International Journal of Geo-Information

MDPI

Article

A Novel Evolutionary Deep Learning Approach for $PM_{2.5}$ Prediction Using Remote Sensing and Spatial–Temporal Data: A Case Study of Tehran

Mehrdad Kaveh [1], Mohammad Saadi Mesgari [1] and Masoud Kaveh [2,*]

[1] Faculty of Geodesy and Geomatics, K. N. Toosi University of Technology, Tehran 19967-15433, Iran; m.kaveh11@email.kntu.ac.ir (M.K.); mesgari@kntu.ac.ir (M.S.M.)
[2] Department of Information and Communication Engineering, Aalto University, 02150 Espoo, Finland
* Correspondence: masoud.kaveh@aalto.fi

Abstract: Forecasting particulate matter with a diameter of 2.5 μm ($PM_{2.5}$) is critical due to its significant effects on both human health and the environment. While ground-based pollution measurement stations provide highly accurate $PM_{2.5}$ data, their limited number and geographic coverage present significant challenges. Recently, the use of aerosol optical depth (AOD) has emerged as a viable alternative for estimating $PM_{2.5}$ levels, offering a broader spatial coverage and higher resolution. Concurrently, long short-term memory (LSTM) models have shown considerable promise in enhancing air quality predictions, often outperforming other prediction techniques. To address these challenges, this study leverages geographic information systems (GIS), remote sensing (RS), and a hybrid LSTM architecture to predict $PM_{2.5}$ concentrations. Training LSTM models, however, is an NP-hard problem, with gradient-based methods facing limitations such as getting trapped in local minima, high computational costs, and the need for continuous objective functions. To overcome these issues, we propose integrating the novel orchard algorithm (OA) with LSTM to optimize air pollution forecasting. This paper utilizes meteorological data, topographical features, $PM_{2.5}$ pollution levels, and satellite imagery from the city of Tehran. Data preparation processes include noise reduction, spatial interpolation, and addressing missing data. The performance of the proposed OA-LSTM model is compared to five advanced machine learning (ML) algorithms. The proposed OA-LSTM model achieved the lowest root mean square error (RMSE) value of 3.01 $\mu g/m^3$ and the highest coefficient of determination (R^2) value of 0.88, underscoring its effectiveness compared to other models. This paper employs a binary OA method for sensitivity analysis, optimizing feature selection by minimizing prediction error while retaining critical predictors through a penalty-based objective function. The generated maps reveal higher $PM_{2.5}$ concentrations in autumn and winter compared to spring and summer, with northern and central areas showing the highest pollution levels.

Keywords: $PM_{2.5}$; aerosol optical depth; geographic information systems; remote sensing; long short-term memory; orchard algorithm

Academic Editors: Godwin Yeboah and Wolfgang Kainz

Received: 10 December 2024
Revised: 15 January 2025
Accepted: 22 January 2025
Published: 23 January 2025

Citation: Kaveh, M.; Mesgari, M.S.; Kaveh, M. A Novel Evolutionary Deep Learning Approach for $PM_{2.5}$ Prediction Using Remote Sensing and Spatial–Temporal Data: A Case Study of Tehran. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 42. https://doi.org/10.3390/ijgi14020042

1. Introduction

1.1. Interest of Studying the $PM_{2.5}$ Works

Particulate matter with a diameter of 2.5 μm ($PM_{2.5}$) is recognized as one of the most harmful components of air pollution due to its microscopic size and far-reaching implications for health, the environment, and the economy [1]. Unlike larger particles that

are often intercepted by the respiratory system's natural defenses, $PM_{2.5}$ can penetrate deep into the lungs and even enter the bloodstream, causing systemic health issues. Its tiny size allows it to remain airborne for extended periods, enabling widespread dispersion and making it a pervasive pollutant [2–4]. The growing urgency of addressing $PM_{2.5}$ pollution is underscored by its role as a leading contributor to the global burden of disease, particularly in urban areas where rapid industrialization, urban sprawl, and high population densities intensify the problem [5]. In cities like Tehran, $PM_{2.5}$ pollution is driven by a combination of vehicular emissions, industrial activities, construction dust, and natural phenomena such as dust storms. These factors, compounded by unfavorable meteorological conditions like temperature inversions, create conditions that trap pollutants near the ground, significantly degrading air quality [6–8].

The health impacts of $PM_{2.5}$ are profound and wide-ranging. These particles are known to cause cardiovascular and respiratory diseases, exacerbate asthma, and even contribute to the development of lung cancer [9]. Studies link long-term exposure to $PM_{2.5}$ to increased mortality rates, making it the 13th leading cause of death worldwide. Economically, this pollution results in higher healthcare costs and reduced labor productivity, imposing a significant burden on both families and national budgets [10,11]. $PM_{2.5}$ also plays a critical role in climate change by affecting the earth's radiation balance and cloud formation processes. Beyond health and economic concerns, $PM_{2.5}$ pollution disrupts ecosystems, damages vegetation, and diminishes the quality of life in urban areas, making it a multidimensional global challenge [12]. Furthermore, its social and psychological effects (such as heightened anxiety, decreased outdoor activities, and diminished urban livability) highlight its far-reaching consequences on community well-being [13].

Efforts to address $PM_{2.5}$ pollution have increasingly focused on leveraging technology to mitigate its effects and raise public awareness. Predictive modeling and mapping of air pollution through advanced techniques, such as deep learning (DL), geographic information systems (GIS), and remote sensing (RS), have become essential tools in combating this issue [14–16]. These technologies not only enable the development of high-resolution $PM_{2.5}$ concentration maps but also help identify pollution hotspots and underlying trends [17]. By forecasting $PM_{2.5}$ levels, policymakers can implement timely interventions, such as regulating industrial emissions or promoting cleaner transportation options. Moreover, sharing these maps with the public empowers citizens to make informed decisions, such as adjusting outdoor activities or adopting air purification measures [18]. This study seeks to contribute to these efforts by leveraging innovative approaches to predict and map $PM_{2.5}$ concentrations, offering a robust framework for tackling one of the most pressing environmental challenges of our time [19].

Today, ground-based air quality monitoring stations have become indispensable for accurately measuring the concentration of $PM_{2.5}$ pollutants in urban environments [20]. These stations provide reliable and high-resolution data on air quality. However, a significant limitation of these ground stations is their sparse distribution and limited spatial coverage. This uneven and insufficient network often results in data gaps, making it challenging to estimate $PM_{2.5}$ concentrations across larger, continuous areas [21]. Consequently, addressing air pollution over vast urban landscapes requires alternative methods that can bridge these gaps and complement ground-based observations. In recent years, aerosol optical depth (AOD) derived from satellite RS has emerged as a tool for estimating $PM_{2.5}$ concentrations, albeit with some limitations [22–24]. AOD, which measures the extent to which aerosols in the atmosphere prevent the transmission of sunlight through scattering and absorption, offers broad spatial and temporal coverage. With advancements in satellite technologies and the increasing accessibility of RS data, AOD provides a promising means to monitor air quality comprehensively [25].

1.2. Related Works

Studies have shown a significant spatial and temporal correlation between AOD and $PM_{2.5}$ concentrations, which can be exploited to develop predictive models for air pollution [25–27]. However, this relationship can be influenced by various meteorological parameters, such as relative humidity, wind speed, and temperature, as well as topographical features like elevation and urban density. These factors necessitate the integration of additional meteorological and spatial–temporal parameters alongside AOD to improve the accuracy of $PM_{2.5}$ predictions [28,29]. Satellite-derived parameters such as the normalized difference vegetation index (NDVI), land surface temperature (LST), and water vapor also contribute valuable insights into air quality dynamics [30–32]. These variables not only complement ground-level data but also provide the broader context necessary for understanding the complex interactions between atmospheric conditions and $PM_{2.5}$ pollution.

Nabavi et al. [24] conducted a comprehensive study on estimating $PM_{2.5}$ concentrations in Tehran, Iran, using AOD data and machine learning (ML) algorithms. The research leveraged high-resolution MAIAC AOD data (1 km) alongside DB_DT AOD data (10 km) retrieved from NASA's Terra and Aqua satellites to improve spatial coverage and resolution. Ground-level $PM_{2.5}$ data were collected from 18 air quality monitoring stations in Tehran between 2011 and 2016. Meteorological parameters, including planetary boundary layer height and relative humidity, were incorporated to adjust the AOD-$PM_{2.5}$ relationship, ensuring a greater alignment with real-world conditions. The study evaluated the performance of four ML algorithms. Among these, the random forest (RF) model exhibited the highest accuracy, achieving a root mean square error (RMSE) of 17.52 $\mu g/m^3$ and a coefficient of determination (R^2) of 0.68. The study identified limitations, including a reduced model performance during summer due to a lack of data on aerosol precursors and secondary formation processes. Additionally, the low spatial resolution of DB_DT AOD data in urban environments restricted its ability to accurately differentiate between bright surfaces, such as densely populated urban areas, roads, and other land features.

Bagheri [25] evaluated the potential of the deep ensemble forest method for estimating $PM_{2.5}$ concentrations in Tehran. The data for this study were collected from 2013 to 2019 and included daily $PM_{2.5}$ measurements (averaged from hourly data) from 23 air quality monitoring stations scattered across Tehran. Additional data incorporated into the study included temperature, dew point temperature, planetary boundary layer height, surface pressure, wind speed, wind direction, leaf area index, solar radiation, relative humidity, AOD, latitude, longitude, and the day of the year. The algorithm was trained using a grid search optimization process and 5-fold cross-validation. The results demonstrated that the proposed method achieved a higher accuracy (R^2 = 0.74) compared to other DL methods (R^2 = 0.67) and classical approaches like RFs (R^2 = 0.68). The study achieved notable advancements but faced certain limitations. The proposed algorithm was computationally intensive and heavily reliant on the quality of AOD and meteorological data. Periods of extreme pollution revealed inaccuracies, likely due to their underrepresentation in the training data. Additionally, the use of ERA5 meteorological data, with its coarse resolution, introduced potential discrepancies in localized predictions.

Bagheri [26] conducted a study to estimate $PM_{2.5}$ concentrations in Tehran using a combination of ground-based observations, satellite RS data, and ML techniques. The study utilized daily $PM_{2.5}$ measurements from 23 monitoring stations in Tehran, recorded between 2013 and 2019. Satellite-derived AOD data were retrieved at a 1 km resolution. Meteorological data were sourced from ERA5 and ERA5-Land datasets provided by ECMWF, including parameters such as 2 m temperature, dew point temperature, planetary boundary layer height, surface pressure, solar radiation, 10 m wind speed, relative humidity, wind

direction, and UV radiation. Temporal variables such as the month and day of the year were also incorporated to account for seasonal and daily variations. The best-performing extreme gradient boosting (XGBoost) model achieved an R^2 of 0.74 and an RMSE of 8.97 $\mu g/m^3$ on test data, outperforming other ML approaches such as RF and deep neural networks (DNNs). The study faced some limitations, such as the computational intensity of processing high-resolution data and the reliance on meteorological inputs with coarser resolutions, which introduced potential discrepancies in localized predictions. Furthermore, the model exhibited underperformance during periods of extreme pollution, likely due to their limited representation in the training dataset.

Faraji et al. [27] proposed an integrated spatiotemporal prediction model combining 3D convolutional neural networks (3D CNN) and gated recurrent units (GRU) for $PM_{2.5}$ concentration forecasting. Their study focused on Tehran, leveraging data from 13 Air Quality Monitoring stations, meteorological parameters, and additional auxiliary data collected from 2016 to 2019. The study reported a high R^2 (0.84) and RMSE (7.15 $\mu g/m^3$) for 3D CNN-GRU, outperforming traditional methods such as support vector regression (SVR), and even DL models like long short-term memory (LSTM) and GRU. Missing data and inconsistencies in auxiliary variables were noted as challenges that could hinder model performance. Despite these issues, the integration of CNN for spatial feature extraction and GRU for temporal forecasting provided a robust framework for air pollution prediction. Zamani Joharestani et al. [28] developed a study utilizing RF, XGBoost, and DNN to predict $PM_{2.5}$ concentrations in Tehran, Iran, for the period 2015–2018. Their methodology incorporated 23 features derived from various sources, including satellite data (AOD at 3 km and 10 km resolutions), meteorological data (e.g., temperature, wind speed, and relative humidity), and geographical information (latitude, longitude, and altitude). In addition to these, temporal variables like the day of the year and lag data for $PM_{2.5}$ and rainfall were included. Their results demonstrated that XGBoost performed the best, achieving an R^2 of 0.81 and an RMSE of 13.58 $\mu g/m^3$, particularly when excluding AOD03, which exhibited a high missing value rate (94%). A key limitation highlighted was the high rate of missing values for AOD03 and AOD10, leading to challenges in model reliability and resolution constraints in urban areas.

Handschuh et al. [29] utilized RF and the Copernicus atmosphere monitoring service (CAMS) model to predict $PM_{2.5}$ concentrations in Germany from 2018 to 2022. They incorporated 17 types of input variables, including elevation, land cover, convective available potential energy, dew point temperature, precipitation, water vapor, relative humidity, LST, wind speed, visibility, surface pressure, and AOD. The daily average $PM_{2.5}$ concentrations from all available monitoring stations across the study area were used, regardless of station type (industrial, traffic) or location (rural, urban). The RF model outperformed CAMS, achieving an R^2 of 0.71 compared to 0.57 for CAMS. The study also assessed model performance across different years, with 2020 yielding the worst results due to environmental changes caused by the COVID-19 lockdown. These changes were attributed to decreased aerosol loads resulting from reduced human activity. Additionally, the researchers suggested that the lower data quality during this period may have further impacted the model's performance. The importance of individual features was analyzed by their contribution to reducing model error. Non-physical variables, such as the day of the week and year, were identified as the most critical predictors due to strong seasonal trends in $PM_{2.5}$ concentrations linked to human activities and weather patterns. Among the physical predictors, AOD was the most influential, followed by LST.

Li et al. [30] employed a geo-intelligent approach incorporating geographical correlations into a deep belief network (Geo-DBN) to estimate $PM_{2.5}$ concentrations across China for the year 2015. The dataset included hourly $PM_{2.5}$ measurements from 1500 monitoring

stations, AOD, relative humidity, 2 m air temperature, 10 m wind speed, surface pressure, planetary boundary layer height, NDVI, population data, and road network data. Following preprocessing and data integration, a total of 65,647 records were compiled for model development. Based on validation results, the Geo-DBN model achieved superior daily prediction accuracy compared to other conventional models, with an R^2 of 0.88 and an RMSE of 13.3 $\mu g/m^3$. The spatial distribution of $PM_{2.5}$ predictions revealed that over 80% of China's population resides in areas where the annual mean $PM_{2.5}$ exceeds 35 $\mu g/m^3$. Interestingly, population and road network data had a minimal impact on the model performance. The authors attributed this to the inability of these predictors to capture temporal variations in the AOD- $PM_{2.5}$ relationship, as well as the overly coarse spatial resolution of these datasets, which limited their contribution. They suggested that incorporating fine-scale, real-time data, such as daily traffic flow at the street level, could significantly enhance model performance.

Chen et al. [32] predicted daily $PM_{2.5}$ concentrations across China from 2005 to 2016 using a RF model alongside two traditional regression models, all implemented with a spatial resolution of 10 km. Ground-level $PM_{2.5}$ data were collected daily from 1479 stations across China, covering more than 300 cities. Meteorological data, including daily mean temperature, relative humidity, air pressure, and wind speed, were obtained from 824 meteorological stations in China over the study period. For regions not covered by these meteorological stations, daily values of meteorological variables were interpolated using the Kriging method. In addition, the study utilized various datasets, including AOD, annual urban coverage data with a spatial resolution of 500 m, monthly mean NDVI data with a spatial resolution of 10 km, daily fire count data, and elevation data with a spatial resolution of approximately 90 m. The RF model significantly outperformed the two traditional regression models in terms of daily prediction accuracy, achieving R^2 = 0.83 and RMSE = 28.1 $\mu g/m^3$. Key predictors in the RF model included the day of the year, AOD, and daily temperature. The study also concluded that adding water coverage data did not enhance the final model's performance. Similarly, population data were excluded due to their high correlation with urban coverage data. Table 1 provides a comparative summary of key studies discussed in the Related Works section, offering a clear and concise overview of their methodologies, accuracy, and limitations. This table highlights the evolution of predictive models for air pollution forecasting and their application in various geographical contexts. It allows readers to quickly compare different approaches, their performance, and the specific challenges each study faced.

Table 1. Comparative summary of air pollution prediction models and their limitations.

Authors	Case Study	Model	R^2	Limitations
Zhang et al. [31], 2016	China	GWR	0.87	- Limited monitoring stations in certain areas - Unequal distribution of monitoring stations - Poor performance in high-altitude regions
Li et al. [30], 2017	China	Geo-DBN	0.88	- Prediction errors at high concentrations (60 $\mu g/m^3$) - Limited use of high-resolution spatial data (population and road) - Challenges with spatial and temporal errors

Table 1. *Cont.*

Authors	Case Study	Model	R^2	Limitations
Chen et al. [32], 2018	China	RF	0.83	- Lack of ground-based data before 2014 - Insufficient spatial data in certain regions - Limited temporal coverage of ground variables
Nabavi et al. [24], 2019	Iran	RF	0.68	- Limited monitoring stations and high error in northern Tehran - Performance in summer predictions and impact of missing data - Challenges in high surface reflectance areas
Zamani Joharestani et al. [28], 2019	Iran	XGBoost	0.81	- High rate of missing AOD data - Low spatial resolution of satellite data - Issues with ground-level data (missing data rate of $PM_{2.5}$)
Kianian et al. [18], 2021	US	Improved RF	0.65	- High rate of missing AOD data - Variable performance across regions - Temporal limitation (only on data from July 2011)
Bagheri [26], 2022	Iran	XGBoost	0.74	- High rate of missing AOD data - Lower accuracy in certain urban areas - Limited monitoring stations
Faraji et al. [27], 2022	Iran	3DCNN-GRU	0.84	- Variable performance in high pollution peaks - Complexity in input data processing - Impact of missing data
Bagheri [25], 2023	Iran	Deep RF	0.74	- Impact of missing data - Limited monitoring stations - Difficulty in modeling pollution peaks
Handschuh et al. [29], 2024	Germany	RF	0.71	- Data gaps in AOD - Poor performance at low pollution levels - Reduced accuracy during anomalous years (e.g., COVID-19)

The development of land use regression (LUR) models has advanced significantly over the past decade, driven by the integration of statistical and ML algorithms. Traditional LUR models, based on linear regression, struggled with modeling localized patterns and short-term temporal changes, particularly in urban areas. Recent studies have increasingly employed advanced ML algorithms such as RF, SVM, and ANNs to overcome these challenges. RF has enhanced predictive accuracy by capturing nonlinear relationships, while DNN and ANN excel in modeling high-dimensional, spatially distributed datasets, effectively detecting intricate air quality patterns for both regulated (e.g., $PM_{2.5}$) and unregulated pollutants. Ma et al. [33] emphasize that integrating ML algorithms into LUR frameworks not only enhances spatial and temporal resolution but also facilitates the incorporation of multi-source observations, such as satellite imagery, low-cost sensors, and mobile monitoring data. This integration allows for a better representation of localized pollution sources and dynamic environmental factors, such as meteorological conditions. For instance, hybrid models combining ML with RS data have proven effective in developing high-resolution air quality maps, enabling more accurate exposure assess-

ments in epidemiological studies. Moreover, the flexibility of ML algorithms in handling complex datasets has expanded the scope of LUR models beyond traditional pollutants. For example, novel predictor variables, including traffic volume, industrial emissions, and socioeconomic factors, are now seamlessly integrated into LUR models using ML techniques. These advancements have also improved the scalability of models, making it possible to develop regional, national, and even global air quality models.

1.3. Research Gaps and Motivations

Despite considerable advancements in air pollution prediction, several critical research gaps persist, offering substantial opportunities for further exploration and innovation. One prominent challenge lies in the reliance on traditional and commonly used ML and DL algorithms for air pollution prediction. These methodologies often fail to deliver the required accuracy for reliable air quality forecasting, particularly at finer spatial and temporal resolutions. However, studies that have focused on refining and optimizing these algorithms have demonstrated significant improvements in prediction accuracy and the generation of more precise pollution maps. This contrast highlights the critical need for further enhancement of predictive models to better capture the complex dynamics of $PM_{2.5}$ pollution and provide actionable insights for air quality management. Additionally, the lack of comprehensive comparisons among various ML and DL algorithms hinders a clear understanding of their respective strengths and limitations, leaving stakeholders without robust guidance for selecting the most appropriate modeling approaches. To address these challenges, we employ LSTM models in this study, leveraging their remarkable capabilities in handling sequential and temporal data. LSTM networks excel at capturing long-term dependencies, making them particularly suitable for forecasting $PM_{2.5}$ pollution levels, where temporal patterns and correlations play a crucial role [34,35]. Unlike traditional recurrent neural networks (RNNs), LSTMs are equipped with memory cells and gating mechanisms that allow them to selectively retain or forget information over multiple time steps [36,37]. This unique architecture empowers LSTMs to model the intricate temporal dynamics of $PM_{2.5}$ concentrations, influenced by meteorological conditions, human activities, and natural events. Their ability to process and analyze sequential data with high fidelity makes them a powerful choice for air quality prediction, ensuring better performance compared to conventional models.

LSTM networks have become essential in forecasting applications, especially for time-series data like air quality predictions. Traditional ML and simpler DL models, while useful in some cases, often struggle to capture the temporal dependencies and sequential patterns inherent in air pollution data. This is particularly problematic for long-term dependencies, where historical pollution levels and meteorological factors from the days or weeks prior can significantly influence current $PM_{2.5}$ concentrations. LSTMs address these challenges through their unique architecture, incorporating memory cells and gating mechanisms that retain important information while discarding irrelevant data. Unlike standard RNNs, which suffer from the vanishing gradient problem, LSTMs effectively propagate learning across extended time steps, ensuring critical historical trends are preserved [34]. This capability is crucial as air quality is influenced by both short-term fluctuations (e.g., weather changes and daily traffic) and long-term trends (e.g., seasonal variations). Additionally, LSTMs excel at capturing nonlinear temporal dependencies, which are common in air pollution dynamics, such as the complex interactions between meteorological factors like wind speed, humidity, and temperature, as well as human-induced emissions. Their robustness in handling noisy measurements, irregular sampling intervals, and missing data ensures reliable predictions, even in challenging datasets. This adaptability makes LSTMs highly effective for $PM_{2.5}$ forecasting [35–37].

Although LSTM networks offer significant strengths, they also exhibit certain limitations, particularly in the realm of hyper-parameter optimization. The performance of LSTMs heavily depends on the fine-tuning of key hyper-parameters such as the number of layers, the number of neurons in each layer, learning rates, and dropout rates. Moreover, the fully connected layers of LSTM networks involve weight and bias parameters that significantly impact the accuracy of predictions. Gradient-based learning methods, like stochastic gradient descent (SGD) or Adam, commonly used for training LSTMs, often face challenges such as getting trapped in local minima, high computational costs, and the need for continuous objective functions. Additionally, issues like vanishing or exploding gradients further complicate the optimization process, especially for deep architectures or long sequences. These limitations can hinder the model's ability to generalize effectively, particularly when faced with diverse and complex datasets [38,39]. To overcome these challenges, this paper introduces a novel meta-heuristic optimization algorithm known as the orchard algorithm (OA) [40]. This algorithm is specifically designed to enhance the performance of LSTM networks by optimizing their hyper-parameters and weight configurations. By incorporating the OA, the LSTM network is better equipped to find optimal parameter configurations, addressing the limitations of gradient-based methods and ensuring improved convergence and predictive accuracy. The integration of the OA with LSTM not only enhances the model's robustness but also reduces computational overhead, making it a promising solution for high-resolution and reliable $PM_{2.5}$ forecasting.

The integration of the OA with LSTM networks overcomes key limitations of traditional optimization methods, enhancing the model's ability to capture complex temporal and nonlinear patterns in $PM_{2.5}$ forecasting. Unlike gradient-based methods like SGD or Adam, which focus on local optimization, the OA employs a meta-heuristic global search strategy inspired by orchard dynamics. This approach avoids local minima and ensures convergence to a more optimal configuration of weights and biases. By dynamically balancing exploration and exploitation, the OA fine-tunes parameters such as weights and gate mechanisms, maximizing the LSTM's capacity to model long-term dependencies while mitigating challenges like vanishing gradients. Additionally, the OA enhances the robustness of LSTMs by efficiently navigating high-dimensional parameter spaces, enabling the network to better capture intricate dependencies in air quality data, such as interactions between meteorological factors and pollutant transport dynamics. This is particularly vital for accurate $PM_{2.5}$ predictions. Another advantage of the OA is its computational efficiency. Unlike traditional methods that require extensive iterations and computational resources, the OA achieves faster convergence with high accuracy, making the OA-LSTM framework both effective and resource-efficient for real-world air quality management systems. To ensure a comprehensive evaluation of the proposed approach, we implement and compare a diverse set of algorithms, including the OA-LSTM, LSTM, RNN, DNN, RF, and support vector machine (SVM). This combination of advanced DL and traditional ML models enables a thorough assessment of the strengths and limitations of each method, providing robust insights into their predictive capabilities for $PM_{2.5}$ forecasting.

Another critical research gap arises from the lack of comprehensive integration of spatial and temporal datasets in air quality studies. High-resolution (spatial and temporal) datasets, which are crucial for producing accurate and detailed air quality predictions, are often scarce or underutilized in current studies. Furthermore, the lack of robust preprocessing and noise reduction methods for input data significantly affects the performance of ML algorithms, impeding their predictive reliability. Comprehensive sensitivity analyses and in-depth studies into these relationships are crucial for uncovering key predictors and advancing the precision of $PM_{2.5}$ modeling. To address these challenges, we aim to collect a comprehensive dataset that integrates spatial, temporal, satellite-derived, and

ground-based parameters. By combining high-resolution and coarse-resolution data, we seek to capture a broad range of environmental and atmospheric factors influencing $PM_{2.5}$ levels, ensuring a holistic approach to modeling air pollution dynamics. We implement a robust data preprocessing pipeline to improve the quality and reliability of input data. This process includes noise reduction, the recovery of missing data, spatial interpolation, and data aggregation, ensuring consistency across datasets with varying spatial and temporal resolutions. These steps are essential for minimizing errors and enhancing the accuracy and dependability of predictive models. Furthermore, we will analyze the impact of each input feature on $PM_{2.5}$ concentrations through sensitivity analyses. This investigation will allow us to determine the most influential predictors and their interactions with meteorological, spatial, temporal, and satellite-derived variables.

1.4. Contributions

- This paper presents a novel approach that integrates satellite-derived, ground-based, spatial, and temporal data with a novel DL architecture to predict $PM_{2.5}$ concentrations, leveraging the complementary strengths of diverse data sources for enhanced accuracy and comprehensive air quality modeling;
- This paper proposes an enhanced OA to optimize the training of LSTM networks for $PM_{2.5}$ prediction. In the enhanced OA, a cutting operator has been introduced, inspired by horticultural propagation techniques;
- This paper employs robust preprocessing techniques, such as noise reduction, temporal and spatial interpolation, and data aggregation, to improve the input data quality. The OA-LSTM model is validated against LSTM, RNN, DNN, RF, and SVM using metrics like RMSE, R^2, standard deviation of prediction errors (SD_{error}) convergence trend, and computational complexity for comprehensive performance evaluation;
- This paper introduces a binary OA (BOA) for feature selection, designed to optimize $PM_{2.5}$ prediction by identifying the optimal feature set. The BOA minimizes the prediction error and input features while retaining critical predictors through a penalty-based objective function, ensuring accurate and efficient modeling;
- The results demonstrate that the proposed OA-LSTM model significantly outperforms other algorithms in predicting $PM_{2.5}$ concentrations, achieving superior accuracy and reliability. The integration of the enhanced OA optimizes the weight and bias configurations in the LSTM, improving the convergence speed, model efficiency, and overall predictive performance.

1.5. Organization

This paper is structured as follows: Section 2 covers the materials and methodology, including the study area, datasets, data preparation, and the proposed OA and OA-LSTM models. Section 3 presents the experimental results and discussion, offering a detailed analysis of model performance, feature selection outcomes, spatial–temporal predictions, and comparisons with other algorithms. Section 4 provides an in-depth discussion, including a comparison of the obtained results with related works, analysis of the model's accuracy across different $PM_{2.5}$ concentration ranges, and an evaluation of its performance in handling extreme pollution events. Section 5 concludes with key findings, limitations, and future directions.

2. Materials and Proposed Methodology

Figure 1 provides an overview of the research workflow, encompassing the entire process from data collection to results generation. The study begins with the collection of diverse datasets, including $PM_{2.5}$ data, satellite-based data, ground-based data, as well as

temporal and spatial parameters. These data undergo a comprehensive preparation phase, which includes noise reduction using the Savitzky–Golay filter, missing data recovery through spline interpolation, and spatial interpolation using the inverse distance weighting (IDW) method. Finally, the data are aggregated for input into predictive models. The predictive modeling phase incorporates traditional ML algorithms (e.g., SVM and RF), DL models (e.g., LSTM, RNN, and DNN), and the proposed OA-LSTM model, which integrates evolutionary optimization to enhance predictive accuracy. The results are evaluated using validation metrics such as RMSE and R^2, and outputs include $PM_{2.5}$ distribution maps, the identification of the best-performing model, and a sensitivity analysis. This workflow outlines the structure of the research, and detailed explanations of each data type, preprocessing step, and predictive model are provided in the following sections for clarity and depth.

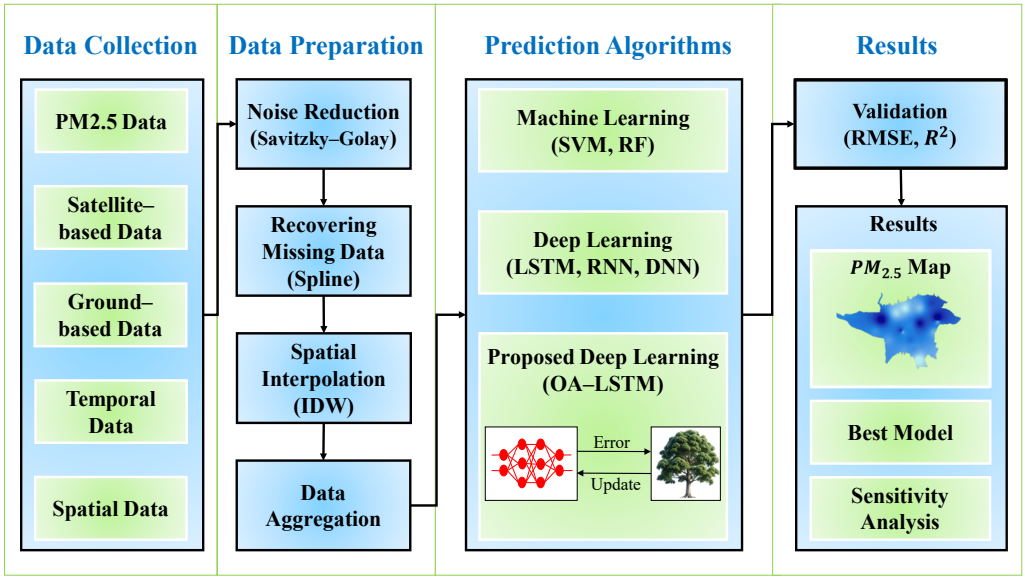


Figure 1. Workflow of the proposed model for $PM_{2.5}$ prediction.

2.1. Study Area

Tehran, the capital of Iran, is located in the northern part of the country, at the southern slopes of the Alborz Mountains. Its geographical position gives it a strategic advantage, serving as a central hub for economic, political, and cultural activities. The city's strategic importance extends to its role as the administrative and economic center of Iran, hosting numerous government institutions, businesses, and industries. Figure 2 illustrates the study area, including Tehran's location in Iran, its province boundaries, and a detailed map of Tehran city, showcasing pollution and meteorological stations along with elevation data. Tehran is the most populous city in Iran, with a population exceeding 8.5 million within the city and more than 15 million in its metropolitan area. Rapid urbanization over the past few decades has transformed Tehran into a sprawling metropolis. The city's development has been characterized by a significant expansion of infrastructure, housing projects, and commercial centers. However, this rapid growth has also led to challenges, including inadequate urban planning, congestion, and strain on public services. Tehran's dense population and urban sprawl have amplified environmental and social issues, particularly air pollution and traffic congestion.

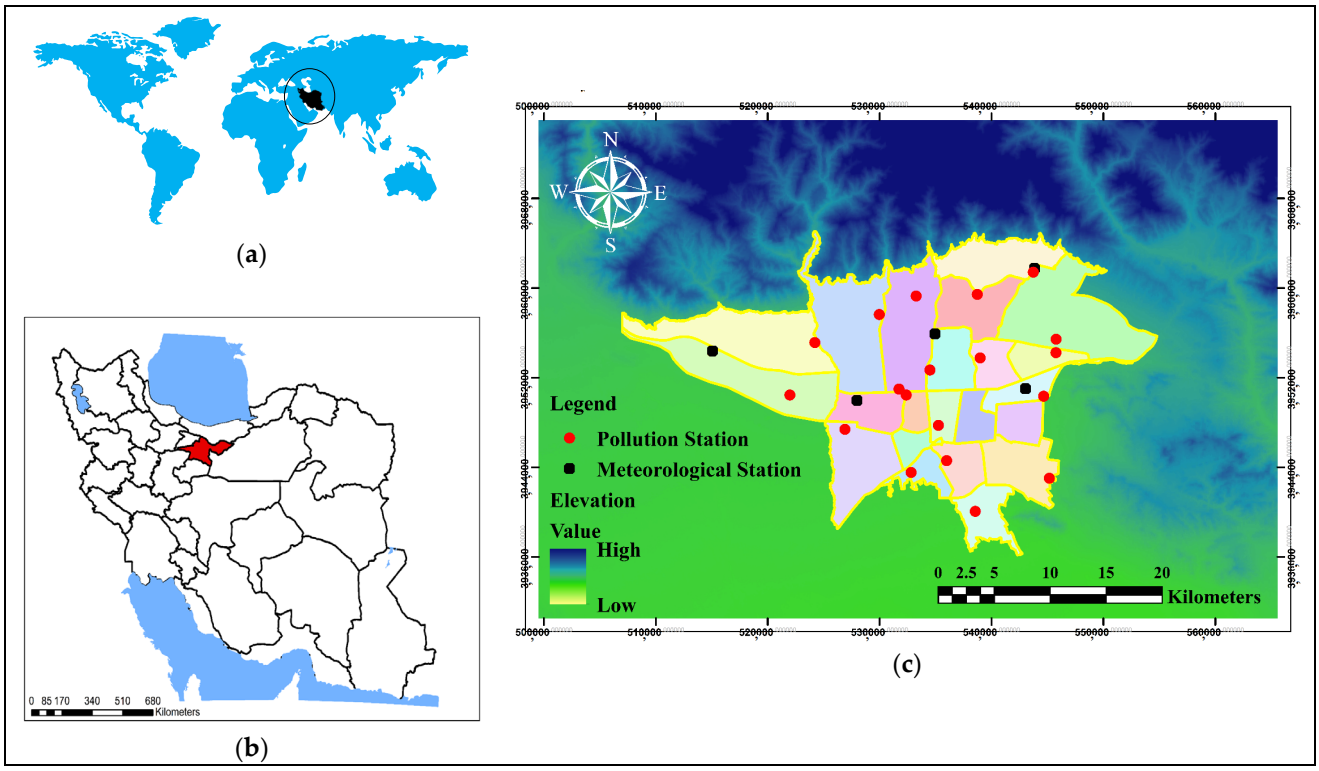


Figure 2. Overview of the study area: (**a**) global map highlighting Iran; (**b**) map of Tehran province indicating the study region; and (**c**) detailed map of Tehran city showing pollution and meteorological monitoring stations along with elevation data.

Tehran's climate is classified as semi-arid, with hot summers and cold winters. The city experiences significant variations in temperature due to its location between the Alborz Mountains and the central desert. Summers are typically dry, with temperatures reaching over 35 °C, while winters are cold, with occasional snowfall. Tehran's air quality is heavily influenced by its geographical features, as the surrounding mountains trap pollutants, particularly during temperature inversions in winter. The lack of sufficient rainfall and high levels of dust exacerbate the city's air pollution problems. Air pollution is one of the most critical environmental challenges faced by Tehran. The city consistently ranks among the most polluted urban areas in the world, with $PM_{2.5}$ and PM_{10} levels often exceeding international safety thresholds. The geographical and climatic conditions, such as the city's location in a basin surrounded by mountains, worsen the accumulation of pollutants. Air pollution in Tehran has serious health implications, contributing to respiratory diseases, cardiovascular problems, and premature deaths.

Industrial activities and traffic congestion are the primary sources of air pollution in Tehran. The city hosts numerous factories and industrial units, many of which rely on outdated technologies and emit significant amounts of pollutants. Additionally, Tehran's traffic is notorious for its intensity, with millions of vehicles on the road daily. A significant proportion of these vehicles are older models with poor emission standards, further exacerbating air quality issues. The combination of industrial emissions and vehicular pollution contributes to the high levels of $PM_{2.5}$ and other harmful pollutants in the city's atmosphere. Given Tehran's severe air pollution challenges, advanced methodologies are required to monitor and predict pollution levels effectively. The integration of RS and spatial–temporal data, combined with novel DL approaches, offers a promising solution. Such methods can provide accurate predictions of $PM_{2.5}$ concentrations, helping policymakers and urban planners implement targeted measures to reduce pollution and improve public health.

Tehran's unique geographical, climatic, and urban characteristics make it an ideal case study for developing and testing innovative approaches to air quality management.

2.2. Dataset and Data Preparation

The selection of variables in this study was carefully guided by a comprehensive review of previous research and the practical availability of data in Tehran. The parameters chosen reflect the multifaceted influences on $PM_{2.5}$ concentration, encompassing pollutant data, RS-derived variables, meteorological parameters, and spatial features. These parameters are integral to understanding the dynamics of $PM_{2.5}$ pollution and were selected based on their demonstrated significance in prior studies and their applicability to Tehran's unique environmental and urban conditions. Meteorological variables, such as temperature, wind speed, wind direction, atmospheric pressure, precipitation, water vapor, and humidity, significantly influence $PM_{2.5}$ dynamics by shaping atmospheric conditions that determine pollutant dispersion, transport, and formation. Temperature affects chemical reaction rates and atmospheric mixing. Wind speed and direction control the horizontal and vertical movement of pollutants, influencing their spatial distribution. Atmospheric pressure plays a role in stabilizing or destabilizing atmospheric layers, which impacts the behavior of particulate matter. Precipitation removes particulate matter from the atmosphere, while water vapor and humidity influence the formation and growth of particles through condensation and hygroscopic processes.

Spatial features, such as elevation and NDVI, are critical due to their influence on pollutant distribution and retention. Elevation impacts airflow patterns and the occurrence of temperature inversions, which can trap pollutants near the ground. NDVI indicates vegetation cover, providing insights into areas where green spaces may influence the interaction of $PM_{2.5}$ with the surrounding environment. RS-derived parameters, including AOD and LST, capture aerosol distributions and thermal conditions, providing broader spatial coverage and complementing localized ground data. AOD serves as a proxy for $PM_{2.5}$ by monitoring aerosol levels, while LST adds insights into surface temperature variations that affect pollutant transport and chemical transformations. Temporal factors, such as the day of the week and seasonal variations, influence $PM_{2.5}$ through changing urban activity patterns and climatic conditions. Traffic volumes, a major $PM_{2.5}$ source, vary with daily and weekly schedules, while seasonal phenomena like dust storms and temperature inversions drive fluctuations in pollution levels. These variables are essential for capturing the dynamic temporal trends of air quality in urban environments.

In this paper, ground-level air quality data for $PM_{2.5}$ concentration in Tehran were obtained from the Tehran Municipality Air Quality Control Company. Data collected daily from 19 selected monitoring stations over a three-year period (2014–2016) provide localized insights into pollution levels. This selection was based on the availability of consistent and reliable data within the study's temporal scope. Some stations were excluded due to insufficient data: three stations were established after 2016 and lacked samples for the target period, while one station had only a single recorded $PM_{2.5}$ sample during 2014–2016. Additionally, certain stations did not provide daily measurements, further limiting their utility. By focusing on these 19 stations, we ensured the integrity and consistency of the dataset, enabling a robust spatial and temporal analysis. The number of monitoring stations plays a crucial role in air quality studies, directly impacting spatial and temporal resolution. A higher density of stations improves the accuracy of pollution distribution mapping, capturing localized sources such as traffic and industrial emissions, especially in cities with significant spatial variability like Tehran. Temporally, more stations enhance data reliability by reducing the impact of gaps or inconsistencies from individual stations. However, logistical, financial, and technical constraints often limit the number of stations,

with placement influenced by accessibility, population density, and governmental priorities, leading to uneven coverage. To address these limitations, advanced modeling techniques (such as DLs) and supplementary data sources, such as satellite imagery, are essential for ensuring comprehensive and reliable air quality predictions.

In addition, meteorological data, including maximum temperature (Max Temp), minimum temperature (Min Temp), wind speed (WS), wind direction (WD), atmospheric pressure (P), and humidity (H), were sourced from the Tehran Meteorological Research Center. These data, collected daily from five stations, play a crucial role in understanding the relationship between weather conditions and $PM_{2.5}$ concentrations. Satellite data extracted through google earth engine provide a powerful tool for analyzing RS data. The satellite parameters used in this study include AOD, elevation (Ele), NDVI, LST, total precipitation (PE), and water vapor (WV). Each of these parameters was extracted through custom coding in the google earth engine, enabling a robust spatial–temporal analysis of $PM_{2.5}$ levels. These satellite-derived metrics complement ground-based data by offering broader spatial coverage and additional atmospheric insights.

Figure 3 presents the histograms and descriptive statistics of the variables used in the study, providing a comprehensive overview of their distribution across the dataset (S = 8106). Each histogram shows the frequency of observed values, highlighting variations in data distribution. The summary statistics (minimum, maximum, mean, and standard deviation) within each plot provide additional insights into the range and variability of these parameters. This visualization aids in understanding the data's characteristics, essential for predictive modeling and analysis. Table 2 provides a detailed overview of the variables included in the case study, outlining their units, and spatial and temporal resolutions. The dataset integrates diverse variables, such as $PM_{2.5}$ concentrations measured at stations with daily resolution, and RS-derived parameters like AOD (1 km, hourly). Vegetation indices such as NDVI are included with a 250 m spatial resolution and a 16-day temporal resolution, while meteorological variables like precipitation, water vapor, temperature, wind speed, and pressure are collected at station-level resolutions on a daily basis. This comprehensive dataset enables a robust spatial–temporal analysis of $PM_{2.5}$ dynamics.

Table 2. Details of each type of data in the case study.

Variables	Unit	Spatial Resolution	Temporal Resolution
$PM_{2.5}$	$\mu g/m^3$	Station	Daily
AOD	N/A	1 km	Hourly
LST	k	1 km	Daily
NDVI	N/A	250 m	16 Day
Total precipitation (PE)	mm/d	1 km	Daily
Water vapor (WV)	kg/m^2	1 km	Daily
Max Temp	°C	Station	Daily
Min Temp	°C	Station	Daily
Wind speed (WS)	m/s	Station	Daily
Wind direction (WD)	°	Station	Daily
Pressure (P)	hPa	Station	Daily
Humidity (H)	%	Station	Daily
Elevation (Ele)	m	90 m	N/A
Longitude	°	N/A	N/A
Latitude	°	N/A	N/A
Day of the week	day	N/A	N/A

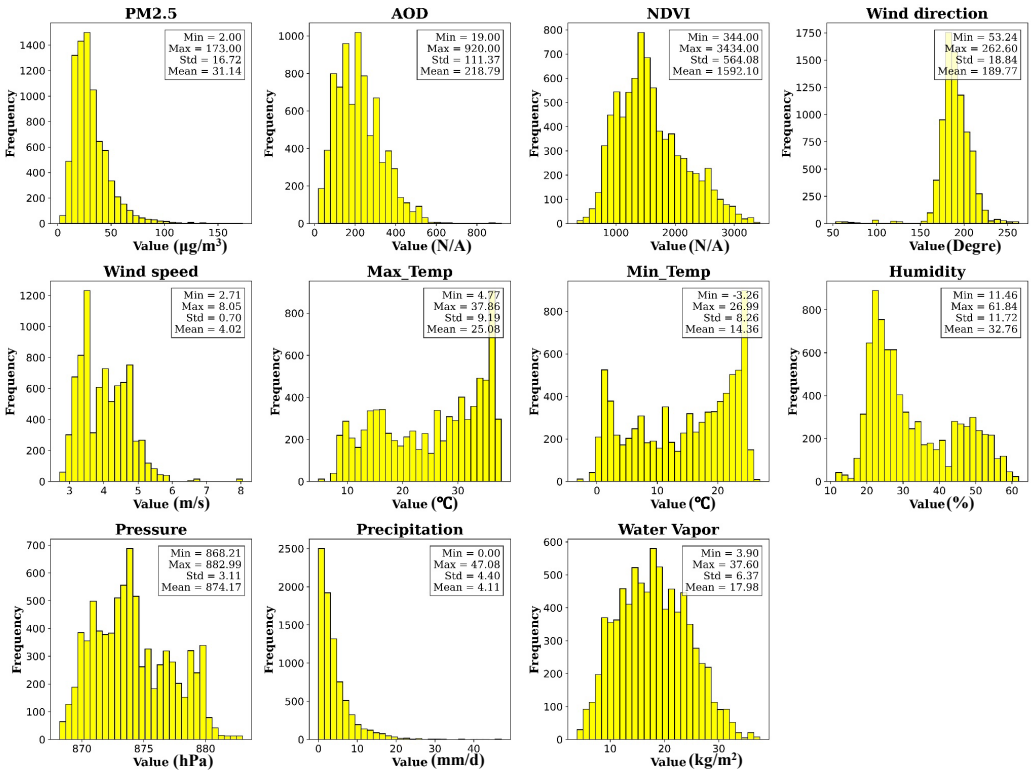


Figure 3. Histogram and summary statistics of variables in the sample dataset (S = 8106).

Data preparation and refinement are critical steps before inputting ground-based and satellite data into predictive models, as they directly influence the accuracy and reliability of the results. Ground-based data, such as $PM_{2.5}$ concentrations and meteorological measurements, often contain missing or inconsistent values due to sensor errors or station downtime, which must be addressed through imputation or filtering techniques. Similarly, satellite data, such as AOD and LST, can suffer from cloud contamination, spatial inconsistencies, or temporal gaps, necessitating preprocessing steps like resampling, interpolation, and noise removal. Proper data refinement ensures that the input dataset is complete, consistent, and representative of real-world conditions, reducing noise and biases that could degrade model performance. Additionally, harmonizing the spatial and temporal resolutions of ground and satellite data is essential for aligning the datasets and enabling effective spatial–temporal analysis. These steps collectively enhance the model's ability to learn patterns and make accurate predictions.

In this study, the training and testing datasets were created based on the availability of $PM_{2.5}$ data. Meteorological data were aligned with the specific days for which $PM_{2.5}$ concentrations were available. Since meteorological data were not consistently available for all days of the year, missing data were imputed using the spline model. However, spline functions alone cannot fit the data accurately due to the presence of noise and irregularities, which hinder proper curve fitting. To address this, a Savitzky–Golay filter was employed for noise reduction prior to data fitting. The Savitzky–Golay filter is one of the powerful methods for data smoothing and noise reduction, introduced by Savitzky and Golay in 1964. This filter is based on fitting low-degree polynomials to small subsets of data and is particularly suitable for preserving key signal features, such as peaks and valleys. While many traditional filters tend to remove details, the Savitzky–Golay filter is designed to smooth data without destroying these details. This filter performs a piecewise polynomial fitting to the data. In each small subset, the existing data are modeled with a polynomial (typically second- or third-degree). These polynomials are adjusted to fit the existing data

well, and their value at the central point of the window is used. This process is repeated for each data point to smooth the output signal.

For our analysis, the selection of the Savitzky–Golay filter parameters was guided by the need to balance noise reduction with the preservation of important signal features such as peaks and valleys. The window size and polynomial order were chosen based on the characteristics of the data and the level of noise observed during preliminary analysis. The window size defines the number of data points considered for fitting the polynomial in each segment. A larger window size results in smoother output but can oversmooth the data and remove critical details. Conversely, a smaller window size retains finer details but may leave some noise in the signal. Through trial and error, supported by visual inspections and quantitative error metrics, a window size of 7 was selected. This provided a good balance between smoothing and maintaining the fidelity of the underlying signal patterns. The polynomial order determines the complexity of the polynomial used to fit the data. Higher-order polynomials allow the filter to model more intricate variations but may also capture noise, leading to overfitting. Lower-order polynomials, on the other hand, simplify the signal and are more robust to noise but may fail to capture subtle trends. After testing various options, a polynomial order of 2 was chosen. This degree effectively captured the overall trends without overfitting the noise. The Savitzky–Golay filter was applied to the $PM_{2.5}$ time-series data to eliminate noise and abrupt fluctuations, as shown in Figure 4. The filter successfully removed irregularities without altering the inherent structure of the data. Unlike meteorological data, which required the use of the spline model for gap-filling, the $PM_{2.5}$ data used in this study were directly obtained from the Tehran Municipality Air Quality Control Company and did not require further interpolation.

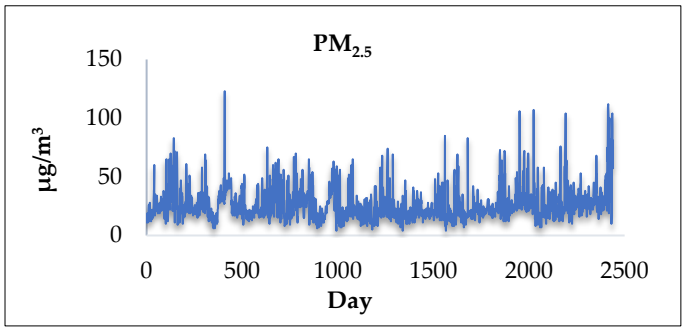


Figure 4. Time series of the refined $PM_{2.5}$ data.

Spline is a powerful technique used to impute missing values in time-series data by fitting smooth and continuous curves to the available data. The method works by dividing the dataset into smaller segments and fitting piecewise polynomials of low degrees to each segment. These polynomials are adjusted to ensure continuity in both the value and derivatives across segment boundaries, preserving the natural flow and structure of the data. This makes spline interpolation particularly suitable for time-series data, where maintaining temporal trends and patterns is critical. The implementation of spline interpolation begins with identifying the valid data points in the time series, excluding the missing values. Using these valid points, a spline curve is constructed to approximate the missing values. The selection of control points is a crucial aspect of the process, as too few knots may result in oversmoothing and a loss of detail, while too many knots can lead to overfitting and an unnecessarily complex curve. The balance is achieved by iterative testing, using visual inspections and error metrics to ensure the imputed values align with the overall data trends. Figure 5 illustrates the time series of AOD, comparing raw and refined datasets. Figure 5a represents the raw data, which contain missing values and noise due to inconsistencies and gaps in the original measurements. Figure 5b shows the refined data,

where noise has been reduced using the Savitzky–Golay filter, and missing values have been imputed using a spline interpolation method. This preprocessing ensures a smoother and more complete dataset for analysis. Figure 6 presents the time series of refined ground-based meteorological data. To prepare the data for algorithm implementation, it is essential to have complete meteorological parameters for all air pollution monitoring stations. Since meteorological measurements are not available for all locations, spatial interpolation is required to transfer information from meteorological stations to air pollution monitoring stations. In this paper, IDW interpolation was performed using ArcGIS 10.4.1 software to spatially align the meteorological data with the air quality monitoring stations, ensuring a consistent and structured dataset for modeling and analysis.

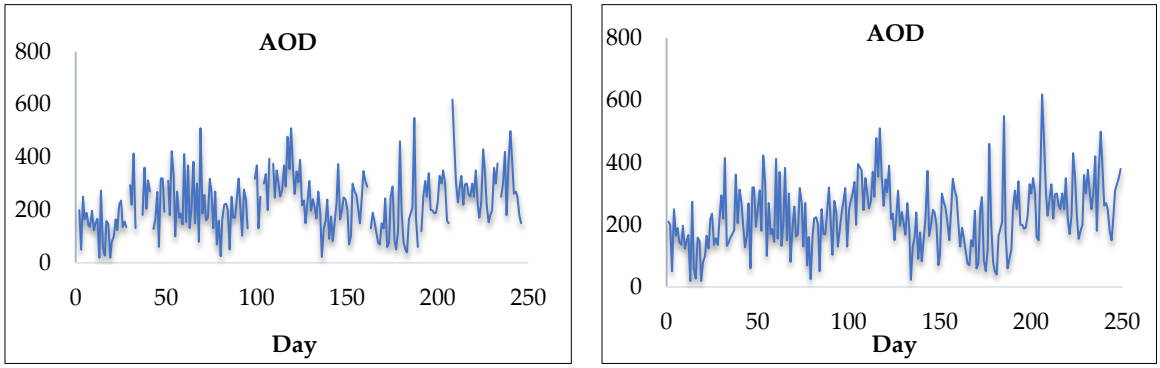


Figure 5. Time series of the AOD data: (**a**) raw data; (**b**) refined data.

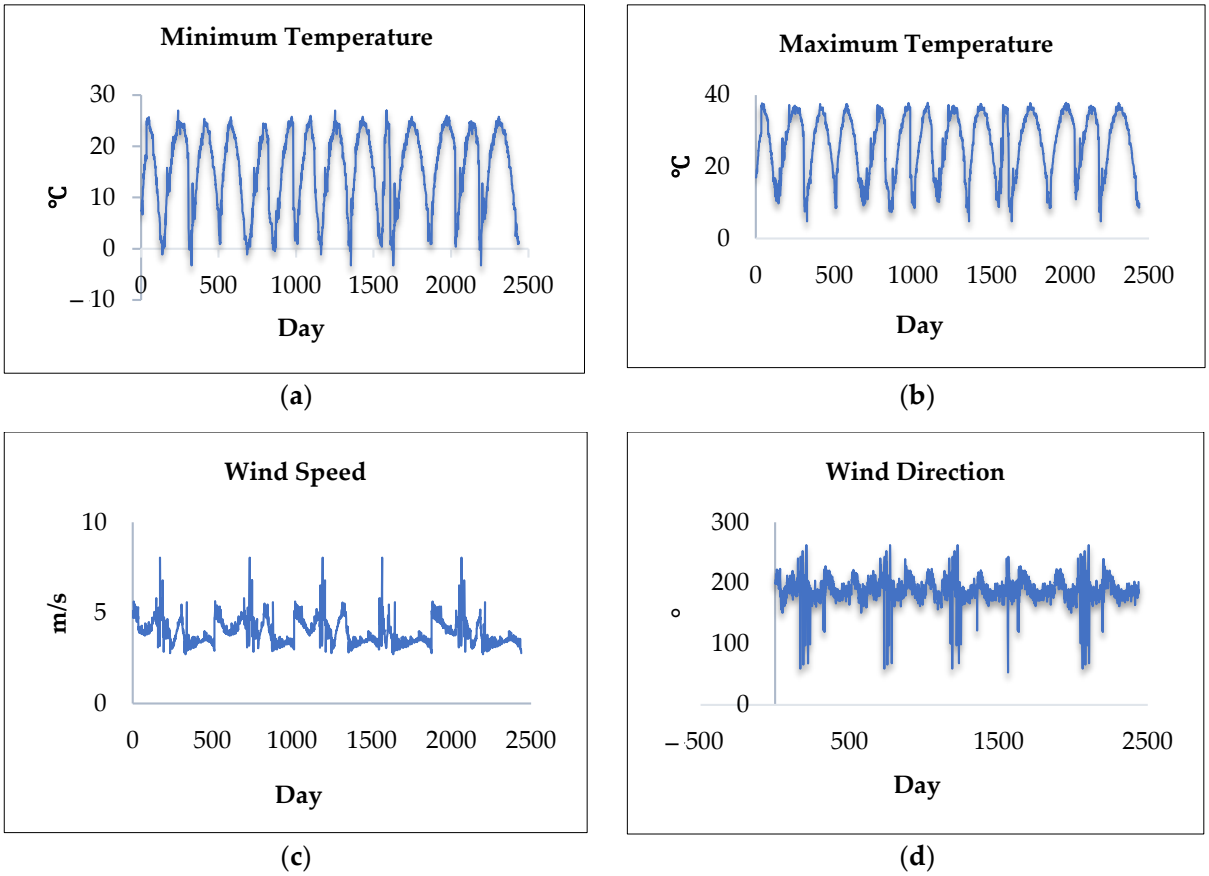
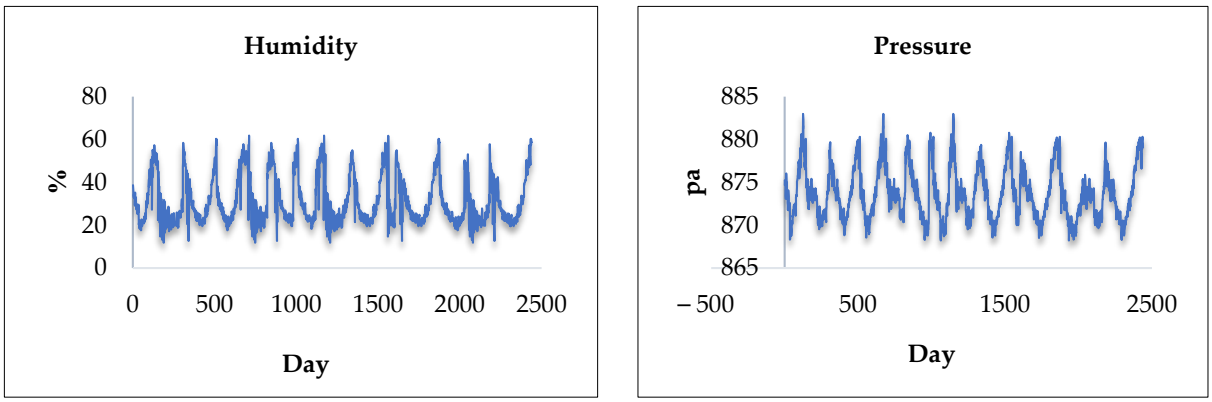


Figure 6. *Cont.*



(**e**) (**f**)

Figure 6. Time series of the refined ground-based meteorological data: (**a**) minimum temperature; (**b**) maximum temperature; (**c**) wind speed; (**d**) wind direction; (**e**) humidity; and (**f**) air pressure.

2.3. Proposed OA

The OA, introduced by Kaveh et al. [40] in 2023, is a novel meta-heuristic optimization approach inspired by the natural growth patterns and processes found in orchards. The OA simulates the behavior of trees growing, competing for resources, and optimizing their positions in an environment to maximize their access to sunlight, water, and nutrients. This nature-inspired approach makes the OA particularly effective for complex, high-dimensional optimization problems, as it focuses on the adaptive exploration and exploitation of the search space. In the OA, the optimization process begins with an initial population of trees, each representing a candidate solution in the search space. These trees grow and adjust their positions iteratively, aiming to improve their fitness values, which measure the quality of each candidate solution. The algorithm incorporates several operators (such as growth, screening, grafting, pruning, and elitism) to mimic the natural selection and growth process within an orchard. These operators enable the OA to refine the population of trees over time, enhancing the quality of solutions as the algorithm progresses [40]. The formulation of the OA is represented through Equations (1)–(7):

$$X_{growth} = X_{current} + \delta R, \tag{1}$$

$$F_j = \alpha f_j' + \beta g_j', \tag{2}$$

$$f_j' = \frac{f_j}{\max\limits_{1 \le j \le n} f_j}, \tag{3}$$

$$g_j = \sum_{i=m-l}^{m} \lambda_i \left(f_j^i - f_j^{i-1} \right), \tag{4}$$

$$g_j' = \frac{g_j}{\max\limits_{1 \le j \le n} g_j}, \tag{5}$$

$$X_{grafted} = \theta \cdot X_{strong} + (1 - \theta) \cdot X_{medium}, \tag{6}$$

$$X_{pruned} = Random(X_{min}, X_{max}) \tag{7}$$

where X_{growth} is the new solution generated through growth operator; $X_{current}$ is the current position of the candidate solution; δ is the growth factor; R is a random variable introducing variability in the direction of growth; F_j is the total objective function of j-th candidate; f_j is the objective function value of j-th candidate; f_j' is the normalized objective function value of j-th candidate; g_j is the growth rate of solution j; g_j' is the normalized growth rate of

solution j; α and β are weighting factors balancing the contributions of fitness and growth; g_j is the growth rate of each solution; m is the total number of growth years before the screening, l is the number of growth years before screening for which the growth rate is considered, and λ is the weight given to those years; $X_{grafted}$ is the new solution generated through grafting; X_{strong} is the position of the stronger candidate; X_{medium} is the position of the medium-quality candidate; θ is a blending coefficient determining the contribution of each parent; $X_{pruning}$ is the new randomly generated solution; and X_{min} and X_{max} are the bounds of the search space.

Equation (1) represents the growth operator, simulating the initial growth phase of trees. Each candidate solution $X_{current}$ adjusts its position based on a small perturbation defined by the growth factor δ and a random direction R. This allows local exploitation to identify better solutions in the vicinity. Equations (2)–(5) define the screening operator, and evaluate and rank candidate solutions based on their fitness and growth rate. Candidates with higher F_j values are considered stronger and are retained for further iterations, while weaker candidates are flagged for replacement or modification. Equation (6) models the grafting operator, where a new solution is generated by blending two parent candidates: a strong candidate and a medium-quality candidate. Equation (7) describes the replacement operator, where weak candidates are replaced by new random solutions within the defined bounds. This introduces fresh diversity into the population, preventing stagnation in local optima.

The standard OA, while effective in balancing exploration and exploitation through its nature-inspired operators, faces certain limitations that can affect its performance in complex, high-dimensional optimization problems. These challenges arise primarily from its reliance on a fixed set of operators and its inability to dynamically adapt to diverse problem landscapes, which can lead to premature convergence or inefficient exploration. One of the main weaknesses of the standard OA is its potential to stagnate in local optima. Although operators like grafting and replacement introduce diversity, they may not be sufficient to escape from local optima in highly rugged or deceptive fitness landscapes. This limitation is particularly pronounced in problems with a large number of local minima, where the algorithm's exploration mechanisms might fail to effectively cover the entire search space. Another notable issue is the algorithm's lack of targeted refinement for individual candidate solutions. The standard OA focuses on improving solutions as a whole, often neglecting the fine-tuning of specific components (genes) within each solution. This can lead to suboptimal performance, especially in cases where only a subset of the solution's parameters requires adjustment. Additionally, the current operators may not fully utilize the potential of high-quality solutions, as they primarily focus on combining or replacing entire solutions rather than selectively enhancing specific attributes. These limitations highlight the need for introducing more adaptive and granular operators, such as the cutting operator, which can address the weaknesses by targeting individual genes for improvement. This not only enhances the algorithm's ability to escape local optima but also enables a more focused exploitation of strong candidate solutions, leading to better convergence and overall performance.

In horticulture, cutting is a widely used propagation technique where a part of the parent plant is cut and cultivated independently to develop roots and grow into a new plant. This method is particularly effective for plants with high rooting potential. Inspired by this, cutting can be introduced as an operator in the OA, where a portion of a strong candidate solution is retained, and the rest is regenerated randomly. This operator allows the algorithm to leverage the strengths of high-quality solutions while introducing diversity by replacing weaker components. The cutting operator involves selecting a strong candidate solution (based on fitness) and dividing it into two parts: a portion of the strong

solution is retained as a "cutting" to preserve its high-quality characteristics; the remaining portion is regenerated randomly to explore new areas in the search space. This hybridization of exploitation (using the strong solution) and exploration (introducing randomness) enhances the algorithm's ability to refine its solutions effectively. The cutting operator can be formulated as Equation (8):

$$X_{cut}[d] = \begin{cases} X_{strong}[d] & if\ d \in SelectedIndices \\ Random(X_{min},\ X_{max}) & if\ d \notin SelectedIndices \end{cases} \quad (8)$$

where $X_{cut}[d]$ is the d-th gene of the new solution after cutting; $X_{strong}[d]$ is the d-th gene of the strong candidate solution; $SelectedIndices$ is a subset of indices corresponding to the retained portion of the strong solution; and $Random(X_{min},\ X_{max})$ is a randomly generated value within the bounds of the search space.

The cutting operator works by first selecting a strong candidate solution X_{strong} based on its fitness value. The solution's genes are then divided into two parts: a portion of the genes, determined by $SelectedIndices$, is retained directly from X_{strong}, while the remaining genes are replaced with random values to introduce diversity. Finally, the retained and randomized genes are combined to form a new candidate solution X_{cut}. This operator enhances exploitation by preserving high-quality features from strong solutions, while also improving exploration by introducing randomized genes, preventing premature convergence and enabling the algorithm to search new areas of the solution space effectively. By incorporating the cutting operator, the OA gains an additional mechanism to refine solutions, enabling it to converge more effectively while maintaining diversity in the search process. Algorithm 1 presents the pseudo-code of the proposed OA.

Figure 7 illustrates the application of the cutting operator within the proposed OA. In this example, a strong candidate solution $X_{strong} = [x_1,\ x_2, x_3, x_4, x_5]$ with an RMSE value of 0.5 $\mu g/m^3$ is selected for improvement. The $SelectedIndices$ are defined as $\{x_1, x_4, x_5\}$, meaning these genes are retained directly from X_{strong}. The remaining genes $\{x_2, x_3\}$ are replaced with random values $\{x_6, x_7\}$ generated within the search space bounds. The resulting new solution $X_{cut} = [x_1,\ x_6, x_7, x_4, x_5]$ demonstrates an improved performance, as indicated by its lower RMSE value of 0.2 $\mu g/m^3$. This example highlights how the cutting operator preserves high-quality components while introducing variability to enhance exploration and improve overall solution quality.

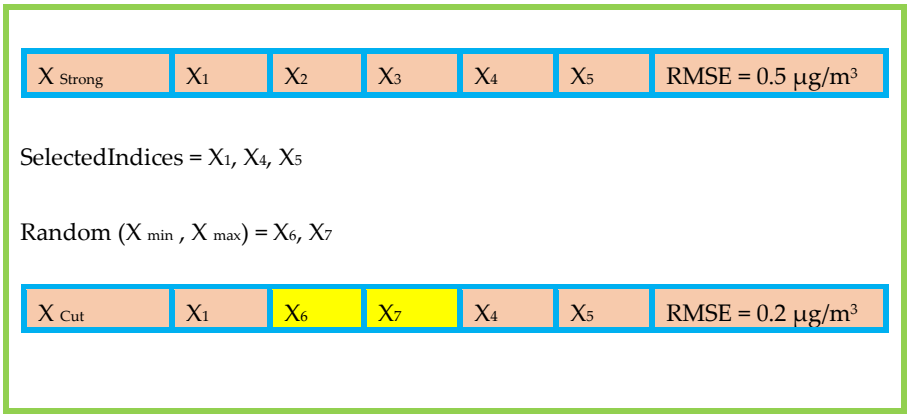


Figure 7. Example of cutting operator in proposed OA.

Algorithm 1 Pseudo-codes of proposed OA

Begin OA
  %%Parameter setting
  Initialization of population size, number of strong/medium/weak trees, α, β, Iteration
  %%Create population
  for n = 1 to population size do
    Create orchard (population)
    Calculate the fitness function $\left(f_j\right)$
  end
  %%Main loop
  for i = 1 to Maximum iteration do
    %% Elitism
    Sort population
    Save elite population
    %% Growth
    for j = 1 to population size do
      $X_{growth} = X_{current} + \delta R$
      Calculate the fitness function $\left(f_j\right)$
    end
    %% Screening
    Save previous populations to calculate growth rate
    for j = 1 to population size do
      Calculate the normalized fitness function $\left(f_j'\right)$ based Equation (3)
      Calculate the normalized growth rate $\left(g_j'\right)$ based Equation (5)
      Calculate the total objective function $\left(F_j\right)$ based Equation (2)
      Divide the seedlings into three groups: strong, medium, and weak
    end
    %% Grafting
    for j = 1 to population size do
      $X_{grafted} = \theta.X_{strong} + (1-\theta).X_{medium}$
      Calculate the fitness function $\left(f_j\right)$
    end
    %% Pruning
    for p = 1 to population size do
      $X_{pruned} = Random(X_{min},\ X_{max})$
      Calculate the fitness function $\left(f_j\right)$
    end
    %% Cutting
    for j = 1 to population size do
      $$X_{cut}[d] = \begin{cases} X_{strong}[d] & if\ d \in SelectedIndices \\ Random(X_{min},\ X_{max}) & if\ d \notin SelectedIndices \end{cases}$$
      Calculate the fitness function $\left(f_j\right)$
    end
    Sort the population
    Show the best solution
  end
End OA

2.4. Proposed OA-LSTM

LSTM networks, introduced by Hochreiter and Schmidhuber in 1997, are a type of RNN designed specifically to handle long-term dependencies in sequential data [34]. LSTMs were developed in response to the limitations of traditional RNNs, which struggle to retain information over extended sequences due to the vanishing gradient problem. This phenomenon results in the exponential decay of gradients during backpropagation, making it challenging for RNNs to learn relationships over long time steps. As a result, standard RNNs often fail in tasks requiring long-term memory, such as language modeling, time-series forecasting, and speech recognition. LSTMs are widely used in applications requiring understanding dependencies across different time steps. Their design, featuring a memory cell and gated structures, allows them to selectively retain relevant information over time, which enables better performance in tasks involving complex sequential patterns. This capability has made LSTMs particularly popular in fields such as time-series analysis, machine translation, and speech-to-text systems, where the ability to capture long-term dependencies significantly enhances the model effectiveness. In a time-series analysis, sequences of data points are often influenced by previous values, making it essential to retain relevant historical information across extended periods. Traditional neural networks and simple RNNs struggle with this requirement due to the vanishing gradient problem, which limits their capacity to remember long-term dependencies. LSTMs, with their gated memory mechanisms, allow models to selectively retain or forget information at each time step, facilitating more accurate predictions by capturing essential temporal patterns [35].

LSTM, like a traditional RNN, is structured as a sequential chain where each cell passes information to the next. Figure 8 illustrates this chain-like structure, where each LSTM cell sequentially passes information to the next. The LSTM network architecture incorporates a cell state c_t, which acts as a memory unit capable of storing long-term information, along with a hidden state h_t that reflects the short-term memory for each time step. In Figure 8, each LSTM cell receives an input x_t, along with the previous cell state c_{t-1} and hidden state h_{t-1}, then outputs an updated cell state c_t and hidden state h_t to the next cell. This sequential structure helps the LSTM retain relevant information across multiple time steps, enabling it to capture complex patterns in sequential data [36].

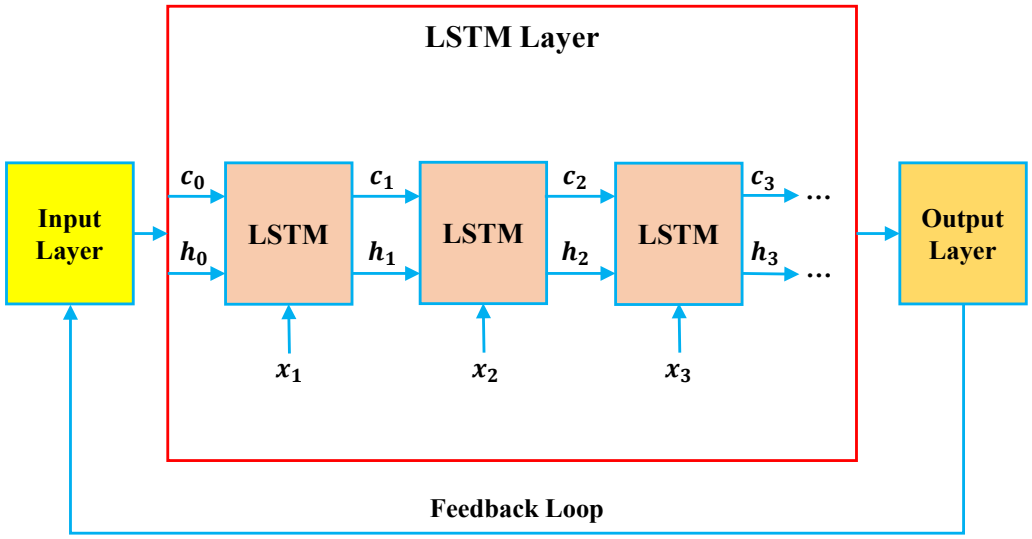


Figure 8. The chain-like structure of the standard LSTM.

Information flow in LSTM is regulated by three key gates: the forget gate, input gate, and output gate. These gates are controlled by trainable weights and biases, allowing the network to retain essential information, discard irrelevant data, and update the cell and

hidden states appropriately. The forget gate is responsible for deciding which information from the previous cell state is no longer relevant and should be "forgotten". This gate enables the LSTM to filter out unnecessary information, ensuring that the cell state remains focused only on the essential data as it progresses through the sequence. By dynamically choosing what to forget, the LSTM prevents irrelevant or outdated information from cluttering the memory, which is particularly useful in long sequences where early inputs may lose significance over time. The input gate determines what new information should be added to the cell state. It evaluates the importance of the current input in the context of the sequence and selectively incorporates it into the memory. This mechanism allows the LSTM to update its memory in a controlled manner, only adding relevant new information that complements the existing context. The input gate, therefore, plays a key role in refining the memory by carefully integrating new data with past knowledge, enhancing the LSTM's ability to capture meaningful patterns [37].

The output gate manages what part of the cell state should be exposed as the hidden state, which serves as the output of the LSTM cell for the current time step. This gate decides how much of the memory should be made available to subsequent cells or layers, balancing the cell's internal state with the need to communicate relevant information. By controlling the output, the LSTM effectively shares only the essential information with the next layer, enabling better learning in deep architectures and sequential processing tasks. In each time step, the cell state and hidden state are updated based on the operations of these gates. The cell state acts as a long-term memory reservoir, retaining crucial information across multiple time steps, while the hidden state functions as a short-term memory that changes at each step to reflect the immediate context. The combination of these two states allows the LSTM to remember relevant information from the past while dynamically adjusting to new inputs, making it adept at handling complex, long-term dependencies in sequential data. Equations (9)–(15) represent the internal computations of an LSTM cell, detailing how information flows through the forget, input, and output gates to update the cell state and hidden state at each time step [22].

$$f_t = \sigma\Big(W_{hf}h_{t-1} + W_{if}x_t + B_{hf} + B_{if}\Big), \tag{9}$$

$$g_t = tanh\Big(W_{hg}h_{t-1} + W_{ig}x_t + B_{hg} + B_{ig}\Big), \tag{10}$$

$$i_t = \sigma(W_{hi}h_{t-1} + W_{ii}x_t + B_{hi} + B_{ii}), \tag{11}$$

$$o_t = \sigma(W_{ho}h_{t-1} + W_{io}x_t + B_{ho} + B_{io}), \tag{12}$$

$$g\prime_t = g_t \odot i_t, \tag{13}$$

$$c_t = (f_t \odot c_{t-1}) + g\prime_t, \tag{14}$$

$$h_t = o_t \odot tanh(c_t) \tag{15}$$

where f_t represents the forget gate; g_t represents the candidate generation; i_t denotes the input gate; o_t represents the output gate; $g\prime_t$ is the modulated candidate state; c_t is the updated cell state that combines retained information from c_{t-1} and new information from $g\prime_t$; h_t is the updated hidden state; W_{hf}, W_{hg}, W_{hi}, W_{ho}, W_{if}, W_{ig}, W_{ii}, W_{io} are weight matrices corresponding to the previous hidden state h_{t-1} and current input x_t for each gate; $B_{hf}, B_{hg}, B_{hi}, B_{ho}, B_{if}$, B_{ig}, B_{ii}, B_{io} are biases terms associated with each gate; σ is the sigmoid activation function, outputting values between 0 and 1; $tanh$ is the hyperbolic tangent function, outputting values between -1 and 1; and $\odot$ is the element-wise multiplication operator, which modulates the interaction between the gates and states.

LSTM networks, despite their strengths, face notable challenges in terms of hyper-parameter optimization, which significantly affects their performance and generalization capabilities. Key hyper-parameters in LSTM models include the weights and biases within the fully connected layers, the number of layers and neurons, the learning rate, and the dropout rate. Optimizing these parameters is critical because each plays a specific role in controlling the behavior of the network. For instance, weights and biases directly influence how input data is transformed as it flows through the network, impacting how well the LSTM captures patterns and dependencies. The learning rate controls the step size in the optimization process; if set too high, it can lead to divergence, while a low learning rate may cause slow convergence and increase the training time.

One of the main challenges with optimizing LSTM networks is the use of gradient-based learning algorithms, such as SGD and Adam. While effective, these methods can be prone to local minima and saddle points, especially in high-dimensional spaces like those encountered in deep LSTM architectures. Additionally, gradient-based methods often struggle to adapt dynamically to the complex landscape of non-convex loss surfaces, which are common in neural networks. The reliance on the gradient descent may also result in issues like vanishing or exploding gradients, further complicating the learning process for LSTM networks, particularly when the model depth or sequence length increases. The optimization of the layer depth and neuron count is another crucial aspect of LSTM configuration. Selecting the right number of layers and neurons is vital to balance model complexity with computational efficiency. An excessive number of layers or neurons can lead to overfitting, where the model performs well on training data but poorly on unseen data. Conversely, too few neurons or layers may hinder the model's ability to capture important patterns, reducing its accuracy on complex sequences. This trade-off highlights the need for effective hyper-parameter tuning to achieve optimal network architecture for a specific task. Furthermore, LSTMs often require the careful tuning of additional parameters, such as the batch size and dropout rates, which influence how well the network generalizes to new data. Together, these hyper-parameters must be finely adjusted to enable the network to learn effectively, avoid overfitting, and maintain computational efficiency.

Given the complexity of LSTM optimization, meta-heuristic algorithms have shown promise in effectively navigating the high-dimensional search space of hyper-parameters. Meta-heuristic approaches, such as genetic algorithms (GAs), particle swarm optimization (PSO), and ant colony optimization (ACO), have proven effective for LSTM training, offering a robust alternative to traditional gradient-based methods. These algorithms can escape local optima and better handle the non-convex optimization landscape of deep networks. In this study, we propose using a novel meta-heuristic called the improved OA to train the LSTM network and optimize its hyperparameters. By applying the OA within the LSTM network, the model can better navigate the high-dimensional parameter space, avoiding issues such as local minima that commonly affect gradient-based optimization methods. This is especially useful in LSTM networks, where the non-convexity of the loss surface can hinder standard optimization techniques like SGD or Adam.

Figure 9 illustrates the proposed architecture, referred to as OA-LSTM, where the standard LSTM network is enhanced using the OA as an optimizer. This approach integrates the OA into the LSTM structure to optimize key hyper-parameters, such as weights and biases, throughout the learning process. The optimizer module, depicted in the figure, dynamically updates these parameters by minimizing errors in each time step, ultimately leading to an improved performance across sequential data tasks. One of the key advantages of using the OA in the LSTM architecture is its ability to balance exploration and exploitation. The OA uses an adaptive approach, allowing the optimizer to explore new parameter spaces when necessary while focusing on fine-tuning existing solutions to

improve convergence. This flexibility helps the LSTM network achieve a more robust and globally optimized set of parameters, ultimately improving accuracy and generalization in tasks such as time-series forecasting, speech recognition, and natural language processing. Furthermore, the OA provides enhanced stability in the optimization process, reducing the likelihood of issues such as vanishing or exploding gradients. By using a global search strategy, the OA can dynamically adjust parameters to maintain stable learning, even as the LSTM model depth and sequence length increase. This stability is crucial for training deeper LSTM architectures that capture more complex temporal dependencies without encountering the computational challenges typical of gradient-based optimizers.

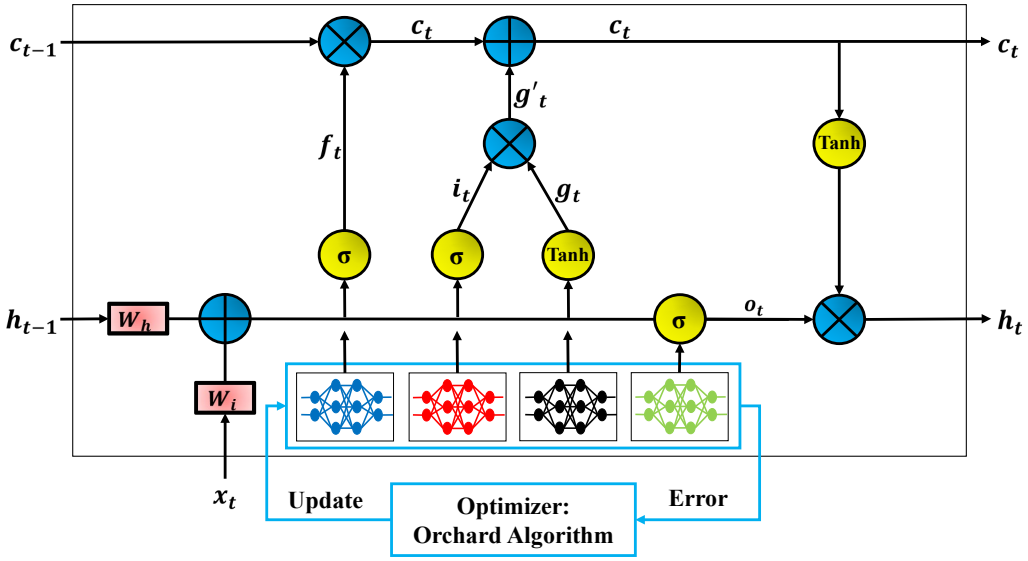


Figure 9. The structure of the proposed OA-LSTM.

Figure 10 demonstrates the implementation of the OA operators for optimizing the weights and biases of an LSTM network. The process begins with the initial population, where each candidate solution contains a set of values representing potential weights and biases, alongside an RMSE value indicating the quality of the solution. The growth phase follows, where the algorithm applies the growth operator to explore the local neighborhood of each seedling by slightly perturbing their values. This results in an updated set of solutions with improved RMSE values, such as 0.5, 0.6, and 0.4 $\mu g/m^3$, demonstrating the effectiveness of local search in improving candidate solutions. Next, the screening phase categorizes the grown seedlings into strong, medium, and weak categories based on their RMSE values. In the final phase, specific operators such as grafting, cutting, and pruning are applied to refine the solutions further. The grafting operator combines features from strong and medium seedlings to generate a new solution with a significantly lower RMSE value (e.g., 0.1 $\mu g/m^3$). Similarly, the cutting operator retains certain parts of a medium seedling and randomizes others, producing a solution with an RMSE of 0.2 $\mu g/m^3$. The pruning operator adjusts specific genes in weak seedlings, resulting in an improved RMSE of 0.3 $\mu g/m^3$.

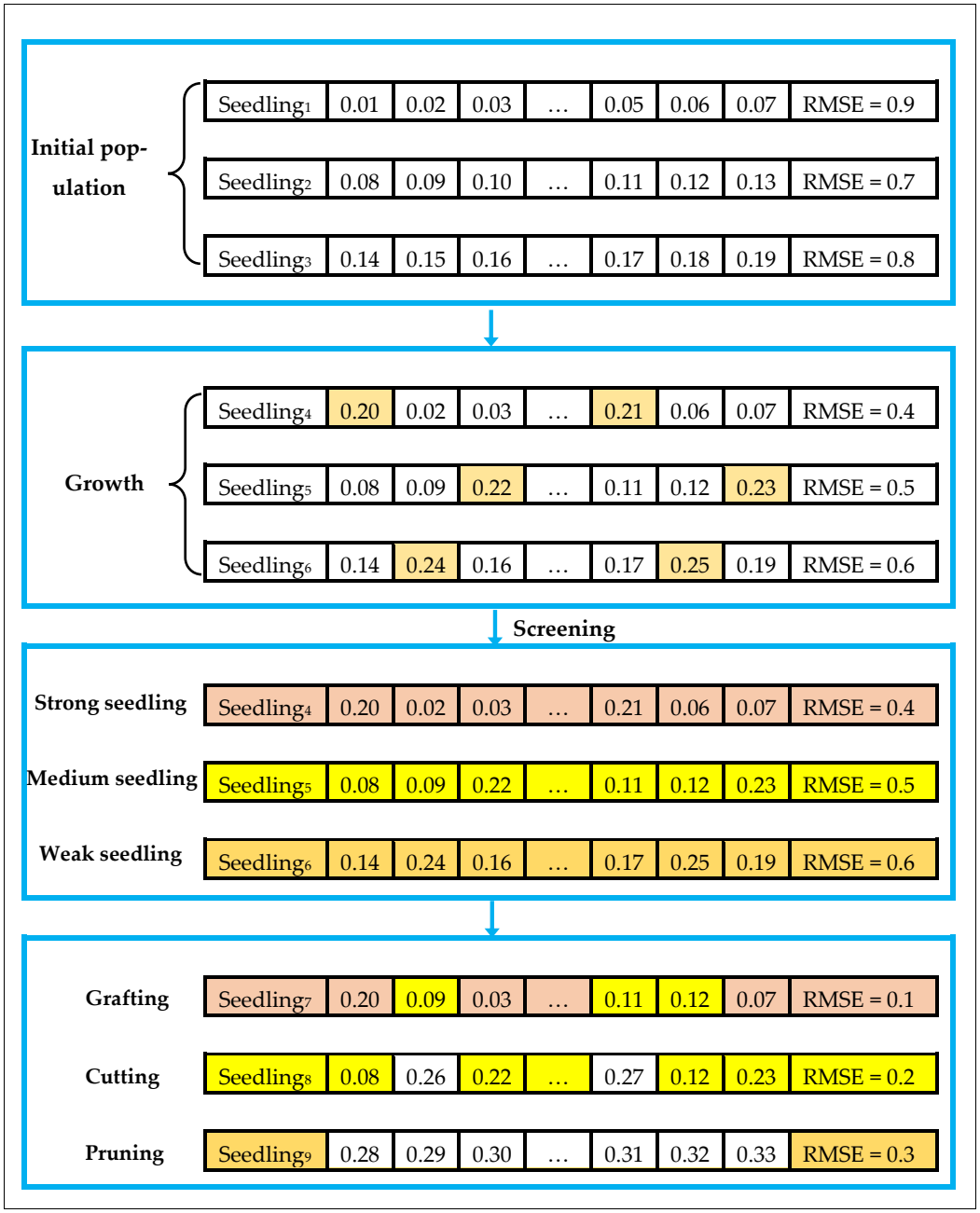


Figure 10. Example of the OA operators for optimizing LSTM weights and biases.

3. Experimental Result

To forecast $PM_{2.5}$ air pollution levels, several advanced algorithms were implemented and evaluated to identify the most effective model for this task. These algorithms include OA-LSTM, LSTM, RNN, DNN, SVM, and RF, and their performance was compared comprehensively. The selection of algorithms for this study was based on their distinct capabilities and relevance to air pollution forecasting. Advanced DL models such as LSTM, RNN, and DNN were chosen for their ability to handle complex, non-linear relationships and sequential data, both of which are critical for $PM_{2.5}$ prediction. LSTM, in particular, excels at capturing long-term dependencies in time series, making it highly effective for modeling the temporal variations of air pollution levels. RNN and DNN complement this by providing alternative approaches to sequential and non-linear modeling, enabling a comprehensive exploration of DL techniques for this problem. In addition to DL models, traditional ML algorithms such as SVM and RF were included as benchmarks due to their

simplicity, robustness, and established success in environmental data modeling. SVM is particularly effective in handling high-dimensional data and non-linear relationships, while RF's ensemble approach provides resilience against noise and complex feature interactions. Comparing these models with the proposed OA-LSTM ensures a thorough evaluation, highlighting its strengths and validating its performance against both classical and advanced techniques. This diverse selection underscores the robustness of OA-LSTM, showcasing its ability to outperform a wide range of predictive methodologies in capturing the intricate dynamics of $PM_{2.5}$ concentrations.

All implementations were conducted in the Python programming environment, leveraging libraries such as TensorFlow, Keras, Scikit-learn, and NumPy for model development, optimization, and evaluation. The dataset used for training and testing comprises meteorological data, topographical features, $PM_{2.5}$ concentrations, and satellite-based parameters such as AOD. In the process of model development and evaluation, proper validation plays a crucial role in determining the model's ability to predict unseen data. In this study, the dataset was split into 70% training and 30% testing. The splitting was designed to account for both temporal and spatial aspects of the data, ensuring that the evaluation process was robust and reflective of real-world scenarios. For the temporal aspect, data from the years 2014 and 2015 were used for training, while data from 2016 were allocated for testing. This chronological split ensures that the model is evaluated on future data that was not available during training, closely mimicking real-world forecasting scenarios. By preserving the temporal sequence, we avoided data leakage and ensured that the model's performance was evaluated on genuinely unseen data. Additionally, the distributions of key variables, such as $PM_{2.5}$ concentrations and meteorological conditions, were analyzed across the training and testing datasets to confirm representativeness and balance.

From a spatial perspective, special care was taken to prevent data leakage by ensuring that individual monitoring stations were exclusively assigned to either the training or testing datasets. For instance, data from two monitoring stations were reserved entirely for the testing set, while the remaining stations were used for training. This strategy ensures that the model is tested on spatially distinct data, representing stations it has never seen during training. This setup mirrors real-world conditions where a trained model may encounter data from new or previously unmonitored locations. The spatial splitting strategy has several advantages. By excluding overlap between training and testing stations, we avoid inflating the model's performance through exposure to similar patterns from the same location. This ensures that the model's predictions are evaluated on entirely new patterns, enhancing its generalization ability. While this approach is more challenging and may lead to slightly lower accuracy on the testing set, it provides a more realistic assessment of the model's performance in scenarios where it encounters new and unseen spatial data.

To evaluate the performance of the implemented proposed models, several key metrics were employed, as defined in Equations (16)–(20). These metrics include RMSE, R^2, SD_{Error}, convergence trend, and computational complexity metrics (such as runtime). Each metric addresses a specific aspect of model performance. RMSE quantifies the average magnitude of prediction errors, with lower values indicating more precise predictions. R^2 evaluates the proportion of variance in the observed data explained by the model, providing insight into the goodness of fit. SD_{Error} measures the variability in prediction errors, highlighting the model's consistency. Metrics related to convergence trends assess the stability and efficiency of optimization, while runtime directly reflects computational demands. The

combination of these metrics ensures a holistic analysis of model performance, covering accuracy, stability, and computational efficiency.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[x_i - \hat{x}_i]^2}, \tag{16}$$

$$R^2 = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \overline{x})(\hat{x}_i - \overline{\hat{x}})}{\sigma_x \sigma_{\hat{x}}}\right]^2, \tag{17}$$

$$SD_{Error} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[e_i - \overline{e}]^2}, \tag{18}$$

$$e_i = x_i - \hat{x}_i\ , \tag{19}$$

$$\overline{e} = \overline{(x - \hat{x})} = \frac{1}{N}\sum_{i=1}^{N} e_i \tag{20}$$

where x_i is the observed parameter; $\hat{x}_i$ is the predicted (calculated) parameter; $\overline{x}$ is the mean of the observed parameter; $\overline{\hat{x}}$ the mean of the predicted parameter; σ_x the standard deviation of the observed parameter; $\sigma_{\hat{x}}$ is the standard deviation of the predicted parameter; e_i is the prediction error for each data point; $\overline{e}$ is the mean of the prediction errors; and N is the number of observations.

Proper parameter calibration is crucial to the success of ML models, as these parameters directly influence model performance. Without optimal tuning, a model may underfit, overfit, or fail to learn the desired patterns in data effectively. This is especially true in a complex problem like $PM_{2.5}$ prediction, where non-linear interactions and temporal dependencies play a significant role. Hence, setting the right parameters ensures the models generalize well to unseen data while maintaining computational efficiency. To determine the optimal parameters for each model, we employed the trial-and-error method, which is a systematic and iterative approach. This process involves testing various combinations of parameters, evaluating model performance for each combination, and gradually narrowing down the parameter ranges based on the results. Metrics such as RMSE and R^2 were used to assess the performance of each parameter setting, allowing us to identify configurations that minimize prediction errors and maximize consistency. Given the complexity of the task, a wide range of parameter values was tested for each model. For instance, learning rates, dropout rates, and batch sizes were varied across several orders of magnitude for DL models, while regularization parameters and kernel coefficients were adjusted for traditional ML models. However, due to the sheer volume of experiments, only the optimized parameter values are presented in Table 3 for brevity. It is worth noting that extensive experimentation was conducted to ensure these values represent the best configurations for each algorithm.

For the OA-LSTM model, critical parameters such as a learning rate of 0.07, a batch size of 64, and a sequence length of 20 were selected. The number of hidden layers and neurons per layer were set to 8 and 64, respectively, ensuring the model could capture complex temporal dynamics without overfitting. Activation functions like *Tanh* and Sigmoid were chosen for their ability to handle non-linear relationships effectively, while the optimizer was calibrated using the novel OA for robust gradient updates. The DNN model was configured with a learning rate of 0.06 and six hidden layers, each containing 36 neurons. The Adam optimizer was employed for its adaptive learning rate properties, while a batch size of 32 ensured efficient training. Similarly, the RNN model utilized a learning rate of 0.08, ten hidden layers, and 64 neurons per layer, with *Tanh* and Sigmoid activation

functions for capturing sequential patterns. The sequence length for RNN was set to 10 to balance the temporal depth and computational efficiency. Traditional ML models like SVM and RF were also fine-tuned. The SVM model featured a regularization parameter of 1, a gamma value of 0.003, and an RBF kernel, chosen for its ability to model non-linear data effectively. The RF model was calibrated with 300 estimators, a maximum depth of 12, and a minimum sample split of 4, ensuring robustness against overfitting while maintaining computational efficiency.

Table 3. Parameter setting of proposed methods through the trial-and-error method.

Proposed Models	Parameters	Value
OA-LSTM	Learning rate	0.07
	Batch size	64
	Recurrent dropout rate	0.3
	Sequence length	20
	Number of Hidden Layers	8
	Number of Neurons per Layer	64
	Activation Function	Tanh and sigmoid
	Optimizer	OA
OA	Number of strong trees (N_1)	38
	Number of medium trees (N_2)	48
	Number of weak trees (N_3)	32
	α	0.72
	β	0.28
	Population size	120
	Iteration	300
DNN	Learning rate	0.06
	Batch size	32
	Number of hidden layers	6
	Number of neurons in hidden layers	36
	Activation	Tanh and sigmoid
	Optimizer	Adam
RNN	Learning rate	0.08
	Batch size	64
	Dropout rate	0.2
	Sequence length	10
	Number of hidden layers	10
	Number of neurons per layer	64
	Activation function	Tanh and sigmoid
	Optimizer	SGD
SVM	Regularization parameter	1
	Number of estimators	300
	Gamma	0.003
	Kernel type	RBF
RF	Criterion	Gini
	Number of estimators	300
	Minimum samples per split	4
	Maximum depth of trees	12

Table 4 presents the performance metrics of various models for predicting $PM_{2.5}$ pollution, focusing on the test dataset. The models are compared based on RMSE, R^2, and SD_{Error}, providing a comprehensive view of their accuracy and stability. Our proposed method, OA-LSTM, significantly outperforms all other models. It achieves the lowest

RMSE (3.01 $\mu g/m^3$), indicating a much higher predictive accuracy compared to traditional LSTM (9.53 $\mu g/m^3$), RNN (9.84 $\mu g/m^3$), and other ML algorithms like DNN, RF, and SVM, which exhibit larger errors. Additionally, the R^2 value of our model (0.88) demonstrates the strongest correlation between the predicted and actual values, reflecting a better model fit. The OA-LSTM model also shows excellent stability, with an SD_{Error} of just 0.05, much lower than the other models, which range from 1.24 to 4.83. These results highlight the robustness and efficiency of the proposed OA-LSTM model in handling spatial–temporal data for air pollution prediction.

Table 4. The results of proposed models for predicting $PM_{2.5}$ pollution (validation dataset).

Algorithms	Metrics		
	RMSE ($\mu g/m^3$)	R^2	SD_{Error}
Ours (OA-LSTM)	3.01	0.88	0.05
LSTM	9.53	0.65	1.24
RNN	9.84	0.63	1.42
DNN	10.41	0.59	1.73
RF	15.39	0.48	3.67
SVM	17.37	0.44	4.83

Figure 11 illustrates scatter plots of the measured versus estimated $PM_{2.5}$ concentrations for six proposed models. These plots visually represent the models' performance in predicting $PM_{2.5}$ values, with the x-axis showing the measured concentrations and the y-axis displaying the predicted values. The color bar on the right of plots represents the frequency of data points, with darker shades (blue) indicating fewer data points and brighter shades (yellow to red) highlighting areas of higher concentration. The density of the points near the diagonal regression line is an indicator of the accuracy and consistency of the predictions. In the OA-LSTM model, the results demonstrate a clear alignment between the predicted and measured values. The regression equation $Y = 1.00X - 0.56$ indicates that the model nearly perfectly follows the trend of the actual data. The clustering of points around the regression line, particularly in the high-frequency areas highlighted in yellow and red, suggests that the OA-LSTM model is not only accurate but also consistent across different ranges of $PM_{2.5}$ concentrations. The superior performance of the OA-LSTM model can be attributed to the integration of the OA, which enhances the LSTM's ability to capture spatial–temporal features in the $PM_{2.5}$ data.

The scatter plots for the LSTM and RNN models reveal a moderate alignment between the measured and estimated values, as evidenced by regression equations $Y = 0.97X + 0.50$ for LSTM and $Y = 0.95X + 1.18$ for RNN. The RMSE values of 9.53 $\mu g/m^3$ (LSTM) and 9.84 $\mu g/m^3$ (RNN) indicate a reasonable performance, although the density of points along the regression line shows some underprediction at higher $PM_{2.5}$ concentrations. These models perform well in capturing trends in low to medium ranges but exhibit limitations in accurately modeling high pollution levels. In contrast, the DNN model shows a weaker correlation between the measured and predicted concentrations, with a regression equation $Y = 0.91X - 2.30$ and an RMSE of 10.41 $\mu g/m^3$. The increased dispersion of points, particularly at higher concentrations, highlights the DNN model's reduced ability to capture complex spatial–temporal patterns, leading to less reliable predictions of peak pollution events. The scatter plots for the RF and SVM display a significantly lower alignment with the measured data, with equations $Y = 0.89X + 2.81$ for RF and $Y = 0.87X + 3.32$ for SVM. The widely scattered points around the regression line, particularly for SVM, indicate a poor predictive performance in both low and high concentration ranges.

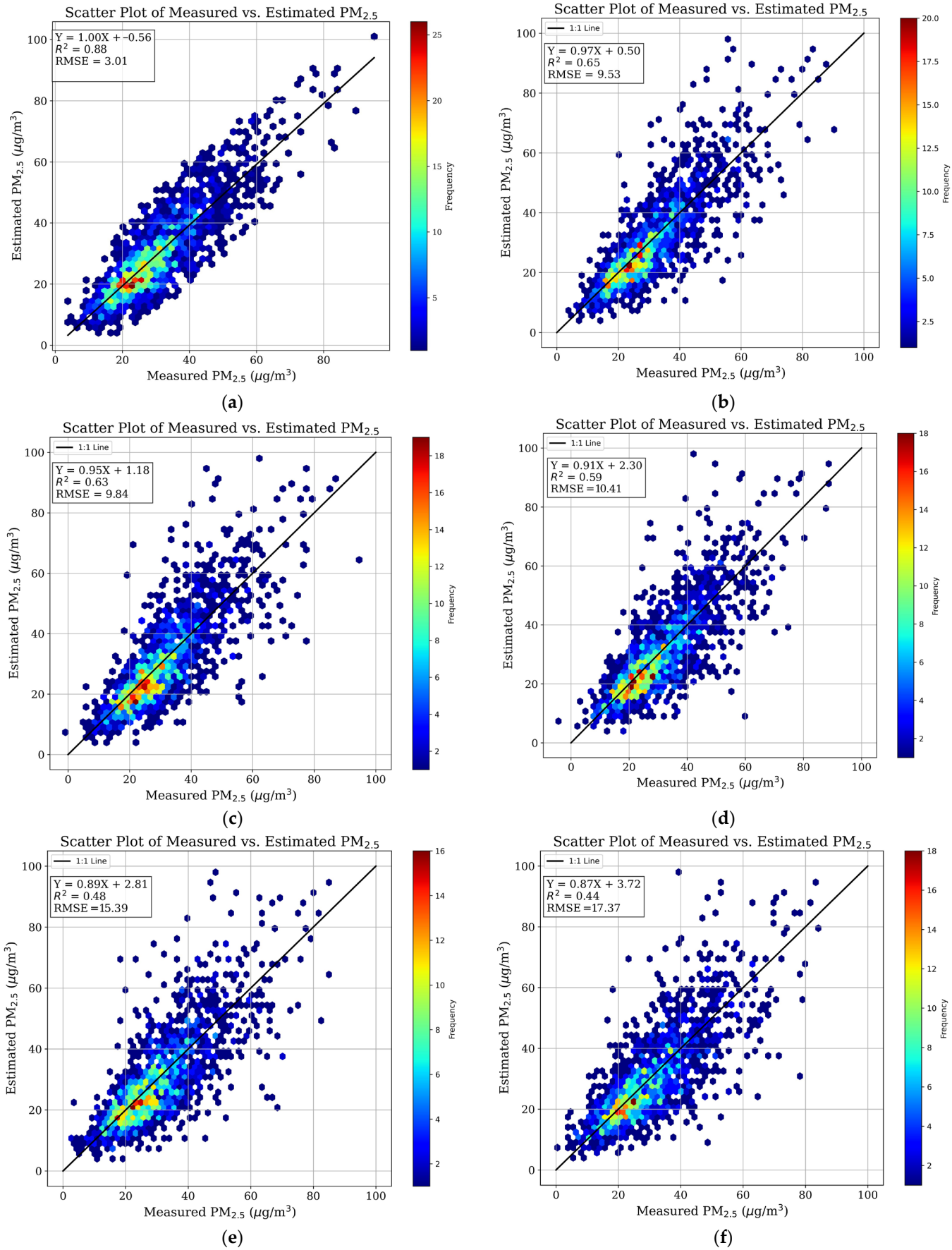


Figure 11. Scatter plot of proposed: (**a**) OA-LSTM; (**b**) LSTM; (**c**) RNN; (**d**) DNN; (**e**) RF; and (**f**) SVM.

Table 5 presents the computational complexity (run time) of various algorithms, measured based on the time required to reach different RMSE thresholds. The OA-LSTM model significantly outperforms other methods in terms of computational efficiency. For instance, OA-LSTM reaches an RMSE of 20 $\mu g/m^3$ in just 32 s, while other models such as LSTM, RNN, and DNN take considerably longer, with 106, 123, and 162 s, respectively. As the RMSE threshold becomes more stringent (e.g., RMSE < 10 $\mu g/m^3$), OA-LSTM continues to demonstrate its advantage by achieving this goal in 173 s, whereas other models either take much longer (RNN: 634 s) or fail to reach that level. This table also highlights the scalability and efficiency of OA-LSTM as the only model capable of reaching an RMSE below 5 $\mu g/m^3$, albeit with a significant increase in runtime (384 s). In contrast, none of the other models manage to reduce the RMSE below 5 $\mu g/m^3$, suggesting a trade-off between complexity and performance. Overall, the results underline OA-LSTM's ability to balance accuracy and computational cost, demonstrating a superior performance in both reducing prediction errors and minimizing computational demands.

Table 5. Comparison of algorithm computational complexity based on RMSE termination criteria.

Methods	Run Time (s)			
	RMSE < 20 $\mu g/m^3$	RMSE < 15 $\mu g/m^3$	RMSE < 10 $\mu g/m^3$	RMSE < 5 $\mu g/m^3$
Ours (OA-LSTM)	32	98	173	384
LSTM	106	189	418	-
RNN	123	206	634	-
DNN	162	236	-	-
RF	149	-	-	-
SVM	201	-	-	-

Figure 12 illustrates the convergence trends of proposed models in terms of their RMSE across different epochs. Among the models, OA-LSTM exhibits the fastest and most effective convergence, with a sharp decline in RMSE within the first 50 epochs, reaching an RMSE value below 5 quite early. This indicates that the OA-LSTM model not only converges more quickly but also maintains a high level of accuracy throughout the training process. In contrast, the other models show a slower convergence rate. For example, while the LSTM model also steadily reduces its RMSE, it does so more slowly, achieving an RMSE around 10 only after approximately 150 epochs. The RNN and DNN models perform similarly but with larger RMSE values, particularly in the earlier epochs. Both models begin with high RMSEs above 30 $\mu g/m^3$ and gradually converge towards values around 10 $\mu g/m^3$ after 200–300 epochs, reflecting less efficient learning compared to OA-LSTM. The RF and SVM models show the slowest convergence trends, with RMSEs above 20 $\mu g/m^3$ even after 300 epochs, indicating that these models struggle to learn effectively and are less suited for this specific prediction task.

Figure 13 illustrates the spatial distribution of the observed $PM_{2.5}$ concentrations for two distinct periods: (a) August 2016 (Wednesday, summer), and (b) December 2016 (Friday, winter). The spatial distribution of $PM_{2.5}$ concentrations in August 2016 (Figure 13a) shows relatively lower levels of pollution compared to the winter map. The highest concentration, marked by darker blue areas, can be seen in the northern and central regions of the map, with values peaking around 58 $\mu g/m^3$. The southern and eastern parts exhibit lower levels, ranging from 15 to 30 $\mu g/m^3$. This pattern of pollution could be attributed to various factors such as the natural topography, dominant wind patterns, traffic, and possibly industrial activities concentrated in the northern areas. During summer, the overall levels of $PM_{2.5}$ are expected to be lower due to better air dispersion caused by warmer temperatures and stronger winds, which help dilute pollutants. In contrast, the map for December 2016

(Figure 13b) presents a significantly higher level of pollution, especially in the northern and central regions, where $PM_{2.5}$ concentrations reach up to 125 $\mu g/m^3$. The eastern and southern regions also exhibit higher pollution compared to the summer map, with values ranging between 40 and 80 $\mu g/m^3$. One possible reason for this stark difference is the phenomenon of temperature inversion, which commonly occurs during the winter. This meteorological condition traps pollutants close to the ground, particularly in urban areas, leading to higher concentrations of $PM_{2.5}$ [25–28]. Additionally, December is a colder month, and higher emissions from residential heating and the reduced dispersion of pollutants likely contribute to the elevated pollution levels. The fact that December 23 was a Friday, a non-working day in many regions, might also affect the data, as reduced traffic could have slightly mitigated $PM_{2.5}$ levels, but still, the impact of inversion and higher winter emissions likely outweighs this factor.

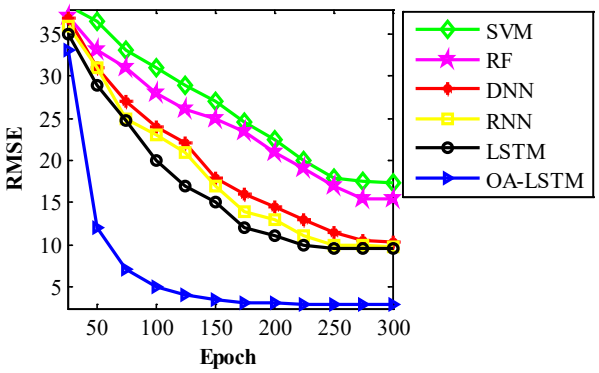
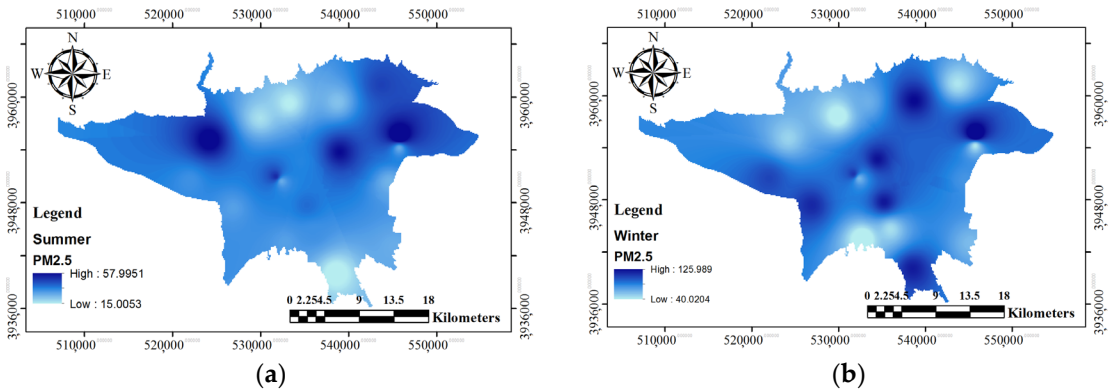


Figure 12. The convergence trend of proposed models based on the RMSE metric.



(**a**) (**b**)

Figure 13. Spatial distribution of observed $PM_{2.5}$ concentrations in (**a**) August 2016 (Wednesday, summer) and (**b**) December 2016 (Friday, winter).

Figure 14 illustrates the spatial distribution of $PM_{2.5}$ concentrations predicted by six different algorithms in August 2016 (Wednesday, summer). These predictions are compared against the ground truth map provided in Figure 13a, enabling a comprehensive evaluation of each model's performance in replicating the observed spatial patterns. The OA-LSTM model (Figure 14a) demonstrates exceptional accuracy, closely aligning with the ground truth map by effectively capturing the peak concentrations in the northern and central regions while maintaining a consistent gradient of lower pollution levels toward the southern and eastern areas. This alignment highlights OA-LSTM's robustness in modeling both high and low concentration zones. In contrast, LSTM (Figure 14b) and RNN (Figure 14c) also perform well but show slight under predictions of maximum values in the northern regions, although their spatial trends remain consistent with the observed data. The DNN model (Figure 14d) provides predictions with reasonable spatial consistency

but tends to slightly overestimate $PM_{2.5}$ levels in the southern areas, deviating from the observed distribution. Traditional ML models like RF (Figure 14e) and SVM (Figure 14f) exhibit comparatively lower predictive accuracy, failing to replicate localized peaks in $PM_{2.5}$ concentrations, particularly in the northern regions. The RF model produces more generalized patterns, while SVM shows greater deviation in both spatial structure and concentration levels, highlighting its limitations in capturing complex spatial relationships. This comparison is not only visual but is also supported by the quantitative evaluation metrics presented in Table 4, which provide a detailed numerical assessment of each model's accuracy in predicting $PM_{2.5}$ concentrations. The reliability of the preprocessing methods used, such as noise reduction and handling missing data, are comprehensively detailed in the manuscript to ensure that the comparisons are robust and trustworthy. By combining visual and numerical evaluations, this study ensures a holistic assessment of model performance, showcasing the capability of OA-LSTM to effectively capture both spatial variability and high-concentration zones more accurately than other models.

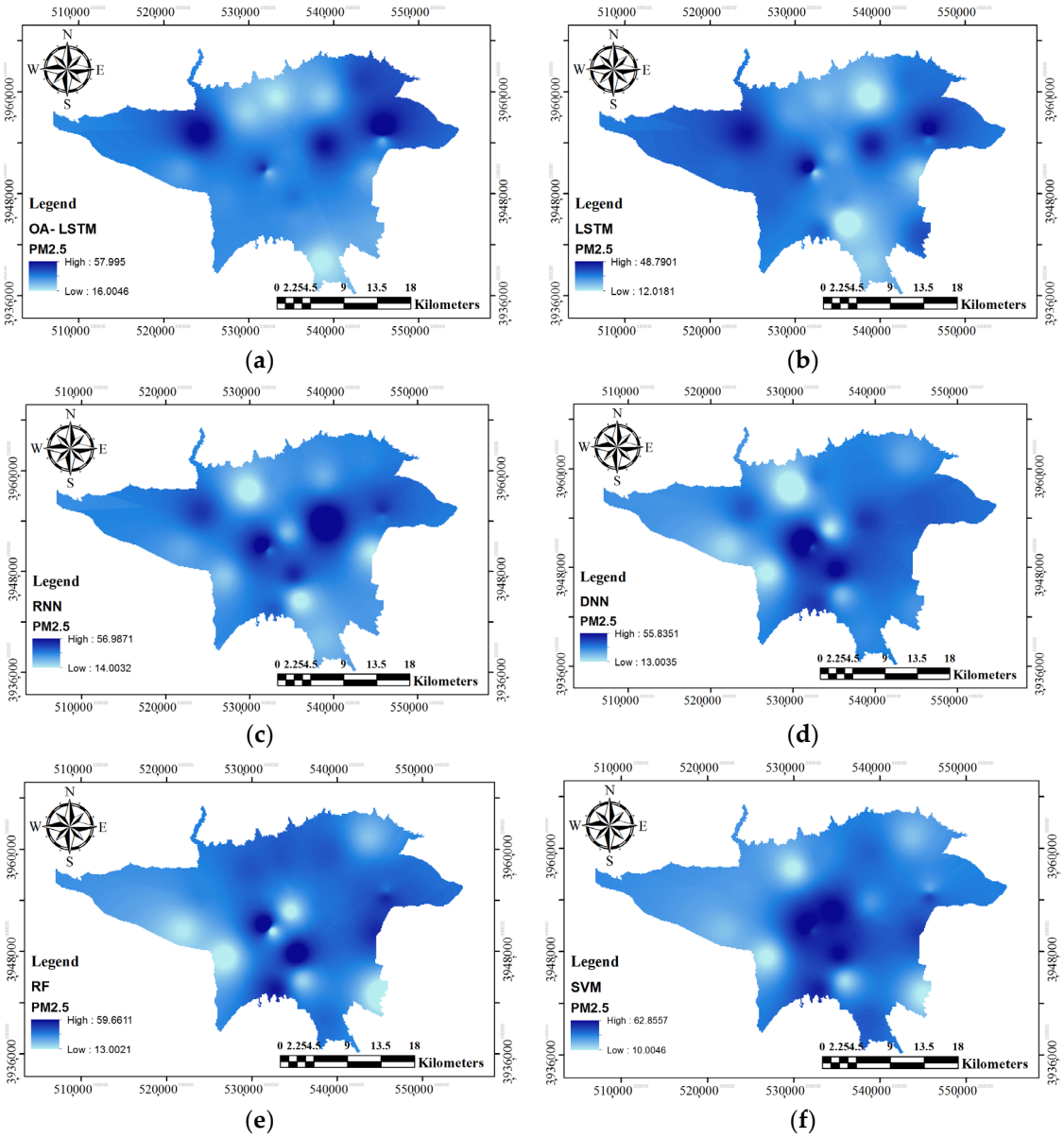


Figure 14. Spatial distribution of estimated $PM_{2.5}$ concentrations in August 2016 (Wednesday, summer): (**a**) OA-LSTM; (**b**) LSTM; (**c**) RNN; (**d**) DNN; (**e**) RF; and (**f**) SVM.

For sensitivity analysis, we propose a binary OA (BOA) for feature selection, designed to achieve two primary objectives: minimizing the number of input features that do not significantly impact $PM_{2.5}$ prediction and simultaneously minimizing the prediction error. To ensure that critical features are not excluded, a penalty term is introduced, which is integrated into the objective function. This penalty accounts for the importance of features in predicting $PM_{2.5}$ concentrations. By integrating the feature reduction term and penalty into a unified framework, the algorithm achieves a balanced trade-off between simplicity and accuracy. Features with a minimal importance and low impact on RMSE are prioritized for exclusion, while highly significant features are retained. Each feature is encoded in binary form within a seedling, where a value of 1 indicates that the feature is included in the selected subset, and 0 indicates its exclusion. This encoding allows the algorithm to explore and select the optimal combination of features. The objective function is defined as Equations (21) and (22):

$$Objective\ Function_i = \varphi \cdot RMSE_i + (1-\varphi) \cdot \left(log_{10} \frac{N}{n_i} + \omega \cdot Penalty_i \right), \tag{21}$$

$$Penalty_i = \sum_{j=1}^{N} Importance_j \cdot \begin{cases} 1, & if\ \ x_j = 0 \\ 0, & if\ \ x_j = 1 \end{cases} \tag{22}$$

where φ is a weighting parameter, with a default value of 0.89; N is the total number of features in the dataset; n_i is the number of selected features in the i-th seedling; ω is the weight assigned to the additional penalty for excluding important features; represents the relative importance of the j-th feature, which can be derived using the Gini index from RF; and x_j is the binary value indicating the inclusion ($x_j = 1$) or exclusion ($x_j = 0$) of the j-th feature in the selected subset.

This objective function ensures that RMSE is minimized, enhancing the predictive accuracy of the model; the number of features is reduced, simplifying the model and improving computational efficiency; and critical features are retained through the penalty term, preventing the exclusion of important predictors of $PM_{2.5}$ concentrations. The BOA starts by initializing a population, where each seedling represents a subset of features. Through iterative operations (elitism, growth, screening, grafting, pruning, and cutting) the algorithm evolves the population to minimize the objective function. The RMSE is computed using predictive models such as OA-LSTM, LSTM, RNN, DNN, RF, and SVM to evaluate the accuracy of the selected features. Table 6 provides a comprehensive analysis of the selected features and their importance in predicting $PM_{2.5}$ concentrations across six predictive models. Each row in the table represents a model, while the columns denote whether a feature is included (1) or excluded (0), alongside the total number of selected features and the resulting RMSE value.

Certain features, such as AOD, LST, NDVI, WS, and the day of the week are consistently selected by all models. This universal inclusion underscores their critical importance in $PM_{2.5}$ prediction. AOD directly relates to particulate matter and air quality, making it indispensable for any $PM_{2.5}$ predictive model. Similarly, LST reflects surface temperature variations, which influence atmospheric stability and pollutant dispersion, while NDVI captures vegetation cover, a vital factor in pollutant absorption and urban heat mitigation. Other features exhibit moderate importance, such as Max T, Min T, and humidity (H). These features are included in most models but not universally. Their occasional exclusion suggests that their importance might be context-dependent, varying with the model's structure or specific interactions between features. For instance, the minimum temperature affects nocturnal cooling and atmospheric stratification, impacting pollutant dispersion, whereas humidity interacts with particulate matter by influencing its hygroscopic growth.

Conversely, some features, P, PE, WV, and WD, are less frequently selected, reflecting their limited role in $PM_{2.5}$ prediction for Tehran. For example, water vapor might have indirect effects on pollution levels, but this effect is not as pronounced as those of primary features like AOD. The wind direction's impact might also be diminished in a densely urbanized setting like Tehran, where pollutant sources are spatially distributed, and localized meteorological factors dominate. Among all models, OA-LSTM demonstrates the best performance, achieving the lowest RMSE (5.12 $\mu g/m^3$) with only nine selected features. This indicates the capability of BOA to identify an optimal subset of features that maximizes predictive accuracy while minimizing redundancy. Models like SVM and RF, on the other hand, require more features and still result in higher RMSE values, indicating a less efficient utilization of the selected features.

Table 6. Feature selection for $PM_{2.5}$ prediction using the BOA across predictive models.

Variables	OA-LSTM	LSTM	RNN	DNN	RF	SVM
AOD	1	1	1	1	1	1
LST	1	1	1	1	1	1
NDVI	1	1	1	1	1	1
Total precipitation (PE)	0	1	0	1	1	1
Water vapor (WV)	0	0	1	0	0	1
Max Temp	1	1	1	1	1	0
Min Temp	1	0	1	1	1	1
Wind speed (WS)	1	1	1	1	1	1
Wind direction (WD)	0	0	1	1	0	1
Pressure (P)	0	1	0	0	1	1
Humidity (H)	1	1	0	0	1	1
Elevation (Ele)	1	1	1	1	0	0
Longitude (X)	0	0	0	1	1	1
Latitude (Y)	0	1	0	0	1	1
Day of the week	1	1	1	1	1	1
Number of features (n_i)	9	11	10	11	12	13
RMSE ($\mu g/m^3$)	5.12	14.34	16.32	18.21	26.38	29.18

4. Discussion

Table 7 presents a comparative analysis of our study's results with previous works conducted in the same study area, Tehran. Although all the listed studies focus on $PM_{2.5}$ prediction, it is essential to note that the conditions of these studies are not entirely identical. Each study used different datasets collected over varying periods and employed distinct predictive models tailored to their respective datasets and objectives. Despite these differences, the comparison highlights the advancements achieved through our proposed methodology. In terms of performance, our model achieves the highest R^2 value of 0.88, outperforming all other studies. The closest result is the 3DCNN-GRU model by Faraji et al. [27], with an R^2 of 0.84, followed by XGBoost models, which achieved R^2 values of 0.81 and 0.74 in the works of Zamani Joharestani et al. [28] and Bagheri [26], respectively. Traditional models such as RF, utilized by Nabavi et al. [24], demonstrate a lower predictive performance with an R^2 of 0.68. The superior performance of OA-LSTM can be attributed to several factors. First, the architectural design of OA-LSTM leverages advanced DL techniques tailored specifically for $PM_{2.5}$ prediction. This architecture effectively captures both spatial and temporal dependencies in the data, offering a significant advantage over conventional ML models. Second, the integration and preprocessing of data, including feature selection using the OA, have likely contributed to the enhanced performance by prioritizing the most relevant predictors and minimizing noise. Finally, the combination of an optimized model and rigorous data preparation underscores the robustness of our model.

This not only improves the accuracy of the predictions but also sets a new benchmark for $PM_{2.5}$ modeling in Tehran, demonstrating the potential of advanced DL frameworks in addressing complex environmental challenges.

Table 7. A comparative analysis of our study's results with previous works in Tehran.

Authors	Publication	Study Period	Model	R^2
Nabavi et al. [24]	2019	2011–2016	RF	0.68
Zamani Joharestani et al. [28]	2019	2015–2018	XGBoost	0.81
Bagheri [26]	2022	2013–2019	XGBoost	0.74
Faraji et al. [27]	2022	2016–2019	3DCNN-GRU	0.84
Bagheri [25]	2023	2013–2019	Deep RF	0.74
Ours	-	2014–2016	OA-LSTM	0.88

Table 8 provides a comprehensive evaluation of the RMSE values for six predictive models across three distinct $PM_{2.5}$ concentration ranges: low, moderate, and high. These ranges were defined based on the distribution of $PM_{2.5}$ concentrations in the dataset, with low concentrations corresponding to values below 35 $\mu g/m^3$, moderate concentrations between 35 and 75 $\mu g/m^3$, and high concentrations above 75 $\mu g/m^3$. This breakdown allows for a detailed analysis of each model's performance under varying pollution levels, offering insights into their strengths and limitations. The OA-LSTM model consistently exhibits the lowest RMSE across all concentration ranges, with values of 2.65, 2.94, and 3.73 $\mu g/m^3$ for low, moderate, and high concentrations, respectively. This highlights the model's robust ability to predict $PM_{2.5}$ concentrations with high accuracy, regardless of the pollution level. In contrast, traditional ML models like RF and SVM demonstrate significantly higher RMSE values, particularly in high-concentration scenarios, where the RMSE reaches 20.40 and 23.46 $\mu g/m^3$, respectively. Among DL models, the LSTM and RNN perform reasonably well, but their RMSE values remain higher than those of OA-LSTM, especially in moderate and high-concentration ranges. This analysis underscores the superior generalizability and precision of the OA-LSTM model, particularly in scenarios involving low and moderate pollution levels. However, at higher concentration ranges, all algorithms, including OA-LSTM, exhibit reduced accuracy. Nonetheless, OA-LSTM demonstrates remarkable consistency, as its RMSE values across low, moderate, and high concentrations are relatively close compared to other models. This consistency highlights the robustness and reliability of OA-LSTM in capturing the intricate patterns of $PM_{2.5}$ concentrations, even under challenging high-pollution scenarios.

Table 8. RMSE for $PM_{2.5}$ concentration across low, moderate, and high concentration ranges.

Methods	RMSE for $PM_{2.5}$ Concentration ($\mu g/m^3$)		
	Low Concentration	Moderate Concentration	High Concentration
OA-LSTM	2.65	2.94	3.73
LSTM	7.12	8.94	11.36
RNN	9.06	10.02	13.70
DNN	9.33	12.64	15.48
RF	12.46	13.76	20.40
SVM	11.37	18.34	23.46

Extreme pollution events are short-term episodes characterized by $PM_{2.5}$ concentrations exceeding standard thresholds, posing severe risks to public health and the environment. These events result from a combination of natural causes, such as dust storms, wildfires, and temperature inversions, and anthropogenic factors, including industrial

emissions, heavy traffic, and increased fossil fuel usage. The irregular and nonlinear nature of these events, coupled with data scarcity and noise, makes their prediction highly challenging. The accurate modeling of such events is critical for timely mitigation strategies, as they lead to acute health issues, economic disruptions, and environmental degradation. LSTM networks are well suited for predicting extreme pollution events due to their ability to capture long-term dependencies and model nonlinear relationships. By leveraging memory cells and gating mechanisms, LSTMs can identify early indicators of these events, such as shifts in meteorological variables or pollutant levels. The integration of the OA further enhances LSTM performance by addressing challenges in optimization, such as avoiding local minima and improving model robustness against noisy and imbalanced datasets. This hybrid approach enables the model to generalize effectively, ensuring higher accuracy in predicting rare and complex pollution events.

To ensure the model's robustness in handling extreme pollution events, effective noise reduction and anomaly management techniques were applied during the data preprocessing stage. Specifically, the Savitzky–Golay filter was utilized to smooth the $PM_{2.5}$ time-series data, reducing irregularities and ensuring that critical signal features, such as sharp peaks and valleys, were preserved. This step was crucial for enabling the model to accurately capture the unique patterns associated with extreme pollution events. By removing noise without compromising the inherent structure of the data, the preprocessing pipeline enhanced the reliability of the input dataset and facilitated the model's ability to identify and predict these rare occurrences. Additionally, the proposed OA-LSTM model was rigorously evaluated using data specifically associated with extreme pollution events, such as days with exceptionally high $PM_{2.5}$ concentrations. The results demonstrated that the model maintained a comparable level of predictive accuracy during extreme pollution events, with error metrics that were consistent with those observed for lower $PM_{2.5}$ concentration levels. This indicates the model's ability to generalize effectively, providing accurate and actionable forecasts even in the presence of anomalies and extreme conditions.

One of the main challenges in this study was the limited number of ground-based air quality monitoring stations, which could reduce prediction accuracy, especially in areas with sparse or no data coverage. To address this, satellite data, such as AOD, were used as complementary sources, providing broad spatial coverage and filling data gaps in unmonitored regions. The model was first trained and validated using satellite data aligned with existing monitoring stations to establish the relationship between satellite-derived variables and $PM_{2.5}$ concentrations. Once validated, the model used satellite data from unmonitored regions to estimate $PM_{2.5}$ levels, enabling comprehensive spatial predictions. Additional spatial features, such as elevation and NDVI, were incorporated to enhance the prediction accuracy by providing environmental context and capturing factors influencing pollution distribution. To harmonize the temporal and spatial resolutions across datasets, specific preprocessing steps were implemented. $PM_{2.5}$ and AOD data, initially recorded at hourly intervals, were aggregated to daily averages to align with daily meteorological data, ensuring a unified temporal resolution. For spatial consistency, resampling and interpolation techniques, such as IDW and Kriging, were applied to adjust satellite and ground-based datasets to matching spatial granularities. These harmonization steps created a cohesive dataset, facilitating accurate spatial–temporal analysis and enabling the model to deliver reliable air quality predictions across diverse regions.

To address the challenge of missing data, we adopted a controlled scenario approach to evaluate the effectiveness of our interpolation methods. In this approach, a certain percentage of complete data was randomly removed, and the missing values were reconstructed using spline interpolation. This method allowed us to assess how accurately the interpolation technique could recover missing data and how it would impact the model's

predictive performance. As missing data are an unavoidable issue in real-world datasets, employing robust interpolation techniques is crucial. The accuracy of these techniques significantly depends on their implementation; therefore, careful calibration and testing of spline parameters were performed to ensure the best possible reconstruction. Table 9 highlights the impact of our preprocessing strategies, showcasing the performance of the OA-LSTM model and other comparative models under different scenarios. For the real dataset, the application of the Savitzky–Golay filter for noise reduction resulted in notable improvements. The RMSE of the OA-LSTM model decreased from 5.29 $\mu g/m^3$ (without Savitzky–Golay) to 4.63 $\mu g/m^3$ (with Savitzky–Golay), demonstrating the filter's effectiveness in maintaining data integrity while eliminating noise. Similar trends were observed across all models, with OA-LSTM consistently outperforming others.

Table 9. Comparison of RMSE values for real and interpolated datasets.

Model	RMSE for Real Data ($\mu g/m^3$)		RMSE for Interpolated Data ($\mu g/m^3$)	
	With Savitzky–Golay	Without Savitzky–Golay	With Spline and Savitzky–Golay	With Spline, Without Savitzky–Golay
OA-LSTM	4.63	5.29	6.26	7.53
LSTM	11.38	13.03	14.54	16.84
RNN	13.02	14.76	15.10	17.32
DNN	15.63	17.20	18.28	20.60
RF	21.76	23.37	24.46	27.99
SVM	20.07	21.95	25.23	28.41

For the interpolated datasets, spline interpolation proved effective, though RMSE values were slightly higher than those of the real dataset. When combined with Savitzky–Golay filtering, spline interpolation yielded reliable results, with the OA-LSTM model achieving an RMSE of 6.26 $\mu g/m^3$. However, when spline interpolation was used without noise reduction, RMSE increased further to 7.53 $\mu g/m^3$. This underscores the critical role of integrating noise reduction methods with interpolation to enhance data quality and improve predictive accuracy. These results emphasize the necessity of employing robust preprocessing strategies in scenarios involving missing data. The OA-LSTM model's superior performance across all scenarios demonstrates its robustness and adaptability, even in challenging conditions. The controlled approach used in this study not only validates the effectiveness of spline interpolation but also highlights the importance of fine-tuning interpolation techniques to maximize accuracy in real-world applications.

The proposed OA-LSTM model holds significant practical potential for integration into existing air quality management systems. By leveraging its high predictive accuracy, the model can provide early warnings for extreme pollution events, enabling policymakers to implement timely mitigation measures such as traffic restrictions, industrial activity adjustments, or public health advisories. The model's compatibility with real-time data streams from ground monitoring stations and satellite sources ensures its relevance in dynamic and rapidly changing urban environments. Furthermore, the model's ability to generalize across regions with varying data densities makes it particularly valuable for areas with sparse monitoring infrastructure, where accurate predictions are crucial for resource allocation and public safety measures.

In addition, this framework can enhance existing systems by serving as a decision-support tool for urban planning and environmental policy development. For example, integrating the OA-LSTM model into mobile applications or online dashboards could provide actionable air quality forecasts directly to the public, increasing awareness and preparedness during high-risk periods. Similarly, policymakers could use model outputs

to design more effective long-term strategies, such as identifying high-emission zones or optimizing the placement of new monitoring stations. Importantly, the model's relatively low computational requirements enable its deployment on cloud-based or local infrastructures, making it accessible to a wide range of governmental and non-governmental organizations. By bridging the gap between predictive modeling and practical implementation, the proposed OA-LSTM model represents a critical advancement in data-driven air quality management.

5. Conclusions

Air pollution, particularly $PM_{2.5}$, poses significant health and environmental challenges in urban areas. Tehran, as a highly populated and industrialized city, has been the focus of numerous studies aiming to model and predict $PM_{2.5}$ concentrations. In this paper, we proposed a novel OA-LSTM model to address the dual challenges of feature selection and predictive accuracy. By analyzing meteorological, environmental, and spatial data collected from Tehran between 2014 and 2016, we aimed to optimize the balance between model simplicity and performance. Our results demonstrated that the proposed OA-LSTM model outperformed all other approaches, achieving the highest R^2 value of 0.88, indicating its robustness and reliability in capturing the complex dynamics of $PM_{2.5}$ concentrations. Key features such as AOD, LST, NDVI, WS, and the day of the week were consistently identified as the most significant predictors across all models. Furthermore, the BOA's ability to reduce the number of features without sacrificing accuracy was evident, as the OA-LSTM model achieved the lowest RMSE of 5.12 $\mu g/m^3$ with only nine selected features, outperforming models like RF and SVM, which required more features and delivered lower predictive performance.

The comparative analysis with previous studies further highlighted the advantages of our proposed framework. While earlier works utilizing models like RF, XGBoost, and 3DCNN-GRU achieved respectable R^2 values, they were constrained by either less sophisticated feature selection techniques or limited temporal and spatial data. In contrast, our approach combined advanced architecture design, optimized data preprocessing, and robust feature selection to set a new benchmark for $PM_{2.5}$ modeling in Tehran. These results underline the importance of integrating state-of-the-art methodologies for tackling complex environmental problems. In conclusion, this paper has successfully demonstrated the potential of combining advanced DL architectures with innovative optimization techniques like the OA method to address air quality prediction challenges. The OA-LSTM model, equipped with optimized feature subsets, provides a scalable and efficient solution for urban air quality management.

Building upon the advancements of this study, several promising directions can enhance the robustness and applicability of the OA-LSTM framework. First, expanding the dataset to include additional years and real-time monitoring data will enable the model to capture temporal variations more comprehensively and detect emerging pollution trends. Integrating high-frequency, real-time data sources can also enhance the framework's ability to provide near-instantaneous predictions, making it suitable for urban air quality management systems. This scalability ensures adaptability to evolving environmental conditions and supports real-time decision-making. Second, the OA-LSTM model demonstrates strong potential for application in various regions and pollutants, beyond the current focus on $PM_{2.5}$. By incorporating region-specific data, such as meteorological variables, pollutant concentrations, and geographical features, the model can adapt to new environmental conditions. For cities or regions with sparse monitoring networks, integrating supplementary data sources like satellite imagery, low-cost sensors, and socioeconomic indicators (e.g., traffic density and industrial activities) can significantly improve spatial coverage and

prediction accuracy. Tailored preprocessing techniques, such as normalization and feature selection, will align these inputs with the unique characteristics of new target regions, ensuring accurate and reliable predictions.

Third, applying transfer learning techniques could streamline the adaptation of the OA-LSTM framework to diverse urban and rural environments. Pretraining the model on comprehensive datasets from one region and fine-tuning it with minimal local data from another can reduce computational costs and enhance its accessibility for resource-constrained areas. Testing the framework across regions with varying environmental conditions and pollutant profiles will validate its generalizability, allowing researchers to refine the model further for global air quality prediction. Lastly, future work could explore ensemble approaches that combine OA-LSTM with complementary models like CNNs to improve predictive accuracy and resilience. Additionally, developing dynamic optimization strategies, such as real-time extensions of the orchard algorithm, can enable continuous feature selection and parameter tuning, ensuring the model remains responsive to new data inputs. Incorporating uncertainty quantification methods into predictions will provide policymakers with reliable tools for risk assessment and targeted interventions. These advancements will position the OA-LSTM framework as a scalable and adaptable tool capable of addressing global challenges in urban air quality management and environmental sustainability.

Author Contributions: Conceptualization, Mehrdad Kaveh and Mohammad Saadi Mesgari; methodology, Mehrdad Kaveh, Mohammad Saadi Mesgari, and Masoud Kaveh; software, Mehrdad Kaveh and Masoud Kaveh; validation, Mehrdad Kaveh, Mohammad Saadi Mesgari, and Masoud Kaveh; formal analysis, Mehrdad Kaveh; investigation, Mehrdad Kaveh; resources, Mehrdad Kaveh and Mohammad Saadi Mesgari; data curation, Mehrdad Kaveh and Mohammad Saadi Mesgari; writing—original draft preparation, Mehrdad Kaveh, Mohammad Saadi Mesgari, and Masoud Kaveh; writing—review and editing, Mehrdad Kaveh, Mohammad Saadi Mesgari, and Masoud Kaveh; visualization, Mehrdad Kaveh and Masoud Kaveh; supervision, Mohammad Saadi Mesgari; project administration, Mehrdad Kaveh and Mohammad Saadi Mesgari; and funding acquisition, Mehrdad Kaveh and Masoud Kaveh. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kumar, R.P.; Prakash, A.; Singh, R.; Kumar, P. Machine learning-based prediction of hazards fine $PM_{2.5}$ concentrations: A case study of Delhi, India. *Discov. Geosci.* **2024**, *2*, 34. [CrossRef]
2. Natsagdorj, N.; Zhou, H. Prediction of $PM_{2.5}$ concentration in Ulaanbaatar with deep learning models. *Urban Clim.* **2023**, *47*, 101357.
3. Aman, N.; Manomaiphiboon, K.; Xian, D.; Gao, L.; Tian, L.; Pala-En, N.; Wangyao, K.; Wang, Y.; Wangyao, K. Spatiotemporal estimation of hourly $PM_{2.5}$ using AOD derived from geostationary satellite Fengyun-4A and machine learning models for Greater Bangkok. *Air Qual. Atmos. Health* **2024**, *17*, 1519–1534. [CrossRef]
4. Gündoğdu, S.; Elbir, T. Elevating hourly $PM_{2.5}$ forecasting in Istanbul, Türkiye: Leveraging ERA5 reanalysis and genetic algorithms in a comparative machine learning model analysis. *Chemosphere* **2024**, *364*, 143096. [CrossRef]
5. Shogrkhodaei, S.Z.; Razavi-Termeh, S.V.; Fathnia, A. Spatio-temporal modeling of $PM_{2.5}$ risk mapping using three machine learning algorithms. *Environ. Pollut.* **2021**, *289*, 117859. [CrossRef]
6. Taghavi, M.; Ghanizadeh, G.; Ghasemi, M.; Fassò, A.; Hoek, G.; Hushmandi, K.; Raei, M. Application of functional principal component analysis in the spatiotemporal land-use regression modeling of $PM_{2.5}$. *Atmosphere* **2023**, *14*, 926. [CrossRef]
7. Habibi, R.; Alesheikh, A.A.; Mohammadinia, A.; Sharif, M. An assessment of spatial pattern characterization of air pollution: A case study of CO and $PM_{2.5}$ in Tehran, Iran. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 270. [CrossRef]

8. Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Choi, S.M. Effects of air pollution in spatio-temporal modeling of asthma-prone areas using a machine learning model. *Environ. Res.* **2021**, *200*, 111344. [CrossRef]
9. Ghajari, Y.E.; Kaveh, M.; Martín, D. Predicting PM10 Concentrations Using Evolutionary Deep Neural Network and Satellite-Derived Aerosol Optical Depth. *Mathematics* **2023**, *11*, 4145. [CrossRef]
10. Delavar, M.R.; Gholami, A.; Shiran, G.R.; Rashidi, Y.; Nakhaeizadeh, G.R.; Fedra, K.; Hatefi Afshar, S. A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 99. [CrossRef]
11. Taghvaee, S.; Sowlat, M.H.; Mousavi, A.; Hassanvand, M.S.; Yunesian, M.; Naddafi, K.; Sioutas, C. Source apportionment of ambient $PM_{2.5}$ in two locations in central Tehran using the Positive Matrix Factorization (PMF) model. *Sci. Total Environ.* **2018**, *628*, 672–686. [CrossRef] [PubMed]
12. Mirzaei, M.; Amanollahi, J.; Tzanis, C.G. Evaluation of linear, nonlinear, and hybrid models for predicting $PM_{2.5}$ based on a GTWR model and MODIS AOD data. *Air Qual. Atmos. Health* **2019**, *12*, 1215–1224. [CrossRef]
13. Wu, Y.; Xu, Z.; Xu, L.; Wei, J. Approach Considering Spatiotemporal Heterogeneity for $PM_{2.5}$ Prediction: A Case Study of Xinjiang, China. *Atmosphere* **2024**, *15*, 460. [CrossRef]
14. Pu, Q.; Yoo, E.H. A gap-filling hybrid approach for hourly $PM_{2.5}$ prediction at high spatial resolution from multi-sourced AOD data. *Environ. Poll.* **2022**, *315*, 120419. [CrossRef] [PubMed]
15. Sun, Y.; Zeng, Q.; Geng, B.; Lin, X.; Sude, B.; Chen, L. Deep learning architecture for estimating hourly ground-level PM 2.5 using satellite remote sensing. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1343–1347. [CrossRef]
16. Nguyen, P.H.; Dao, N.K.; Nguyen, L.S.P. evelopment of machine learning and deep learning prediction models for $PM_{2.5}$ in Ho Chi Minh City, Vietnam. *Atmosphere* **2024**, *15*, 1163. [CrossRef]
17. Karimian, H.; Li, Y.; Chen, Y.; Wang, Z. Evaluation of different machine learning approaches and aerosol optical depth in $PM_{2.5}$ prediction. *Environ. Res.* **2023**, *216*, 114465. [CrossRef]
18. Kianian, B.; Liu, Y.; Chang, H.H. Imputing satellite-derived aerosol optical depth using a multi-resolution spatial model and random forest for $PM_{2.5}$ prediction. *Remote Sens.* **2021**, *13*, 126. [CrossRef]
19. Feng, H.; Zou, B.; Tang, Y. Scale-and region-dependence in landscape-PM2. 5 correlation: Implications for urban planning. *Remote Sens.* **2017**, *9*, 918. [CrossRef]
20. Wei, J.; Huang, W.; Li, Z.; Xue, W.; Peng, Y.; Sun, L.; Cribb, M. Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach. *Remote Sens. Environ.* **2019**, *231*, 111221. [CrossRef]
21. Zuo, X.; Guo, H.; Shi, S.; Zhang, X. Comparison of six machine learning methods for estimating $PM_{2.5}$ concentration using the Himawari-8 aerosol optical depth. *J. Indian Soc. Remote Sens.* **2020**, *48*, 1277–1287. [CrossRef]
22. Yang, Y.; Wang, Z.; Cao, C.; Xu, M.; Yang, X.; Wang, K.; Guo, H.; Gao, X.; Li, J.; Shi, Z. Estimation of $PM_{2.5}$ concentration across china based on multi-source remote sensing data and machine learning methods. *Remote Sens.* **2024**, *16*, 467. [CrossRef]
23. Fang, X.; Zou, B.; Liu, X.; Sternberg, T.; Zhai, L. Satellite-based ground $PM_{2.5}$ estimation using timely structure adaptive modeling. *Remote Sens. Environ.* **2016**, *186*, 152–163. [CrossRef]
24. Nabavi, S.O.; Haimberger, L.; Abbasi, E. Assessing $PM_{2.5}$ concentrations in Tehran, Iran, from space using MAIAC, deep blue, and dark target AOD and machine learning algorithms. *Atmos. Pollut. Res.* **2019**, *10*, 889–903. [CrossRef]
25. Bagheri, H. Using deep ensemble forest for high-resolution mapping of $PM_{2.5}$ from MODIS MAIAC AOD in Tehran, Iran. *Environ. Monit. Assess.* **2023**, *195*, 377. [CrossRef] [PubMed]
26. Bagheri, H. A machine learning-based framework for high resolution mapping of $PM_{2.5}$ in Tehran, Iran, using MAIAC AOD data. *Adv. Space Res.* **2022**, *69*, 3333–3349. [CrossRef]
27. Faraji, M.; Nadi, S.; Ghaffarpasand, O.; Homayoni, S.; Downey, K. An integrated 3D CNN-GRU deep learning method for short-term prediction of $PM_{2.5}$ concentration in urban environment. *Sci. Total Environ.* **2022**, *834*, 155324. [CrossRef]
28. Zamani Joharestani, M.; Cao, C.; Ni, X.; Bashir, B.; Talebiesfandarani, S. $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere* **2019**, *10*, 373. [CrossRef]
29. Handschuh, J.; Erbertseder, T.; Baier, F. On the added value of satellite AOD for the investigation of ground-level $PM_{2.5}$ variability. *Atmos. Environ.* **2024**, *331*, 120601. [CrossRef]
30. Li, T.; Shen, H.; Yuan, Q.; Zhang, X.; Zhang, L. Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: A geo-intelligent deep learning approach. *Geophys. Res. Lett.* **2017**, *44*, 11–985. [CrossRef]
31. Zhang, T.; Gong, W.; Wang, W.; Ji, Y.; Zhu, Z.; Huang, Y. Ground level $PM_{2.5}$ estimates over China using satellite-based geographically weighted regression (GWR) models are improved by including NO2 and enhanced vegetation index (EVI). *Int. J. Environ. Res. Public Health* **2016**, *13*, 1215. [CrossRef]
32. Chen, G.; Li, S.; Knibbs, L.D.; Hamm, N.A.; Cao, W.; Li, T.; Guo, J.; Ren, H.; Abramson, M.; Guo, Y. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. *Sci. Total Environ.* **2018**, *636*, 52–60. [CrossRef] [PubMed]

33. Ma, X.; Zou, B.; Deng, J.; Gao, J.; Longley, I.; Xiao, S.; Guo, B.; Wu, Y.; Xu, T.; Xu, X.; et al. A comprehensive review of the development of land use regression approaches for modeling spatiotemporal variations of ambient air pollution: A perspective from 2011 to 2023. *Environ. Int.* **2024**, *183*, 108430. [CrossRef] [PubMed]
34. Al-Selwi, S.M.; Hassan, M.F.; Abdulkadir, S.J.; Muneer, A.; Sumiea, E.H.; Alqushaibi, A.; Ragab, M.G. RNN-LSTM: From applications to modeling techniques and beyond—Systematic review. *J. King Saud Univ.-Comput. Inf. Sci.* **2024**, *321*, 114698. [CrossRef]
35. Abou Houran, M.; Bukhari, S.M.S.; Zafar, M.H.; Mansoor, M.; Chen, W. COA-CNN-LSTM: Coati optimization algorithm-based hybrid deep learning model for PV/wind power forecasting in smart grid applications. *Appl. Energy* **2023**, *349*, 121638. [CrossRef]
36. Wan, A.; Chang, Q.; Khalil, A.B.; He, J. Short-term power load forecasting for combined heat and power using CNN-LSTM enhanced by attention mechanism. *Energy* **2023**, *282*, 128274. [CrossRef]
37. Zhang, Q.; Kong, W.; Zhou, G.; Wu, C.; Cui, E. ISSA-LSTM: A new data-driven method of heat load forecasting for building air conditioning. *Energy Build.* **2024**, *36*, 102068.
38. Kaveh, M.; Mesgari, M.S. Application of meta-heuristic algorithms for training neural networks and deep learning archi-tec-tures: A comprehensive review. *Neural Process. Lett.* **2022**, *55*, 4519–4622. [CrossRef]
39. Shoeibi, M.; Nevisi, M.M.S.; Khatami, S.S.; Martín, D.; Soltani, S.; Aghakhani, S. Improved IChOA-Based Reinforcement Learning for Secrecy Rate Optimization in Smart Grid Communications. *Comput. Mater. Contin.* **2024**, *81*, 2819–2843. [CrossRef]
40. Kaveh, M.; Mesgari, M.S.; Saeidian, B. Orchard Algorithm (OA): A new meta-heuristic algorithm for solving discrete and continuous optimization problems. *Math. Comput. Simul.* **2023**, *208*, 91–135. [CrossRef]