

Article

An Optimization-Driven Fuzzy Transformer–Deep Belief Network for PM_{2.5} Air Pollution Prediction: A Spatio-Temporal Framework Based on Aerosol Optical Depth

Mohammad Mehdi Sharifi Nevisi ¹, Pardis Sadatian Moghaddam ², Mehrdad Kaveh ¹, Diego Martín ^{1,*}, Nuria Serrano ¹ and José Vicente Álvarez-Bravo ¹

¹ Department of Computer Science, Escuela de Ingeniería Informática de Segovia, Universidad de Valladolid, 40005 Segovia, Spain; mahdi.shn1374@gmail.com (M.M.S.N.); mehrdadkaveh11@gmail.com (M.K.); nuria.serrano@uva.es (N.S.); jvalvarez@uva.es (J.V.Á.-B.)

² Department of Computer Science, University of Georgia, Athens, GA 30602, USA; ps24068@uga.edu

* Correspondence: diego.martin.andres@uva.es

Abstract

Forecasting fine particulate matter with a diameter of 2.5 μm (PM_{2.5}) is critically important due to its adverse effects on human health and environmental sustainability. Although ground-based monitoring stations provide accurate measurements, their limited spatial coverage restricts large-scale PM_{2.5} assessment, especially in complex urban regions. Consequently, aerosol optical depth (AOD) derived from satellite imagery, combined with advanced deep learning (DL) techniques, has emerged as an effective alternative by offering wide spatial coverage and rich spatio-temporal information. This paper proposed an optimization-driven fuzzy transformer–deep belief network (ODFT-DBN) for accurate PM_{2.5} air pollution prediction. The proposed framework integrates a fuzzy inference module to model uncertainty and nonlinear environmental relationships, a transformer encoder to capture long-range spatio-temporal dependencies, and a DBN to extract hierarchical features and improve prediction robustness. In addition, a novel multi-objective gray wolf optimizer (NMOGWO) is employed to jointly optimize the model hyper-parameters and fuzzy membership functions. The proposed approach is implemented for the city of Tehran, Iran, using meteorological variables, topographical features, ground-based PM_{2.5} measurements, and satellite-derived AOD data. The ODFT-DBN model is compared with several benchmark methods, including bidirectional encoder representations from transformers (BERT), transformer, long short-term memory (LSTM), gated recurrent unit (GRU), convolutional neural network (CNN), DBN, and extreme gradient boosting (XGBoost). Experimental results demonstrate that the proposed framework achieves superior predictive performance, attaining an R^2 value of 0.94 and root mean square error (RMSE) of 0.8 $\mu\text{g}/\text{m}^3$. Scatter plot analyses indicate a strong agreement between predicted and observed PM_{2.5} values, while the proposed model exhibits low variance, stable convergence behavior, and acceptable computational time. Overall, the results confirm the effectiveness, robustness, and practical applicability of the proposed ODFT-DBN framework for spatio-temporal PM_{2.5} forecasting.

Academic Editor: Gaige Wang

Received: 29 May 2026

Revised: 30 June 2026

Accepted: 1 July 2026

Published: 5 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: PM_{2.5} air pollution; aerosol optical depth; fuzzy transformer; deep belief network; novel multi-objective gray wolf optimizer

MSC: 68T07

1. Introduction

Rapid urbanization, industrial expansion, population growth, and increasing energy consumption have significantly intensified air-pollution problems across modern metropolitan regions [1]. In recent decades, large urban areas have experienced substantial increases in vehicular traffic, industrial emissions, fossil-fuel combustion, and construction activities, all of which contribute directly to atmospheric pollution accumulation. In addition, climate change, reduced green-space coverage, and unfavorable meteorological conditions have further aggravated urban air-quality deterioration. As a result, air pollution has become one of the most serious environmental challenges affecting both developed and developing countries, particularly in densely populated cities where pollutant emissions continuously exceed safe atmospheric thresholds [2–4].

The environmental impacts of urban air pollution extend far beyond atmospheric degradation alone and directly affect public health, economic productivity, transportation systems, and overall quality of life. High pollution levels can reduce visibility, damage ecosystems, increase healthcare expenditures, and negatively affect workforce productivity due to pollution-related diseases [5]. Furthermore, urban air pollution imposes considerable economic burdens on governments and healthcare systems through increased medical costs, environmental mitigation expenses, and reduced labor efficiency. Consequently, accurate monitoring and forecasting of air pollution have become critically important for supporting environmental management policies, smart-city planning, and sustainable urban development strategies [6–8].

Among different atmospheric pollutants, particulate matter with an aerodynamic diameter smaller than $2.5\ \mu\text{m}$ ($\text{PM}_{2.5}$) is considered one of the most hazardous pollutants due to its extremely small particle size and ability to penetrate deeply into the human respiratory system [6–8]. Unlike larger particulate pollutants, $\text{PM}_{2.5}$ particles can enter the lungs and bloodstream, leading to severe respiratory illnesses, cardiovascular diseases, lung cancer, asthma aggravation, and premature mortality. Long-term exposure to elevated $\text{PM}_{2.5}$ concentrations has additionally been associated with neurological disorders, reduced life expectancy, and increased vulnerability among children, elderly populations, and individuals with pre-existing health conditions [9–11]. Due to these severe health implications, $\text{PM}_{2.5}$ has become one of the primary pollutants monitored by environmental protection agencies and public-health organizations worldwide [12].

Tehran, the capital and largest metropolitan city of Iran, represents a prominent example of a highly polluted urban environment facing severe $\text{PM}_{2.5}$ challenges [4]. The city experiences persistent air-quality problems due to a combination of heavy traffic congestion, dense population, industrial emissions, rapid urbanization, and unfavorable geographical conditions. Tehran is surrounded by the Alborz Mountains, which restrict atmospheric circulation and intensify pollutant accumulation, particularly during cold seasons characterized by temperature inversion phenomena. During winter months, stable atmospheric layers trap pollutants near the ground, causing severe $\text{PM}_{2.5}$ concentration increases across central and southern urban regions [8]. These environmental and geographical conditions make Tehran an important case study for developing accurate and robust $\text{PM}_{2.5}$ forecasting frameworks capable of supporting real-world environmental monitoring and pollution-management applications [4,8].

One of the major challenges in urban air-quality monitoring is the limited spatial coverage of ground-based pollution monitoring stations. Although these stations provide highly accurate pollutant measurements, their installation, maintenance, and operational costs are typically expensive, resulting in sparse and uneven station distributions across large metropolitan areas [13–15]. Consequently, many urban regions lack sufficient monitoring coverage, making it difficult to accurately capture the spatial variability of $\text{PM}_{2.5}$ concentrations throughout the city. This limitation becomes particularly problematic in

highly dynamic urban environments where pollution levels may vary significantly between industrial, residential, commercial, and traffic-intensive regions. Therefore, accurate PM_{2.5} prediction and spatial estimation frameworks have become increasingly important for compensating for the insufficient spatial density of ground-monitoring infrastructures [16].

To address these limitations, remote-sensing (RS) technologies and satellite-derived environmental parameters have recently attracted considerable attention in air-pollution forecasting applications [4]. In particular, aerosol optical depth (AOD) has emerged as one of the most important satellite-derived indicators for PM_{2.5} estimation due to its strong relationship with atmospheric aerosol concentration. Unlike ground-monitoring stations with limited spatial coverage, satellite imagery provides continuous large-scale observations across entire urban regions, enabling more comprehensive spatial monitoring of atmospheric conditions [17–19]. AOD measurements obtained from satellite sensors reflect the degree of light extinction caused by suspended aerosol particles in the atmosphere and therefore provide valuable information regarding pollutant distributions and aerosol loading conditions [20–22].

Recent studies have demonstrated that integrating AOD data with meteorological variables and deep learning (DL) techniques can substantially improve PM_{2.5} prediction accuracy and spatial coverage [23–25]. Since PM_{2.5} particles constitute a major component of atmospheric aerosols, strong correlations are frequently observed between AOD measurements and PM_{2.5} concentrations, particularly under stable atmospheric conditions. Furthermore, satellite-derived AOD enables continuous monitoring even in urban regions where ground stations are unavailable or sparsely distributed. The combination of RS data and intelligent learning frameworks therefore provides a highly promising solution for developing accurate, scalable, and cost-effective air-quality forecasting systems capable of supporting modern smart-city environmental-management applications [26–28].

In recent years, machine learning (ML) and DL techniques have played an increasingly important role in air-pollution forecasting due to their strong capability for modeling complex nonlinear relationships within environmental datasets [29,30]. Unlike conventional statistical approaches, ML and DL architectures can automatically learn hidden patterns, temporal dependencies, and spatial interactions from large-scale heterogeneous atmospheric data. As a result, these intelligent learning methods have significantly improved the accuracy and robustness of PM_{2.5} prediction systems. Traditional ML algorithms such as extreme gradient boosting (XGBoost) and random forests have demonstrated effective regression capability for environmental forecasting, while advanced DL architectures including bidirectional encoder representations from transformers (BERT), convolutional neural networks (CNNs), long short-term memory networks (LSTMs), gated recurrent units (GRUs), transformers, and deep belief networks (DBNs) have further enhanced the capability of pollution prediction frameworks by enabling hierarchical feature extraction and long-range temporal dependency learning [31–33].

Among modern DL approaches, transformer-based architectures have recently attracted substantial attention because of their self-attention mechanisms and ability to model complex spatio-temporal dependencies more effectively than conventional recurrent structures [25–27]. Similarly, DBN-based architectures have demonstrated strong capability for hierarchical probabilistic feature learning and nonlinear environmental representation [28,29]. In addition, recent studies increasingly suggest hybrid intelligent frameworks that combine multiple learning paradigms to exploit their complementary advantages. For example, integrating fuzzy reasoning with DL architectures can improve uncertainty representation, while combining attention-based learning with hierarchical feature extraction can significantly enhance environmental pattern recognition capability. These hybrid architectures have demonstrated promising performance improvements in

complex environmental forecasting applications and indicate a growing research trend toward cooperative intelligent learning systems for accurate and robust PM_{2.5} prediction.

1.1. Related Works and Research Gaps

In recent years, extensive research efforts have focused on developing intelligent frameworks for PM_{2.5} prediction using heterogeneous environmental datasets and advanced ML/DL techniques. Due to the limited spatial coverage of ground-based monitoring stations, many studies have increasingly incorporated satellite-derived remote-sensing variables, particularly AOD, together with meteorological observations, geographical information, and air-quality measurements to improve spatial pollution estimation and forecasting capability [30–32]. The strong relationship between AOD and atmospheric particulate concentration has made satellite imagery an important complementary data source for large-scale PM_{2.5} monitoring applications. In parallel, the rapid advancement of computational intelligence and DL architectures has significantly improved the capability of environmental forecasting systems to model nonlinear atmospheric dynamics and complex spatio-temporal pollutant relationships [33–35]. Consequently, various ML and DL approaches, including convolutional neural networks, recurrent neural networks, transformers, graph-based architectures, deep belief networks, and hybrid optimization-driven frameworks, have been proposed in the literature to enhance PM_{2.5} prediction accuracy, robustness, and generalization capability under diverse environmental conditions [36].

In [30], the authors proposed a decoder-only air transformer network (D-ATN) framework for forecasting PM_{2.5}, PM₁₀, and air quality index (AQI) levels in several major Indian metropolitan cities, including Delhi, Mumbai, Bengaluru, Chennai, and Kolkata. The proposed approach integrated historical pollutant observations with meteorological variables to improve spatio-temporal air-pollution prediction capability. The study mainly focused on exploiting transformer-based attention mechanisms to capture long-range temporal dependencies within atmospheric data. The proposed D-ATN model was evaluated against several conventional and DL architectures, including autoregressive integrated moving average (ARIMA), XGBoost, random forest (RF), LSTM, and standard transformer models. Experimental results demonstrated that the proposed transformer-based framework achieved improved prediction accuracy compared with conventional approaches. However, the study mainly relied on attention-based temporal learning and did not incorporate uncertainty-aware fuzzy reasoning, hierarchical probabilistic feature extraction, or adaptive multi-objective optimization mechanisms, which may limit its capability to fully model highly nonlinear and uncertain environmental relationships associated with complex PM_{2.5} dynamics.

In [31], the authors proposed a multi-view ML framework for estimating PM_{2.5} concentrations using smartphone photographs and rigorously spatiotemporally registered environmental datasets collected from Hebei and Fujian provinces in China between 2022 and 2024. Unlike conventional air-pollution forecasting studies that rely mainly on meteorological and monitoring-station data, the proposed approach utilized semantic image features extracted from multiple regions within smartphone images, including vegetation and sky regions. An XGBoost model was employed to establish the relationship between extracted visual features and PM_{2.5} concentrations. Experimental results demonstrated strong prediction performance, achieving R^2 values of 0.96 and 0.89, with root mean square error (RMSE) values of 6.82 $\mu\text{g}/\text{m}^3$ and 3.63 $\mu\text{g}/\text{m}^3$ for the Hebei and Fujian datasets, respectively. The study highlighted the importance of multi-view image analysis and semantic feature extraction for improving image-based air-quality estimation.

In [32], the authors investigated PM_{2.5} prediction in Jaipur city using air-pollution measurements and meteorological variables collected between 2019 and 2023. The study evaluated several ML and DL approaches, including multiple linear regression (MLR),

support vector regression (SVR), artificial neural network (ANN), RF, k-nearest neighbors (KNN), GRU, and CNN. A total of 39,645 environmental observations were analyzed after preprocessing and multicollinearity assessment. The results revealed strong relationships between PM_{2.5} concentrations and gaseous pollutants such as NO₂, SO₂, and NH₃, while sensitivity analysis indicated that SO₂ and O₃ exerted significant influence on PM_{2.5} behavior. Among all comparative models, the CNN architecture achieved the best performance with an R^2 value of 0.98 and lower prediction errors compared with the other methods. The study demonstrated the effectiveness of DL-based environmental prediction frameworks for urban air-quality analysis.

In [33], the authors investigated monthly PM_{2.5} concentration modeling and continuous spatial-distribution estimation using several DL architectures under the environmental conditions of Iran. The study integrated satellite-derived variables, ground-based monitoring observations, and meteorological parameters to improve spatial air-pollution estimation capability. Several DL approaches, including multilayer perceptron (MLP), CNN, LSTM, and convolutional long short-term memory (ConvLSTM), were implemented and comparatively evaluated using RMSE and R^2 metrics. Experimental results demonstrated that the ConvLSTM architecture achieved the best overall prediction performance with an RMSE of 4.95 $\mu\text{g}/\text{m}^3$ and an R^2 value of 91.24%. In addition, sensitivity analysis revealed that population density, AOD, NO₂, and rainfall were among the most influential environmental variables affecting PM_{2.5} concentrations. The study highlighted the effectiveness of combining satellite observations and DL techniques for large-scale air-pollution estimation.

In [34], the authors proposed a hybrid PM_{2.5} prediction framework combining the whale optimization algorithm (WOA), CNN, LSTM, and an attention mechanism (AM) for daily air-pollution forecasting. The model was trained using meteorological and air-pollution data collected between 2014 and 2018, aiming to improve both short- and medium-term PM_{2.5} prediction performance. The proposed WOA-CNN-LSTM-AM architecture integrated convolutional feature extraction, temporal sequence learning, and attention-based contextual weighting, while the WOA optimizer was employed to enhance parameter tuning and prediction efficiency. Experimental results demonstrated strong forecasting capability, achieving mean absolute percentage error (MAPE), RMSE, and R^2 values of 14.29, 21.96, and 0.93, respectively. In addition, the model reduced prediction errors by 39% compared with CNN and 34% compared with standalone LSTM architectures.

In [35], the authors proposed a spatio-temporal PM_{2.5} prediction framework based on the integration of ConvLSTM and graph convolutional network (GCN) architectures for hourly air-pollution forecasting across the Los Angeles region. The proposed model utilized multisource remote-sensing imagery and environmental sensor data instead of explicit atmospheric physics or chemical mechanism modeling. Specifically, satellite observations from MODIS, TROPOMI, and wildfire-related fire radiative power data were incorporated to capture the major environmental factors affecting PM_{2.5} concentrations. The combination of ConvLSTM and GCN enabled the framework to jointly model temporal pollutant evolution and spatial correlations among different monitoring regions. Experimental results demonstrated significant improvements in RMSE and normalized RMSE compared with existing spatio-temporal prediction approaches, highlighting the effectiveness of integrating graph-based learning with sequential DL architectures for large-scale environmental forecasting.

In [36], the authors investigated PM_{2.5} concentration estimation in Tehran, Iran, by integrating AOD observations with several ML techniques. The study utilized high-resolution MAIAC AOD products with a spatial resolution of 1 km together with DB_DT AOD products with a coarser spatial resolution of 10 km derived from NASA Terra and Aqua

satellite platforms. Ground-based PM_{2.5} observations collected from 18 monitoring stations between 2011 and 2016 were employed for model development and validation. To improve the relationship between satellite-derived aerosol measurements and ground-level PM_{2.5} concentrations, several meteorological variables, including relative humidity and planetary boundary layer height, were additionally incorporated into the prediction framework. Among the evaluated ML approaches, the random forest (RF) model achieved the best overall performance, obtaining an RMSE value of 17.52 µg/m³ and an R^2 value of 0.68. The study demonstrated the effectiveness of integrating satellite observations and meteorological information for urban PM_{2.5} estimation. However, the authors reported several important limitations, including reduced prediction accuracy during summer conditions due to insufficient representation of aerosol precursor dynamics and secondary pollutant formation processes. In addition, the relatively low spatial resolution of the DB_DT AOD products limited the capability of the framework to accurately distinguish complex urban land-cover characteristics and highly heterogeneous pollution patterns across densely populated metropolitan regions.

In [37], the authors introduced a hybrid spatio-temporal PM_{2.5} forecasting framework based on the integration of three-dimensional CNN (3D CNNs) and GRUs for air-quality prediction in Tehran, Iran. The study utilized environmental data collected between 2016 and 2019 from 13 air-quality monitoring stations together with meteorological and auxiliary environmental variables. The proposed architecture employed 3D CNN layers to extract spatial pollutant features and GRUs to model temporal pollutant evolution and sequential atmospheric dependencies. Experimental results demonstrated that the proposed 3D CNN-GRU framework achieved superior prediction performance compared with traditional ML approaches such as SVR as well as several standalone DL architectures, including LSTM and GRU models. The proposed framework obtained an R^2 value of 0.84 and an RMSE of 7.15 µg/m³, highlighting the effectiveness of combining spatial feature extraction and temporal sequence learning for PM_{2.5} forecasting. Nevertheless, the study reported challenges associated with missing environmental observations and inconsistencies within auxiliary datasets, which could negatively affect model robustness and prediction stability under complex atmospheric conditions.

In [38], the authors proposed a GRU-based encoder–decoder (GRU-ED) framework for forecasting hourly PM_{2.5} concentrations in New Delhi, India. The study investigated short-term air-pollution prediction at 1 h, 8 h, and 24 h forecasting horizons using environmental data collected between 2008 and 2010. To analyze the influence of different environmental factors, four distinct input configurations were evaluated, incorporating combinations of meteorological variables, vehicle-population information, and emission-related data. The proposed GRU-ED architecture was comparatively assessed against several conventional ML and DL approaches. Experimental results demonstrated that the GRU-ED framework achieved superior forecasting capability, particularly when all environmental factors were jointly utilized. The best-performing configuration obtained an R^2 value of 0.959 together with strong performance across multiple evaluation metrics. The study additionally highlighted that meteorological variables exerted the strongest influence on PM_{2.5} prediction accuracy among the evaluated environmental factors.

In [39], the authors investigated satellite-based PM_{2.5} estimation in Tehran, Iran, using AOD observations retrieved from the MODIS sensor with a spatial resolution of 1 km over a ten-year period spanning 2012 to 2023. The study focused on estimating ground-level PM_{2.5} concentrations for six air-quality monitoring stations and comparing the predicted values with measured observations obtained from the Tehran Air Quality Control Company. To improve the reliability of the AOD–PM_{2.5} relationship, several environmental and meteorological conditioning variables were incorporated into the framework, includ-

ing wind speed, wind direction, temperature, precipitation, NDVI, and visibility. In addition, genetic algorithm optimization and recursive feature elimination (RFE) were utilized to identify the most influential environmental features affecting PM_{2.5} estimation accuracy. Three ML approaches, namely RF, SVM, and Gaussian process regression (GPR), were comparatively evaluated. Experimental results demonstrated that the RF model achieved the best overall performance, providing prediction accuracies ranging from 94% to 98% across the studied monitoring stations. The study highlighted the effectiveness of combining high-resolution satellite observations with environmental feature-selection strategies for urban air-quality estimation. However, the proposed framework mainly relied on conventional ML regression techniques and did not investigate advanced hybrid DL architectures capable of modeling complex spatio-temporal dependencies, uncertainty-aware environmental interactions, or adaptive multi-objective optimization mechanisms for robust PM_{2.5} forecasting under highly dynamic atmospheric conditions.

Despite the significant progress achieved in PM_{2.5} forecasting studies, several important research gaps still remain unresolved in the existing literature. One major limitation is the sparse spatial distribution of ground-based air-quality monitoring stations across large urban regions. The high installation, maintenance, and calibration costs associated with monitoring infrastructures restrict their deployment density, leading to insufficient spatial coverage in many metropolitan areas. Consequently, accurately capturing localized pollution variations and heterogeneous urban atmospheric conditions remains challenging. In addition, many existing forecasting frameworks still rely on conventional or shallow learning strategies that cannot fully model highly nonlinear atmospheric interactions and evolving pollutant dynamics under complex urban conditions.

Another important research gap is related to the limited capability of many existing forecasting models to effectively exploit advanced intelligent learning paradigms and hybrid optimization-driven architectures. Although DL approaches such as CNNs, LSTMs, and transformers have demonstrated promising performance improvements, many current studies still employ standalone learning structures without integrating complementary learning mechanisms. As a result, the simultaneous modeling of uncertainty, hierarchical feature abstraction, and long-range spatio-temporal dependencies remains insufficiently explored. Furthermore, comprehensive comparative investigations among multiple advanced ML and DL architectures are still limited in the literature, making it difficult to fully evaluate the strengths and weaknesses of different environmental forecasting approaches under identical atmospheric conditions.

The quality and structure of environmental datasets additionally represent another critical challenge in PM_{2.5} forecasting applications. Many environmental datasets suffer from missing observations, temporal inconsistencies, sensor noise, and heterogeneous spatial resolutions, which can significantly degrade prediction accuracy and model stability. However, several existing studies provide limited discussion regarding comprehensive data preprocessing strategies and robust refinement procedures prior to model implementation. Moreover, the influence of spatially and temporally aligned heterogeneous datasets on forecasting performance remains insufficiently investigated, particularly when integrating satellite-derived observations with ground-based environmental measurements.

Although satellite-derived environmental parameters such as AOD have recently attracted substantial attention for air-pollution estimation, their effective integration within intelligent forecasting systems still faces important challenges. Many existing studies utilize AOD as an auxiliary feature without comprehensively investigating its nonlinear interaction with PM_{2.5} concentrations under varying meteorological and atmospheric conditions. Furthermore, limited research has explored uncertainty-aware modeling mecha-

nisms capable of improving the robustness of AOD-based environmental forecasting under noisy or incomplete observational conditions. As a result, the full potential of remote-sensing variables for high-resolution PM_{2.5} prediction remains underexploited.

Finally, many current air-pollution forecasting frameworks still exhibit limited generalization capability when applied to dynamic real-world urban environments. Existing models are often optimized for specific datasets or restricted geographical conditions and may experience performance degradation when exposed to unseen atmospheric patterns or spatially distinct regions. Moreover, the integration of auxiliary information sources such as socioeconomic indicators, urban activity patterns, and unknown hidden environmental factors remains relatively unexplored in current PM_{2.5} prediction studies. Therefore, developing robust, adaptive, and uncertainty-aware hybrid forecasting frameworks capable of handling heterogeneous environmental conditions and evolving real-world atmospheric dynamics remains an important open research direction.

1.2. Motivation and Contribution

The motivation behind utilizing satellite-derived AOD data in this study originates from the limited spatial coverage and insufficient density of conventional ground-based air-quality monitoring stations in large metropolitan regions such as Tehran. Although monitoring stations provide highly accurate pollutant measurements, they are geographically sparse and incapable of comprehensively capturing fine-grained spatial variations in PM_{2.5} concentrations across highly heterogeneous urban environments. Satellite-derived AOD products offer extensive spatial coverage and continuous atmospheric observations, enabling large-scale environmental monitoring with significantly higher spatial representativeness. In recent years, numerous studies have demonstrated the strong correlation between AOD and ground-level PM_{2.5} concentrations, making AOD an effective complementary environmental indicator for air-pollution estimation and forecasting applications.

In addition to AOD, this study integrates heterogeneous environmental datasets, including meteorological observations, geographical variables, vegetation indicators, and remote-sensing parameters, to construct a more comprehensive environmental representation framework. Variables such as temperature, humidity, pressure, wind speed, wind direction, precipitation, land surface temperature (LST), water vapor, normalized difference vegetation index (NDVI), elevation, and spatial coordinates provide important contextual information regarding atmospheric dynamics, pollutant transportation, aerosol accumulation, and seasonal environmental variability. Since PM_{2.5} behavior is strongly affected by complex nonlinear interactions among meteorological, geographical, and anthropogenic factors, relying on a single environmental variable may lead to limited forecasting robustness and insufficient generalization capability. Therefore, integrating multisource environmental observations improves the capability of the proposed framework to capture hidden atmospheric dependencies and heterogeneous spatio-temporal pollution patterns.

The motivation for integrating transformer and DBN architectures arises from the complementary learning capabilities of these two DL paradigms. Transformer architectures are highly effective in modeling long-range temporal dependencies and sequential atmospheric relationships through self-attention mechanisms. In complex environmental forecasting problems, pollutant evolution may depend on distant temporal observations and nonlinear temporal interactions that conventional recurrent architectures often fail to capture efficiently. Therefore, transformers provide a powerful mechanism for extracting global contextual relationships and temporal attention patterns from heterogeneous environmental sequences.

On the other hand, DBN architectures provide strong hierarchical probabilistic feature extraction capability and are highly effective for discovering hidden nonlinear environmental representations from complex multidimensional datasets. While transformers primarily focus on long-range dependency modeling, DBNs are capable of generating deep latent feature abstractions and enhancing representation robustness under noisy environmental conditions. Motivated by these complementary strengths, the proposed ODFT-DBN framework combines transformer-based temporal attention learning with DBN-based hierarchical feature abstraction to simultaneously improve contextual sequence modeling, nonlinear environmental representation learning, and forecasting stability. This hybrid integration addresses several limitations observed in standalone DL architectures commonly used in previous PM_{2.5} forecasting studies.

Another major motivation of this study is the incorporation of fuzzy logic within the proposed framework. Environmental systems and atmospheric pollutant dynamics inherently contain uncertainty, vagueness, incomplete observations, and nonlinear relationships that are difficult to model using crisp deterministic learning approaches alone. Fuzzy systems provide uncertainty-aware reasoning capability by representing environmental relationships using soft membership functions and fuzzy inference mechanisms. Integrating fuzzy reasoning into the proposed architecture improves the capability of the framework to handle ambiguous atmospheric conditions, noisy environmental observations, and uncertain pollutant behaviors more effectively. Furthermore, the fuzzy module enhances interpretability and adaptive environmental representation within the hybrid forecasting framework.

The motivation for employing the novel multi-objective gray wolf optimizer (NMOGWO) optimization strategy originates from the highly nonlinear and high-dimensional optimization characteristics of the proposed hybrid architecture. The simultaneous optimization of transformer, DBN, and fuzzy-system hyper-parameters using conventional gradient-based or trial-and-error approaches may lead to local optima, unstable convergence behavior, and insufficient exploration of the solution space. Therefore, an enhanced multi-objective gray wolf optimizer was utilized to improve global search capability, convergence stability, and optimization robustness. The proposed NMOGWO framework enables simultaneous optimization of multiple conflicting objective functions related to prediction accuracy, network complexity, hidden neuron configuration, and fuzzy-system performance, thereby improving both model efficiency and forecasting reliability.

Finally, another important motivation of this work is the implementation of a comprehensive data preprocessing and refinement procedure prior to model development. Environmental datasets frequently contain missing observations, inconsistent temporal resolutions, noisy measurements, and heterogeneous spatial representations, which can significantly degrade model performance and forecasting reliability. Therefore, extensive preprocessing procedures, including normalization, temporal alignment, spatial synchronization, feature refinement, and missing-data handling, were employed to improve data consistency and environmental representation quality. This preprocessing pipeline enhances the capability of the proposed ODFT-DBN framework to learn robust atmospheric relationships from heterogeneous multisource environmental observations while reducing the influence of noise and incomplete measurements on forecasting performance.

Unlike existing PM_{2.5} forecasting studies that primarily rely on standalone DL models or simple hybrid architectures, the proposed ODFT-DBN framework introduces several methodological and optimization innovations within a unified uncertainty-aware spatio-temporal prediction framework. First, a fuzzy-guided attention mechanism is developed to incorporate fuzzy relevance information directly into the transformer attention process,

enabling uncertainty-aware temporal dependency learning and improving the representation of complex atmospheric interactions. Second, an adaptive transformer–DBN fusion strategy is proposed to jointly exploit transformer-based contextual representations and DBN-based hierarchical latent features, allowing the model to simultaneously capture long-range spatio-temporal dependencies and deep nonlinear environmental patterns. Third, a novel NMOGWO is introduced, which incorporates a theta-guided leadership mechanism, Pareto-based non-dominated sorting, crowding-distance preservation, and an adaptive nonlinear exploration–exploitation control strategy to improve convergence stability, population diversity, and optimization robustness. Furthermore, the proposed NMOGWO simultaneously optimizes transformer hyper-parameters, DBN structural parameters, fuzzy membership functions, and fuzzy-rule configurations within a unified multi-objective framework. To the best of our knowledge, the integration of fuzzy-guided transformer learning, DBN-based hierarchical feature abstraction, and the proposed enhanced NMOGWO for AOD-assisted PM_{2.5} forecasting has not been previously investigated for Tehran’s highly heterogeneous atmospheric environment. Therefore, the proposed framework provides a novel and comprehensive solution for uncertainty-aware, robust, and accurate PM_{2.5} prediction using heterogeneous environmental datasets. Therefore, the main contributions of this study are summarized as follows:

- This paper introduces a novel hybrid ODFT-DBN framework for PM_{2.5} concentration prediction by integrating fuzzy logic, transformer networks, DBNs, and NMOGWO;
- This paper introduces an integrated fuzzy-transformer-DBN learning architecture that combines transformer-based temporal attention mechanisms, DBN-based hierarchical probabilistic feature extraction, and fuzzy reasoning strategies to simultaneously capture long-range spatio-temporal atmospheric dependencies, nonlinear environmental relationships, hidden latent representations, and uncertainty-aware pollutant dynamics for robust PM_{2.5} forecasting under complex urban atmospheric conditions;
- This paper proposes an enhanced NMOGWO optimization framework with improved exploration and exploitation capability through modified leadership updating and multi-objective optimization strategies for simultaneous hyper-parameter optimization of the proposed hybrid architecture;
- This paper utilizes heterogeneous multisource environmental datasets, including ground-based PM_{2.5} measurements, meteorological observations, satellite-derived AOD, NDVI, LST, precipitation, water vapor, elevation, and spatial information collected over Tehran, Iran, as the study region;
- This paper performs extensive comparative analysis against several ML and DL baseline models, including BERT, transformer, DBN, GRU, LSTM, CNN, and XGBoost, using multiple evaluation metrics such as RMSE, R^2 , MAPE, convergence analysis, variance analysis, inference latency, and statistical t -tests;
- This paper demonstrates that the proposed ODFT-DBN framework achieves superior prediction accuracy, convergence stability, robustness, and generalization capability compared with existing baseline approaches, highlighting its effectiveness and potential applicability for future intelligent environmental monitoring and urban air-quality forecasting systems.

1.3. Paper Organization

The remainder of this paper is organized as follows. Section 2 presents the overall workflow of the proposed framework, including the study area, dataset description, data preprocessing procedures, and the detailed mathematical formulation of each module within the proposed ODFT-DBN architecture. This section additionally explains the transformer, DBN, fuzzy reasoning, and NMOGWO optimization components together with

their interaction mechanisms. Section 3 presents the experimental results and comparative performance analysis using multiple statistical and computational evaluation metrics under different environmental conditions. Section 4 provides further discussion regarding model robustness, computational complexity, inference behavior, and real-world generalization capability for practical air-quality forecasting applications. Finally, Section 5 concludes the paper and outlines several future research directions for improving intelligent environmental forecasting systems and extending the proposed framework to other complex spatio-temporal prediction applications.

2. Materials and Methods

Figure 1 presents the overall workflow of the proposed ODFT-DBN framework for $PM_{2.5}$ air pollution prediction. The proposed framework integrates satellite-derived AOD data, ground-based $PM_{2.5}$ measurements, meteorological information, and topographical features within a unified spatio-temporal DL architecture. The primary objective of the proposed model is to accurately estimate $PM_{2.5}$ concentration levels by simultaneously capturing nonlinear environmental relationships, spatial correlations, and temporal dependencies. As illustrated in Figure 1, the framework consists of several sequential stages, including data collection, preprocessing, fuzzy representation, transformer-based feature extraction, DBN-based hierarchical learning, and evolutionary optimization. The detailed mathematical formulations and implementation procedures of each stage are presented in the following subsections.

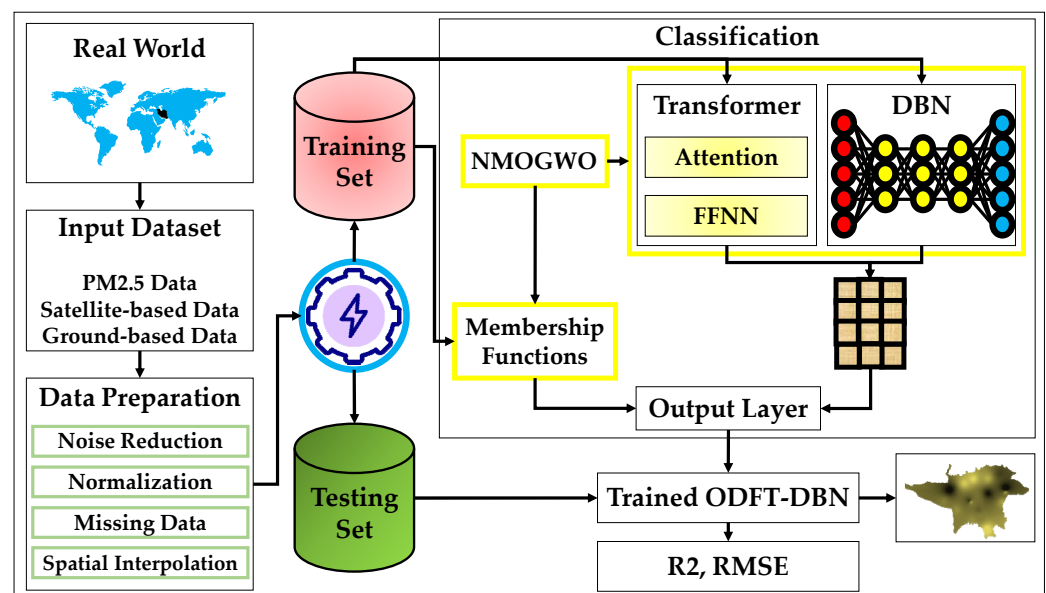


Figure 1. Overall workflow of the proposed framework for $PM_{2.5}$ Air Pollution Prediction.

Initially, real-world environmental datasets are collected from multiple heterogeneous sources. These datasets include satellite-based AOD observations, ground monitoring station $PM_{2.5}$ measurements, meteorological variables, and topographical information associated with the study area. Since $PM_{2.5}$ concentration is affected by complex atmospheric conditions and environmental interactions, combining multiple data sources provides more comprehensive information regarding pollutant behavior across both spatial and temporal dimensions. This multi-source integration improves the capability of the proposed framework to model complex pollution patterns and enhances prediction robustness. After the data acquisition stage, several preprocessing operations are performed to improve data quality and ensure reliable model training. As shown in Figure 1, the pre-

processing stage includes noise reduction, normalization, missing data handling, and spatial interpolation. Noise reduction is applied to eliminate abnormal fluctuations and measurement inconsistencies from the environmental observations. Subsequently, normalization is utilized to scale the variables into a unified numerical range, which facilitates stable convergence during the training process. Missing observations are reconstructed using interpolation-based approaches, while spatial interpolation techniques are employed to preserve the spatial continuity of the environmental information across different geographical regions. Finally, the processed dataset is divided into training and testing subsets for model development and performance evaluation.

Following preprocessing, the proposed framework employs a fuzzy inference mechanism to model uncertainty and nonlinear environmental relationships. PM_{2.5} prediction is inherently associated with uncertain atmospheric dynamics and ambiguous interactions among variables such as temperature, humidity, wind speed, and AOD. Therefore, fuzzy membership functions are incorporated to transform the input variables into fuzzy representations, allowing the framework to better handle uncertainty and improve prediction stability. In addition, the parameters of the fuzzy membership functions are adaptively optimized using the proposed NMOGWO algorithm to further enhance learning capability and prediction accuracy.

The extracted fuzzy representations are subsequently processed through the transformer encoder and DBN architecture. The transformer encoder utilizes a self-attention mechanism and feed-forward neural network (FFNN) layers to capture long-range spatio-temporal dependencies and contextual environmental relationships within sequential observations. This enables the framework to effectively learn complex temporal patterns associated with PM_{2.5} variations. The transformer-generated features are then forwarded to the DBN module, which performs hierarchical feature extraction and nonlinear representation learning. By integrating transformer-based temporal modeling with DBN-based deep feature abstraction, the proposed ODFT-DBN framework can generate more informative representations for accurate PM_{2.5} prediction.

Finally, the proposed NMOGWO algorithm jointly optimizes the hyper-parameters of the transformer-DBN framework and the fuzzy membership functions to improve convergence behavior and predictive performance. After completing the training phase, the optimized ODFT-DBN model is evaluated using the testing dataset. The prediction performance of the framework is assessed using statistical evaluation metrics such as the R^2 and RMSE. Overall, the workflow presented in Figure 1 demonstrates how the proposed framework combines data preprocessing, fuzzy reasoning, transformer-based spatio-temporal learning, hierarchical feature extraction, and evolutionary optimization within a unified prediction system for robust and accurate PM_{2.5} forecasting.

2.1. Case Study

The proposed framework is implemented and evaluated using environmental and atmospheric data collected from Tehran, Iran. As the political, economic, and industrial center of the country, Tehran represents one of the most complex urban environments in the Middle East in terms of air quality dynamics and pollutant distribution. The city is located along the southern edge of the Alborz mountain range and extends toward the semi-arid plains of central Iran. This geographical configuration creates unique atmospheric conditions that strongly influence pollutant accumulation and dispersion processes. Figure 2 illustrates the geographical characteristics of the study region, including the location of Iran on the global map, the position of Tehran province within the country, and the detailed spatial distribution of pollution monitoring stations, meteorological stations, and elevation information across Tehran city [4].

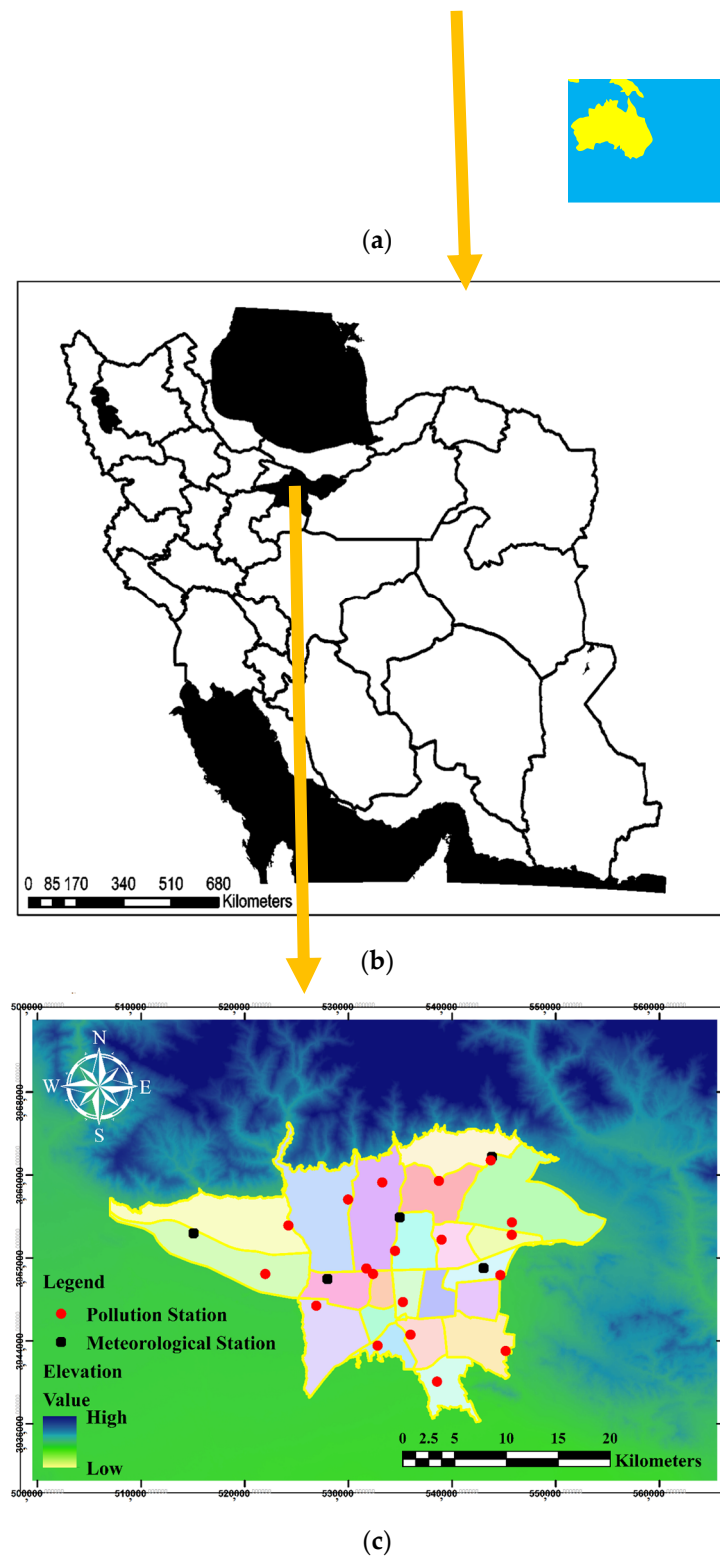


Figure 2. Geographical overview of the study area used in this work: (a) Global map indicating the location of Iran; (b) map of Tehran province within the national boundaries; and (c) detailed spatial representation of Tehran city showing pollution monitoring stations, meteorological stations, and elevation distribution across the study region.

Over the past several decades, extensive urbanization and population growth have transformed Tehran into a highly populated megacity with dense residential, commercial, and industrial districts. The metropolitan region currently hosts millions of residents and experiences substantial daily transportation activity. The continuous expansion of highways, industrial infrastructures, construction projects, and energy consumption has significantly increased environmental pressure on the city. In addition, the concentration of economic and governmental activities within Tehran has intensified traffic flow and fuel consumption, both of which are major contributors to atmospheric pollution. The atmospheric behavior of Tehran is heavily affected by its local climate and topographical conditions. The city experiences a semi-arid climate with dry summers and relatively cold winters. Seasonal meteorological variations strongly impact air quality conditions throughout the year. During colder months, temperature inversion events frequently occur, forming stable atmospheric layers that prevent the vertical movement of pollutants. Consequently, suspended particles and gaseous pollutants remain trapped close to the ground surface for extended periods. Furthermore, the mountainous terrain surrounding the northern part of the city limits natural air circulation and reduces pollutant dispersion efficiency.

Among different air contaminants, PM_{2.5} has emerged as one of the most hazardous pollutants affecting public health and environmental sustainability in Tehran. Fine particulate matter with a diameter smaller than 2.5 μm is particularly dangerous because it can penetrate deeply into the respiratory tract and bloodstream. Long-term exposure to elevated PM_{2.5} concentrations is associated with respiratory illnesses, cardiovascular diseases, reduced lung function, and increased mortality risk. In Tehran, PM_{2.5} concentration levels frequently exceed international air quality standards due to the combined impact of urban activities, industrial emissions, and adverse meteorological conditions. Vehicular emissions are considered one of the primary sources of PM_{2.5} generation in the study area. The city experiences severe traffic congestion on a daily basis, and a considerable portion of the transportation fleet consists of older vehicles with insufficient emission control technologies. In parallel, industrial facilities located inside and around the metropolitan area continuously release particulate pollutants into the atmosphere. Additional pollution sources, including fossil fuel combustion, construction activities, and dust transport from nearby dry regions, further contribute to the deterioration of air quality conditions across Tehran [4].

Because PM_{2.5} concentration in Tehran is influenced by highly nonlinear interactions among meteorological variables, topographical characteristics, urban emissions, and atmospheric processes, accurate prediction remains a challenging research problem. Therefore, integrating satellite-derived AOD observations, ground-based monitoring data, and advanced DL models provides a promising solution for improving air pollution forecasting accuracy. The severe pollution conditions, diverse environmental characteristics, and complex spatio-temporal pollutant behavior observed in Tehran make this region an appropriate case study for evaluating the effectiveness and robustness of the proposed ODFT-DBN framework [4].

2.2. Dataset and Data Preparation

Accurate prediction of PM_{2.5} concentration requires the integration of multiple environmental, atmospheric, meteorological, and spatial variables due to the highly dynamic nature of air pollution formation and dispersion. PM_{2.5} levels are not governed by a single factor; instead, they are influenced by complex interactions among weather conditions, aerosol distributions, land surface characteristics, urban activities, and geographical properties. Therefore, this study utilizes a comprehensive multi-source dataset composed of ground-based observations, satellite-derived remote sensing products, meteorological

measurements, and spatial-temporal descriptors to provide a robust representation of the environmental conditions affecting PM_{2.5} concentration across Tehran.

The selection of variables was primarily motivated by their demonstrated effectiveness in previous air pollution studies and their direct physical relationship with particulate matter behavior. Meteorological variables, including maximum temperature, minimum temperature, atmospheric pressure, humidity, wind speed, and wind direction, strongly influence atmospheric stability, pollutant transport, chemical reactions, and particle accumulation processes. Temperature affects atmospheric mixing conditions and pollutant formation rates, while wind characteristics determine the movement and dispersion of suspended particles throughout urban regions. Humidity and water vapor contribute to aerosol growth and condensation mechanisms, whereas atmospheric pressure is associated with stable atmospheric layers and temperature inversion events that frequently intensify pollution episodes in Tehran.

In addition to meteorological information, several satellite-derived variables were incorporated to capture large-scale atmospheric and land-surface characteristics that cannot be adequately represented using ground stations alone. AOD was selected as one of the most important variables because it provides valuable information regarding aerosol loading and atmospheric particle concentration. Since ground monitoring stations are spatially sparse and unevenly distributed, AOD measurements extracted from satellite imagery provide broader spatial coverage and improve the ability of the model to identify regional pollution patterns. Similarly, LST was included to characterize surface thermal conditions that influence atmospheric turbulence and pollutant transport mechanisms. Total precipitation and water vapor were additionally considered because precipitation removes particulate matter from the atmosphere, while atmospheric moisture affects aerosol formation and chemical interactions.

Spatial and geographical variables were also integrated into the dataset to represent terrain characteristics and urban environmental conditions. Elevation was included because topographical structures significantly influence airflow circulation, pollutant accumulation, and inversion formation. Tehran's mountainous northern regions and lower southern plains create heterogeneous atmospheric conditions that directly impact PM_{2.5} distribution. Furthermore, the NDVI was incorporated to quantify vegetation density and green-space distribution across the study area. Vegetation coverage plays an important role in pollutant absorption, airflow modification, and local environmental regulation. Temporal descriptors, including the day of the week, were also considered to capture variations associated with traffic intensity, industrial activities, and urban mobility patterns during different days.

To improve the spatial representativeness of the prediction framework, both ground-based and satellite-derived datasets were jointly utilized in this study. Ground monitoring stations provide highly accurate local measurements of PM_{2.5} concentration and meteorological conditions; however, their spatial coverage is limited. In contrast, satellite observations offer extensive regional coverage and continuous spatial information but may lack localized precision. Therefore, combining these complementary data sources enables the proposed framework to simultaneously exploit the accuracy of ground observations and the large-scale spatial capability of remote sensing products. This hybrid data integration strategy enhances the ability of the proposed model to capture both local pollution dynamics and broader atmospheric patterns.

Ground-based PM_{2.5} concentration measurements were obtained from the Tehran Air Quality Control Company. The dataset consisted of daily observations collected from 19 air-quality monitoring stations across Tehran during the period from 2014 to 2016. The selection of these stations was based on data availability and temporal consistency throughout the study period. Several monitoring stations were excluded from the analysis

due to insufficient records. Specifically, three stations became operational after 2016 and therefore did not provide observations within the target time window, while another station contained only a single PM_{2.5} measurement during the entire study period. In addition, some stations lacked continuous daily observations, which limited their suitability for long-term spatio-temporal modeling. Consequently, only stations with sufficiently complete and reliable records were retained to ensure the consistency and quality of the final dataset.

The selected monitoring network provides representative coverage of different urban regions of Tehran and supports the investigation of spatial and temporal variations in PM_{2.5} concentrations. The availability and distribution of monitoring stations are important factors in air-quality studies because they directly influence the ability to characterize local pollution patterns and capture the effects of different emission sources. Nevertheless, practical limitations associated with station deployment, operation, and maintenance often result in incomplete spatial coverage. Therefore, the integration of station measurements with satellite-derived environmental variables offers an effective approach for improving the comprehensiveness and reliability of air-pollution assessment and prediction.

Meteorological information, including maximum temperature, minimum temperature, wind speed, wind direction, atmospheric pressure, and relative humidity, was obtained from five meteorological stations operating within the study area. These observations provide valuable information regarding the influence of atmospheric conditions on PM_{2.5} dynamics. In addition, several remote-sensing variables were extracted through google earth engine (GEE), including AOD, elevation, NDVI, LST, total precipitation, and water vapor. All satellite-derived variables were processed using dedicated GEE scripts and subsequently integrated with the ground-based observations to establish a comprehensive spatio-temporal dataset for PM_{2.5} prediction (see Table 1).

Table 1. Summary of the spatial-temporal variables utilized in the proposed prediction framework.

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

The reliability of PM_{2.5} prediction models strongly depends on the quality, consistency, and completeness of the input dataset. Since the proposed framework integrates heterogeneous data sources obtained from ground monitoring stations and satellite observations, several preprocessing and refinement operations were required prior to model

implementation. Environmental and atmospheric datasets frequently contain noise, missing observations, inconsistent measurements, and spatial discontinuities caused by sensor limitations, atmospheric interference, and irregular acquisition intervals. Therefore, a comprehensive data preparation pipeline was designed in this study to improve the robustness and learning capability of the proposed ODFT-DBN framework. As illustrated in Figure 1, the preprocessing stage consists of four primary operations, including noise reduction, missing data reconstruction, normalization, and spatial interpolation.

Initially, a noise reduction procedure was applied to the collected time-series datasets in order to eliminate abrupt fluctuations and measurement inconsistencies while preserving the underlying temporal patterns of the environmental variables. In this study, the Savitzky–Golay filter was utilized due to its effectiveness in smoothing nonlinear signals without destroying important structural characteristics such as peaks, trends, and local variations. Unlike conventional moving-average filters, the Savitzky–Golay approach performs local polynomial fitting within sliding windows, allowing the refined signal to maintain its original temporal behavior. After several experimental evaluations, a window size of 7 and a second-order polynomial were selected to provide a suitable balance between smoothing performance and feature preservation. Figure 3 illustrates the refined $\text{PM}_{2.5}$ time-series data after the noise reduction stage, where irregular fluctuations are effectively suppressed while the general temporal behavior and pollution peaks remain preserved. The selected filtering configuration was intentionally chosen to suppress measurement noise while preserving major $\text{PM}_{2.5}$ pollution episodes and extreme atmospheric variations, thereby reducing the risk of over-smoothing.

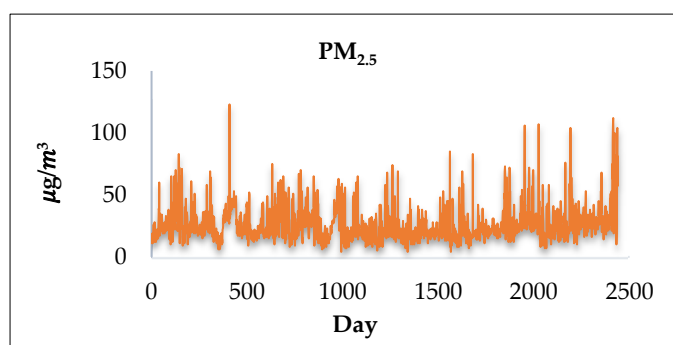


Figure 3. Refined temporal variation in the $\text{PM}_{2.5}$ concentration data after the preprocessing.

Following the denoising stage, missing observations within the environmental and satellite datasets were reconstructed to ensure temporal continuity and improve data completeness. In practical environmental monitoring systems, missing values frequently occur due to sensor malfunction, communication interruptions, atmospheric contamination, or inconsistent acquisition schedules. To address this issue, a Fourier-series-based reconstruction approach was employed in this study. The Fourier series method approximates temporal signals using combinations of sinusoidal functions with different frequencies, enabling effective reconstruction of periodic and nonlinear environmental patterns. Compared with conventional interpolation techniques, Fourier-based reconstruction is particularly suitable for atmospheric and meteorological time-series data because many environmental variables exhibit seasonal and cyclic behavior. By utilizing the temporal frequency characteristics of the datasets, the missing values were estimated while preserving the natural variability and periodic trends of the original observations.

In addition to temporal refinement, spatial alignment of the environmental variables was also required because meteorological stations and air pollution monitoring stations are not spatially colocated throughout the study region. Therefore, spatial interpolation was performed to estimate meteorological conditions at pollution monitoring locations.

In this work, ordinary Kriging interpolation was adopted due to its strong capability in modeling spatial autocorrelation and generating smooth spatial estimations. Kriging utilizes the statistical relationships among neighboring observations through semivariogram analysis to estimate unknown spatial values with minimized estimation error. Compared with deterministic interpolation methods, Kriging provides improved spatial estimation accuracy in heterogeneous urban environments where atmospheric conditions vary significantly across different regions of Tehran. This interpolation stage enabled the creation of a spatially consistent environmental dataset suitable for spatio-temporal learning.

After completing the reconstruction and interpolation stages, all variables were normalized before being introduced into the proposed DL framework. Since the dataset contains variables with different units and numerical scales, direct utilization of raw values may lead to unstable convergence and biased learning behavior during training. To overcome this issue, Min–Max normalization was applied to scale all variables into a unified numerical range between 0 and 1. The normalized value of each variable was computed as Equation (1):

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x_{min} and x_{max} denote the minimum and maximum load values within each household profile.

This normalization strategy improves numerical stability, accelerates convergence, and prevents variables with large magnitudes from dominating the optimization process. Figure 4 presents the effect of the refinement process on the AOD time-series data. Figure 4a shows the original raw AOD observations, which contain noise, abrupt variations, and incomplete measurements caused by atmospheric interference and satellite acquisition limitations. In contrast, Figure 4b illustrates the refined AOD dataset after applying the preprocessing pipeline, including noise reduction and missing data reconstruction. The refined signal exhibits smoother temporal behavior and improved continuity while preserving the primary atmospheric trends and aerosol variation patterns. These preprocessing operations collectively improve the consistency and reliability of the integrated environmental dataset, enabling the proposed ODFT-DBN framework to achieve more accurate and stable PM_{2.5} prediction performance.

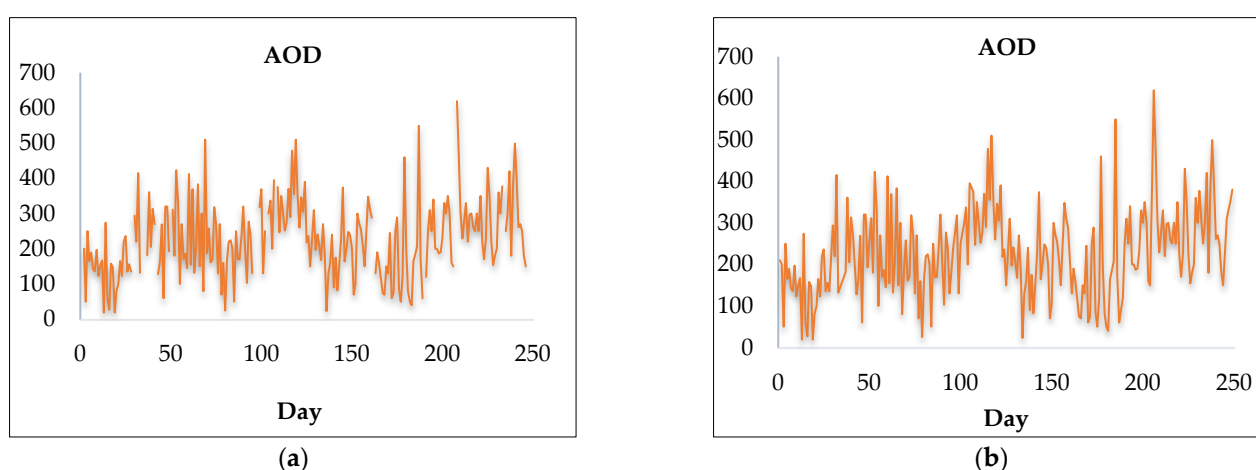


Figure 4. Time series representation of AOD dataset: (a) Raw observations; (b) Refined observations.

Since the PM_{2.5} observations and the majority of the environmental variables were organized at a daily temporal resolution, the hourly AOD observations were aggregated to daily values before the data integration stage. Specifically, for each day, the daily AOD value was calculated as the arithmetic mean of all available hourly AOD measurements acquired during that day. This aggregation procedure ensured temporal consistency

among the different environmental variables and facilitated the construction of a unified daily spatio-temporal dataset for model development and evaluation.

To ensure reliable model evaluation and realistic forecasting assessment, the dataset was divided into training and testing subsets using both temporal and spatial separation strategies. Approximately 70% of the available samples were allocated for training, while the remaining 30% were reserved for testing and performance validation. From the temporal perspective, environmental observations collected during 2014 and 2015 were utilized for model training, whereas the data corresponding to 2016 were exclusively employed for testing. This chronological separation allows the proposed framework to be evaluated on future unseen observations, closely simulating practical real-world PM_{2.5} forecasting scenarios. Such a strategy additionally minimizes the possibility of temporal data leakage and ensures that the predictive capability of the model is assessed under realistic atmospheric evolution conditions.

In addition to temporal separation, spatial independence between the training and testing datasets was also considered to improve the robustness and generalization capability of the proposed framework. Specifically, several monitoring stations were completely excluded from the training process and reserved only for testing. Consequently, the model was evaluated using spatial locations that were never observed during training. This strategy provides a more challenging but realistic validation setup because environmental forecasting systems in practical applications frequently encounter new urban regions and previously unseen monitoring locations. Although spatially independent validation may slightly reduce testing accuracy compared with random data splitting approaches, it provides a more reliable assessment of the real-world generalization capability of the prediction framework.

Furthermore, the distributions of major environmental variables, including PM_{2.5} concentration levels and meteorological parameters, were analyzed across both training and testing subsets to ensure balanced representation and prevent distribution bias. This balanced validation framework improves the reliability of the obtained experimental results and provides a more comprehensive evaluation of the capability of the proposed ODFT-DBN architecture to model highly dynamic and heterogeneous spatio-temporal atmospheric conditions.

2.3. Transformer Encoder

Transformer networks have recently emerged as one of the most powerful DL architectures for modeling sequential and spatio-temporal data due to their strong capability in capturing long-range dependencies and contextual relationships. Unlike conventional recurrent architectures, such as recurrent neural networks (RNNs), LSTM, and GRU, transformer models process the entire input sequence simultaneously through an attention mechanism rather than sequential recursion. This parallel learning structure significantly improves computational efficiency and enables the model to better capture complex nonlinear interactions among environmental variables distributed across different temporal intervals. Because PM_{2.5} concentration is highly dependent on historical atmospheric conditions and long-term environmental patterns, transformer architectures are particularly suitable for air pollution forecasting applications. The core component of the transformer encoder is the self-attention mechanism, which allows the model to dynamically identify the relative importance of different temporal observations within the input sequence. In PM_{2.5} prediction problems, environmental variables observed at earlier time intervals may still significantly influence future pollution concentrations due to delayed atmospheric reactions, pollutant transport, and meteorological interactions. Therefore, the self-attention mechanism enables the proposed framework to effectively learn both short-

term and long-term dependencies within the environmental dataset. In addition, transformer encoders exhibit strong feature extraction capability in high-dimensional heterogeneous datasets, making them highly effective for integrating meteorological, satellite-derived, and spatial variables simultaneously. Figure 5 illustrates the overall architecture of the basic transformer encoder utilized in the proposed framework.

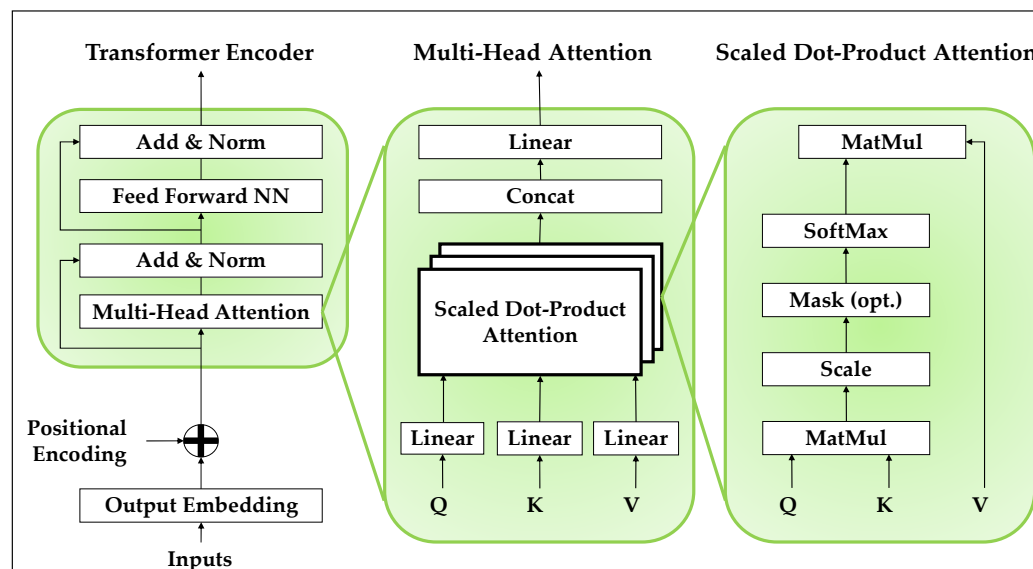


Figure 5. Structural representation of the transformer encoder architecture.

As shown in the figure, the input environmental features are first transformed into embedding representations and combined with positional encoding information. Since transformer networks do not inherently preserve temporal ordering, positional encoding is incorporated to provide sequential information regarding the relative positions of observations within the input time series. The embedded representations are subsequently forwarded into the multi-head attention module, which represents the primary learning component of the transformer encoder.

Within the multi-head attention mechanism, the input sequence is projected into three different representations, namely query (Q), key (K), and value (V) matrices through linear transformation layers. These matrices are utilized to compute the scaled dot-product attention, which determines the relationship and dependency strength among different temporal observations. The attention operation enables the model to focus on the most informative environmental features and temporal intervals relevant to PM_{2.5} prediction. Multiple attention heads are simultaneously employed to learn different contextual relationships from various representation subspaces. The outputs generated by all attention heads are concatenated and passed through an additional linear transformation layer to produce the final attention representation. After the attention stage, the extracted contextual features are processed through a FFNN layer to further enhance nonlinear feature learning capability. Residual connections and normalization layers are additionally incorporated after both the attention and FFNN blocks to stabilize the training process, improve convergence behavior, and mitigate gradient degradation problems in deep architectures.

Equations (2) and (3) define the positional encoding mechanism used in the transformer encoder. Since the transformer architecture processes the input sequence in parallel, it does not inherently preserve temporal ordering information. Therefore, sinusoidal positional encoding is incorporated to provide sequential information regarding the rela-

tive position of each observation within the input time series. By combining sine and cosine functions with different frequencies, the model can effectively capture both short-range and long-range temporal dependencies among environmental observations.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{1000^{2i/d}}\right), \quad (2)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{1000^{2i/d}}\right) \quad (3)$$

where pos is the position index; i is the dimension index; d is the embedding size.

Equations (4)–(6) represent the linear transformation process used to generate the query, key, and value matrices from the input feature representation. These three matrices form the core components of the self-attention mechanism. The query matrix determines which temporal information should be emphasized, the key matrix represents the contextual characteristics of the input sequence, and the value matrix contains the corresponding feature information utilized during the attention computation process.

$$\vec{Q} = \vec{Z}W^Q, \quad (4)$$

$$\vec{K} = \vec{Z}W^K, \quad (5)$$

$$\vec{V} = \vec{Z}W^V \quad (6)$$

where W^Q , W^K , and W^V are learnable weight matrices responsible for transforming the input into distinct subspaces.

Equation (7) describes the scaled dot-product attention mechanism employed within the transformer encoder. In this operation, the similarity between the query and key matrices is first calculated to determine the dependency strength among different temporal observations. The resulting attention weights are then normalized using the softmax function and multiplied by the value matrix to generate contextual feature representations. The scaling factor is additionally incorporated to stabilize the gradient behavior and prevent excessively large attention values during training.

$$Attention(\vec{Q}, \vec{K}, \vec{V}) = softmax\left(\frac{\vec{Q}\vec{K}^T}{\sqrt{d_K}}\right)\vec{V} \quad (7)$$

where d_K is the dimensionality of each attention head.

Equations (8) and (9) formulate the multi-head attention mechanism. Instead of learning a single attention representation, multiple attention heads are simultaneously utilized to extract different contextual relationships from various feature subspaces. Each attention head independently performs the self-attention operation using separate learnable projection matrices. The outputs of all attention heads are subsequently concatenated and linearly transformed to generate the final multi-head attention representation. This structure significantly improves the capability of the transformer encoder to capture complex spatio-temporal dependencies within environmental datasets.

$$MultiHead(\vec{Q}, \vec{K}, \vec{V}) = Concat(head_1, head_2, \dots, head_h)W^O, \quad (8)$$

$$head_1 = Attention(\vec{Q}W_i^Q, \vec{K}W_i^K, \vec{V}W_i^V) \quad (9)$$

Equation (10) defines the FFNN layer used after the attention mechanism. The FFNN performs nonlinear feature transformation through fully connected layers and activation functions, enabling the model to further enhance high-level feature representations extracted from the attention stage. This operation improves the discriminative capability of the learned environmental features associated with PM_{2.5} prediction.

$$FFNN(\vec{x}) = \text{ReLU}(0, \vec{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (10)$$

where $\vec{x}$ is the input vector corresponding to a single token or position in the sequence; $\mathbf{W}$ is the weight matrix; $\mathbf{b}$ is bias vector.

Finally, Equations (11) and (12) represent the residual connection and layer normalization operations integrated into the transformer encoder architecture. Residual connections help preserve the original feature information and mitigate gradient degradation during deep network training, while layer normalization improves numerical stability and accelerates convergence. Together, these operations enable the transformer encoder to learn robust and stable spatio-temporal representations from complex atmospheric and environmental datasets.

$$\hat{\vec{Z}} = \text{LayerNorm}(\vec{Z} + \text{MultiHead}(\vec{Q}, \vec{K}, \vec{V})), \quad (11)$$

$$\vec{Z}^{out} = \text{LayerNorm}(\hat{\vec{Z}} + FFNN(\hat{\vec{Z}})) \quad (12)$$

2.4. DBN

DBNs are probabilistic DL architectures designed to learn hierarchical and high-level feature representations from complex datasets. A DBN is generally constructed by stacking multiple restricted Boltzmann machines (RBMs), where the output of each RBM serves as the input of the subsequent layer. Unlike conventional shallow neural networks, DBNs can effectively extract hidden nonlinear patterns and representative features through layer-wise unsupervised learning. This hierarchical learning capability makes DBNs highly suitable for environmental and atmospheric prediction problems characterized by high-dimensional, nonlinear, and heterogeneous data distributions.

In PM_{2.5} forecasting applications, pollutant concentration is influenced by numerous interacting environmental variables, including meteorological conditions, aerosol dynamics, spatial characteristics, and temporal dependencies. These variables often exhibit highly nonlinear relationships that are difficult to capture using traditional machine learning approaches. Therefore, DBNs provide an efficient solution by progressively learning abstract feature representations from lower-level environmental observations. The layered structure of DBNs enables the framework to identify hidden pollutant patterns, complex feature interactions, and latent environmental structures associated with PM_{2.5} concentration variations. Figure 6 illustrates the generic structure of the DBN utilized in the proposed framework.

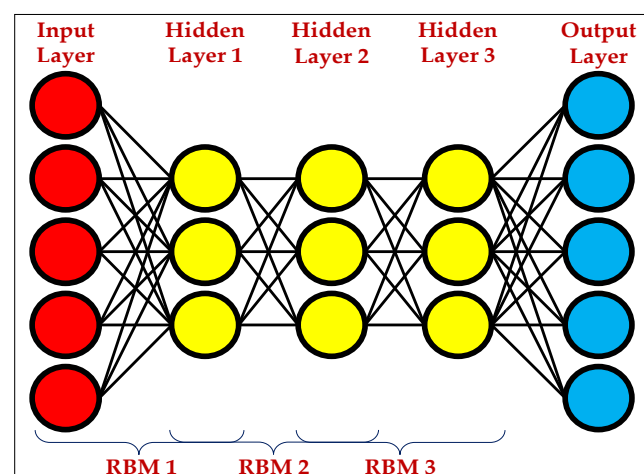


Figure 6. Overall structural representation of the DBN architecture.

As shown in the figure, the DBN consists of an input layer, multiple hidden layers formed by stacked RBMs, and an output layer responsible for final prediction generation. Initially, the environmental features extracted by the transformer encoder are forwarded into the first RBM layer, where low-level feature representations are learned. The learned representations are then propagated through deeper RBM layers to gradually capture more complex and abstract environmental relationships. Finally, the extracted high-level features are delivered to the output layer to generate the final PM_{2.5} prediction results. Equation (13) defines the energy function of the RBM layer within the DBN architecture. This energy-based formulation measures the relationship between the visible units and hidden units of the network. Lower energy values indicate stronger compatibility between the input environmental features and the learned hidden representations. Through this mechanism, the RBM can effectively capture nonlinear dependencies and hidden structures within the PM_{2.5} dataset.

$$E(v, h) = -v^T W h - b^T v - c^T h \quad (13)$$

Here, v denotes the visible vector; h the hidden vector; W the weight matrix connecting visible and hidden units; b and c represent the bias vectors for visible and hidden layers, respectively.

Equations (14) and (15) represent the joint probability distribution and partition function of the RBM model, respectively. Equation (14) computes the probability associated with a specific configuration of visible and hidden units based on the energy function, while Equation (15) defines the normalization term required to ensure a valid probabilistic distribution. These probabilistic formulations allow the DBN to model complex environmental feature distributions and improve representation learning capability.

$$P(v, h) = \frac{1}{Z} \exp(-E(v, h)) \quad (14)$$

$$Z = \sum_{v, h} \exp(-E(v, h)) \quad (15)$$

In this expression, Z is the partition function, which ensures that probabilities are properly normalized. From the joint distribution, the marginal probability of a visible vector can be derived.

Equation (16) formulates the marginal probability of the visible vector by summing over all hidden unit configurations. This operation enables the DBN to estimate the probability distribution of the observed environmental data while incorporating latent hidden representations extracted from the RBM structure. Consequently, the network can better model the statistical characteristics of PM_{2.5}-related variables.

$$P(v) = \frac{1}{Z} \sum_h \exp(-E(v, h)) \quad (16)$$

Equations (17) and (18) describe the conditional activation probabilities of the hidden and visible units, respectively. Equation (17) computes the probability of activating hidden neurons based on the observed visible input features, while Equation (18) reconstructs the visible units from the hidden representations. These bidirectional probabilistic interactions enable the RBM to iteratively learn representative latent environmental features and preserve important nonlinear relationships among atmospheric variables.

$$P(h_j = 1 | v) = \sigma \left(\sum_i v_i W_{ij} + c_j \right) \quad (17)$$

$$P(v_i = 1 | h) = \sigma \left(\sum_j h_j W_{ij} + b_i \right) \quad (18)$$

where $\sigma(\cdot)$ is the logistic sigmoid function.

Finally, Equation (19) represents the hierarchical probabilistic formulation of the DBN architecture. In this structure, multiple hidden layers are stacked sequentially, where each layer learns increasingly abstract feature representations from the output of the previous layer. As the information propagates through deeper layers, the DBN gradually captures high-level spatio-temporal environmental patterns associated with PM_{2.5} concentration dynamics, thereby improving the overall prediction capability of the proposed framework.

$$P(v, h^1, h^2, \dots, h^L) = P(v | h^1) \prod_{l=1}^{L-1} P(h^l | h^{l+1}) P(h^L) \quad (19)$$

In this formulation, $h^1, h^2, \dots, h^L$ represent the hidden layers of the DBN, with higher layers modeling increasingly abstract features.

2.5. NMOGWO

The GWO is a population-based meta-heuristic optimization algorithm inspired by the social hierarchy and cooperative hunting behavior of gray wolves in nature. Due to its strong exploration capability, simple mathematical structure, and efficient convergence behavior, GWO has become one of the widely adopted optimization techniques in engineering, ML, and DL applications. Unlike traditional gradient-based optimization methods, GWO does not require differentiable objective functions and can effectively solve highly nonlinear, nonconvex, and multi-dimensional optimization problems. These characteristics make GWO particularly suitable for optimizing the hyper-parameters of complex DL architectures and fuzzy systems used in PM_{2.5} prediction problems [40,41].

In environmental forecasting applications, optimization plays a critical role because the predictive performance of DL models is highly dependent on parameter selection, network configuration, and fuzzy membership tuning. Improper parameter initialization may lead to slow convergence, local optimum stagnation, and unstable prediction performance. Therefore, evolutionary optimization methods such as GWO provide an efficient mechanism for adaptively searching the solution space and identifying near-optimal parameter configurations. Compared with several conventional optimization algorithms, GWO offers a balanced tradeoff between global exploration and local exploitation while maintaining relatively low computational complexity.

The optimization process of GWO is inspired by the hierarchical leadership structure observed in gray wolf packs. In this hierarchy, wolves are divided into four main categories, namely alpha, beta, delta, and omega. The alpha wolf represents the best candidate solution obtained during the optimization process and is considered the primary leader responsible for guiding the search behavior of the population. The beta and delta wolves correspond to the second and third best candidate solutions, respectively, and assist the alpha wolf in directing the hunting process. The remaining wolves are classified as omega wolves, which update their positions according to the guidance provided by the dominant wolves [40].

The hunting mechanism of GWO mainly consists of three stages: searching, encircling, and attacking the prey. During the searching stage, gray wolves explore different regions of the solution space to identify promising candidate solutions. Once potential prey locations are identified, the wolves begin to encircle the prey by updating their positions relative to the dominant wolves. Finally, during the attacking stage, the wolves gradually converge toward the prey location to refine the solution and improve optimization accuracy. This adaptive cooperative behavior enables GWO to effectively balance exploration and exploitation throughout the optimization process, thereby reducing the probability of premature convergence and improving solution quality.

Equations (20) and (21) describe the encircling behavior of gray wolves during the optimization process. In the standard GWO algorithm, wolves gradually update their positions relative to the estimated prey location, which corresponds to the current optimal solution in the search space. Equation (20) computes the distance between the prey position and the current wolf position, while Equation (21) updates the position of the wolf according to the estimated prey location and adaptive coefficient vectors. Through this mechanism, candidate solutions iteratively move toward promising regions of the search space. Equations (22) and (23) define the adaptive coefficient vectors utilized during the position updating process. These coefficients control the exploration and exploitation behavior of the algorithm throughout the optimization iterations. The coefficient vector gradually decreases during the optimization process, enabling the wolves to initially explore the search space broadly and later focus on local exploitation around the best candidate solutions. Meanwhile, the coefficient vector introduces stochastic variations into the search behavior, improving population diversity and reducing the probability of premature convergence [41].

$$\vec{D} = \vec{C}\vec{X}_p(t) - \vec{X}(t), \quad (20)$$

$$\vec{X}(t+1) = \vec{C}\vec{X}_p(t) - \vec{A} \cdot \vec{D}, \quad (21)$$

$$\vec{A} = 2\vec{a}\vec{r}_1 - \vec{a}, \quad (22)$$

$$\vec{C} = 2\vec{r}_2 \quad (23)$$

where $\vec{X}$ is the position vector of a gray wolf, $\vec{X}_p$ is hunting position vector, $\vec{A}$ and $\vec{C}$ are coefficient vectors, $\vec{r}_1$ and $\vec{r}_2$ are random vectors in the interval $[0, 1]$; the $\vec{a}$ vector is linearly reduced from 2 to 0 during the repetition.

Equations (24)–(26) formulate the cooperative hunting mechanism of the GWO algorithm. In this stage, the three dominant wolves, namely alpha, beta, and delta, guide the movement of the remaining wolves in the population. Equation (24) computes the distances between the current wolf and each of the three leading wolves. Subsequently, Equation (25) estimates three candidate positions corresponding to the guidance provided by the alpha, beta, and delta wolves. Finally, Equation (26) updates the position of the current wolf by averaging these candidate positions. This cooperative updating strategy improves the balance between global exploration and local exploitation while enhancing optimization stability [29].

$$\vec{D}_\alpha = |\vec{C}_1\vec{X}_\alpha - \vec{X}|, \quad \vec{D}_\beta = |\vec{C}_2\vec{X}_\beta - \vec{X}|, \quad \vec{D}_\delta = |\vec{C}_3\vec{X}_\delta - \vec{X}|, \quad (24)$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot (\vec{D}_\alpha) \quad \vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot (\vec{D}_\beta), \quad \vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot (\vec{D}_\delta), \quad (25)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (26)$$

where $\vec{X}_\alpha, \vec{X}_\beta, \vec{X}_\delta$ are the positions of the three leading wolves, $\vec{D}_\alpha, \vec{D}_\beta, \vec{D}_\delta$ are respective distances to each leader, $\vec{A}_1, \vec{A}_2, \vec{A}_3, \vec{C}_1, \vec{C}_2, \vec{C}_3$ updated coefficient vectors, and $\vec{X}_1, \vec{X}_2, \vec{X}_3$ are intermediate candidate positions.

Figure 7 illustrates the position updating process in the standard GWO. As shown in the figure, the current wolf updates its movement according to the relative positions of the alpha, beta, and delta wolves surrounding the prey region. The adaptive coefficient vectors dynamically modify the movement direction and search radius during the optimization process. In the early iterations, the wolves perform broader exploration across the solution space, while in later iterations the search gradually concentrates around the

optimal region. This adaptive hunting behavior enables GWO to efficiently search complex nonlinear optimization spaces and improve convergence performance [29].

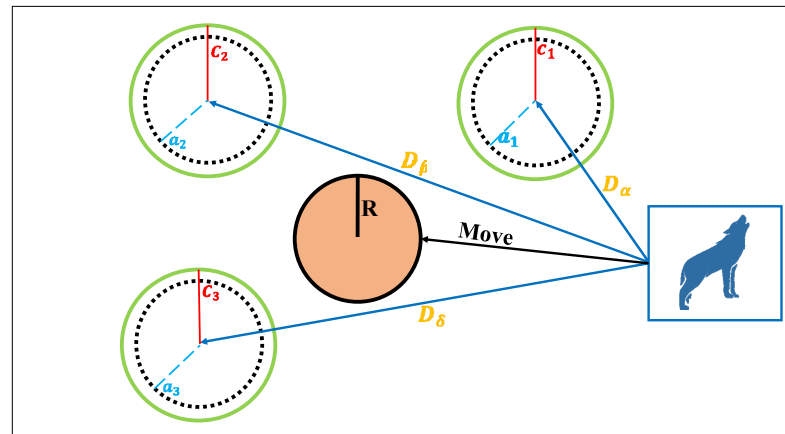


Figure 7. Illustration of the position updating mechanism in the conventional GWO algorithm.

Although the standard GWO algorithm demonstrates strong optimization capability in many engineering applications, it still suffers from several limitations when dealing with highly nonlinear and high-dimensional optimization problems. One of the primary weaknesses of the conventional GWO is the imbalance between exploration and exploitation during the optimization process. In the early iterations, the wolves may excessively explore the search space without sufficiently refining promising candidate solutions, while in later stages the population may converge prematurely around suboptimal regions. This issue can reduce optimization accuracy and limit the capability of the algorithm to identify the global optimum. Another limitation of the standard GWO is its heavy dependence on only three dominant wolves, namely alpha, beta, and delta, for guiding the entire population. Although this leadership mechanism improves convergence speed, relying solely on three leaders may reduce population diversity and increase the probability of local optimum stagnation, particularly in complex optimization landscapes with multiple local minima. Furthermore, the averaging strategy utilized in the standard position updating process may weaken exploitation capability near the optimal region and slow down the final convergence behavior of the algorithm.

In the proposed PM_{2.5} prediction framework, the optimization problem involves simultaneous tuning of transformer hyper-parameters, DBN parameters, and fuzzy membership functions within a highly nonlinear spatio-temporal learning environment. Such optimization tasks contain complex multidimensional search spaces with strong parameter interdependencies. Therefore, utilizing a more robust optimization strategy is essential to improve convergence stability, enhance exploitation capability, and prevent premature convergence during training. To address these limitations, an enhanced optimization mechanism is introduced in this study by incorporating an additional dominant wolf into the conventional GWO hierarchy.

Specifically, a new leadership wolf referred to as theta is integrated into the optimization framework to strengthen the guidance mechanism during the position updating process. The proposed theta wolf acts as an additional high-quality candidate solution positioned immediately after the alpha wolf in terms of fitness ranking. By introducing this supplementary leader, the optimizer can utilize richer directional information during population updating and improve the exploitation capability around promising search regions. The inclusion of the theta wolf additionally increases population diversity and reduces the risk of convergence toward locally optimal solutions.

Equations (27)–(29) describe the proposed theta-based position updating mechanism integrated into the enhanced GWO framework. Equation (27) computes the distance between the current wolf and the theta wolf position, allowing the optimizer to estimate an additional guidance direction during the hunting process. Subsequently, Equation (28) determines a new candidate position generated according to the theta wolf guidance strategy. Finally, Equation (29) updates the position of the current wolf using the averaged contributions of four dominant wolves, namely alpha, beta, delta, and theta, instead of the conventional three-leader formulation.

$$\vec{D}_\theta = |\vec{C}_4 \vec{X}_\theta - \vec{X}|, \quad (27)$$

$$\vec{X}_4 = \vec{X}_\theta - \vec{A}_4 \cdot (\vec{D}_\theta), \quad (28)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3 + \vec{X}_4}{4} \quad (29)$$

where $\vec{X}_\theta$ is the positions of the three leading wolves, $\vec{D}_\theta$ is the respective distances to each leader, $\vec{A}_4, \vec{C}_4$ updated coefficient vectors, and $\vec{X}_4$ is intermediate candidate positions.

The incorporation of the theta wolf significantly improves the exploitation behavior of the optimizer by introducing additional directional refinement around high-quality regions of the search space. In the standard GWO algorithm, the search process may become unstable when the three dominant wolves are located far from the global optimum or concentrated around local optima. However, the proposed theta-guided mechanism provides an additional adaptive search direction that stabilizes convergence and improves local search capability. As a result, the enhanced optimizer can achieve better convergence accuracy, stronger optimization robustness, and improved search efficiency for complex PM_{2.5} prediction problems involving high-dimensional DL parameter optimization.

In practical DL optimization problems, relying on a single objective function is often insufficient because multiple performance criteria must be optimized simultaneously. In the proposed PM_{2.5} prediction framework, improving prediction accuracy alone may lead to increased model complexity, unstable convergence, or excessive computational cost. Therefore, a multi-objective optimization strategy is adopted in this study to simultaneously balance several optimization goals, including prediction accuracy, convergence stability, and generalization capability. To achieve this objective, the proposed optimizer is extended into a novel multi-objective gray wolf optimizer referred to as NMOGWO.

The proposed NMOGWO integrates the hunting behavior of GWO with non-dominated sorting and crowding-distance mechanisms inspired by multi-objective evolutionary optimization. Instead of searching for a single optimal solution, the optimizer attempts to identify a set of Pareto-optimal solutions representing different tradeoffs among conflicting objectives. During the optimization process, candidate solutions are ranked according to Pareto dominance relationships. Solutions that are not dominated by any other candidates are assigned to the first Pareto front, while the remaining solutions are categorized into subsequent fronts according to their dominance level. This non-dominated sorting strategy enables the optimizer to preserve high-quality candidate solutions while maintaining optimization diversity across the search space.

In addition to Pareto ranking, crowding-distance analysis is employed to preserve population diversity and avoid excessive concentration of solutions around limited regions of the Pareto front. Candidate solutions with larger crowding distances are preferred because they contribute to broader search-space coverage and improved exploration capability. Consequently, the combination of non-dominated sorting and crowding-distance mechanisms improves both convergence quality and population diversity

throughout the optimization process. This multi-objective optimization strategy is particularly beneficial for the proposed PM_{2.5} forecasting framework, where multiple hyper-parameters and fuzzy membership functions must be simultaneously optimized within a highly nonlinear learning environment.

To further enhance the optimization capability of NMOGWO, an adaptive nonlinear exploration–exploitation control strategy is additionally incorporated into the proposed framework. In the standard GWO algorithm, the parameter controlling the balance between exploration and exploitation decreases linearly during the optimization iterations. However, linear reduction may not provide sufficient flexibility for complex optimization landscapes and may cause premature convergence or insufficient local refinement. Therefore, a nonlinear adaptive updating mechanism is introduced to dynamically regulate the search behavior according to the optimization state. Equation (30) defines the proposed nonlinear adaptive control parameter used to regulate the exploration and exploitation behavior of the optimizer during different optimization stages.

$$a(t) = 2 \left(1 - \left(\frac{t}{T} \right)^\eta \right) \quad (30)$$

where t denotes the current iteration number; T represents the maximum number of iterations; η is the adaptive nonlinear control coefficient.

Unlike the conventional linear updating strategy, the proposed formulation enables stronger exploration during early iterations and gradually increases exploitation capability near the final optimization stages. Equation (31) defines the adaptive updating mechanism of the nonlinear control coefficient.

$$\eta = 1 - \frac{f_{best}(t) - f_{avg}(t)}{f_{best}(t) + \varepsilon} \quad (31)$$

where $f_{best}(t)$ denotes the best fitness value at iteration t ; $f_{avg}(t)$ represents the average population fitness; ε is a small positive constant introduced to avoid numerical instability.

An additional advantage of the proposed NMOGWO emerges from the interaction between the theta-guided leadership mechanism and the adaptive exploration–exploitation controller. In conventional GWO, the population movement is primarily determined by the geometric average of three elite wolves, which may lead to directional bias when these leaders become clustered around a locally attractive region. In contrast, the proposed four-leader formulation introduces an additional search reference that modifies the geometric distribution of elite solutions in the search space. Consequently, the position-update process becomes less sensitive to misleading local fitness gradients and better reflects the overall structure of the optimization landscape. From an optimization perspective, the theta wolf acts as an intermediate guidance source that bridges global exploration and local exploitation. During the transition from exploration to exploitation, it preserves alternative search directions that might otherwise disappear due to rapid population convergence. This mechanism becomes particularly beneficial in the considered PM_{2.5} forecasting problem, where transformer hyper-parameters, DBN structures, and fuzzy-system parameters exhibit strong nonlinear dependencies. By maintaining richer directional diversity among elite solutions while simultaneously refining promising regions, the proposed NMOGWO achieves a more stable search trajectory and a lower probability of premature convergence than the conventional GWO framework.

This adaptive mechanism dynamically adjusts the exploration–exploitation balance according to the population distribution. When the population diversity is high, the optimizer performs broader exploration across the search space. Conversely, when the candidate solutions gradually converge, the optimizer strengthens local exploitation around promising regions. By integrating theta-guided leadership, non-dominated sorting,

crowding-distance preservation, and adaptive nonlinear exploration–exploitation control, the proposed NMOGWO achieves improved convergence stability, stronger optimization robustness, and enhanced global search capability compared with the conventional GWO algorithm. These improvements enable the optimizer to more effectively tune the transformer–DBN hyper-parameters and fuzzy membership functions within the proposed PM_{2.5} prediction framework.

2.6. Proposed ODFT-DBN

The overall architecture of the proposed ODFT-DBN framework is illustrated in Figure 8. The proposed model integrates fuzzy reasoning, transformer-based spatio-temporal learning, DBN-based hierarchical feature extraction, and multi-objective evolutionary optimization within a unified prediction architecture for PM_{2.5} forecasting. The primary objective of the framework is to accurately model the highly nonlinear relationships among atmospheric variables, aerosol distributions, meteorological conditions, and spatial environmental characteristics. By combining multiple intelligent learning components, the proposed framework can effectively capture both local and global environmental dependencies associated with PM_{2.5} concentration dynamics.

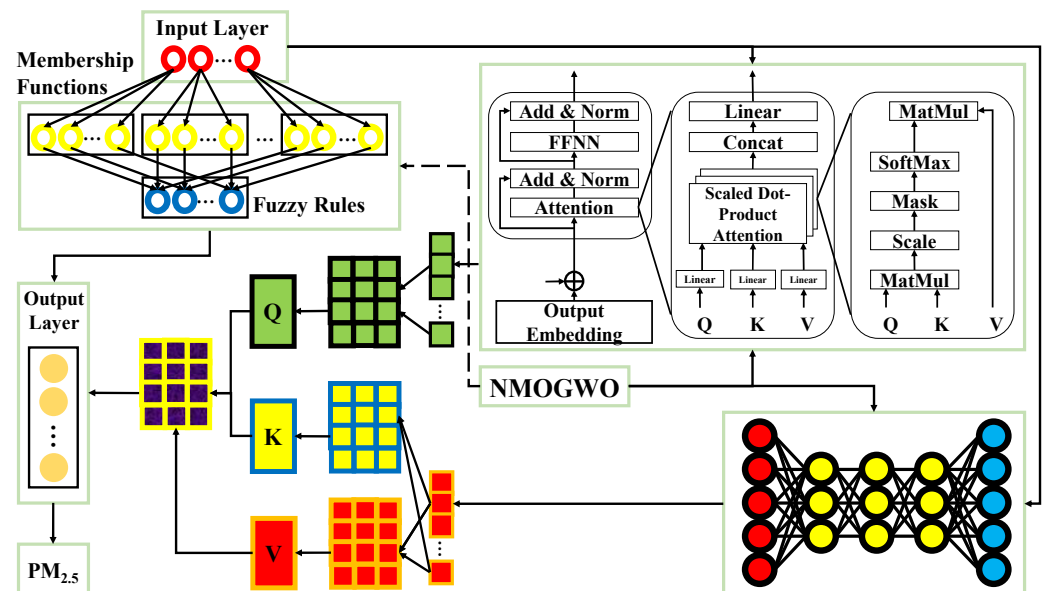


Figure 8. Overall architecture of the proposed ODFT-DBN framework for PM_{2.5} prediction.

As shown in Figure 8, the environmental input variables are initially processed through a fuzzy inference layer consisting of membership functions and fuzzy rules. The fuzzy module transforms the raw environmental observations into fuzzy representations capable of handling uncertainty and ambiguity within atmospheric data. The resulting fuzzy outputs are subsequently embedded and transferred into the transformer encoder, where the self-attention mechanism extracts long-range spatio-temporal dependencies from the environmental sequences. The transformer-generated contextual features are then forwarded into the DBN architecture for hierarchical feature abstraction and nonlinear representation learning. Finally, the optimized output layer generates the predicted PM_{2.5} concentration values.

The entire learning framework is jointly optimized using the proposed NMOGWO algorithm. Unlike conventional optimization approaches that focus on a single objective, the proposed optimizer simultaneously considers multiple optimization criteria associated with prediction accuracy, network complexity, hidden neuron configuration, fuzzy

rule adaptation, and output stability. Through this multi-objective optimization mechanism, the proposed framework can achieve improved convergence behavior, stronger generalization capability, and enhanced robustness for complex air pollution forecasting tasks. The interaction among the fuzzy system, transformer encoder, DBN structure, and NMOGWO enables the framework to adaptively learn representative atmospheric patterns from heterogeneous environmental datasets.

The fuzzy module represents the first intelligent processing stage of the proposed framework and plays a critical role in modeling uncertainty within environmental observations. Atmospheric variables associated with PM_{2.5} prediction often exhibit ambiguous and nonlinear relationships due to changing weather conditions, seasonal effects, and pollutant interactions. Conventional deterministic models may struggle to accurately represent these uncertainties. Therefore, fuzzy membership functions are employed to convert environmental variables into fuzzy linguistic representations capable of preserving gradual transitions and uncertain relationships among atmospheric features. In addition to uncertainty modeling, the fuzzy inference system improves the interpretability and adaptability of the prediction framework by incorporating fuzzy rules that describe environmental interactions in a flexible manner. The membership functions and fuzzy rule parameters are dynamically optimized through the proposed NMOGWO algorithm to improve the quality of the extracted fuzzy representations. Consequently, the fuzzy module enhances the capability of the proposed architecture to process noisy and highly variable atmospheric datasets associated with PM_{2.5} concentration prediction.

Gaussian membership functions were adopted for all input variables because of their smooth nonlinear representation capability and suitability for gradient-free optimization. The proposed framework uses ($M = 16$) input features, and each normalized input variable is independently partitioned into three linguistic fuzzy sets, namely low, medium, and high. The initial membership-function centers are uniformly distributed over the normalized interval $([0,1])$, with approximate center locations of 0.25, 0.50, and 0.75. The corresponding membership widths are initialized to provide moderate overlap between adjacent fuzzy regions, thereby ensuring smooth transitions among the linguistic states. Importantly, the proposed fuzzy module does not explicitly construct the complete Cartesian rule base formed by all possible combinations of the input fuzzy sets. Such a construction would generate (3^{16}) potential rules and would therefore be computationally impractical. Instead, a factorized feature-wise fuzzy representation is employed, in which the three membership degrees of each input variable are evaluated independently. This procedure produces ($3M = 48$) elementary fuzzy components rather than an exponentially large combinatorial rule set. Accordingly, no separate feature-screening or hard rule-pruning procedure is required, because the exhaustive rule base is never generated. Each elementary fuzzy component is assigned an adaptive weight, allowing the contributions of different input-variable-membership-set pairs to be adjusted according to their predictive relevance. During the optimization stage, the proposed NMOGWO jointly optimizes the membership-function centers, membership widths, and adaptive fuzzy-component weights together with the transformer and DBN hyper-parameters. Consequently, the initial fuzzy partitions serve only as starting configurations, whereas the final fuzzy representation is adaptively learned from the environmental data to improve uncertainty modeling and PM_{2.5} prediction performance.

After fuzzy processing, the extracted environmental representations are transferred into the transformer encoder for spatio-temporal feature learning. The transformer module utilizes self-attention mechanisms to identify the relative importance of different environmental observations distributed across the temporal sequence. Since PM_{2.5} concentration is highly dependent on both short-term atmospheric fluctuations and long-term environmental conditions, the transformer encoder enables the framework to effectively

capture temporal correlations and contextual dependencies among heterogeneous variables. Unlike recurrent architectures, the transformer can process the entire environmental sequence simultaneously, thereby improving computational efficiency and feature extraction capability. The multi-head attention mechanism further improves the learning process by extracting multiple contextual representations from different feature subspaces. This allows the transformer encoder to simultaneously focus on different environmental interactions, including aerosol transport, meteorological variability, and seasonal atmospheric behavior. Through this attention-driven representation learning mechanism, the transformer module generates highly informative contextual features suitable for subsequent deep hierarchical analysis within the DBN architecture.

The DBN module performs hierarchical feature abstraction using multiple stacked hidden layers constructed from RBM structures. The transformer-generated features are progressively transformed into increasingly abstract representations capable of modeling complex nonlinear pollutant relationships. This multilayer learning structure enables the DBN to identify latent environmental patterns and hidden dependencies that may not be directly observable within the original environmental variables. Furthermore, the DBN improves the robustness and generalization capability of the proposed framework by refining high-level environmental representations before final prediction generation. Through layer-wise probabilistic learning, the DBN can effectively preserve important atmospheric characteristics while suppressing redundant or irrelevant feature information. Consequently, the integration of transformer-based contextual learning and DBN-based hierarchical representation extraction significantly enhances the predictive capability of the proposed ODFT-DBN framework for PM_{2.5} forecasting applications.

The objective of incorporating a DBN after the transformer encoder is not to increase model depth, but to provide a complementary learning mechanism. The transformer encoder is primarily responsible for capturing long-range temporal dependencies and contextual relationships through the self-attention mechanism. However, its output representations are still contextual sequence features. The DBN is subsequently employed to perform hierarchical probabilistic feature abstraction and latent representation learning. In particular, the DBN progressively transforms transformer-generated contextual features into higher-level nonlinear representations that are more suitable for PM_{2.5} prediction under complex atmospheric conditions. This hierarchical abstraction capability is fundamentally different from the temporal dependency modeling performed by the transformer. Furthermore, using a deeper transformer would mainly increase the model capacity within the same attention-based learning paradigm, leading to higher computational complexity and a greater risk of overfitting. In contrast, the DBN introduces a complementary representation-learning stage that enhances feature robustness, suppresses redundant information, and improves nonlinear environmental pattern extraction. Therefore, the transformer and DBN modules play different but complementary roles within the proposed framework.

The NMOGWO serves as the adaptive optimization engine of the proposed architecture and is responsible for jointly optimizing multiple components of the framework. Specifically, the optimizer simultaneously tunes the transformer hyper-parameters, DBN structural parameters, fuzzy membership functions, and fuzzy rule configurations using a multi-objective optimization strategy. The optimization process considers multiple objective functions associated with prediction accuracy, hidden-layer complexity, fuzzy-rule efficiency, reconstruction quality, and output stability to achieve balanced model performance. By integrating theta-guided leadership, non-dominated sorting, crowding-distance preservation, and adaptive nonlinear exploration–exploitation control, the proposed NMOGWO algorithm can effectively balance global search capability and local so-

lution refinement during optimization. This adaptive optimization mechanism significantly improves convergence stability and prevents premature convergence in highly nonlinear optimization landscapes. As a result, the proposed ODFT-DBN framework achieves enhanced prediction accuracy, stronger optimization robustness, and improved learning efficiency for spatio-temporal PM_{2.5} forecasting problems.

Equation (32) defines the input representation of the proposed ODFT-DBN framework. At each time step, the input vector is constructed by combining ground-based PM_{2.5} observations, meteorological variables, satellite-derived AOD-related features, and spatial descriptors into a unified environmental feature vector.

$$x_t = [x_t^{PM}, x_t^{met}, x_t^{sat}, x_t^{sp}] \quad (32)$$

where x_t denotes the input feature vector at time step t ; x_t^{PM} represents the PM_{2.5}-related observation; x_t^{met} denotes the meteorological variables; x_t^{sat} represents satellite-derived features such as AOD; x_t^{sp} denotes spatial descriptors.

Equation (33) describes the fuzzy transformation of the input variables. Each input feature is mapped into several fuzzy membership degrees to represent uncertainty and gradual transitions in environmental conditions.

$$h_{j,k}(x_{t,j}) = \exp\left(-\frac{(x_{t,j} - c_{j,k})^2}{2\sigma_{j,k}^2}\right) \quad (33)$$

where $h_{j,k}(x_{t,j})$ denotes the membership degree of the j -th input variable in the k -th fuzzy set; $c_{j,k}$ is the membership center; $\sigma_{j,k}^2$ is the membership width; $j = 1, 2, \dots, M$; $M = 16$; $K = 3$.

This formulation enables the model to represent uncertain atmospheric patterns more flexibly than direct numerical inputs. In the adopted factorized fuzzy representation, the membership degrees are independently calculated for each input variable and fuzzy set, without explicitly generating all possible combinations of the input fuzzy partitions. Equation (34) formulates the elementary fuzzy activation strength.

$$r_{t,j,k} = h_{j,k}(x_{t,j}) \quad (34)$$

where $r_{t,j,k}$ denotes the activation strength of the k -th fuzzy set at time step t ; and M represents the total number of input variables. A higher value of $r_{t,k}$ indicates that the current environmental condition strongly matches the corresponding fuzzy rule.

Equation (35) defines the normalized fuzzy rule weight. This normalization ensures that all fuzzy rules contribute proportionally to the final fuzzy representation.

$$\bar{r}_{t,k} = \frac{r_{t,k}}{\sum_{k=1}^K r_{t,q,k} + \varepsilon} \quad (35)$$

where $\bar{r}_{t,k}$ is the normalized activation strength of the k -th fuzzy rule; K denotes the number of fuzzy sets assigned to each input variable; q is the summation index over the fuzzy sets; ε is a small constant used to avoid division by zero.

Equation (36) presents the fuzzy-enhanced feature representation generated from the normalized fuzzy rules. This output becomes the input representation forwarded to the transformer encoder.

$$f_t = \sum_{j=1}^M \sum_{k=1}^K \bar{r}_{t,j,k} (W_{j,k}^f x_{t,j} + b_{j,k}^f) \quad (36)$$

where f_t denotes the fuzzy-enhanced feature vector; $W_{j,k}^f$ and $b_{j,k}^f$ represent the feature- and fuzzy-set-specific weight matrix and bias vector, respectively. This equation establishes the first interaction stage, where uncertain environmental observations are transformed into structured fuzzy features.

Equation (37) defines the sequence-level fuzzy representation used as the input of the transformer encoder. The fuzzy-enhanced vectors from consecutive time steps are arranged as a temporal sequence.

$$F = [f_1, f_2, \dots, f_T] \quad (37)$$

where F denotes the fuzzy temporal feature matrix and T is the length of the input sequence. This representation allows the transformer to model temporal dependencies based on uncertainty-aware environmental features rather than raw observations.

Equation (38) introduces the proposed fuzzy-guided attention mechanism. In this formulation, the transformer attention scores are adaptively weighted using fuzzy rule information, allowing the attention mechanism to focus more strongly on environmentally meaningful and uncertainty-aware temporal patterns.

$$A_F = \text{softmax}\left(\frac{\vec{Q}\vec{K}^T}{\sqrt{d_K}}\right)\lambda_f R \quad (38)$$

where A_F denotes the fuzzy-guided attention matrix; λ_f is a learnable fuzzy-attention balancing coefficient; R represents the fuzzy relevance matrix derived from normalized fuzzy rule activations. This equation is the first proposed interaction innovation between the fuzzy system and the transformer encoder.

Equation (39) computes the fuzzy-guided transformer output by applying the fuzzy-guided attention matrix to the value representation.

$$H_T = A_F \vec{V} \quad (39)$$

where H_T denotes the transformer-generated contextual representation. This operation produces attention-enhanced features that preserve both temporal dependencies and fuzzy uncertainty information.

Equation (40) defines the DBN input mapping. The contextual features generated by the transformer encoder are projected into the input space of the DBN for hierarchical feature abstraction.

$$h^0 = \phi(W_{td}H_T + b_{td}) \quad (40)$$

where h^0 denotes the initial DBN input representation; W_{td} and b_{td} are the transformer-to-DBN mapping parameters; $\phi(\cdot)$ is a nonlinear activation function. This equation defines how transformer-based contextual features are transferred into the DBN module.

Equation (41) formulates the hierarchical feature propagation within the DBN. Each hidden layer receives the representation from the previous layer and transforms it into a more abstract feature space.

$$h^l = \sigma(W^l h^{l-1} + b^l) \quad (41)$$

where h^l denotes the hidden representation at layer l ; W^l and b^l are the weight matrix and bias vector of the l -th DBN layer; l is the number of hidden layers, and $\sigma(\cdot)$ denotes the activation function. This process enables the DBN to learn increasingly abstract representations from transformer-enhanced environmental features.

Equation (42) introduces the proposed attention-DBN fusion representation. This formulation combines the final DBN hidden representation with the transformer contextual representation to preserve both deep hierarchical features and attention-based temporal information.

$$Z_{fuse} = \gamma h^l + (1 - \gamma) \text{Pool}(H_T) \quad (42)$$

where Z_{fuse} denotes the fused representation; h^L is the final DBN hidden output; $Pool(.)$ represents temporal pooling; γ is a learnable fusion coefficient. This equation provides the second proposed interaction innovation by adaptively integrating transformer-based temporal context with DBN-based hierarchical abstraction.

Equation (43) defines the final PM_{2.5} prediction output generated from the fused feature representation.

$$\hat{Y}_t = W_0 Z_{fuse} + b_0 \quad (43)$$

where $\hat{Y}_t$ denotes the predicted PM_{2.5} concentration; W_0 is the output-layer weight matrix; b_0 is the output bias.

The transformer output is not directly fed into the DBN. Instead, the contextual features generated by the transformer encoder are first passed through the projection layer defined in Equation (40), which maps the transformer feature space into a latent representation compatible with the DBN input structure. The resulting feature vector is then used as the visible input of the DBN, where hierarchical feature learning is performed through multiple hidden layers. Consequently, the transformer is responsible for extracting temporal dependencies and contextual information, while the DBN further refines these representations through hierarchical nonlinear abstraction, enabling the proposed framework to benefit from both attention-based temporal learning and deep feature representation within a unified architecture.

Through Equations (32)–(43), the proposed ODFT-DBN framework establishes a sequential interaction pathway in which raw environmental variables are first converted into fuzzy uncertainty-aware features, then processed by the transformer encoder for temporal dependency learning, and finally refined by the DBN for hierarchical prediction generation. Equations (44)–(48) define the multi-objective optimization functions utilized within the proposed NMOGWO framework. During the optimization process, the proposed optimizer simultaneously tunes the hyper-parameters of the transformer encoder, DBN architecture, and fuzzy inference system in order to achieve balanced prediction performance, reduced model complexity, and improved learning stability. Unlike single-objective optimization approaches, the proposed NMOGWO searches for Pareto-optimal solutions capable of satisfying multiple optimization criteria simultaneously.

$$F_1 = \text{Minimize} \sqrt{\frac{1}{N} \sum_{i=1}^N [x_i - \hat{x}_i]^2}, \quad (44)$$

$$F_2 = \text{Minimize} N_l, \quad (45)$$

$$F_3 = \text{Minimize} \sum_{l=1}^L N_l \quad (46)$$

$$F_4 = \text{Minimize} \sum_{i=1}^R \left(\sum_{a=1}^{W_b} Y_d(i) - Y_a(i) \right)^2 \quad (47)$$

$$F_5 = \sum_{i=1}^T \sum_{j=1}^U (Y_{jb} - Y_{ji})^2 - \sum_{i=1}^T \sum_{j=1}^U \max[(Y_{jb} - Y_{jmin}), (Y_{jmax} - Y_{jb})]^2 \quad (48)$$

where x_i is the observed parameter; N is the number of samples; $\hat{x}_i$ is the calculated parameter; N_l is a binary indicator for the presence of a hidden neuron in layer l ; L is the total number of hidden layers; $Y_d(i)$ is the desired output; $Y_a(i)$ is the actual output; Y_{ji}

corresponds to the i -th input for the j -th output; Y_{jb} serving as a reference for the membership center; R indicates the length of the simulation sequence; W_b shows the number of fuzzy rules; T is the number of input–output data; U is the number of outputs.

Equation (44) represents the primary prediction objective function associated with the transformer–DBN learning framework. This objective minimizes the RMSE between the observed PM_{2.5} concentration values and the predicted outputs generated by the proposed model. Through minimizing this objective, the optimizer improves the prediction accuracy and overall regression capability of the integrated transformer–DBN architecture. Equation (45) defines the hidden-layer complexity objective associated with the DBN structure. In this formulation, the optimizer attempts to minimize the number of active hidden neurons within each layer of the network. Reducing unnecessary hidden neurons improves computational efficiency, decreases overfitting risk, and enhances the generalization capability of the proposed framework. Consequently, the optimizer searches for a compact DBN structure capable of preserving high prediction accuracy while reducing architectural redundancy.

Equation (46) formulates the global structural complexity minimization objective of the hierarchical transformer–DBN framework. Unlike Equation (45), which focuses on neuron-level optimization, this objective considers the cumulative complexity of all hidden layers within the network. By minimizing the total hidden-layer complexity, the proposed NMOGWO improves convergence stability and prevents excessive network growth during training. The first three objective functions are therefore directly associated with the transformer encoder and DBN modules. Specifically, Equation (44) focuses on prediction accuracy, while Equations (45) and (46) regulate structural complexity and hierarchical network optimization. Through jointly optimizing these objectives, the proposed NMOGWO can effectively balance model accuracy and computational efficiency within the DL framework.

Equation (47) defines the fuzzy inference optimization objective. This objective minimizes the squared error between the desired fuzzy output and the actual fuzzy output generated by the fuzzy inference system. During optimization, the proposed NMOGWO adaptively tunes the fuzzy membership functions and fuzzy-rule parameters to improve the consistency and reliability of the fuzzy reasoning process. As a result, the fuzzy module becomes more capable of representing uncertain atmospheric relationships associated with PM_{2.5} prediction. Equation (48) represents the fuzzy membership regularization objective utilized to optimize the distribution and positioning of fuzzy membership centers. This objective controls the overlap and separation among fuzzy membership functions to avoid excessively narrow or highly overlapping fuzzy regions. Proper membership distribution is essential for maintaining stable fuzzy inference behavior and improving uncertainty representation capability within the environmental prediction framework.

The final two objective functions are therefore specifically associated with the fuzzy subsystem of the proposed ODFT–DBN architecture. Equation (47) improves fuzzy inference accuracy, while Equation (48) regulates membership-function stability and distribution quality. By simultaneously optimizing both fuzzy-system objectives and transformer–DBN objectives, the proposed NMOGWO establishes a cooperative optimization mechanism capable of jointly improving uncertainty modeling, temporal feature extraction, hierarchical representation learning, and final PM_{2.5} prediction performance.

During the optimization process, the proposed NMOGWO continuously evaluates candidate solutions according to all five objective functions and applies non-dominated sorting to identify Pareto-optimal populations. Candidate wolves with superior tradeoffs among prediction accuracy, network complexity, and fuzzy-system quality are preserved and propagated throughout successive iterations. Meanwhile, the adaptive nonlinear exploration–exploitation mechanism dynamically regulates the search behavior to improve

convergence stability and prevent premature convergence. Consequently, the proposed optimizer can effectively coordinate the interaction among the fuzzy module, transformer encoder, and DBN architecture within a unified multi-objective learning framework. After the non-dominated sorting stage, the final solution was selected from the first Pareto front according to the minimum prediction error criterion. Since prediction accuracy is the primary objective of the proposed PM_{2.5} forecasting framework, the Pareto-optimal solution providing the lowest RMSE while maintaining acceptable model complexity and fuzzy-system quality was adopted as the final network configuration.

The training procedure of the proposed ODFT-DBN framework is performed in two sequential stages to establish a clear connection between transformer-based contextual learning and DBN-based hierarchical representation learning. Initially, the fuzzy-enhanced input sequences are forwarded through the transformer encoder and the projection layer. The transformer is first trained for a supervised warm-up stage using the PM_{2.5} prediction loss, allowing it to generate stable contextual representations from the environmental sequences. The resulting projected transformer features are subsequently used as the visible inputs of the DBN rather than directly applying DBN pre-training to the original environmental observations. During the DBN pre-training stage, the transformer and projection-layer parameters are temporarily fixed, and the RBM layers are trained sequentially using a greedy layer-wise procedure. The first RBM is trained using the projected transformer representations, after which its hidden activations are used as the input of the second RBM. The same procedure is repeated for the remaining RBM layers. This unsupervised pre-training stage provides an informed initialization for the DBN parameters and enables each layer to progressively capture higher-level latent patterns from the transformer-generated contextual features.

After completing layer-wise pre-training, the RBM layers are stacked to form the complete DBN and integrated with the transformer encoder, projection layer, attention-DBN fusion layer, and final prediction layer. All trainable parameters are then unfrozen and jointly fine-tuned in an end-to-end supervised manner using the final PM_{2.5} prediction loss. During this stage, the prediction error is back-propagated through the output layer, fusion module, DBN, projection layer, and transformer encoder, allowing both modules to adapt cooperatively to the forecasting objective. The NMOGWO algorithm operates as an outer optimization mechanism for selecting fuzzy parameters and architectural hyperparameters, whereas the internal network weights are learned through the layer-wise pre-training and end-to-end gradient-based fine-tuning procedures.

The proposed ODFT-DBN framework introduces several important architectural and optimization innovations compared with conventional hybrid PM_{2.5} prediction models. Unlike many existing studies that directly combine DL modules in a sequential manner, the proposed framework establishes an adaptive interaction mechanism among fuzzy reasoning, transformer-based attention learning, DBN hierarchical abstraction, and multi-objective evolutionary optimization. This cooperative integration enables the framework to simultaneously address uncertainty modeling, long-range temporal dependency extraction, nonlinear feature abstraction, and optimization stability within a unified prediction architecture.

One of the primary innovations of the proposed framework is the introduction of the fuzzy-guided attention mechanism formulated in Equation (38). In conventional transformer models, the attention mechanism mainly relies on statistical relationships among query and key representations. However, environmental and atmospheric datasets are inherently uncertain and often contain ambiguous nonlinear relationships that cannot be adequately represented through purely deterministic attention structures. To address this limitation, the proposed framework integrates fuzzy relevance information directly into the attention computation process. This allows the transformer encoder to dynamically

focus on environmentally meaningful temporal patterns while incorporating uncertainty-aware contextual information during feature extraction. Consequently, the proposed fuzzy-guided attention strategy improves the robustness and interpretability of temporal dependency learning compared with standard transformer architectures.

Another important contribution of the proposed framework is the adaptive interaction established between the transformer encoder and the DBN architecture. In many conventional hybrid DL models, transformer and DBN modules operate independently, where one module simply forwards its output to the next stage without adaptive feature refinement. In contrast, the proposed framework introduces an attention–DBN fusion mechanism formulated in Equation (42), which adaptively combines transformer-based contextual representations with DBN hierarchical feature abstractions. This fusion strategy preserves both global temporal dependencies extracted through self-attention and deep nonlinear latent features learned by the DBN. As a result, the proposed architecture can simultaneously capture long-range atmospheric interactions and hidden nonlinear pollutant structures more effectively than conventional hybrid models.

The proposed framework also provides significant advantages in uncertainty representation and environmental adaptability through the integration of fuzzy reasoning with DL-based spatio-temporal learning. Unlike traditional DL architectures that process raw numerical observations directly, the fuzzy subsystem transforms atmospheric variables into uncertainty-aware fuzzy representations before temporal learning begins. This pre-processing stage improves the capability of the model to handle noisy observations, gradual environmental transitions, and ambiguous atmospheric conditions commonly observed in PM_{2.5} prediction tasks. Furthermore, the fuzzy membership functions and rule structures are not statically defined; instead, they are adaptively optimized by the proposed NMOGWO algorithm, enabling the fuzzy system to dynamically adapt to complex environmental behaviors.

From an optimization perspective, the proposed NMOGWO introduces several improvements over the conventional GWO algorithm. First, the incorporation of the theta-guided leadership mechanism enhances exploitation capability and improves convergence stability by introducing an additional dominant search direction during population updating. Second, the integration of non-dominated sorting and crowding-distance preservation enables simultaneous optimization of multiple conflicting objectives while maintaining population diversity across the Pareto front. Third, the proposed nonlinear adaptive exploration–exploitation control mechanism dynamically regulates the search behavior according to the optimization state, thereby improving global search capability during early iterations and strengthening local refinement near convergence stages. These optimization enhancements significantly reduce the probability of premature convergence and improve solution quality for high-dimensional DL optimization problems.

Another major advantage of the proposed architecture is its balanced optimization strategy across all learning modules. In many previous hybrid PM_{2.5} prediction studies, optimization is primarily focused on prediction accuracy alone, often neglecting model complexity, fuzzy-system quality, and hidden-layer efficiency. In contrast, the proposed framework simultaneously optimizes five objective functions associated with prediction accuracy, DBN structural complexity, hidden-neuron configuration, fuzzy inference accuracy, and fuzzy membership stability. This multi-objective optimization strategy allows the proposed architecture to achieve improved generalization capability and computational efficiency while maintaining high prediction performance. Algorithm 1 presents the complete training and optimization procedure of the proposed ODFT-DBN framework. It summarizes the transformer warm-up stage, layer-wise DBN pre-training, end-to-end supervised fine-tuning, and NMOGWO-based multi-objective optimization.

Algorithm 1. Pseudo-codes of the proposed ODFT-DBN**Begin**

1. Construct four-day input sequences and divide into training and testing subsets.
2. Initialize wolves (fuzzy parameters, transformer parameters, and DBN parameters).
3. for each wolf (X_i) do
 - Decode the candidate configuration.
 - Generate fuzzy-enhanced input features using Equations (33)–(37).
 - Initialize the transformer, projection layer, DBN, fusion layer, and output layer.
 - for ($e = 1$) to (E_T) do
 - Train the fuzzy–transformer module using the $PM_{2.5}$ prediction loss.
 - if validation RMSE improves then
 - Save the transformer parameters.
 - end if
 - end for
 - Freeze the transformer and extract projected features using Equations (38)–(40).
 - for ($l = 1$) to (L_DBN) do
 - for ($e = 1$) to (E_R) do
 - Train RBM (l) using Equations (13)–(18).
 - end for
 - Use the hidden output of RBM (l) as the input of RBM ($l + 1$).
 - end for
 - Stack the pre-trained RBMs to construct the DBN and unfreeze all modules.
 - for ($e = 1$) to (E_F) do
 - Generate transformer features using Equations (38) and (39).
 - Project and propagate the features through DBN using Equations (40) and (41).
 - Generate the fused representation and $PM_{2.5}$ prediction using Equations (42) and (43).
 - if the early-stopping condition is satisfied then
 - Break.
 - end if
 - end for
 - Calculate the five objective functions using Equations (44)–(48).
- end for
4. Apply non-dominated sorting and crowding-distance evaluation.
5. for ($t = 1$) to (T_max) do
 - Select the alpha, beta, delta, and theta leaders.
 - Update the adaptive control parameters using Equations (30) and (31).
 - for each wolf (X_i) do
 - Update its position using Equations (22), (23), and (27)–(29).
 - Train and evaluate the corresponding ODFT-DBN configuration.
 - Recalculate the objectives using Equations (44)–(48).
 - end for
 - Apply non-dominated sorting and update the Pareto archive.
 - if the termination condition is satisfied then
 - Break.
 - end if
- end for
6. Select the final Pareto-optimal configuration.
7. Retrain the selected ODFT-DBN model using the training and validation data.
8. Evaluate the final model on the testing dataset.
9. Return the predicted $PM_{2.5}$ values and evaluation metrics.

End

3. Results

All experiments and simulations in this study were implemented using Python 3.10 within the Anaconda development environment. The proposed ODFT-DBN framework, benchmark DL models, optimization algorithms, and preprocessing procedures were developed using several scientific computing and ML libraries. In particular, TensorFlow 2.15 and Keras 2.15 were employed for implementing the transformer-based architectures, hierarchical DBN structures, and neural-network training procedures. PyTorch 2.2 was additionally utilized for several attention-based operations and comparative deep learning implementations. Data preprocessing, normalization, interpolation, statistical analysis, and matrix operations were performed using NumPy 1.26, Pandas 2.2, SciPy 1.12, and Scikit-learn 1.4. Spatial interpolation and geospatial processing tasks associated with environmental variables and monitoring stations were conducted using ArcGIS 10.4.1 and GeoPandas 0.14. In addition, satellite-derived environmental parameters were extracted and processed through GEE.

The proposed NMOGWO optimization framework and the associated multi-objective evolutionary procedures were implemented using custom Python-based optimization modules. All experiments were executed on a high-performance workstation equipped with an Intel Core i9-13900K processor, 64 GB DDR5 RAM, and an NVIDIA RTX 4090 GPU with 24 GB memory. The operating system used for all implementations was Windows 11 Pro 64-bit. This hardware configuration provided sufficient computational capability for large-scale spatio-temporal environmental processing, transformer training, multi-objective optimization, and repeated simulation experiments associated with PM_{2.5} forecasting. To comprehensively evaluate the effectiveness of the proposed ODFT-DBN framework, several well-established ML and DL models were selected as benchmark methods, including BERT, Transformer, DBN, GRU, LSTM, CNN, and XGBoost. These comparative models were chosen because they represent different learning paradigms commonly utilized in environmental prediction and time-series forecasting applications. Specifically, Transformer- and BERT-based models provide strong attention-driven temporal learning capability, while GRU and LSTM represent recurrent architectures specifically designed for sequential dependency modeling. CNN-based approaches are effective in extracting local spatial and temporal patterns, DBN provides hierarchical probabilistic feature learning capability, and XGBoost serves as a powerful tree-based ensemble learning baseline widely used in environmental regression tasks. By comparing the proposed framework against models with fundamentally different learning mechanisms, a more reliable and comprehensive performance evaluation can be achieved.

In addition to comparative benchmarking, an ablation study was also conducted to evaluate the contribution of each major component within the proposed architecture. Several reduced variants of the proposed framework were implemented by selectively removing or modifying individual modules, including the fuzzy inference system, transformer encoder, DBN structure, fuzzy-guided attention mechanism, and the proposed NMOGWO optimization strategy. This analysis enables investigation of the individual impact of each module on prediction accuracy, convergence behavior, and overall model robustness. Through both comparative evaluation and ablation analysis, the effectiveness of the proposed cooperative interaction among fuzzy reasoning, attention-based temporal learning, hierarchical feature extraction, and multi-objective optimization can be systematically validated.

To comprehensively evaluate the prediction capability, robustness, convergence behavior, and statistical reliability of the proposed ODFT-DBN framework, several quantitative and qualitative evaluation metrics were utilized in this study, including R^2 , MAPE, RMSE, scatter-plot analysis, convergence analysis, variance analysis, runtime evaluation, and paired *t*-test statistical analysis. These metrics were selected because they collectively

evaluate prediction accuracy, stability, computational efficiency, correlation consistency, optimization behavior, and statistical significance of the obtained results.

Equation (49) defines the R^2 , which measures the correlation and agreement between the observed PM_{2.5} values and the predicted outputs generated by the model. R^2 evaluates how effectively the proposed framework can explain the variability of the observed environmental data. Higher R^2 values indicate stronger predictive capability and better consistency between predicted and actual observations. An R^2 value close to 1 demonstrates that the model can accurately capture the underlying spatio-temporal pollutant dynamics and environmental relationships associated with PM_{2.5} concentration. Equation (50) represents the MAPE, which measures the average percentage deviation between predicted and observed PM_{2.5} values. Unlike absolute error metrics, MAPE provides a scale-independent evaluation of prediction performance, making it useful for comparing different models under varying environmental conditions. Lower MAPE values indicate smaller prediction deviations and improved forecasting reliability. Therefore, MAPE was selected to evaluate the relative prediction accuracy and practical forecasting consistency of the proposed framework.

$$R^2 = \left[\frac{1}{N} \sum_{i=1}^N \frac{(x_i - \bar{x})(\hat{x}_i - \bar{\hat{x}})}{\sigma_x \sigma_{\hat{x}}} \right]^2, \quad (49)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{x_i - \hat{x}_i}{x_i} \right| \times 100 \quad (50)$$

where x_i is the observed parameter; $\hat{x}_i$ is the calculated parameter; $\bar{x}$ is the mean of the observed parameter; $\bar{\hat{x}}$ the mean of the predicted parameter; N is the number of observations; σ_x the standard deviation of the observed parameter; $\sigma_{\hat{x}}$ is the standard deviation of the predicted parameter.

Equation (51) defines the RMSE, which measures the average magnitude of prediction errors by assigning larger penalties to significant deviations between predicted and observed values. Consequently, RMSE is particularly effective for evaluating the capability of the model to handle extreme pollution events and abrupt PM_{2.5} fluctuations. Lower RMSE values indicate better prediction precision and stronger robustness against large forecasting errors.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N [x_i - \hat{x}_i]^2} \quad (51)$$

Scatter-plot analysis was additionally utilized to visually evaluate the agreement between predicted and observed PM_{2.5} concentrations. In these plots, prediction performance can be analyzed according to the distribution of data points around the ideal regression line. A denser concentration of points near the diagonal line indicates stronger prediction consistency and improved regression capability. Scatter plots also provide intuitive insights regarding prediction bias, variance distribution, and outlier behavior among different models.

Convergence analysis was employed to investigate the optimization behavior and training stability of the proposed NMOGWO algorithm. This analysis illustrates how rapidly the optimization framework approaches the optimal solution during successive iterations. Faster and smoother convergence curves indicate improved optimization efficiency, stronger exploitation capability, and better search stability. Convergence analysis is particularly important for validating the effectiveness of the proposed theta-guided and adaptive nonlinear optimization strategies. Variance analysis was conducted to evaluate the stability and robustness of the proposed framework under repeated simulation runs.

Models exhibiting lower variance demonstrate more stable prediction behavior and reduced sensitivity to initialization conditions or stochastic optimization processes. This analysis is important because DL-based optimization frameworks may occasionally suffer from unstable learning behavior and inconsistent prediction performance across different executions.

Runtime analysis was additionally performed to assess the computational efficiency and practical applicability of the proposed framework. Since the proposed ODFT-DBN architecture integrates multiple learning modules and a multi-objective optimization strategy, evaluating computational complexity and execution time is essential for determining real-world feasibility. Runtime evaluation enables comparison between prediction accuracy improvements and associated computational costs among different benchmark models. Finally, paired *t*-test statistical analysis was employed to verify the statistical significance of the performance improvements achieved by the proposed framework compared with baseline methods. The *t*-test evaluates whether the observed prediction improvements are statistically meaningful rather than resulting from random variations. A low *p*-value indicates that the superiority of the proposed framework is statistically significant and reliable. Therefore, this analysis strengthens the validity and credibility of the experimental results reported in this study.

Before presenting the main PM_{2.5} prediction results, it is important to carefully determine the optimal hyper-parameter configurations of the implemented DL architectures. Hyper-parameters play a critical role in controlling the learning behavior, convergence stability, generalization capability, and computational efficiency of DL models. Improper hyper-parameter selection may lead to overfitting, underfitting, unstable convergence, excessive computational complexity, or poor prediction accuracy. In environmental forecasting problems, where atmospheric variables exhibit highly nonlinear and dynamic behavior, the influence of hyper-parameter settings becomes even more significant because small parameter variations may substantially affect the capability of the model to capture complex spatio-temporal pollutant patterns.

In addition, the proposed ODFT-DBN framework integrates multiple intelligent modules, including fuzzy reasoning, transformer-based temporal learning, DBN hierarchical abstraction, and multi-objective optimization. Consequently, the interaction among different architectural parameters further increases the complexity of the optimization problem. Parameters such as learning rate, hidden-layer configuration, attention-head number, dropout rate, batch size, and optimizer settings directly affect both the prediction performance and convergence behavior of the framework. Therefore, identifying suitable hyper-parameter configurations is essential for ensuring stable training behavior and achieving reliable PM_{2.5} prediction performance.

Table 2 summarizes the final hyper-parameter configurations utilized for the proposed framework and all comparative benchmark models. The table presents the selected learning settings, architectural parameters, optimization configurations, and structural properties associated with each implemented model. For the proposed ODFT-DBN architecture, the hyper-parameters were adaptively optimized using the proposed NMOGWO algorithm. Unlike conventional tuning approaches, NMOGWO simultaneously optimized multiple learning objectives associated with prediction accuracy, network complexity, hidden-layer configuration, and fuzzy-system stability. This adaptive optimization process enabled the proposed framework to identify balanced parameter configurations capable of improving convergence stability and overall prediction robustness. For the proposed ODFT-DBN framework, the optimizer selected a learning rate of 0.004, batch size of 64, feed-forward hidden dimension of 2048, and four attention heads within the transformer encoder.

Table 2. Hyper-parameter configurations for the proposed algorithms.

Model	Parameter	Value
ODFT-DBN (NMOGWO-Fuzzy- Transformer-DBN)	Learning rate	0.004
	Batch size	64
	Feed forward hidden size	2048
	Weight decay	0.01
	Dropout rate	0.2
	Number of attention heads	4
	Number of encoder layers	4
	Momentum	0.4
	Number of hidden layers	3
	Activation function	GELU, tanh
	Optimizer	NMOGWO
	C	0.7
	A	0.3
	α	[0, 2]
	Population size	40
	Iteration	300
	Learning rate	0.002
	Batch size	32
	Dropout rate	0.2
BERT	Number of self-attention heads per layer	6
	Number of transformer encoder layers	6
	Activation Function	GELU
	Optimizer	Adam
	Learning rate	0.005
LSTM	Batch size	64
	Sequence length	4
	Activation function	Tanh & sigmoid
	Number of hidden layers	3
	Optimizer	Adam
GRU	Learning rate	0.003
	Sequence length	6
	Hidden units per layer	64
	Number of GRU layers	4
	Dropout rate	0.2
CNN	Optimizer	Adam
	Number of convolution layers	6
	Kernel size	5 × 5
	Pooling type	Max pooling (2 × 2)
	Number of hidden layers	3
XGBoost	Activation	Tanh
	Optimizer	Adam
	Learning rate	0.08
	Max depth	4
	Number of estimators	300

In addition, four encoder layers and three DBN hidden layers were employed to provide sufficient hierarchical representation capability for complex spatio-temporal environmental learning. A dropout rate of 0.2 and weight decay of 0.01 were utilized to improve generalization and reduce overfitting risk during training. The GELU and tanh ac-

tivation functions were selected because of their strong nonlinear representation capability and stable gradient behavior. Furthermore, the NMOGWO employed a population size of 40 wolves and 300 optimization iterations, enabling efficient exploration and exploitation of the high-dimensional hyper-parameter search space.

The proposed framework performs one-day-ahead PM_{2.5} forecasting. A sliding temporal window consisting of four historical daily observations is used as the model input. Consequently, the transformer sequence length is set to 4, meaning that the model learns temporal dependencies from the previous four days to predict the PM_{2.5} concentration of the subsequent day. For the comparative benchmark architectures, the hyper-parameters were also carefully optimized using a grid-search strategy to ensure fair and reliable comparison conditions. The BERT-based model utilized a learning rate of 0.002, batch size of 32, dropout rate of 0.2, six self-attention heads per layer, and six transformer encoder layers. In addition, the GELU activation function and Adam optimizer were employed to improve nonlinear representation capability and convergence stability during sequential environmental learning.

For the recurrent architectures, the LSTM employed a learning rate of 0.005, batch size of 64, sequence length of 4, and three hidden layers using tanh and sigmoid activation functions together with the Adam optimizer. Similarly, the GRU utilized a learning rate of 0.003, sequence length of 6, four GRU layers, 64 hidden units per layer, and a dropout rate of 0.2. These recurrent configurations were selected to effectively capture temporal dependencies and sequential pollutant variations within the PM_{2.5} dataset while maintaining stable training behavior. For the CNN, six convolutional layers with a kernel size of 5×5 and max-pooling operations of size 2×2 were employed to extract local spatial and temporal patterns from the environmental data. The CNN model additionally utilized three hidden layers, tanh activation, and Adam optimization. Meanwhile, the XGBoost regression model was configured using a learning rate of 0.08, maximum tree depth of 4, and 300 estimators to provide strong ensemble-based nonlinear regression capability. These optimized configurations enabled all benchmark models to operate under near-optimal learning conditions during the comparative PM_{2.5} forecasting experiments.

Table 3 presents the comparative prediction performance of the proposed ODFT-DBN framework and the benchmark ML/DL models for PM_{2.5} forecasting. The evaluation is conducted using three regression metrics, namely RMSE, R^2 , and MAPE. The reported results clearly demonstrate that the proposed ODFT-DBN framework achieves the best overall prediction performance among all comparative methods. Specifically, the proposed model obtains the lowest RMSE value of 0.80 $\mu\text{g}/\text{m}^3$, the highest R^2 value of 0.94, and the lowest MAPE value of 0.35%, indicating superior prediction accuracy, stronger correlation consistency, and significantly lower forecasting error compared with the benchmark architectures.

Among the comparative models, BERT achieved the second-best overall performance with an RMSE of 9.24 $\mu\text{g}/\text{m}^3$, R^2 of 0.81, and MAPE of 8.23%. Although the attention-based structure of BERT enables effective temporal dependency extraction, its performance remains considerably lower than the proposed framework because it lacks uncertainty-aware fuzzy reasoning and hierarchical probabilistic feature abstraction. Similarly, the standard Transformer model achieved relatively competitive results compared with recurrent architectures, obtaining an RMSE of 12.35 $\mu\text{g}/\text{m}^3$ and R^2 of 0.79, confirming the effectiveness of self-attention mechanisms for spatio-temporal environmental learning. However, the absence of fuzzy-guided attention and DBN-based hierarchical refinement limits its capability to capture highly nonlinear atmospheric relationships associated with PM_{2.5} concentration dynamics.

The DBN, LSTM, and GRU models achieved moderate prediction performance, with progressively higher RMSE and lower R^2 values compared with transformer-based approaches. Although recurrent architectures such as LSTM and GRU are effective for sequential learning, their capability to model long-range temporal dependencies and complex heterogeneous environmental interactions is relatively limited compared with attention-driven architectures. The CNN model achieved lower performance due to its primary focus on local feature extraction, which restricts its ability to capture long-term atmospheric dependencies and uncertainty-aware environmental patterns. XGBoost produced the weakest overall performance, with an RMSE of 24.19 $\mu\text{g}/\text{m}^3$, R^2 of 0.58, and MAPE of 21.50%, indicating the limitations of conventional ensemble-learning approaches when handling highly nonlinear spatio-temporal air pollution data.

Table 3. Comparative prediction performance of the proposed models for PM_{2.5} forecasting.

Algorithms	Metrics		
	RMSE ($\mu\text{g}/\text{m}^3$)	R^2	MAPE (%)
ODFT-DBN	0.80	0.94	0.35
BERT	9.24	0.81	8.23
Transformer	12.35	0.79	10.14
DBN	14.93	0.75	11.55
LSTM	16.12	0.71	12.35
GRU	18.36	0.68	14.96
CNN	19.84	0.63	16.87
XGBoost	24.19	0.58	21.50

The superior performance of the proposed ODFT-DBN framework can be attributed to the cooperative interaction among its intelligent modules. First, the fuzzy inference system effectively models uncertainty and ambiguous atmospheric relationships before temporal learning begins. Second, the fuzzy-guided transformer encoder extracts long-range spatio-temporal dependencies while adaptively emphasizing environmentally meaningful temporal patterns through the proposed fuzzy-attention mechanism. Third, the DBN module performs hierarchical feature abstraction capable of learning latent nonlinear pollutant representations from transformer-enhanced contextual features. Finally, the proposed NMOGWO optimization strategy jointly optimizes multiple architectural and fuzzy-system objectives, enabling balanced learning, improved convergence stability, and enhanced generalization capability. These complementary mechanisms collectively enable the proposed framework to significantly outperform existing hybrid and standalone DL architectures for PM_{2.5} prediction tasks.

Scatter-plot analysis provides a visual evaluation of the agreement between measured and estimated PM_{2.5} concentrations generated by prediction models. In these plots, each point represents the relationship between an observed PM_{2.5} value and its corresponding predicted output. The closer the data points are distributed around the ideal diagonal regression line, the stronger the prediction consistency and regression capability of the model. Scatter plots additionally provide important insights regarding prediction dispersion, variance distribution, outlier behavior, and model stability under different pollution conditions. Therefore, this analysis is widely utilized to visually assess the reliability and robustness of environmental forecasting models.

Figure 9 illustrates the scatter-plot comparison between the proposed ODFT-DBN framework and the BERT model, which represents the strongest competing baseline among the benchmark architectures. The scatter plots provide a visual assessment of the agreement between measured and predicted PM_{2.5} concentrations. As shown in Figure 9a,

the scatter points associated with the proposed ODFT-DBN framework are densely concentrated around the regression line, indicating a strong correspondence between measured and estimated PM_{2.5} values across different pollution levels. In comparison, Figure 9b shows a noticeably larger dispersion of scatter points for the BERT model, particularly at moderate and high pollution concentrations, suggesting reduced prediction consistency and larger forecasting deviations.

For the proposed ODFT-DBN framework, the fitted regression equation is approximately $Y = 1.00X + 0.55$, with an R^2 value of 0.94 and RMSE of 0.80 $\mu\text{g}/\text{m}^3$. The regression slope being extremely close to 1 indicates that the proposed framework can accurately preserve the proportional relationship between measured and predicted PM_{2.5} concentrations across different pollution levels. Furthermore, the very small intercept value demonstrates low systematic prediction bias. The high R^2 value confirms that the proposed framework successfully captures the dominant spatio-temporal environmental relationships associated with PM_{2.5} concentration dynamics. In addition, the low RMSE value indicates strong robustness against large prediction errors and high reliability under both low- and high-pollution conditions.

In contrast, the BERT model shown in Figure 9b produces a regression equation of approximately ($Y = 0.53X + 11.58$), with an R^2 value of 0.81 and an RMSE of 9.24 $\mu\text{g}/\text{m}^3$. The substantially smaller regression slope indicates that the BERT model tends to underestimate higher PM_{2.5} concentrations and exhibits weaker sensitivity to severe pollution events. Moreover, the larger intercept value and broader scatter-point distribution reveal increased prediction bias and reduced stability compared with the proposed framework. Although BERT benefits from attention-based sequence learning, its predictions remain more dispersed around the ideal regression line, particularly at elevated concentration levels. The superior behavior of the proposed ODFT-DBN framework can be attributed to the cooperative interaction among fuzzy uncertainty modeling, transformer-based temporal attention learning, DBN hierarchical feature abstraction, and NMOGWO-driven adaptive optimization. These components collectively improve environmental feature representation, uncertainty handling, and learning stability, enabling the proposed framework to achieve substantially higher prediction accuracy and stronger agreement with the measured PM_{2.5} observations than the BERT baseline.

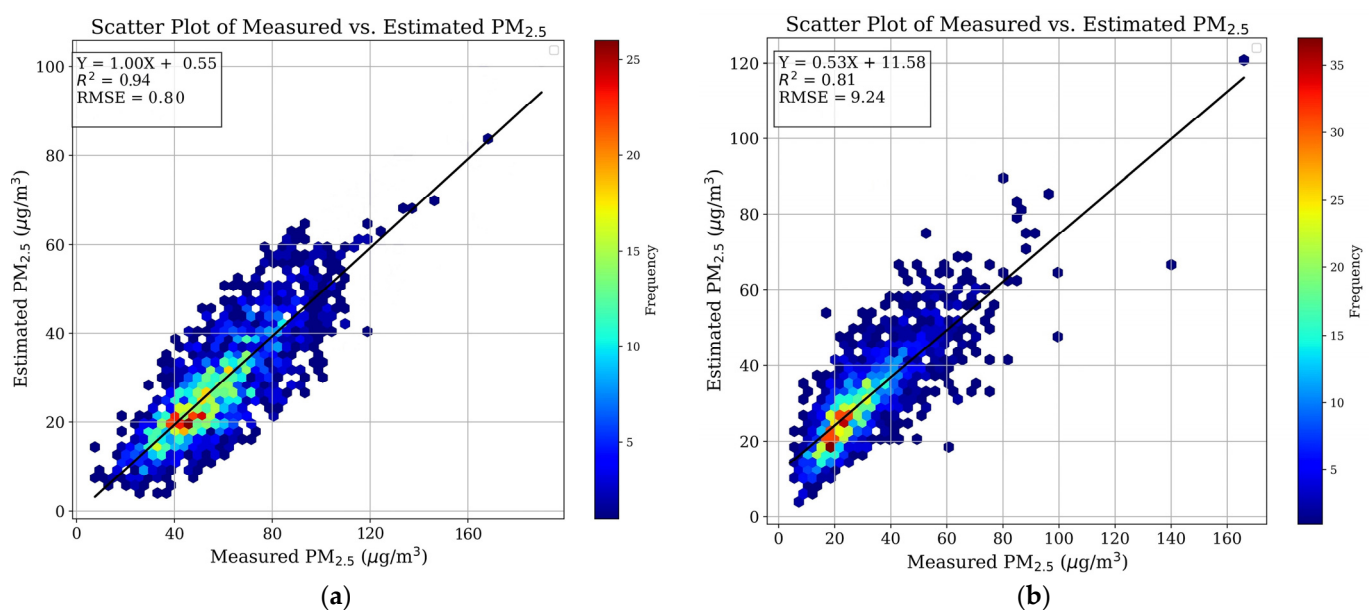


Figure 9. Comparison of scatter plots between measured and predicted PM_{2.5} concentrations for: (a) ODFT-DBN; (b) BERT.

Figure 10 illustrates the convergence trends of the proposed ODFT-DBN framework and the comparative benchmark models based on the RMSE metric over 300 training epochs. Convergence analysis is important for evaluating the optimization stability, learning efficiency, and training behavior of DL architectures. In general, faster convergence toward lower RMSE values indicates stronger optimization capability and improved learning efficiency. Furthermore, smoother convergence curves demonstrate better stability during training and reduced sensitivity to local minima or optimization fluctuations.

As shown in Figure 10, the proposed ODFT-DBN framework achieves significantly faster and more stable convergence compared with all benchmark models. The proposed model rapidly reduces the RMSE value during the early training epochs and reaches near-optimal convergence after approximately 120–150 epochs. After this stage, the convergence curve becomes highly stable with only very small fluctuations, indicating strong optimization stability and effective parameter adaptation. In contrast, the benchmark models exhibit slower convergence behavior and require substantially larger numbers of epochs to approach their final performance levels. Among the comparative methods, BERT demonstrates the second-fastest convergence trend, followed by Transformer and DBN architectures, while CNN and XGBoost exhibit the slowest convergence behavior and maintain relatively high RMSE values throughout the training process.

The superior convergence behavior of the proposed ODFT-DBN framework can be attributed to several cooperative architectural and optimization mechanisms. First, the fuzzy inference module improves the quality and stability of the input feature representations before temporal learning begins, thereby reducing uncertainty and noise propagation during training. Second, the transformer encoder efficiently captures long-range spatio-temporal dependencies using the proposed fuzzy-guided attention mechanism, enabling more informative contextual learning during early training stages. Third, the DBN architecture performs hierarchical feature abstraction that progressively refines latent environmental representations and stabilizes deep feature learning. Finally, the proposed NMOGWO, together with the theta-guided leadership mechanism and adaptive nonlinear exploration–exploitation strategy, significantly improves optimization efficiency and prevents premature convergence.

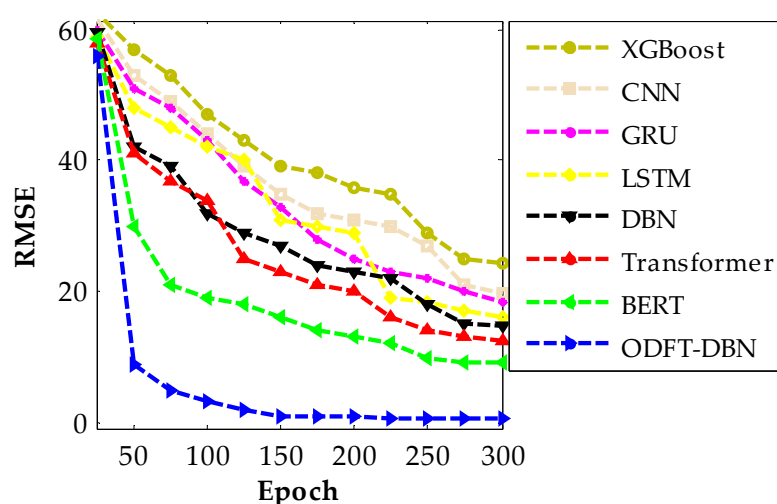


Figure 10. The convergence trend of proposed models based on the RMSE metric.

Another important observation from Figure 10 is the difference in convergence smoothness among the compared models. The CNN, GRU, and LSTM architectures exhibit larger fluctuations and slower RMSE reduction throughout the training process, indicating weaker optimization stability and lower capability for modeling highly nonlinear

atmospheric dependencies. Similarly, the XGBoost model demonstrates limited convergence improvement after approximately 200 epochs, suggesting reduced adaptability for complex spatio-temporal environmental forecasting tasks. In contrast, the proposed ODFT-DBN maintains a consistently smooth convergence curve with minimal oscillations after the initial optimization stage. This behavior confirms the effectiveness of the proposed multi-objective optimization framework in balancing exploration and exploitation while adaptively tuning the transformer, DBN, and fuzzy-system parameters during training.

Figure 11 presents the spatial distribution maps of PM_{2.5} concentrations across Tehran for a selected day in November 2016. The figure compares the ground-truth pollution map obtained from observed environmental measurements with the spatial prediction maps generated by the proposed ODFT-DBN framework and several benchmark DL architectures. These maps provide an intuitive visualization of the capability of different models to reconstruct the spatial behavior of PM_{2.5} concentrations under real atmospheric conditions. Unlike numerical metrics alone, spatial distribution analysis enables evaluation of how accurately the models preserve pollutant hotspots, regional gradients, and localized pollution variations across different urban regions.

As shown in Figure 11a, the observed PM_{2.5} distribution demonstrates significant spatial variability throughout Tehran during the selected winter period. Higher pollution concentrations are mainly concentrated in the central, southeastern, and southwestern regions of the city, while comparatively lower pollution levels are observed in northern and northeastern areas. This spatial behavior is strongly associated with Tehran's geographical and urban characteristics. The southern and central districts are typically characterized by higher traffic density, industrial activity, dense urbanization, and limited air circulation, leading to increased pollutant accumulation. In contrast, northern regions are located closer to the Alborz mountain slopes and generally benefit from slightly improved ventilation and lower urban density. Furthermore, November corresponds to the cold season in Tehran, during which temperature inversion phenomena frequently occur and trap pollutants near the surface, significantly intensifying PM_{2.5} accumulation across the metropolitan area.

Figure 11b demonstrates that the proposed ODFT-DBN framework can reconstruct the spatial pollution distribution with high consistency relative to the ground-truth observations. The proposed model successfully preserves the major pollution hotspots observed in the central and southern regions while simultaneously capturing smoother transitional pollutant gradients across the urban area. In particular, the model accurately reproduces the severe pollution concentrations observed in the southeastern districts and the comparatively cleaner regions in the northwestern parts of the city. The close visual agreement between the predicted and observed maps confirms the strong spatial learning capability of the proposed architecture. This superior performance can be attributed to the cooperative integration of fuzzy uncertainty modeling, transformer-based spatio-temporal dependency extraction, DBN hierarchical feature learning, and NMOGWO-driven optimization.

The comparative benchmark models exhibit noticeably weaker spatial reconstruction capability. Although the BERT and Transformer models shown in Figure 11c,d can partially capture the general spatial distribution trends, several local pollution hotspots are either underestimated or spatially smoothed. Their predicted pollutant gradients appear less consistent with the observed ground-truth map, particularly in highly polluted central regions. This behavior indicates that attention-based architectures alone may struggle to fully model nonlinear environmental uncertainty and localized atmospheric interactions without fuzzy reasoning and hierarchical feature refinement mechanisms. The DBN and LSTM models shown in Figure 11e,f exhibit even larger deviations from the observed

pollution patterns. In particular, these models demonstrate weaker spatial sensitivity and reduced capability in preserving localized $\text{PM}_{2.5}$ concentration peaks, resulting in over-smoothed pollution distributions across multiple urban regions.

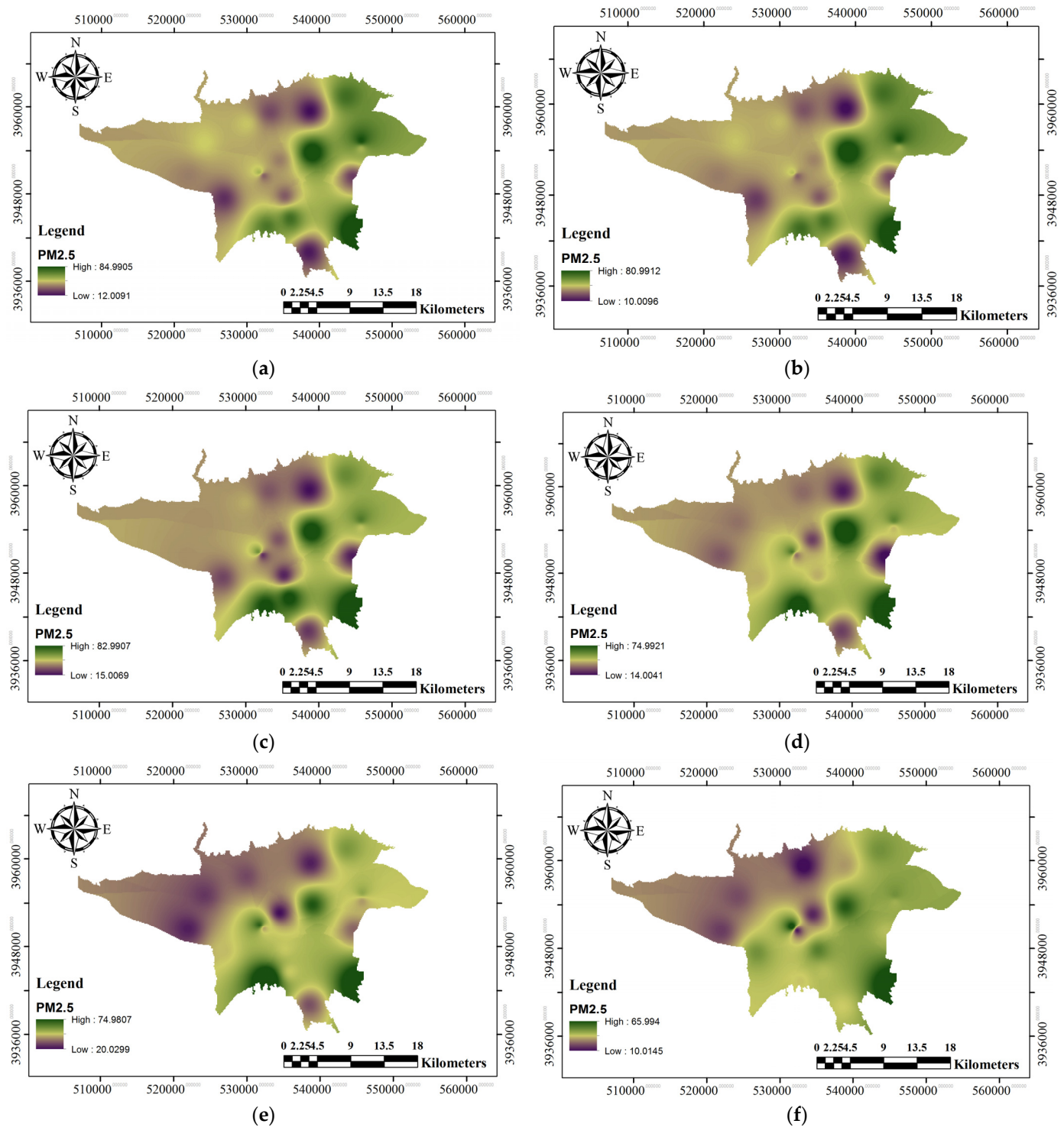


Figure 11. Spatial comparison of observed and predicted $\text{PM}_{2.5}$ distributions in November 2016 using: (a) Measured ground-truth data; (b) ODFT-DBN; (c) BERT; (d) Transformer; (e) DBN; (f) LSTM.

Overall, the spatial distribution analysis further validates the superiority of the proposed ODFT-DBN framework beyond conventional numerical regression metrics. The proposed architecture not only achieves improved RMSE and R^2 performance but also demonstrates stronger capability in preserving realistic spatial pollutant structures under

complex winter atmospheric conditions. The ability to accurately capture both high-pollution hotspots and low-concentration regions is critically important for practical air-quality monitoring, environmental management, and urban pollution mitigation strategies. These results confirm that the proposed cooperative learning framework can effectively model the complex spatio-temporal and environmental dynamics governing PM_{2.5} behavior in large metropolitan regions such as Tehran.

Table 4 presents the RMSE performance of the proposed ODFT-DBN framework and the comparative benchmark models across different PM_{2.5} concentration ranges, including low, moderate, and high pollution conditions. This analysis is important because air pollution forecasting models may exhibit different prediction behaviors depending on pollution severity and atmospheric complexity. In practical environmental monitoring applications, accurately predicting high-pollution events is particularly critical due to their severe impacts on public health, transportation visibility, and urban environmental quality. Therefore, evaluating model performance across multiple pollution ranges provides deeper insight into the robustness and adaptability of the forecasting architectures under varying atmospheric conditions.

The results demonstrate that the proposed ODFT-DBN framework consistently achieves the lowest RMSE values across all concentration ranges. Specifically, the proposed model achieves RMSE values of 0.42 µg/m³, 0.64 µg/m³, and 0.91 µg/m³ under low, moderate, and high-concentration conditions, respectively. These results indicate that the proposed framework maintains highly stable prediction performance even under severe pollution scenarios characterized by strong atmospheric nonlinearity and rapid pollutant accumulation. In contrast, all comparative benchmark models exhibit substantially larger prediction errors, particularly under high-pollution conditions. For example, the RMSE of XGBoost increases from 22.19 µg/m³ under low pollution to 26.37 µg/m³ under high pollution levels, while recurrent and CNN-based models also demonstrate noticeable performance degradation as pollution intensity increases.

The increasing prediction difficulty observed under high PM_{2.5} concentration levels is strongly associated with the complex environmental and meteorological processes governing pollutant accumulation in Tehran. High-pollution conditions are typically influenced by multiple interacting factors, including traffic density, industrial emissions, atmospheric pressure, humidity, temperature inversion, weak wind circulation, and seasonal climatic conditions. During severe winter pollution episodes, pollutants become trapped near the ground due to inversion layers and limited atmospheric mixing, leading to highly nonlinear spatial and temporal pollutant behavior. Such environmental complexity significantly increases forecasting difficulty because pollutant distributions become more dynamic, heterogeneous, and sensitive to local atmospheric variations.

Table 4. RMSE comparison of the proposed models under different PM_{2.5} concentration levels.

Methods	RMSE (µg/m ³)		
	Low Concentration	Moderate Concentration	High Concentration
ODFT-DBN	0.42	0.64	0.91
BERT	7.36	8.14	10.42
Transformer	10.75	11.37	12.33
DBN	12.41	13.18	15.87
LSTM	14.26	15.37	17.49
GRU	16.10	17.44	18.05
CNN	17.83	18.63	20.48
XGBoost	22.19	23.25	26.37

Among the benchmark architectures, BERT demonstrates the strongest comparative performance, achieving RMSE values of $7.36 \mu\text{g}/\text{m}^3$, $8.14 \mu\text{g}/\text{m}^3$, and $10.42 \mu\text{g}/\text{m}^3$ across low-, moderate-, and high-pollution conditions, respectively. This result confirms the effectiveness of attention-based temporal learning for environmental forecasting tasks. However, the performance of BERT still remains considerably lower than the proposed framework because it lacks explicit uncertainty modeling and hierarchical feature abstraction mechanisms. Similarly, the Transformer architecture exhibits reasonable adaptability but shows larger performance degradation at higher pollution levels due to the absence of fuzzy-guided contextual learning and DBN-based nonlinear feature refinement.

Table 5 presents the ablation-study results of the proposed ODFT-DBN framework under different architectural combinations and module configurations. The primary objective of this analysis is to evaluate the individual contribution of each major component, including the fuzzy inference system, transformer encoder, DBN architecture, and NMOGWO optimization strategy. By systematically removing or combining different modules, the effectiveness of each component and their cooperative interactions can be quantitatively investigated.

The complete ODFT-DBN framework achieves the best overall performance, obtaining an RMSE of $0.80 \mu\text{g}/\text{m}^3$, R^2 of 0.94, and MAPE of 0.35%. These results clearly confirm that the cooperative integration of fuzzy reasoning, transformer-based temporal learning, DBN hierarchical abstraction, and NMOGWO optimization significantly improves $\text{PM}_{2.5}$ prediction capability. Among the partial hybrid configurations, the NMOGWO-Fuzzy-Transformer model achieves the strongest comparative performance with an RMSE of $1.52 \mu\text{g}/\text{m}^3$ and R^2 of 0.90, indicating that the interaction between uncertainty-aware fuzzy reasoning and transformer-based temporal attention learning plays a highly important role in environmental forecasting. The strong performance of this configuration demonstrates the capability of the proposed fuzzy-guided attention mechanism to improve contextual temporal representation under nonlinear atmospheric conditions.

Table 5. Ablation study of ODFT-DBN components.

Algorithms	Metrics		
	RMSE ($\mu\text{g}/\text{m}^3$)	R^2	MAPE (%)
ODFT-DBN (NMOGWO-Fuzzy-Transformer-DBN)	0.80	0.94	0.35
NMOGWO-Fuzzy-Transformer	1.52	0.90	1.72
NMOGWO-Fuzzy-DBN	2.14	0.89	1.89
NMOGWO-Transformer -DBN	2.90	0.89	2.95
Fuzzy-Transformer-DBN	3.70	0.88	3.08
NMOGWO-Transformer	4.10	0.87	4.15
NMOGWO-DBN	5.01	0.86	5.21
Fuzzy-Transformer	6.42	0.85	5.90
Fuzzy-DBN	6.95	0.84	6.50
Transformer-DBN	7.07	0.83	7.01
Transformer	12.35	0.79	10.14
DBN	14.93	0.75	11.55

The NMOGWO-Fuzzy-DBN and NMOGWO-Transformer-DBN architectures also achieve relatively competitive results compared with simpler combinations, obtaining RMSE values of $2.14 \mu\text{g}/\text{m}^3$ and $2.90 \mu\text{g}/\text{m}^3$, respectively. The stronger performance of NMOGWO-Fuzzy-DBN compared with NMOGWO-Transformer-DBN suggests that uncertainty modeling through fuzzy reasoning contributes more significantly to $\text{PM}_{2.5}$ forecasting stability than standalone temporal attention learning when hierarchical feature

abstraction is already present. Since atmospheric pollution behavior is strongly influenced by ambiguous environmental transitions, seasonal variability, and nonlinear meteorological interactions, fuzzy-based uncertainty representation substantially improves environmental feature quality before deep learning begins.

Another important observation is that removing the NMOGWO optimization module noticeably degrades prediction performance across all architectural combinations. For example, the Fuzzy-Transformer-DBN architecture without NMOGWO produces an RMSE of $3.70 \mu\text{g}/\text{m}^3$ and R^2 of 0.88, which is considerably weaker than the fully optimized ODFT-DBN framework. This result confirms the importance of adaptive multi-objective optimization for jointly tuning transformer hyper-parameters, DBN configurations, and fuzzy membership functions. Without the proposed NMOGWO, the interaction among different modules becomes less balanced, leading to weaker convergence stability and reduced learning efficiency.

The simpler dual-module architectures, including Fuzzy-Transformer, Fuzzy-DBN, and Transformer-DBN, achieve moderate prediction performance but remain substantially weaker than the full hybrid framework. Specifically, Transformer-DBN achieves an RMSE of $7.07 \mu\text{g}/\text{m}^3$ and R^2 of 0.83, while Fuzzy-DBN and Fuzzy-Transformer produce RMSE values of $6.95 \mu\text{g}/\text{m}^3$ and $6.42 \mu\text{g}/\text{m}^3$, respectively. These results indicate that although each pairwise module combination contributes useful environmental representation capability, none of them alone can fully capture the complex spatio-temporal and uncertainty-aware pollutant relationships governing $\text{PM}_{2.5}$ dynamics. The transformer primarily improves long-range temporal dependency extraction, the DBN strengthens nonlinear hierarchical abstraction, and the fuzzy subsystem enhances uncertainty modeling. Therefore, the simultaneous integration of all three modules is necessary to maximize forecasting performance.

Finally, the standalone Transformer and DBN architectures achieve the weakest performance among the ablation configurations, with RMSE values of $12.35 \mu\text{g}/\text{m}^3$ and $14.93 \mu\text{g}/\text{m}^3$, respectively. These results further demonstrate that single-module DL architectures are insufficient for modeling the highly nonlinear, uncertain, and heterogeneous environmental processes associated with $\text{PM}_{2.5}$ pollution forecasting. Overall, the ablation analysis clearly validates that the superior performance of the proposed ODFT-DBN framework is not attributed to a single component alone, but rather to the adaptive cooperative interaction among fuzzy reasoning, transformer attention learning, DBN hierarchical feature extraction, and NMOGWO-based multi-objective optimization.

To evaluate the effectiveness of the proposed NMOGWO, seven representative state-of-the-art multi-objective meta-heuristic algorithms were additionally integrated into the same ODFT-DBN framework and compared under identical experimental conditions. The considered optimizers include multi-objective orchard optimization algorithm (MOOA), multi-objective chimp optimization algorithm (MOChOA), MOGWO, multi-objective artificial bee colony (MOABC), multi-objective biogeography-based optimization (MOBBO), multi-objective particle swarm optimization (MOPSO), and non-dominated sorting genetic algorithm (NSGA-II). In all cases, only the optimization engine was replaced, while the fuzzy system, transformer encoder, DBN architecture, dataset, and training configuration remained unchanged. This experimental design enables a fair assessment of the contribution of the proposed NMOGWO algorithm.

The results presented in Table 6 clearly demonstrate the superiority of the proposed NMOGWO. Specifically, the proposed ODFT-DBN framework achieves the lowest RMSE value of $0.80 \mu\text{g}/\text{m}^3$, the highest coefficient of determination (R^2) of 0.94, and the lowest MAPE value of 0.35%. The closest competitor is MOOA, which achieves an RMSE of $1.19 \mu\text{g}/\text{m}^3$ and an MAPE of 1.67%. Compared with MOOA, the proposed optimizer reduces RMSE by approximately 32.8% and decreases MAPE by nearly 79%, while simultaneously

improving the coefficient of determination. These results indicate that the proposed optimization strategy is more effective in identifying high-quality parameter configurations for the fuzzy, transformer, and DBN modules.

Table 6. Comparison of NMOGWO with other optimizers within the FT-DBN framework.

Algorithms	Metrics		
	RMSE ($\mu\text{g}/\text{m}^3$)	R^2	MAPE (%)
ODFT-DBN (NMOGWO-Fuzzy-Transformer-DBN)	0.80	0.94	0.35
MOOA-Fuzzy-Transformer-DBN	1.19	0.91	1.67
MOChOA-Fuzzy-Transformer-DBN	1.27	0.91	1.81
MOGWO-Fuzzy-Transformer-DBN	1.56	0.90	2.05
MOABC-Fuzzy-Transformer-DBN	1.69	0.90	2.15
MOBBO-Fuzzy-Transformer-DBN	2.55	0.89	2.71
MOPSO-Fuzzy-Transformer-DBN	2.68	0.89	2.83
NSGA_II-Fuzzy-Transformer-DBN	2.83	0.89	2.97

The superior performance of NMOGWO can be attributed to several enhancements introduced into the conventional GWO framework. First, the proposed theta-guided leadership mechanism enriches the search process by incorporating an additional dominant wolf, thereby providing more informative search directions and reducing the probability of stagnation around local optima. Second, the adaptive nonlinear exploration–exploitation control mechanism dynamically regulates the search behavior according to the optimization state, allowing broader exploration during early iterations and stronger exploitation near convergence. Third, the integration of non-dominated sorting and crowding-distance preservation improves population diversity and enables more effective exploration of the Pareto-optimal search space.

Although algorithms such as MOPSO, MOABC, and MOBBO possess strong global-search capabilities, they frequently exhibit unstable convergence behavior when dealing with highly nonlinear and strongly coupled optimization problems. Similarly, NSGA-II provides effective Pareto-front generation but often requires a larger number of evolutionary updates to identify highly competitive solutions in complex high-dimensional search spaces. Furthermore, conventional MOGWO relies only on three dominant wolves and employs a linear exploration–exploitation transition mechanism, which may reduce diversity and increase the risk of premature convergence. These limitations become particularly evident in the simultaneous optimization of transformer hyper-parameters, DBN structural parameters, and fuzzy-system configurations.

Figure 12 illustrates the convergence behavior of the proposed NMOGWO and the benchmark multi-objective optimization algorithms when integrated into the ODFT-DBN framework. The horizontal axis represents the optimization iterations (epochs), while the vertical axis denotes the corresponding RMSE values. It can be observed that the proposed NMOGWO exhibits the fastest convergence rate among all compared optimizers. During the early optimization stages, NMOGWO rapidly reduces the RMSE value, indicating strong global exploration capability and efficient identification of promising search regions. As the optimization progresses, the proposed optimizer continues refining the candidate solutions and converges toward the lowest RMSE values, whereas most competing algorithms require substantially more iterations to achieve comparable performance. In particular, NSGA-II, MOPSO, MOBBO, and MOABC exhibit slower convergence behavior and remain trapped at relatively high RMSE levels for a considerable number of iterations before approaching their final solutions.

Another important observation is the convergence stability achieved by the proposed optimizer. While several benchmark algorithms exhibit noticeable fluctuations and irregular convergence trajectories, NMOGWO demonstrates a smoother and more consistent optimization process throughout the search procedure. This behavior can be attributed to the combined effects of the theta-guided leadership mechanism, adaptive nonlinear exploration–exploitation control, and Pareto-based diversity preservation. These enhancements improve search-direction quality, reduce premature convergence, and maintain population diversity during optimization. Consequently, the proposed NMOGWO achieves a more favorable balance between exploration and exploitation, allowing it to converge faster and reach higher-quality solutions than the competing multi-objective optimization algorithms.

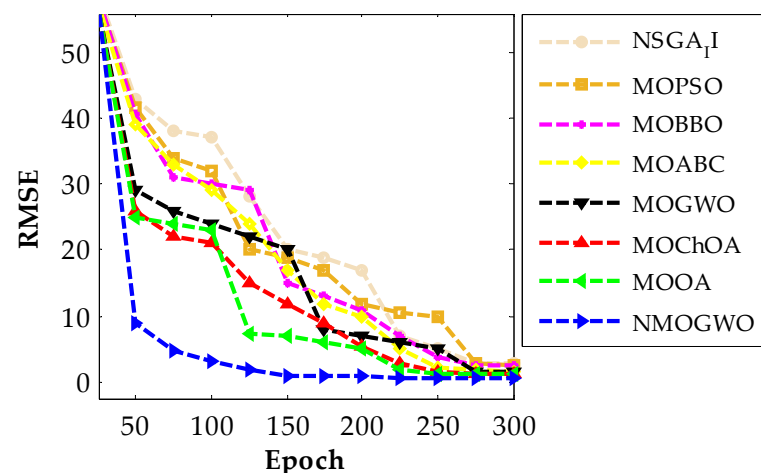


Figure 12. Convergence comparison of the proposed NMOGWO and benchmark algorithms.

Another important observation concerns the swarm-based optimization algorithms, including MOPSO and several population-driven search strategies. Although these methods are capable of performing effective global exploration during the early stages of optimization, their search behavior is often highly dependent on the quality and distribution of the initial population. When the initial particles or agents are generated far from promising regions of the search space, the population may require a large number of iterations to gradually move toward high-quality solutions. This limitation becomes more pronounced in high-dimensional optimization problems, where the distance between candidate solutions and the global optimum can be substantial. Furthermore, most swarm-based algorithms rely on gradual position updates driven by collective population behavior. As a result, they generally lack strong mechanisms for performing abrupt directional corrections when the search process becomes trapped around locally optimal regions. Consequently, the population may continue exploring suboptimal areas for extended periods before identifying more promising search directions. This issue often leads to slower convergence rates and increased computational cost, particularly in complex optimization landscapes characterized by multiple local optima and strong parameter interactions.

In addition, maintaining a proper balance between exploration and exploitation remains a challenging task for many swarm-intelligence algorithms. Excessive exploration may delay convergence and reduce optimization efficiency, whereas aggressive exploitation can rapidly reduce population diversity and increase the risk of premature convergence. These challenges are especially critical in the proposed optimization problem, where transformer hyper-parameters, DBN parameters, fuzzy membership functions, and fuzzy-rule configurations must be optimized simultaneously within a highly nonlinear and strongly coupled search space. In contrast, the proposed NMOGWO addresses these

limitations through the integration of the theta-guided leadership mechanism, adaptive nonlinear exploration–exploitation control, and Pareto-based diversity preservation. The additional theta leader provides richer directional guidance during population updates, while the adaptive search mechanism dynamically adjusts exploration and exploitation according to the current optimization state. Together with non-dominated sorting and crowding-distance preservation, these improvements enhance convergence stability, reduce sensitivity to the initial population distribution, and improve the optimizer’s ability to escape local optima, ultimately leading to superior optimization performance.

To provide a more rigorous evaluation against state-of-the-art PM_{2.5} forecasting approaches, the proposed ODFT-DBN framework was compared with several advanced spatio-temporal prediction models, including the temporal fusion transformer (TFT), attention-based spatial–temporal graph convolutional network (ASTGCN), spatio-temporal graph convolutional network (ST-GCN), and diffusion convolutional recurrent neural network (DCRNN). These architectures were selected because they represent some of the most widely adopted DL frameworks for modeling complex spatial and temporal dependencies in environmental forecasting applications. While TFT focuses on attention-based temporal dependency learning, ASTGCN, ST-GCN, and DCRNN utilize graph-based representations to capture spatial interactions among monitoring locations.

The results presented in Table 7 demonstrate that the proposed ODFT-DBN framework consistently outperforms all competing models across all evaluation metrics. Specifically, ODFT-DBN achieves an RMSE of 0.80 $\mu\text{g}/\text{m}^3$, an R^2 value of 0.94, and a MAPE of only 0.35%, whereas the strongest competing baseline, TFT, achieves an RMSE of 6.30 $\mu\text{g}/\text{m}^3$, an R^2 value of 0.85, and a MAPE of 5.82%. Similar performance gaps can be observed for ASTGCN, ST-GCN, and DCRNN. These results indicate that the proposed framework is capable of capturing the underlying PM_{2.5} dynamics with substantially higher accuracy and lower prediction error than existing state-of-the-art spatio-temporal forecasting architectures.

The superior performance of the proposed framework can be attributed to its multi-stage learning strategy. Unlike graph-based architectures that primarily focus on spatial dependency modeling or transformer-based architectures that mainly emphasize temporal attention learning, the proposed ODFT-DBN simultaneously integrates fuzzy uncertainty modeling, transformer-based contextual feature extraction, DBN-based hierarchical representation learning, and NMOGWO-driven adaptive optimization. Consequently, the framework is able to model uncertainty, nonlinear feature interactions, long-range temporal dependencies, and hierarchical environmental patterns within a unified learning architecture. Another important observation is that although graph-based models such as ASTGCN, ST-GCN, and DCRNN effectively exploit spatial relationships among monitoring stations, their performance remains limited by the predefined graph structure and neighborhood aggregation mechanisms. Similarly, TFT provides strong temporal modeling capability but does not explicitly address uncertainty representation or hierarchical feature abstraction.

Table 7. Performance comparison of the ODFT-DBN and state-of-the-art spatio-temporal models.

Algorithms	Metrics		
	RMSE ($\mu\text{g}/\text{m}^3$)	R^2	MAPE (%)
ODFT-DBN	0.80	0.94	0.35
TFT	6.30	0.85	5.82
ASTGCN	6.76	0.84	6.42
ST-GCN	7.29	0.83	6.93
DCRNN	8.24	0.82	7.53

4. Discussion

In the previous section, the prediction capability of the proposed ODFT-DBN framework was comprehensively evaluated using several quantitative and qualitative performance metrics, including RMSE, R^2 , MAPE, scatter-plot analysis, convergence behavior, and spatial distribution analysis. The obtained results demonstrated that the proposed framework achieves significantly improved PM_{2.5} forecasting accuracy and stronger convergence stability compared with conventional ML and DL architectures. Furthermore, the ablation-study results confirmed the effectiveness of the cooperative interaction among fuzzy reasoning, transformer-based temporal learning, DBN hierarchical abstraction, and NMOGWO-driven multi-objective optimization.

Although prediction accuracy is one of the most important evaluation criteria in environmental forecasting applications, relying solely on accuracy-based metrics is insufficient for determining the practical applicability and generalization capability of a prediction framework. In real-world atmospheric monitoring systems, models are additionally required to exhibit stable prediction behavior, computational efficiency, low inference latency, and strong robustness under different operational conditions. Therefore, beyond conventional regression metrics, it is essential to investigate additional evaluation criteria capable of analyzing the reliability, scalability, and deployment feasibility of the proposed framework in practical environmental monitoring scenarios.

In this section, several additional analytical criteria are investigated to further evaluate the robustness and real-world applicability of the proposed ODFT-DBN architecture. Specifically, variance analysis is conducted to examine the stability and consistency of the prediction framework under repeated simulation runs and stochastic training conditions. Statistical significance analysis based on paired *t*-test evaluation is additionally performed to verify whether the observed performance improvements are statistically meaningful rather than resulting from random optimization variations. These analyses provide deeper insight into the reliability and reproducibility of the proposed framework across multiple executions.

Moreover, computational efficiency and deployment feasibility are analyzed through runtime evaluation, inference-latency analysis, and computational complexity assessment. Runtime analysis investigates the total computational cost associated with model training and optimization procedures, while inference latency evaluates the time required to generate PM_{2.5} predictions after training. In practical smart-city and environmental-monitoring applications, low inference latency is critically important because pollution forecasting systems often require near real-time prediction capability for rapid environmental response and decision-making. Consequently, evaluating both runtime complexity and inference efficiency is necessary for determining the suitability of the proposed framework for large-scale operational deployment. Therefore, the analyses presented in this section aim not only to evaluate the predictive superiority of the proposed framework, but also to determine its practical feasibility, robustness, and adaptability for real-world PM_{2.5} monitoring and forecasting applications.

Table 8 presents the inference latency comparison of the proposed ODFT-DBN architecture and the comparative baseline models. Inference latency represents the time required by a trained model to generate a PM_{2.5} prediction after receiving the input environmental data. This metric is particularly important for practical smart-city and air-quality monitoring applications, where prediction systems are often expected to operate in near real-time conditions. Lower inference latency indicates faster prediction response and improved suitability for online environmental monitoring and rapid decision-making systems.

The results indicate that the proposed ODFT-DBN framework achieves an inference latency of 9.1 ms, which is slightly higher than the benchmark architectures due to its

more sophisticated hybrid structure integrating fuzzy reasoning, transformer-based attention learning, DBN hierarchical abstraction, and NMOGWO-driven optimization. Nevertheless, the latency difference remains relatively small and computationally practical for real-world deployment scenarios. Simpler models such as XGBoost and CNN achieve lower latency values of 7.2 ms and 7.5 ms, respectively, mainly because of their reduced architectural complexity and smaller computational overhead. However, despite their slightly faster inference speed, these models demonstrated substantially weaker prediction accuracy and lower robustness in previous evaluations. Therefore, the proposed ODFT-DBN framework provides a highly favorable tradeoff between prediction performance and computational efficiency, making it suitable for practical PM_{2.5} forecasting applications requiring both high accuracy and near real-time inference capability.

Table 8. Comparison of inference latency between the ODFT-DBN framework and other models.

Proposed Methods	Inference Latency (ms)
ODFT-DBN	9.1
BERT	8.7
Transformer	8.3
DBN	7.8
LSTM	8.5
GRU	8.2
CNN	7.5
XGBoost	7.2

Table 9 presents the runtime analysis of the proposed ODFT-DBN framework and the comparative benchmark models under different predefined RMSE convergence thresholds. Unlike conventional runtime evaluations that simply report total execution time, this analysis measures the time required for each architecture to reach specific prediction-error levels, namely RMSE < 20, RMSE < 15, RMSE < 10, and RMSE < 5. Therefore, the table provides a more meaningful evaluation of optimization efficiency and convergence capability by directly linking runtime performance to prediction quality. Models that achieve lower RMSE thresholds within shorter execution times demonstrate stronger learning efficiency and more effective optimization behavior.

Table 9. Computational time required by the models to achieve predefined RMSE thresholds.

Proposed Methods	Run Time (s)			
	RMSE < 20	RMSE < 15	RMSE < 10	RMSE < 5
ODFT-DBN	23	91	166	324
BERT	55	305	518	-
Transformer	93	372	-	-
DBN	121	415	-	-
LSTM	173	-	-	-
GRU	201	-	-	-
CNN	183	-	-	-
XGBoost	-	-	-	-

The results indicate that the proposed ODFT-DBN framework achieves substantially faster convergence toward all predefined RMSE thresholds compared with the benchmark models. Specifically, the proposed architecture reaches RMSE < 20 in only 23 s, RMSE < 15 in 91 s, RMSE < 10 in 166 s, and RMSE < 5 in 324 s. In contrast, BERT requires 55 s, 305 s, and 518 s to reach RMSE thresholds of 20, 15, and 10, respectively, while failing to achieve RMSE < 5 during the training process. Similarly, the Transformer model requires 93 s and

372 s to achieve $RMSE < 20$ and $RMSE < 15$, respectively, but does not converge toward lower RMSE levels. The DBN, LSTM, GRU, CNN, and XGBoost models demonstrate even slower convergence behavior and are unable to achieve several of the lower RMSE thresholds within the training duration.

An important observation from Table 9 is that several comparative architectures completely fail to achieve lower RMSE thresholds despite prolonged training. For example, LSTM, GRU, and CNN are only able to reach $RMSE < 20$, while XGBoost fails to achieve even this initial convergence condition. This behavior highlights the difficulty of modeling highly nonlinear spatio-temporal $PM_{2.5}$ dynamics using conventional sequential or shallow-learning architectures. In contrast, the proposed ODFT-DBN framework consistently achieves all predefined RMSE targets with significantly lower runtime requirements. The superior convergence efficiency of the proposed framework can be attributed to the cooperative interaction among fuzzy uncertainty modeling, transformer-based temporal dependency extraction, DBN hierarchical feature learning, and NMOGWO-driven adaptive optimization. These mechanisms collectively accelerate feature learning and improve optimization stability throughout the training process.

When jointly analyzing the results of Tables 8 and 9, it can be concluded that the proposed ODFT-DBN framework achieves a highly favorable balance between computational complexity, inference efficiency, and prediction performance. Although the proposed architecture exhibits slightly higher inference latency than simpler models due to its hybrid multi-module structure, the latency remains within practical real-time operational limits. At the same time, the runtime analysis demonstrates that the proposed framework converges substantially faster toward high-accuracy prediction states compared with benchmark architectures. Therefore, despite the increased architectural sophistication, the computational complexity of the proposed framework remains practically manageable and suitable for real-world $PM_{2.5}$ monitoring applications, particularly in smart-city systems and large-scale environmental forecasting platforms requiring both high prediction accuracy and reliable near real-time operation.

Although the proposed ODFT-DBN framework utilizes a multi-objective optimization procedure based on NMOGWO, it should be noted that this optimization process is performed only during the offline model-development stage. The computational cost associated with population-based optimization is therefore incurred once during model construction and parameter tuning, whereas the final deployed model operates using only the optimized network parameters. Consequently, the optimization overhead does not affect the real-time forecasting stage, where predictions are generated using the trained ODFT-DBN model. The inference latency analysis presented in Table 8 further confirms that the final framework remains suitable for practical near real-time $PM_{2.5}$ forecasting applications. From the optimization perspective, the selected configuration of 40 wolves and 300 iterations was found to provide a suitable balance between prediction performance and computational cost. Increasing the population size or the maximum number of iterations beyond these values produced only marginal improvements in prediction accuracy while substantially increasing the optimization runtime. Conversely, smaller populations or fewer iterations reduced the computational burden but often resulted in premature convergence and less stable solutions. Therefore, the adopted configuration was selected as a practical compromise between optimization quality and computational efficiency.

Nevertheless, reducing the computational complexity of population-based hyper-parameter optimization remains an important research direction. Several strategies may be employed to further improve efficiency, including adaptive population-size control, early-stopping mechanisms, surrogate-assisted fitness evaluation, and parallel execution

of candidate-solution evaluations. Such approaches can significantly reduce the optimization time while maintaining comparable prediction performance. Therefore, future work will investigate computationally efficient optimization strategies that preserve the forecasting accuracy of the proposed framework while further improving scalability for large-scale environmental prediction applications.

Table 10 presents the variance analysis of the proposed ODFT-DBN framework and the comparative benchmark models across 15 independent simulation runs. Variance analysis is important for evaluating the stability, robustness, and consistency of prediction models under repeated training and stochastic optimization conditions. The results clearly demonstrate that the proposed ODFT-DBN framework achieves the lowest variance among all compared architectures, with a variance value of only 0.000417. In contrast, the benchmark models exhibit substantially larger variance values, including 3.140256 for BERT, 5.365280 for Transformer, 7.014589 for DBN, and progressively higher values for recurrent and conventional architectures such as LSTM, GRU, CNN, and XGBoost. These results indicate that the proposed framework maintains highly consistent prediction behavior across repeated executions, while the benchmark models suffer from greater instability and sensitivity to stochastic training conditions. The significantly larger variance values observed in CNN and XGBoost further demonstrate the limitations of conventional architectures in handling highly nonlinear and uncertain PM_{2.5} forecasting tasks.

The superior stability of the proposed ODFT-DBN framework can be attributed to the cooperative integration of fuzzy uncertainty modeling, transformer-based contextual learning, DBN hierarchical abstraction, and NMOGWO-driven adaptive optimization. The fuzzy subsystem reduces uncertainty propagation during learning, while the transformer encoder stabilizes temporal dependency extraction through adaptive attention learning. In addition, the DBN module improves hierarchical feature consistency, and the proposed NMOGWO optimization strategy prevents unstable parameter updates and premature convergence during training. These mechanisms collectively contribute to smoother optimization behavior and more reliable prediction performance across multiple independent runs.

From a practical perspective, low variance is critically important for real-world environmental forecasting systems because operational PM_{2.5} monitoring platforms require stable and reproducible prediction behavior under changing atmospheric and computational conditions. Models with high variance may produce inconsistent forecasts across repeated deployments, reducing their reliability for environmental management and decision-making applications. In contrast, the extremely low variance achieved by the proposed ODFT-DBN framework indicates strong generalization capability and high robustness against stochastic variations, making the framework highly suitable for practical smart-city deployment, long-term air-quality monitoring, and real-time environmental forecasting applications.

Table 10. Variance comparison of the proposed models over 15 independent experimental runs.

Proposed Methods	Variance
ODFT-DBN	0.000417
BERT	3.140256
Transformer	5.365280
DBN	7.014589
LSTM	8.145236
GRU	9.102548
CNN	10.902143
XGBoost	13.554012

Table 11 presents the statistical paired *t*-test results comparing the proposed ODFT-DBN framework with the benchmark ML/DL architectures. The purpose of this analysis is to determine whether the prediction improvements achieved by the proposed framework are statistically significant rather than resulting from random variations during training or testing procedures. In this study, all comparisons were conducted using a significance level of $\alpha = 0.01$. The reported *p*-values quantify the probability that the observed performance differences occurred randomly. Lower *p*-values indicate stronger statistical confidence regarding the superiority of the proposed framework.

As shown in Table 11, all pairwise comparisons between the proposed ODFT-DBN architecture and the benchmark models produce extremely small *p*-values, including 0.0005 for BERT, 0.0003 for Transformer, 0.00009 for DBN, 0.00006 for LSTM, 0.00005 for GRU, 0.00003 for CNN, and 0.000001 for XGBoost. Since all *p*-values are substantially lower than the predefined significance threshold of $\alpha = 0.01$, the null hypothesis can be rejected for every comparison case. Consequently, the prediction improvements achieved by the proposed framework are statistically significant with high confidence. These results confirm that the superior RMSE, R^2 , MAPE, convergence, and stability performances observed in previous analyses are not caused by random optimization behavior or stochastic training fluctuations.

Table 11. Statistical paired *t*-test analysis of the proposed models.

Algorithm	<i>t</i> -Tests		
	<i>p</i> -Value	Significance Level	Results
ODFT-DBN vs. BERT	0.0005	$\alpha = 0.01$	Significant
ODFT-DBN vs. Transformer	0.0003	$\alpha = 0.01$	Significant
ODFT-DBN vs. DBN	0.00009	$\alpha = 0.01$	Significant
ODFT-DBN vs. LSTM	0.00006	$\alpha = 0.01$	Significant
ODFT-DBN vs. GRU	0.00005	$\alpha = 0.01$	Significant
ODFT-DBN vs. CNN	0.00003	$\alpha = 0.01$	Significant
ODFT-DBN vs. XGBoost	0.000001	$\alpha = 0.01$	Significant

Another important observation is that the smallest *p*-values are obtained when comparing the proposed framework with conventional architectures such as CNN and XGBoost, indicating very large statistical performance differences between these methods and the proposed ODFT-DBN architecture. Although attention-based architectures such as BERT and Transformer demonstrate comparatively stronger forecasting capability, their *p*-values still remain significantly below the $\alpha = 0.01$ threshold, confirming that the proposed framework consistently outperforms even advanced transformer-based benchmark models.

Table 12 summarizes the representative optimized centers and widths of the Gaussian membership functions for six influential environmental variables. For each variable, the Low, Medium, and High fuzzy regions are reported in the normalized input domain together with their corresponding environmental interpretation. Because Gaussian membership functions overlap smoothly, the reported intervals should be interpreted as approximate dominant regions rather than fixed crisp thresholds. The optimized AOD membership functions exhibit a clear separation among low, transitional, and high aerosol-loading conditions. The High AOD center is shifted toward the upper part of the normalized domain, indicating that the model assigns the strongest high-pollution membership only when aerosol loading becomes distinctly elevated. This behavior is environmentally meaningful because AOD provides an indirect measure of the total atmospheric aerosol burden and is therefore closely related to the likelihood of increased PM_{2.5} concentrations. The moderate overlap between the AOD membership functions also allows the fuzzy

module to represent transitional atmospheric conditions in which the relationship between column-integrated aerosol loading and ground-level PM_{2.5} is affected by humidity, boundary-layer height, and vertical aerosol distribution.

Table 12. Representative optimized Gaussian function parameters and physical interpretation.

Variable	Optimized Centers: Low Medium High	Optimized Widths: Low Medium High	Approximate Normalized Intervals: Low Medium High	Environmental Interpretation
AOD	0.18	0.12	0–0.34	Low aerosol loading
	0.49	0.15	0.34–0.66	Transitional aerosol conditions
	0.82	0.13	0.66–1.00	High atmospheric aerosol burden associated with elevated PM _{2.5}
Maximum temperature	0.20	0.13	0–0.36	Cooler and relatively stable conditions
	0.52	0.16	0.36–0.68	Moderate thermal conditions
	0.83	0.14	0.68–1.00	High-temperature conditions affecting atmospheric mixing and aerosol formation
LST	0.17	0.12	0–0.33	Weak surface heating
	0.48	0.15	0.33–0.64	Moderate land-surface heating
	0.80	0.13	0.64–1.00	Intense surface heating influencing boundary-layer development and dry-surface particle resuspension
Wind speed	0.14	0.10	0–0.28	Atmospheric stagnation and pollutant accumulation
	0.42	0.14	0.28–0.59	Partial ventilation and dispersion
	0.76	0.15	0.59–1.00	Strong air circulation and pollutant dilution
NDVI	0.22	0.14	0–0.36	Sparse vegetation and limited particle capture
	0.50	0.15	0.36–0.64	Moderate vegetation coverage
	0.78	0.14	0.64–1.00	Dense vegetation associated with enhanced deposition and reduced dust resuspension
Elevation	0.16	0.11	0–0.31	Low-lying regions susceptible to pollutant accumulation
	0.45	0.15	0.31–0.62	Intermediate terrain conditions
	0.79	0.14	0.62–1.00	Elevated regions generally associated with improved ventilation and lower local pollutant retention

Maximum temperature and LST show similar but not identical fuzzy patterns. The optimized maximum-temperature functions distinguish cooler stable conditions from warmer conditions that alter atmospheric mixing, pollutant transport, and secondary aerosol formation. In comparison, the LST membership functions specifically represent the intensity of surface heating. High LST may strengthen boundary-layer development and vertical mixing, but it may also indicate dry surface conditions that promote dust and particle resuspension. The separate optimization of maximum temperature and LST therefore enables the fuzzy system to distinguish between atmospheric thermal conditions and

land-surface energy effects rather than treating them as equivalent temperature indicators.

The optimized wind-speed functions provide the clearest physical separation between accumulation and dispersion regimes. Low wind speed is represented by a relatively narrow membership function centered near the lower end of the normalized range, reflecting atmospheric stagnation conditions under which $\text{PM}_{2.5}$ can accumulate. The wider High wind-speed function captures a broader range of ventilation conditions associated with pollutant transport and dilution. The NDVI and elevation functions additionally describe the spatial environmental context of the prediction problem. Higher NDVI generally indicates greater vegetation coverage, which can enhance particle deposition and reduce surface-dust resuspension, whereas higher elevation is commonly associated with improved ventilation and reduced pollutant retention compared with low-lying urban areas. Collectively, these optimized fuzzy intervals demonstrate that the proposed model learns environmentally interpretable transitions among aerosol, meteorological, land-surface, vegetation, and topographical conditions.

Figure 13 presents the averaged self-attention weights learned by the proposed ODFT-DBN model over the four-day historical input sequence. The heatmap shows how each query day distributes its attention across the available historical observations, while the bar chart summarizes the mean attention assigned to each lag after averaging across the test samples and attention heads of the final transformer encoder layer. The heatmap reveals a clear concentration of attention along the main diagonal, indicating that each temporal position retains a strong dependence on its corresponding historical observation. The diagonal weights are 0.38, 0.36, 0.39, and 0.53 for (t-4), (t-3), (t-2), and (t-1), respectively. The highest individual attention weight is observed at (t-1), where the most recent historical observation receives a weight of 0.53. In the same query row, the model assigns weights of 0.28, 0.13, and 0.06 to (t-2), (t-3), and (t-4), respectively. Therefore, 81% of the attention associated with the final historical position is concentrated on the two most recent days, demonstrating a strong short-term temporal persistence in $\text{PM}_{2.5}$ dynamics.

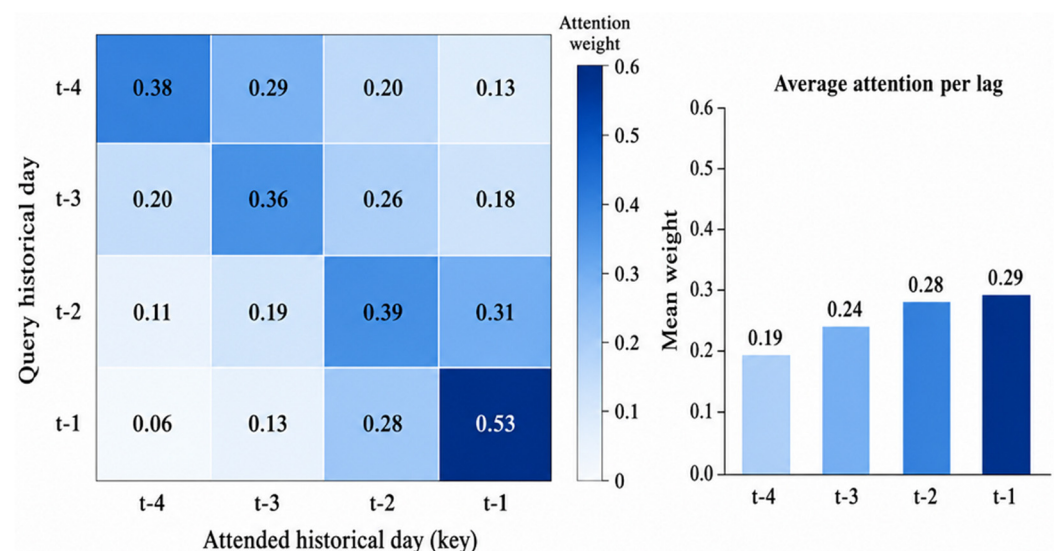


Figure 13. Averaged self-attention weights of ODFT-DBN over four-day historical input sequence.

A similar recency-oriented pattern is observed for the preceding query positions. For the (t-2) query, the model assigns weights of 0.39 and 0.31 to (t-2) and (t-1), respectively, meaning that 70% of its attention is concentrated on these two neighboring positions. For (t-3), the largest weight is assigned to (t-3) at 0.36, followed by (t-2) at 0.26. The earliest

query position, (t-4), similarly places the highest weights on (t-4) and (t-3), with values of 0.38 and 0.29. These results indicate that the transformer does not treat all historical observations equally; instead, it emphasizes temporally adjacent information while gradually reducing the contribution of more distant observations. The mean attention weights shown in the bar chart further confirm this temporal pattern. The average importance values increase from 0.19 for (t-4) and 0.24 for (t-3) to 0.28 for (t-2) and 0.29 for (t-1). Accordingly, the two most recent days jointly account for 57% of the overall attention, whereas the third and fourth preceding days account for the remaining 43%. This distribution indicates that one-day-ahead PM_{2.5} prediction is primarily governed by short-term atmospheric persistence and recent environmental conditions, while older observations still provide complementary contextual information. Because the adopted input window contains only four historical days, the results should be interpreted as evidence of short-term temporal dependency rather than weekly, monthly, or seasonal periodicity.

Figure 14 presents the normalized feature-importance ranking of the 16 input variables used by the proposed ODFT-DBN framework. The contribution ratios sum to 100% and quantify the relative influence of historical pollution observations, satellite-derived variables, meteorological conditions, temporal information, and topographical descriptors on PM_{2.5} prediction. Historical PM_{2.5} is identified as the most influential individual predictor, with a contribution ratio of 24%. This result confirms the strong short-term persistence of PM_{2.5} concentrations, as the pollution level of the following day is substantially affected by the conditions observed during the preceding days. AOD ranks second with a contribution of 19%, demonstrating the importance of satellite-derived aerosol information for representing the atmospheric particle burden. Together, historical PM_{2.5} and AOD account for 43% of the total model contribution, indicating that both ground-level pollution persistence and column-integrated aerosol loading play dominant and complementary roles in the prediction process.

Among the land-surface variables, LST and NDVI contribute 8% and 7%, respectively. The relatively high importance of LST reflects the influence of surface heating on boundary-layer development, atmospheric stability, vertical mixing, and particle resuspension. NDVI provides an additional contribution by representing vegetation coverage and land-surface characteristics that may affect particle deposition, dust generation, and local pollutant dispersion. Collectively, LST and NDVI account for 15% of the total importance, confirming that satellite-derived surface conditions provide information that cannot be fully represented by AOD or ground-based meteorological measurements alone. The meteorological variables collectively contribute 30% of the total feature importance. Wind speed and the day-independent temperature variables show notable contributions, with wind speed accounting for 6%, while maximum and minimum temperatures each contribute 5%. Relative humidity contributes 4%, followed by wind direction and precipitation at 3% each, and water vapor and atmospheric pressure at 2% each. Wind speed is the most influential meteorological variable because it directly controls pollutant transport, dilution, and accumulation. Temperature and humidity additionally influence atmospheric stability, secondary aerosol formation, and hygroscopic particle growth, whereas precipitation contributes through wet scavenging and the removal of suspended particles.

The day of the week contributes 6%, indicating that recurrent human-activity patterns, including traffic intensity, industrial operations, and weekday–weekend variations, provide useful contextual information for PM_{2.5} forecasting. Elevation contributes 4%, while latitude and longitude each contribute 1%. Therefore, the combined contribution of the topographical and spatial variables is 6%. Although their individual effects are lower than those of historical PM_{2.5}, AOD, and meteorological variables, they remain relevant

because they characterize persistent spatial differences in terrain, ventilation, urban structure, and pollutant retention across the study area. Overall, the ranking demonstrates that the proposed model does not depend on a single source of environmental information. The principal contributions arise from historical PM_{2.5} at 24%, meteorological variables at 30%, AOD at 19%, land-surface variables at 15%, temporal context at 6%, and topographical and spatial variables at 6%. This distribution confirms that the improved prediction performance of the proposed framework results from integrating complementary pollution, atmospheric, surface, temporal, and geographical information within a unified learning architecture.

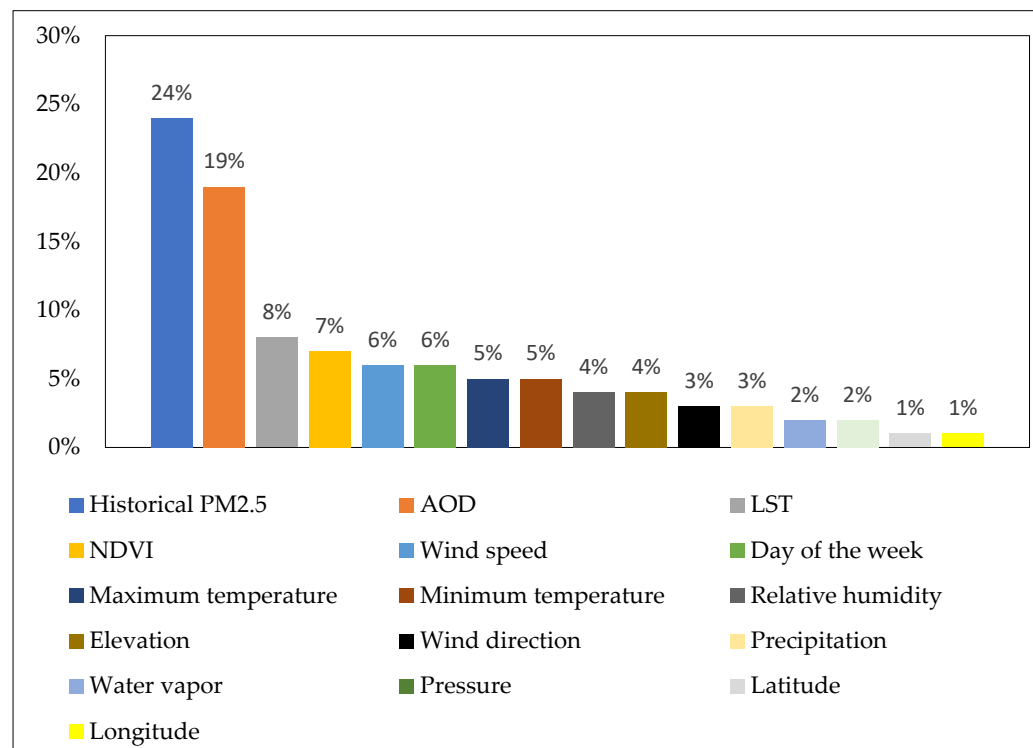


Figure 14. Normalized contribution ratios of the input variables obtained from the feature-importance analysis of the proposed ODFT-DBN model.

5. Conclusions

PM_{2.5} air pollution has become one of the most critical environmental and public-health challenges in modern urban areas due to its severe impacts on human health, atmospheric quality, and environmental sustainability. High concentrations of PM_{2.5} are strongly associated with respiratory diseases, cardiovascular disorders, reduced visibility, and increased mortality rates, particularly in densely populated metropolitan regions. Therefore, developing accurate and robust PM_{2.5} forecasting models is essential for improving environmental monitoring systems, supporting urban pollution management, and enabling timely decision-making for public-health protection. In this paper, an ODFT-DBN framework was proposed for accurate PM_{2.5} prediction using heterogeneous spatio-temporal environmental data.

The proposed study was conducted for Tehran, Iran, which is considered one of the most polluted metropolitan regions due to severe traffic congestion, industrial emissions, rapid urbanization, and unfavorable geographical conditions that intensify pollutant accumulation. The utilized dataset integrated multiple environmental and remote-sensing variables, including ground-based PM_{2.5} observations, meteorological measurements,

AOD, LST, NDVI, elevation, precipitation, humidity, water vapor, and spatial geographical information. To improve data quality and consistency, several preprocessing procedures including noise reduction, missing-data reconstruction, normalization, and spatial interpolation were applied prior to model training. The proposed ODFT-DBN framework combined fuzzy uncertainty modeling, transformer-based spatio-temporal attention learning, DBN hierarchical feature extraction, and a novel NMOGWO optimization strategy to jointly optimize the learning architecture and fuzzy-system parameters for robust PM_{2.5} forecasting.

The experimental results demonstrated that the proposed ODFT-DBN framework significantly outperformed all comparative ML and DL architectures across multiple quantitative and qualitative evaluation criteria. The proposed model achieved the best overall prediction performance with an RMSE value of 0.80 µg/m³, an R^2 value of 0.94, and a MAPE of 0.35%, indicating highly accurate and reliable PM_{2.5} forecasting capability. Scatter-plot analysis additionally confirmed the strong agreement between measured and predicted pollutant concentrations, while spatial-distribution analysis demonstrated that the proposed framework could accurately reconstruct realistic pollution hotspots and spatial pollutant gradients across Tehran under complex winter atmospheric conditions. Furthermore, the ablation-study results verified that the cooperative integration of fuzzy reasoning, transformer-based attention learning, DBN hierarchical abstraction, and NMOGWO-driven optimization collectively contributed to the superior forecasting capability of the proposed framework.

Additional analyses further validated the robustness, stability, and practical applicability of the proposed architecture. The convergence analysis showed that the proposed ODFT-DBN framework achieved substantially faster and smoother convergence behavior compared with benchmark models, reaching lower RMSE levels within significantly fewer training epochs. Runtime analysis demonstrated that the proposed framework reached RMSE < 20, RMSE < 15, RMSE < 10, and RMSE < 5 within 23 s, 91 s, 166 s, and 324 s, respectively, outperforming all comparative architectures in optimization efficiency. Moreover, despite the sophisticated hybrid architecture, the proposed framework maintained a practical inference latency of only 9.1 ms, confirming its suitability for near real-time environmental monitoring applications. Variance analysis further demonstrated highly stable prediction behavior, where the proposed model achieved an extremely low variance value of 0.000417 across 15 independent runs, substantially lower than all benchmark models. Finally, paired *t*-test analysis confirmed that the performance improvements achieved by the proposed framework were statistically significant under $\alpha = 0.01$, validating the reliability and superiority of the proposed ODFT-DBN framework for practical PM_{2.5} forecasting applications in complex urban environments.

The findings of this study demonstrate that accurate PM_{2.5} forecasting in complex urban environments cannot be effectively achieved using standalone DL architectures alone. Air-pollution dynamics are inherently nonlinear, uncertain, heterogeneous, and strongly dependent on long-range spatial and temporal atmospheric interactions. Consequently, combining complementary intelligent learning mechanisms within a unified cooperative framework can substantially improve environmental prediction capability. The proposed ODFT-DBN architecture showed that integrating fuzzy uncertainty modeling with transformer-based attention learning and DBN hierarchical feature extraction enables more reliable representation of atmospheric behavior under varying meteorological and pollution conditions. In addition, the proposed NMOGWO optimization strategy demonstrated that adaptive multi-objective optimization can significantly improve convergence efficiency, training stability, and generalization capability when optimizing complex hybrid environmental forecasting systems. Therefore, one important conclusion

of this work is that future smart-environment prediction systems should increasingly rely on cooperative hybrid DL frameworks rather than isolated single-module architectures.

Another important conclusion is that the proposed framework demonstrates strong practical potential for real-world environmental-monitoring and smart-city applications. The obtained results confirmed that the proposed architecture not only achieves high prediction accuracy, but also maintains low variance, stable convergence behavior, statistically significant performance improvements, and practically acceptable runtime and inference latency characteristics. These properties are critically important for operational environmental systems that require reliable and near real-time pollution forecasting under continuously changing atmospheric conditions. Furthermore, the strong spatial reconstruction capability of the proposed framework indicates that hybrid attention-based environmental models can effectively support urban pollution management, early-warning systems, traffic-emission control strategies, and public-health decision-making processes. Overall, this study suggests that combining uncertainty-aware fuzzy reasoning, attention-driven temporal learning, hierarchical deep representation learning, and adaptive evolutionary optimization provides a highly promising direction for developing next-generation intelligent environmental forecasting systems capable of handling large-scale, heterogeneous, and dynamically evolving real-world data.

Future research directions can further enhance both the methodological and environmental aspects of the proposed framework. From the modeling perspective, future studies may investigate lightweight transformer architectures, adaptive online learning strategies, federated learning frameworks, and graph-based spatio-temporal networks to further improve scalability and real-time deployment capability in large smart-city systems. In addition, integrating explainable artificial intelligence (AI) mechanisms could improve the interpretability of the forecasting process and provide more transparent environmental decision support for policymakers. From the environmental perspective, future work may extend the proposed framework toward multi-pollutant forecasting scenarios involving PM₁₀, NO₂, SO₂, O₃, and CO simultaneously, while also incorporating additional urban factors such as traffic density, industrial activity, population mobility, and IoT sensor networks. Furthermore, the proposed ODFT-DBN architecture can potentially be adapted to other complex real-world prediction applications beyond air pollution forecasting, including climate analysis, smart-energy management, hydrological prediction, intelligent transportation systems, healthcare monitoring, and large-scale environmental risk assessment problems involving highly nonlinear and uncertain spatio-temporal data.

Author Contributions: Conceptualization, M.M.S.N., P.S.M., M.K., D.M., N.S. and J.V.Á.-B.; methodology, M.M.S.N., P.S.M., M.K. and D.M.; software, M.M.S.N., M.K. and N.S.; validation, M.M.S.N., P.S.M. and D.M.; formal analysis, J.V.Á.-B.; investigation, M.M.S.N., P.S.M., M.K., D.M., N.S. and J.V.Á.-B.; resources, D.M.; data curation, P.S.M., M.K., N.S. and J.V.Á.-B.; writing—original draft preparation, M.M.S.N., P.S.M. and M.K.; writing—review and editing, M.M.S.N., D.M. and N.S.; visualization, M.K. and J.V.Á.-B.; supervision, D.M.; project administration, D.M.; funding acquisition, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study.

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

References

- Patel, S.; Shah, M.; Patel, K.; Prajapati, M. A General Review on the Applications of Machine Learning to PM_{2.5} Air Pollution Forecasting. *Mach. Learn. Comput. Sci. Eng.* **2025**, *1*, 33.
- 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.
- 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.
- 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.
- Mirzaei, A.; Bagheri, H.; Sattari, M. Data Level and Decision Level Fusion of Satellite Multi-Sensor AOD Retrievals for Improving PM_{2.5} Estimations, a Study on Tehran. *Earth Sci. Inform.* **2023**, *16*, 753–771.
- Kermani, M.; Norzaee, S.; Shahsavani, A.; Ghorbanian, A. Assessing Relative Population Exposure and Non-Carcinogenic Health Risks of PM_{2.5} in Tehran Through Machine Learning, Satellite Observations, and Uncertainty Analysis. *J. Hazard. Mater. Adv.* **2026**, *21*, 101059.
- 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.
- Shakerdonyavi, A.; Yeganeh, B. High-Resolution PM_{2.5} Mapping and Evaluation in Tehran Using Mobile Monitoring, Land-Use Regression, and Explainable Hybrid Models. *Urban Clim.* **2026**, *65*, 102773.
- Anbari, K.; Omid Khaniabadi, Y.; Abolghasem Gorji, H.; Sicard, P.; Abdollahi, A.; Baharvand, P.; Rashidi, R. Tracking Atmospheric PM_{2.5} and O₃ in Tehran Metropolitan and Related Burden of Mortality. *Hum. Ecol. Risk Assess. Int. J.* **2026**, *32*, 392–406.
- Pak, U.; Ma, J.; Ryu, U.; Ryom, K.; Juhyok, U.; Pak, K.; Pak, C. Deep Learning-Based PM_{2.5} Prediction Considering the Spatio-temporal Correlations: A Case Study of Beijing, China. *Sci. Total Environ.* **2020**, *699*, 133561.
- 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.
- Bhatta, J.; Acharya, S.R.; Yang, K.M. Machine Learning-Enhanced Air Quality Forecasting and Trend Analysis: A Five-Year Comprehensive Assessment of PM_{2.5} Concentrations in Bangkok, Thailand. *Environ. Chall.* **2026**, *22*, 101442.
- Jia, H.; Yao, S.; Tang, X.; Zheng, M.; Dong, Y.; Fan, F.; Yang, S.; Zhang, H. Multi-Objective Machine Learning for Health-Oriented O₃ and PM_{2.5} Control: Integrating VOC Photochemical Consumption and Source Apportionment. *J. Hazard. Mater.* **2026**, *505*, 141483.
- Zhang, G.; Wang, S.; Xu, J.; McVicar, T.R.; Cheng, Y.; Azorin-Molina, C.; Minola, L.; Wang, N.; Zhang, J.; Shi, P. A Deep Learning Approach Predicts O₃ Increase and PM_{2.5} Declines under High Emission Scenario across the Northern China Urban Agglomeration. *Urban Clim.* **2026**, *65*, 102752.
- 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.
- Yohanna, E.; San, L.H. Comparative Analysis of Machine Learning Models for Predicting PM₁₀ and PM_{2.5} Concentrations in Pulau Pinang, Malaysia. *Air Qual. Atmos. Health* **2026**, *19*, 21.
- Patel, P.; Patel, S.; Shah, K.; Desai, K.; Patel, S.; Shah, M.; Patel, S. A Systematic Study on PM_{2.5} and PM₁₀ Concentration Prediction in Air Pollution Using Machine Learning and Deep Learning Model. *Environ. Chem. Ecotoxicol.* **2025**, *7*, 1401–1415.
- Suthar, G.; Singh, S. Application of Machine Learning Models for PM_{2.5} Prediction in Bengaluru Using Precursor Air Pollutants and Meteorological Data. *Theor. Appl. Climatol.* **2025**, *156*, 181.
- Hou, Y.; Wang, Q.; Tan, T. Evaluating Drivers of PM_{2.5} Air Pollution at Urban Scales Using Interpretable Machine Learning. *Waste Manag.* **2025**, *192*, 114–124.
- Zandi, I.; Jafari, A.; Lotfata, A. Enhancing PM_{2.5} Air Pollution Prediction Performance by Optimizing the Echo State Network Deep Learning Model Using New Metaheuristic Algorithms. *Urban Sci.* **2025**, *9*, 138.
- Zhang, Z.; Zhang, S. Modeling Air Quality PM_{2.5} Forecasting Using Deep Sparse Attention-Based Transformer Networks. *Int. J. Environ. Sci. Technol.* **2023**, *20*, 13535–13550.
- Yu, M.; Masrur, A.; Blaszczyk-Boxe, C. Predicting Hourly PM_{2.5} Concentrations in Wildfire-Prone Areas Using a SpatioTemporal Transformer Model. *Sci. Total Environ.* **2023**, *860*, 160446.
- Zeng, Q.; Wang, L.; Zhu, S.; Gao, Y.; Qiu, X.; Chen, L. Long-Term PM_{2.5} Concentrations Forecasting Using CEEMDAN and Deep Transformer Neural Network. *Atmos. Pollut. Res.* **2023**, *14*, 101839.

24. 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.
25. Zou, R.; Huang, H.; Lu, X.; Zeng, F.; Ren, C.; Wang, W.; Zhou, L.; Dai, X. PD-LL-Transformer: An Hourly PM_{2.5} Forecasting Method over the Yangtze River Delta Urban Agglomeration, China. *Remote Sens.* **2024**, *16*, 1915.
26. Al-Qaness, M.A.; Dahou, A.; Ewees, A.A.; Abualigah, L.; Huai, J.; Abd Elaziz, M.; Helmi, A.M. ResInformer: Residual Transformer-Based Artificial Time-Series Forecasting Model for PM_{2.5} Concentration in Three Major Chinese Cities. *Mathematics* **2023**, *11*, 476.
27. Huang, Y.; Han, F.; Feng, Q. A Novel Model for Predicting PM_{2.5} Concentrations Utilizing Graph Convolutional Networks and Transformer. *IEEE Access* **2025**, *13*, 30613–30622.
28. Shen, H.; Li, T.; Yuan, Q.; Zhang, L. Estimating Regional Ground-Level PM_{2.5} Directly from Satellite Top-of-Atmosphere Reflectance Using Deep Belief Networks. *J. Geophys. Res. Atmos.* **2018**, *123*, 13875–13886.
29. Xing, Y.; Yue, J.; Chen, C.; Xiang, Y.; Chen, Y.; Shi, M. A Deep Belief Network Combined with Modified Grey Wolf Optimization Algorithm for PM_{2.5} Concentration Prediction. *Appl. Sci.* **2019**, *9*, 3765.
30. Rana, R.; Kumar, N. Smart Air: A Spatiotemporal Attention Based Deep Learning Approach for Accurate PM_{2.5} and PM₁₀ Forecasting. *Earth Syst. Environ.* **2026**, *10*, 101–116.
31. Liu, J.; Zheng, Z.; Yao, F.; Li, W. A Multi-View Machine Learning Approach for Estimating PM_{2.5} Concentrations from Smartphone Photographs. *J. Hazard. Mater.* **2026**, *511*, 142172.
32. Singh, S.; Suthar, G. Machine Learning and Deep Learning Approaches for PM_{2.5} Prediction: A Study on Urban Air Quality in Jaipur, India. *Earth Sci. Inform.* **2025**, *18*, 97.
33. Sohrabi, Z.; Maleki, J. Fusing Satellite Imagery and Ground-Based Observations for PM_{2.5} Air Pollution Modeling in Iran Using a Deep Learning Approach. *Sci. Rep.* **2025**, *15*, 21449.
34. Wei, Q.; Zhang, H.; Yang, J.; Niu, B.; Xu, Z. PM_{2.5} Concentration Prediction Using a Whale Optimization Algorithm Based Hybrid Deep Learning Model in Beijing, China. *Environ. Pollut.* **2025**, *371*, 125953.
35. Muthukumar, P.; Nagrecha, K.; Comer, D.; Calvert, C.F.; Amini, N.; Holm, J.; Pourhomayoun, M. PM_{2.5} Air Pollution Prediction through Deep Learning Using Multisource Meteorological, Wildfire, and Heat Data. *Atmosphere* **2022**, *13*, 822.
36. 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.
37. 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.
38. Shakya, D.; Deshpande, V.; Goyal, M.K.; Agarwal, M. PM_{2.5} Air Pollution Prediction through Deep Learning Using Meteorological, Vehicular, and Emission Data: A Case Study of New Delhi, India. *J. Clean. Prod.* **2023**, *427*, 139278.
39. Amiri, Z.; Shahne, M.Z. Modeling PM_{2.5} Concentration in Tehran Using Satellite-Based Aerosol Optical Depth (AOD) and Machine Learning: Assessing Input Contributions and Prediction Accuracy. *Remote Sens. Appl. Soc. Environ.* **2025**, *38*, 101549.
40. Niu, M.; Wang, Y.; Sun, S.; Li, Y. A Novel Hybrid Decomposition-and-Ensemble Model Based on CEEMD and GWO for Short-Term PM_{2.5} Concentration Forecasting. *Atmos. Environ.* **2016**, *134*, 168–180.
41. Zhang, C.; Wang, S.; Wu, Y.; Zhu, X.; Shen, W. A Long-Term Prediction Method for PM_{2.5} Concentration Based on Spatiotemporal Graph Attention Recurrent Neural Network and Grey Wolf Optimization Algorithm. *J. Environ. Chem. Eng.* **2024**, *12*, 111716.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.