Comparative Investigation of Machine Learning and Deep Learning Approaches for Air Quality Prediction

Borui Zhang*

Faculty of Data Science, City University of Macau, 999078 Macau, China

Abstract. Air pollution is a critical environmental issue with significant impacts on human health and ecosystems, exacerbated by urbanization and industrialization, leading to increased emissions. Forecasting air quality accurately is crucial for risk mitigation and policy direction. Recent advancements in deep learning have enhanced prediction capabilities by automatically extracting features and managing complex data. This paper compares machine learning and deep learning approaches in air quality forecasting, highlighting their strengths and weaknesses. Machine learning offers easier interpretability with limited data but struggles with complex data relationships. Deep learning captures nonlinear patterns more effectively but lacks interpretability and requires more data. Challenges in the field of air quality forecasting include feature selection, model interpretability, and applicability across regions. Future directions involve introducing feedback mechanisms, interpretability methods, and transfer learning to improve model performance and generalization. This review provides valuable insights into existing methodologies and guides future research for effective air quality management.

1 Introduction

Air pollution is a major environmental problem that affects both ecological systems and human health on a global scale [1]. Growing urbanization, industrialization, and energy consumption have made this problem more severe by significantly increasing the number of pollutants consisting of carbon monoxide, ozone, nitrogen dioxide, sulfur dioxide, and particulate matter that are released into the atmosphere [2]. In light of these critical circumstances, Forecasting air quality accurately is increasingly crucial [3]. Precise air quality predictions offer several benefits: 1) Mitigation of health risks related to air pollution by prompting public preventive actions and timely dissemination of air quality warnings. 2) Facilitation of informed environmental policy-making, guiding industrial transformation and modernization, promoting the adoption of clean energy, and reducing pollution emissions at their source. 3) Enhancement of urban environments through the promotion of sustainable travel, improved traffic management, and overall improvement of living standards.

*Corresponding author: D21090102101@cityu.edu.mo

The evolution of deep learning has revolutionized the processing of complex data, offering capabilities previously unseen. By surpassing the limitations of traditional models regarding data features, deep learning exhibits the ability to handle nonlinear relationships and multidimensional data, automatically extracting intricate features from large datasets [4]. In recent years, this capability has been increasingly applied to environmental sciences, including meteorological forecasting, hydrological simulation, and ecological prediction. This technique has also facilitated significant advancements in fields for instance speech recognition, image recognition, and natural language processing [5, 6]. This remarkable progress in deep learning has opened up new possibilities for addressing complex environmental challenges, including air quality prediction. The challenges encountered in air quality predictions are where the strengths of deep learning distinctly apply. Air quality data typically includes multiple pollution indicators from various sources, characterized by complex nonlinear interactions and influenced by both temporal and spatial factors. Due to its robust feature extraction and nonlinear relationship learning capabilities, deep learning is a valuable tool for addressing the complexities of air quality forecasting—challenges that traditional forecasting methods often struggle to resolve effectively. In order to offer a thorough comprehension of this subject, the paper is organized as follows: Section 2 reviews the methodologies employed in recent studies on air quality prediction, detailing the various data sources, modeling techniques, and evaluation metrics utilized. Section 3 discusses the key findings from the literature, highlighting trends, challenges, and implications for public health and policy. In conclusion, Section 4 offers a summary of the paper's key findings as well as recommendations for future directions in the field of air quality forecasting research.

2 Method

2.1 Introduction of the machine learning workflow

The following components make up the typical machine learning workflow to forecast air quality: 1) Gathering important data affecting air quality from various sources, for instance satellite imagery, government monitoring stations, and mobile communication data. 2) Preparing and cleaning the data gathered. Due to air quality data can frequently be associated with meteorological data, additional preprocessing steps are required, especially for transforming meteorological data into a machine-readable format. 3) selecting a suitable machine learning model such as random forest and neural network according to the data's characteristics and the specific prediction mission. 4) Training the model selected on a subset of the collected data to learn underlying relationships and patterns. 5) Evaluating the precision and generalizability of the trained model through assessing its performance on a distinct portion of the data.

2.2 Traditional machine learning methods

2.2.1 Support vector regression (SVR) model

SVR is a machine learning regression method based on support vector machines. The regression problem's objective function is estimated by finding a hyperplane where the data points are as near to it as feasible.

SVR models were used by Castelli et al. in their research to forecast California's air quality, with an emphasis on AQI and pollutant concentrations [7]. The use of the Radial Basis Function (RBF) kernel effectively addressed nonlinear data relationships, resulting in enhanced model accuracy. This research relied on data from the U.S. Environmental

Protection Agency (EPA). However, this model requires extensive experimentation to find the optimal hyperparameter combination, making it both time-consuming and computationally demanding.

2.2.2 Extreme gradient boosting (XGBoost) model

XGBoost employs decision trees as the fundamental classifier to progressively improve the model performance. It is based on the Gradient Boosting framework. Its precision and efficiency make it highly effective in prediction tasks.

Kumar et al. employed the XGBoost model to predict the Air Quality Index (AQI) in 23 Indian cities [8], comparing it with other machine learning models to evaluate their performance across diverse regions. To address data imbalance, they utilized the Synthetic Minority Over-sampling Technique (SMOTE), which helped improve model accuracy. This method enhances model performance by augmenting the dataset with synthetic instances for minority classes. However, the study does not fully explore the influence of meteorological factors on air quality.

2.2.3 Random forest (RF) model

Due to its ability to handle high-dimensional data and avoid overfitting, the Random Forest model is a decision tree-based ensemble machine learning method.

Gariazzo et al. combined a Chemical Transport Model (CTM) with the Random Forest (RF) model to predict air pollutant concentrations in six Italian metropolitan areas and estimate population exposure using dynamic population data [9]. The model's predictive accuracy was enhanced through parameter tuning and the use of Geographic Information System (GIS) data to describe road networks in detail. Despite the increased accuracy from using GIS data, the study's limitation lies in its focus on only a few pollutants.

2.2.4 Categorical boosting (CatBoost) model

CatBoost is a machine learning algorithm known for its capability to handle categorical features automatically, making it particularly suitable for datasets with diverse feature types.

Ravindiran et al. utilized the CatBoost model for forecasting the AQI in the coastal city of Visakhapatnam [10], India, with a focus on evaluating the capacity for prediction of different machine learning models. Their research demonstrated the potential of the CatBoost model at the level of air quality forecasting.

2.2.5 Stacking ensemble model

Liang et al. employed a stacking ensemble model to predict the AQI in Taiwan and evaluated the model's predictive accuracy and robustness [11]. The innovation in this study is the use of stacking, where predictions from multiple models are used as new features to train an additional model, thereby enhancing the overall predictive accuracy.

2.3 Deep learning methods

2.3.1 Long short-term memory (LSTM) model

Due to it could capture temporal relationships, LSTM network is a sort of recurrent neural network that is ideally suited for sequence forecasting tasks.

Seng et al. developed a Multi-output Multi-indicator Supervised Learning (MMSL) model using LSTM neural networks to predict air quality [12]. The model integrates data from various monitoring stations alongside meteorological parameters to enhance prediction accuracy.

The innovation in this study involves using LSTM to build the MMSL model, which is a relatively novel application in the field of air quality prediction. Additionally, the integration of data from different monitoring stations with meteorological parameters improves the model's understanding and prediction of spatiotemporal variations in air quality.

2.3.2 Gated recurrent unit (GRU) model

The vanishing gradient problem and other drawbacks of conventional recurrent neural network (RNN) architecture are addressed by the GRU type of RNN architecture, which is intended to capture temporal dependencies in sequential data. GRUs have the capability to carry out flow of information via gating mechanisms to either selectively retain or discard information at each time step.

Espinosa et al. employed a multi-criteria methodology grounded in time series forecasting [13], utilizing a deep learning architecture known as the Gated Recurrent Unit (GRU) to analyze hourly nitrogen oxides (NOx) concentration data over a three-year period at a heavily trafficked intersection in Wrocław, Poland. The analysis considered the influence of various meteorological and traffic-related factors, as well as ozone (O3) levels, on the predictive modeling of NOx and nitrogen dioxide (NO2). A sliding window transformation is applied in conjunction with multi-index decision-making processes to identify the optimal model. Furthermore, a novel multi-index approach, which assesses models based on both accuracy and robustness criteria, is proposed for evaluating the prediction performance of various forecasting models. This methodological framework offers a more comprehensive assessment than traditional single indicator evaluation methods, thereby providing a more nuanced understanding of the models' real-world performance.

2.3.3 Convolutional neural networks (CNN) model

CNN is powerful for capturing spatial features, while LSTM networks excel in handling temporal data. Combining these architectures can enhance predictive performance when dealing with spatiotemporal data.

Bekkar et al. employed a CNN-LSTM model to predict PM2.5 concentrations in Beijing [14], achieving higher accuracy compared to traditional methods. The innovative aspect of this study is the combination of CNN and LSTM models, effectively utilizing the CNN's strength in extracting spatial features alongside the LSTM's proficiency in managing temporal sequences. Additionally, the research enhanced the predictability and interpretability of the model by integrating meteorological data, air quality data, and PM2.5 concentration data from nearby monitoring stations into the CNN-LSTM framework.

2.3.4 Hierarchical model

Abirami et al. introduced a hierarchical deep learning model named DL-Air, designed for air quality prediction [15]. The encoder encodes spatial relationships in the provided data, mapping them into an undiscovered space. The STAA-LSTM component discovers temporal and spatiotemporal relationships, forecasting future relations within this latent space. Finally, the decoder translates these predicted relationships into actual forecasts.

The DL-Air model excels in predictive accuracy, Additionally, its training efficiency makes it suitable for rapid prediction needs. The hierarchical structure and weight matrices

enhance interpretability, allowing for better insights into the model's decision process. When compared to traditional LSTM models, DL-Air demonstrates superior capability in capturing spatiotemporal data features, yielding more accurate predictions.

3 Discussion

3.1 Comparison of deep learning and machine learning-based prediction models

In the application of air quality prediction, machine learning models and deep learning models possess their own strengths and drawbacks respectively. Machine learning models demand limited data and provide stronger interpretability of prediction results, nevertheless due to the low complexity of the model, machine learning models may not be able to capture all the relationships in the data, which leads to lower prediction accuracy. Simultaneously, manual feature extraction and selection must be carried out, which is typically a time-consuming procedure.

In contrast to traditional machine learning models, deep learning architectures display the ability to autonomously extract features from unprocessed data, thereby diminishing the necessity for manual feature engineering. Furthermore, the sophisticated model architecture helps it more effectively grasp complicated nonlinear interactions in the data. despite this, poor interpretability causes things tricky for politicians to determine targeted emission reduction strategies or intervention measures. Consequently, in real-world applications, it frequently becomes crucial when selecting between machine learning and deep learning models.

3.2 Limitations and challenges

1) Difficulty in feature selection: Estimating air quality is a multifaceted, intricate, and cross-domain problem. The sources of pollutants that affect air quality are numerous as well as complex. Numerous variables, for instance time series, geographic location, meteorological data, and data for human activities, possess an impact on air quality.

Numerous variables, for instance time series, geographic location, meteorological data, and data for human activities, possess an impact on air quality. While deep learning is employed to automatically acquire characteristics from raw data, researchers cannot put all relevant features into the model, resulting in lower the model's predictive ability.

2) Lack of interpretability: Due to deep learning models' potent predictive power, an increasing number of researchers have used them to predict air quality in the past few years, with impressive results. However, deep learning models' interpretability is typically lacking. This is due to the reality of deep learning models typically consist of millions or hundreds of millions of parameters, plenty hidden layers, and neurons, rendering it challenging to decipher the model's internal decision-making process utilizing simple mathematical or logical reasoning.

3) Limitations in Model Applicability： The majority air quality prediction models in this field are constructed employing datasets that are limited to specific regions. These models may perform well in prediction within their original context, nevertheless this does not mean that other areas' distinct urban environments can simply apply the same techniques to them. Consequently, fresh modeling and parameter adjustments tend to be required in different areas, indicating poor portability. Furthermore, factors influencing air quality can vary over time and with environmental changes; for instance, new traffic policies, temporary industrial activities, and natural disasters can significantly impact the model's original predictive

capacity. With this reason, it might be crucial to update current models often in order to keep them functional. These issues ultimately affect the applicability and reusability of air quality prediction models.

3.3 Future prospects

1) Introduction of Feedback Mechanism: In view of the difficulty of feature selection, feedback mechanisms should be appropriately introduced by researchers. By continuously monitoring air quality, the monitoring data is compared with the model prediction results to determine the prediction accuracy of the model. Based on this feedback information, the effectiveness of the selected features is evaluated to input new features or delete irrelevant features.

2) Interpretability methods :Future research can incorporate different interpretability algorithms, including Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), to support air quality prediction models in response to the growing demand for interpretability in models. These algorithms assistance in the explanation of intricate model outputs, enabling researchers and managers to comprehend the process by which the model generates certain predictions.

3) Introduction of transfer learning technology :To address the challenge of limited model transferability, it is essential to incorporate advanced transfer learning technologies. For instance, the adoption of them facilitates the transfer of knowledge acquired by a model from one task to another related but distinct task. Additionally, domain adaptation technologies are particularly relevant in scenarios where there is a discrepancy in data distribution between two domains. By employing these two methodologies to facilitate the migration of models to other cities that exhibit similar characteristics yet possess different data distributions, substantial improvements in model generalization can be achieved.

4 Conclusion

This paper has carried out an extensive analysis of the several deep learning and machine learning techniques applied for air quality forecasting. The discussed methodologies include traditional machine learning techniques, for instance SVR and RF models, alongside advanced deep learning approaches, including MMSL and CNN-LSTM models. Through the analysis, it can be identified several challenges and limitations, including difficulties in feature selection, lack of model interpretability, and constraints related to model applicability across diverse regions.

This review presents valuable insights for researchers in the field of air quality forecasting, clarifying the benefits and drawbacks of the existing approaches and guiding future research directions. While this paper has thoroughly examined relevant methodologies, it should be acknowledged that this review did not explore certain emerging technologies and hybrid models that could further enhance predictive accuracy and interpretability. Thus, future research should aim to bridge these gaps by investigating innovative strategies that improve model performance while ensuring transparency, thereby fostering more effective air quality management interventions.

References

1. World Health Organization, Ambient (outdoor) air quality and health, (2022). https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

2. IPCC, Climate Change 2021: The Physical Science Basis, Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, (Cambridge University Press, 2021).
3. United Nations Environment Programme, Global Air Quality: An urgent call for action, United Nations Environment Programme. (2022), https://www.unep.org/resources/publication/global-air-quality-urgent-call-action
4. I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, (2016).
5. J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks, 61, 85-117 (2015).
6. M. Reichstein et al., Deep learning and process understanding for data-driven Earth system science, Nature, 566(7743), 195-204 (2019).
7. M. Castelli, F. M. Clemente, A. Popovič, S. Silva, L. Vanneschi, A machine learning approach to predict air quality in California, Complexity, 2020(1), 8049504 (2020).
8. K. Kumar, B. P. Pande, Air pollution prediction with machine learning: a case study of Indian cities, International Journal of Environmental Science and Technology, 20(5), 5333-5348 (2023).
9. C. Gariazzo, G. Carlino, C. Silibello, M. Renzi, S. Finardi, N. Pepe, et al., A multi-city air pollution population exposure study: Combined use of chemical-transport and random-Forest models with dynamic population data, Science of The Total Environment, 724, 138102 (2020).
10. G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, C. Sonne, Air quality prediction by machine learning models: A predictive study on the Indian coastal city of Visakhapatnam, Chemosphere, 338, 139518 (2023).
11. Y. C. Liang, Y. Maimury, A. H. L. Chen, J. R. C. Juarez, Machine learning-based prediction of air quality, Applied Sciences, 10(24), 9151 (2020).
12. D. Seng, Q. Zhang, X. Zhang, G. Chen, X. Chen, Spatiotemporal prediction of air quality based on LSTM neural network, Alexandria Engineering Journal, 60(2) (2021).
13. R. Espinosa, J. Palma, F. Jiménez, J. Kamińska, G. Sciavicco, E. Lucena-Sánchez, A time series forecasting-based multi-criteria methodology for air quality prediction, Applied Soft Computing, 113, 107850 (2021).
14. A. Bekkar, B. Hssina, S. Douzi, K. Douzi, Air pollution prediction in smart city: Deep learning approach, Journal of Big Data, 8, 1-21 (2021).
15. S. Abirami, P. Chitra, Regional air quality forecasting using spatiotemporal deep learning, Journal of Cleaner Production, 283, 125341 (2021).