, and high computation cost impede air pollution prediction employing ML and DL approaches. This part This section addresses the potential challenges in balancing the practical application of AQI prediction with resource efficiency.

5.1.1 Data Quality: Missing data can impede the training process, which may result in skewed or incorrect predictions. Missing records can occur due to equipment failure, adverse weather conditions, or human errors. For example, maintenance problems and noisy data can hinder model performance. Similarly, sensor readings can be affected by technical failures, environmental influences, and calibration problems that may result in noisy datasets (Liu, Y., & Zhang, L. 2022).

5.1.2 Scalability: DL models require significant resources for both training and inference, and they are mathematically intensive. The cost of training grows exponentially with the amount of data and its complexity. This issue is worsened by high-dimensional environmental data such as spatiotemporal AQI

datasets. High-performance Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs) are needed for training advanced DL models (Kumar, R., et al. 2023); however, both of these may not be available in an environment with limited resources.

5.1.3 Interpretability: DL techniques that are involved in complex design and lack interpretability are considered "black boxes" (Doshi, T., & Joshi, H. 2022). They may restrict their openness and reliability in practical applications. Although these techniques achieve improved accuracy, they provide an incomplete picture of why particular forecasts are made. Interpretable models assist decision-makers in supporting actions and investments in environmental policymaking and human health.

5.1.4 Ethical Implications: AQI predictions raise some ethical concerns, such as data bias and fairness. Fair data collection in rural areas is required because there is limited monitoring coverage as compared to denser metropolitan areas with larger air monitoring stations. It is essential to minimize disparities in air quality evaluations and public health measures. Another ethical issue is the misuse of AQI predictions by governments and other organizations, who manipulate air quality reports to downplay pollution levels. To prevent such misuse, AQI prediction models must be transparent, independent, and thoroughly validated by environmental authorities.

5.1.5 Real-World Deployment Challenges: In the real world, using air quality index (AQI) prediction algorithms presents a variety of technological, infrastructure, and operational difficulties (Houdou, et al.). Two of the most critical problems are data dependability and availability. Standardization and interoperability suffer when several data sources—such as satellite photos, IoT devices, and weather data—are applied. Two further major issues are scalability and computer performance. Though they require a lot of processing power and cannot be employed in real time if resources are restricted, deep learning (DL) models—especially those that use transformers and LSTMs—are rather accurate.

5.2. Future Directions:

Improving model explainability, integrating hybrid techniques, and leveraging multi-source data for improved accuracy and dependability are the key components of the future of AQI predictions. Explainable AI (XAI) techniques, like SHAP and LIME, can make ML/DL models more interpretable, empowering trust in the decision-making process for policymakers and the public. Prediction accuracy can be increased by hybrid models such as the integration of RF with LSTMs, which strikes a balance between feature interpretability and temporal pattern recognition. Transformers are much better than LSTMs at capturing long-range dependencies, which makes them useful for modeling seasonal trends and changes in AQI over time and space.

Sustainability is crucial because deep learning models, especially transformers, need a lot of computing power. Future research should look into energy-efficient AI that uses pre-trained models, model pruning, and quantization to cut down on carbon footprints. Real-world deployment challenges, including sensor errors, data availability, and latency, need powerful models capable of handling missing data and adapting across regions. AQI predictions will be more intensive when multi-source datasets such as traffic, meteorological, and industrial data are incorporated with state-of-the-art models like Graph Neural Networks (GNNs), improving spatial relationships.

Beyond accuracy, models' robustness, generalizability, and computational efficiency must be assessed. Hybrid physics-AI models may represent an encouraging direction, blending scientific knowledge with data-driven adaptability. Future efforts should concentrate on scalable, interpretable, and sustainable AQI forecasting systems that empower real-time environmental decision-making.

6. CONCLUSIONS

This study attempts to highlight the strengths and weaknesses of various ML and DL methods used for Air Quality Index (AQI) predictions. ML techniques such as RF and XGBoost models have become popular due to their effectiveness and usability. DL models such as LSTM and transformers are capable of handling spatiotemporal complexities. Understanding how input features affect predictions and possibly attain improved predictive accuracy is difficult due to the "black-box" nature of DL models. Accuracy and interpretability represent a significant trade-off in AQI prediction. Future research should focus on improving prediction accuracy by enhancing the scalability, interpretability, and efficiency of ML and DL models. It must focus on incorporating explainability and the development of hybrid techniques by combining ML and DL models. To optimize both accuracy and efficiency and the utilization of modern computational resources, there is a need to develop systems that are scalable and interpretable for AQI prediction. Real-time AQI monitoring can be made possible by integrating ML and DL techniques with IoT sensors. However, issues like sensor calibration, data inconsistencies, and communication delays must be resolved. Improved data quality can further refine AQI prediction. The ultimate goal of future AQI prediction models is not only to provide predictions but also to provide actionable insights to support environmental interventions and mitigation strategies.

7. REFERENCES

1. Agarwal, A., & Sharma, S. (2021). PM2.5 Prediction in Delhi Using Long Short-Term Memory Networks. *IEEE Transactions on Environmental Computing*, *15*(3), 112–125.
2. Alghamdi, A. S. (2020). Gradient Boosting Feature Selection with Machine Learning Classifiers. *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, 1–6
3. Chen, H., & Zhao, Y. (2018). Machine Learning in Predicting Air Pollution Levels: A Case Study. *Environmental Monitoring and Assessment*, *190*(), 205–216.
4. Chen, M., et al. (2023). Multivariate AQI forecasting with transformer networks. *Journal of AI Applications in Environmental Science, 9*(2), 89–102.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–79.
6. Chen, X. (2022). Deep Learning for Air Quality Prediction: A Comprehensive Review. *Environmental Informatics Letters*, *2*(3), 1–12.
7. Cortes, C., & Vapnik, V. (1995). Support Vector Networks. *Machine Learning*, *20*(3), 273–297.
8. Doshi, T., & Joshi, H. (2022). Interpretability of Black Box Models in Air Quality Prediction Using SHAP. *IEEE Environmental Applications Journal*, *5*(1), 5–57.
9. Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.
10. Green, A., & Patel, M. (2023). Energy-Efficient Deep Learning Models for Environmental Applications. *AI for Sustainability Journal*, *3*(1), 15–28.
11. Gupta, A., & Singh, S. (2020). Air Quality Prediction in Indian Cities Using Random Forest Models. *Environmental Science Research Journal*, *19*(3), 98–110.
12. Hettige KH, Ji J, Xiang S, Long C, Cong G, Wang J (2024) AirPhyNet: Harnessing Physics-Guided Neural Networks for Air Quality Prediction. arXiv:2402.03784
13. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory: A Deep Learning Model for Sequential Data. *Neural Computation*, *9*(8), 1735–1780.

14. Houdou, A., El Badisy, I., Khomsi, K., Abdala, S.A., Abdulla, F., Najmi, H., Obtel, M., Belyamani, L., Ibrahimi, A., & Khalis, M. (2024). Interpretable Machine Learning Approaches for Forecasting and Predicting Air Pollution: A Systematic Review. *Aerosol and Air Quality Research*, 24, 230151.Huang, A., et al. (2021). Evaluating Memory and Speed Performance of XGBoost in AQI Forecasting. *Environmental Data Science*, *10*(3), 112–126.
15. Kumar, A., Gupta, P., & Sharma, S. K. (2020). Impact of Air Pollution on Human Health: A Review. *Environmental Research Letters*, *15*(), 1–9.
16. Kumar, R., et al. (2023). Scalability Issues in Deep Learning for Urban Air Quality Prediction. *Journal of Environmental Monitoring Systems*, *12*(3), 115–128.
17. LeCun, Y., et al. (2015). Deep Learning with Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.
18. Lee, J., et al. (2020). Spatial AQI Prediction in Seoul Using Deep CNN Models. *Atmospheric Science Research Journal*, *15*(3), 200–215.
19. Li, B., & Huang, Y. (2022). Computational Efficiency and Accuracy Trade-Offs in Machine Learning for AQI Prediction. *Journal of Atmospheric Modelling*, *15*(2), 50–67.
20. Li, M., & Feng, X. (2022). Challenges in Overfitting of LSTMs for Air Quality Prediction. *Journal of Environmental Informatics*, *1*(3), 78–9.
21. Li, S., Zhou, K., & He, J. (2020). Meteorological Influences on Air Pollution and Its Prediction Using Random Forest. *Atmospheric Research*, *28*(5), 15–27.
22. Li, Y., & Wang, Z. (2021). Fast and Accurate AQI Forecasting Using XGBoost. *Environmental Science Journal*, *17*(5), 551–563.
23. Li, Y., et al. (2022). Transformer-Based Air Quality Prediction Across Europe. *Atmospheric Science Advances*, *18*(5), 78–91.
24. Li, W., et al. (2021). Integrating GIS and CNN for air quality mapping. *Journal of Environmental Informatics, 35*(2), 102–115.
25. Liu, H., & Ma, J. (2019). A Feature Selection Approach for AQI Prediction Using Random Forest. *Environmental Modelling and Software*, *78*, 1–10.
26. Liu, Y., & Zhang, L. (2022). Addressing Missing Data Challenges in Air Quality Datasets Using ML Techniques. *Environmental Informatics Letters*, *19*(3), 205–219.
27. Ma, J., et al. (2020). Computational Efficiency of Random Forest for AQI Prediction. *Environmental Modelling Letters*, *15*(2), 50–67.
28. Ribeiro, M.T., Singh, S., Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144.
29. Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proceedings of the International Conference on Learning Representations (ICLR)*.
30. Singh, R., & Sharma, A. (2021). Scalable Air Quality Prediction with XGBoost. *Journal of Environmental Monitoring Systems*, *11*(2), 205–220.
31. Singh, R., & Verma, M. (2022). Multi-Step AQI Prediction Using LSTM: A Case Study in Delhi. *Atmospheric Research Letters*, *20*(6), 155–168.
32. Singh, M., & Verma, S. (2023). Balancing Computational Cost in Deep Learning Models for Air Pollution Forecasting. *IEEE Transactions on Environmental Computing*, *17*(3), 112–125.

33. Sharma, A. (2020). Urban Air Quality Management: Machine Learning for Decision Support. *Urban Ecology Research*, *3*(2), 85–102.
34. Sun, L., et al. (2022). Improving PM2.5 Prediction with SVM and Air Pollution Data Fusion. *Atmospheric Environment*, *207*, 19–28.
35. Sun, L., & Zhao, P. (2023). Resource Requirements of Transformer Models in AQI Prediction. *AI Applications in Environmental Science*, *6*(), 205–219.
36. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.
37. Vaswani, A., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NIPS)*, 5998–6008.
38. Wang, J., & Song, M. (2018). A Review of Traditional and Machine Learning Approaches for AQI Prediction. *Atmospheric Environment*, *195*, 11–21.
39. Wang, Z., & Luo, M. (2020). Limitations of CNNs in Sequential Air Quality Prediction Tasks. *Atmospheric Analytics Letters*, *8*(2), 123–136.
40. Wang, F., et al. (2022). Edge Computing for Real-Time AQI Prediction in Smart Cities. *IEEE Internet of Things Journal*, *9*(), 5602–5615.
41. Wong, K., & Zhang, S. (2022). Transformer-Based Models for Regional AQI Prediction. *Environmental Computing Journal*, *5*(1), 11–23.
42. World Health Organization. (2021). *Air Quality Guidelines: Global Update 2021*. World Health Organization.
43. Yang, L., et al. (2021). Real-Time AQI Prediction Using IoT and Deep Learning. *IEEE Transactions on Industrial Informatics*, *17*(5), 501–510.
44. Yu, F., & Zhao, H. (2021). Forecasting AQI with LSTM Networks. *Environmental Modelling and Software*, *27*(2), 5–56.
45. Zhou, P., & Feng, X. (2020). Classifying air pollution levels using SVM: A study in Beijing. *Environmental Data Analytics, 16*(3), 315–323.
46. Zhao, H., & Tan, K. (2021). PM2.5 prediction using kernel-based support vector regression. *Environmental Monitoring Letters, 6*(), 32–.
47. Zhang, D., Li, X., & Yang, J. (2021). Predicting Air Quality Using Machine Learning Techniques. *Journal of Atmospheric Pollution Research*, *12*(1), 112–123.
48. Zhang, M., Zhou, H., & Wang, L. (2020). Hyperparameter Optimization in SVM for Air Quality Prediction. *Environmental Informatics*, *12*(), 55–67.
49. Zhang, X., & Yang, F. (2022). Feature importance analysis for AQI prediction using XGBoost. *Environmental Research Letters, 18*(3), 120–133.
50. Zhang, Y., Li, X., Wang, J., & Chen, L. (2022). Hybrid RF-LSTM model for air quality prediction: Integrating feature selection with deep learning. *IEEE Transactions on Computational Intelligence and AI in Environmental Science, 19*(), 256–267.
51. Zhu, L., et al. (2020). Spatiotemporal air quality prediction using CNNs. *Environmental Data Science Letters, 10*(), 8–60.
52. Zhou, J., & Chen, W. (2023). Hybrid ML-DL models for AQI prediction: A comprehensive review. *Atmospheric Modelling Advances, 22*(), 178–198.