Atmospheric Pollution Research 16 (2025) 102654

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: www.elsevier.com/locate/apr

Uncertainty assessment based on data decomposition and Boruta-driven extreme gradient boosting to predict spatiotemporal urban air dust heavy metal index

Akram Seifi[a], Somayeh Soltani-Gerdefaramarzi[b], Mumtaz Ali[c,d,*]

[a] Department of Water Science & Engineering, Vali-e-Asr University of Rafsanjan, Iran
[b] Department of Water Sciences and Engineering, Collage of Agriculture and Natural Resource, Ardakan University, Ardakan, 89518- 95491, Iran
[c] UniSQ College, University of Southern Queensland, 4350, QLD, Australia
[d] Scientific Research Centre, Al-Ayen University, Nasiriyah 64001, Thi-Qar, Iraq

ARTICLE INFO

Keywords:
Air pollution
Boruta feature selection
Readily available data
Uncertainty
Extreme gradient boosting

ABSTRACT

Accurate prediction of urban air dust pollutants is essential for public health and environmental management. Achieving reliable predictions of the air pollution due to heavy metals existence in these areas is extremely important. This study for the first time develop an ensemble approach based on multivariate variational model decomposition (MVMD) and extreme gradient boosting (XGBoost) integrated with Bayesian optimizer of Optuna and different feature selection techniques to predict the spatiotemporal distribution of pollution load index (PLI) in Yazd urban area, Iran. For comparison, gated recurrent unit (GRU) network, adaptives neuro-fuzzy-inference system (ANFIS), and multilayer perceptron (MLP) models were are develpoed. Variables including meteorological data, heavy metals concentration of roof dust, and distance to pollution sources were gathered. The seasonal data of variables were analyzed using Boruta feature selection approach (BFSA), SHapley additive explanations (SHAP), and Wavelet methods to identify valuable and easily accessible variables to predict PLI index. The results confirmed that the BFSA has high capability for selecting the most important features over SHAP, and wavelet techniques, that provides cost-effective input vector of Max WD, Min RH, Cd, and Zn with readily available variables. Morover, the XGBoost model shows high prediction accuracy for PLI in terms of $R^2 = 0.90$, RMSE = 0.08, and MAE = 0.06. Furthermore, by stationarity test of multivariate variational mode decomposition (MVMD) method applied to all input variables, the Max WD and Min RH were decompossed into three intrinsic mode functions (IMFs). These IMFs along with Cd and Zn were used as input vector in the XGBoost to create the final model for predicting temporal uncertainty and generate seasonal urban spatiotemporal maps. The evaluation of uncertainties demonstrated that the MVMD-XGBoost effectively captured 83.33 %, 96.67 %, 63.33 %, and 68.97 % of observed data within the 95 % confidence interval in spring, summer, autumn, and winter seasons, respectively. Findings from this study allow decision-makers to reduce air pollution monitoring costs and enhance control measures by leveraging readily available variables.

1. Introduction

Air pollution poses a significant global threat to environmental and human health across various regions. Long-term exposure, particularly to heavy metals, is linked to severe health issues such as lung diseases, heart disease, lung tumors, and strokes (Wang et al., 2024). Beyond health, polluted air negatively impacts human performance, increases healthcare costs, and hinders economic activity (Tao et al., 2023). Highly toxic pollutants like As, Pb, and Cr present severe health risks (McCartor and Becker, 2010), while elevated concentrations of Zn and Cu contribute significantly to the toxicity of airborne particulate matter (PM) in urban areas (Khanal et al., 2015). The deposition of heavy metals in soil and water also leads to toxic consequences for land ecosystems, affecting plant growth and reducing productivity (Chen et al., 2016). Despite considerable economic losses, global air pollution levels are steadily rising, especially in developing regions. Therefore, robust pollution monitoring and predictive modeling are essential to mitigate these health risks, design exposure minimization strategies, and enable

* Corresponding author. UniSQ College, University of Southern Queensland, 4350, QLD, Australia.
E-mail addresses: a.seifi@vru.ac.ir (A. Seifi), ssoltani@ardakan.ac.ir (S. Soltani-Gerdefaramarzi), mumtaz.ali@unisq.edu.au (M. Ali).

https://doi.org/10.1016/j.apr.2025.102654
Received 23 January 2025; Received in revised form 14 July 2025; Accepted 14 July 2025
Available online 16 July 2025

timely public health decisions (Tao et al., 2024).

Predicting air pollutants is challenging due to the complicated processes involved in their generation and the numerous influencing factors, including weather conditions, emission levels, and urban structural features (Tian et al., 2014; Li et al., 2019a,b; Wu et al., 2021). Conventional approaches for detecting airborne heavy metals, typically involving laboratory-based chemical analyses, are accurate but demand substantial time and financial resources. The high costs and technical demands often render these analyses impractical, particularly in resource-limited regions. Consequently, there is growing interest in developing cost-effective methods to accurately predict heavy metal concentrations using readily available parameters. For instance, meteorological data has shown great potential in estimating PM pollution levels, reducing dependence on costly laboratory testing, and permitting more frequent air quality assessments (Olawoyin et al., 2018; Leng et al., 2018). Studies indicate that meteorological parameters, such as rainfall (De Nevers, 2010), temperature, and wind (Yang et al., 2020; Li et al., 2019a,b), significantly influence air pollution dispersion and removal. High humidity is also positively associated with ambient PM2.5 levels (Tao et al., 2023). Furthermore, factors related to traffic, regional geography, and time are crucial variables in air pollution research (Yang et al., 2024; He et al., 2022). Therefore, accurate air pollution forecasting necessitates careful selection of key input variables, as irrelevant or redundant data can compromise model accuracy and computational efficiency (Ebrahimi-Khusfi et al., 2021). To address this, this study employs Boruta Feature Selection Approach (BFSA), Shapley Additive Explanations (SHAP), and Wavelet techniques, along with a collinearity test, to prioritize variables with the greatest impact on pollution prediction. By focusing on easily accessible, measurable, and low-cost variables, these methods enhance model efficiency and interpretability. Additionally, to overcome issues of non-stationarity and non-linearity often present in environmental datasets, multivariate variational mode decomposition (MVMD) is applied in this study. MVMD effectively decomposes complex datasets into simpler intrinsic mode functions (IMFs), providing a robust framework for handling intricate signal characteristics (Seifi et al., 2024). This preprocessing step ensures that machine learning models can more effectively capture underlying patterns and make reliable predictions in dynamic environmental systems.

Recent advancements in machine learning (ML) have revolutionized air pollution prediction by offering robust models capable of understanding the complex, non-linear connections within environmental data (Wang et al., 2024). Evolved deep learning (DL) frameworks, such as the Gated Recurrent Unit (GRU), effectively handle complex data, improve multivariable forecasting, and provide deeper insights into interdependencies (Tao et al., 2023). GRUs are known for their ability to minimize uncertainty by extracting information through training and applying non-linear transformations (Kow et al., 2020). Similarly, the XGBoost algorithm, a decision-tree-based technique, offers superior prediction accuracy and reduced computational cost by iteratively combining weak learners and focusing on bias reduction (Chen and Guestrin, 2016; Ling et al., 2024; Tao et al., 2024). Wang et al. (2024) used convolutional neural network (CNN) model to estimate the spatial distribution of CO concentrations in Nanjing with a resolution of 10 m. The model utilized variables like building height, terrain features, and emission data and achieved impressive accuracy ($R^2 > 0.8$) when validated against PALM simulation outputs. Leng et al. (2018) developed a support vector machine (SVM)-based predictive model to estimate heavy metal levels in $PM_{2.5}$ utilizing the magnetic properties of tree leaves. The model demonstrated superior accuracy by achieving correlation coefficients above 0.7 for Fe and Pb. Results indicated higher pollutant levels in winter and near industrial/traffic sources. Tao et al. (2024) incorporated GRU and XGBoost models for time-sensitive prediction of NO_2. The results demonstrated the model produced great prediction performance by 4.1 % ± 1.0 % lower root mean square error over XGBoost and had low spatial uncertainty. Hu et al. (2023) recently utilized a hybrid approach combining CNN-LSTM-GRU to forecast $PM_{2.5}$ and O_3 levels. The results indicated that the proposed model achieved higher accuracy compared to single ML algorithms. Wang et al. (2023a, b) used CNN-LSTM model to predict the air quality index (AQI) and O_3 and $PM_{2.5}$ levels in the atmosphere. The proposed model shows high correlation coefficients with values more than 0.90 compared with SVM and random forest models. Despite advancements, significant challenges remain in predicting air heavy metal pollution, particularly in identifying and analyzing various influencing factors and addressing spatiotemporal dependencies with readily available variables. While GRU and XGBoost have demonstrated remarkable performance in various environmental studies, their potential in predicting spatiotemporal patterns of heavy metals in air dust remains relatively underexplored. Despite these advancements, significant challenges persist in predicting air heavy metal pollution, particularly in identifying and analyzing diverse influencing factors and addressing spatiotemporal dependencies with readily available variables.

1.1. *Research gap and motivation*

Previous studies have predominantly focused on PM prediction, often utilizing meteorological variables while neglecting the influence of factors such as distance to pollution sources and potential dust sources. This study is motivated by the pressing need for cost-effective, efficient, and accurate methods to predict airborne heavy metal concentrations. The integration of advanced ML techniques like GRU and XGBoost, coupled with robust variable selection approaches such as Boruta, SHAP, and Wavelet analysis, offers a unique opportunity to address the constraints of conventional methods and propose robust techniques for accurate prediction. By focusing on accessible and measurable variables, this research aims to enhance the interpretability and computational efficiency of predictive models, thereby contributing to more effective air pollution management strategies. The study seeks to fill the gap in understanding and forecasting spatiotemporal patterns of heavy metal pollution, providing insights critical for mitigating associated health and environmental risks.

1.2. *Research objectives*

Despite advancements in pollution prediction, the use of low-cost, easily accessible parameters for air dust pollution monitoring and forecasting is especially necessary in urban locales with limited resources. This study aims to address these gaps by developing an ML framework to predict the Pollution Load Index (PLI) using optimized features and accessible parameters. By integrating meteorological, spatial, and pollution-related inputs, the current study aims to create a reliable and scalable framework that supports cost-effective air quality assessments and facilitates proactive health risk management in urban environments. Accordingly, this study proposes a model to examine the air pollution status in Yazd city, with the ability to predict the PLI pollution map for the upcoming season using minimal data from the previous three seasons. This approach assists in predicting the spatial pattern of seasonal PLI concentrations before each season begins. Therefore, GRU, XGBoost, and other ML models are employed, combined with BFSA and MVMD to address limitations such as data scarcity, uncertainty, and time constraints. The ultimate goal is to provide an efficient model for urban air quality management and monitoring.

2. Material and metods

2.1. *Study area*

To develop a model for predicting air pollutant due to heavy metals (using PLI), the current study chosen Yazd city, Iran, as the targeted research area (Fig. 1). Yazd city is located in the central part of Iran, between 31° 46′ to 31° 58′ north latitudes and 54° 16′ to 54° 26′ east

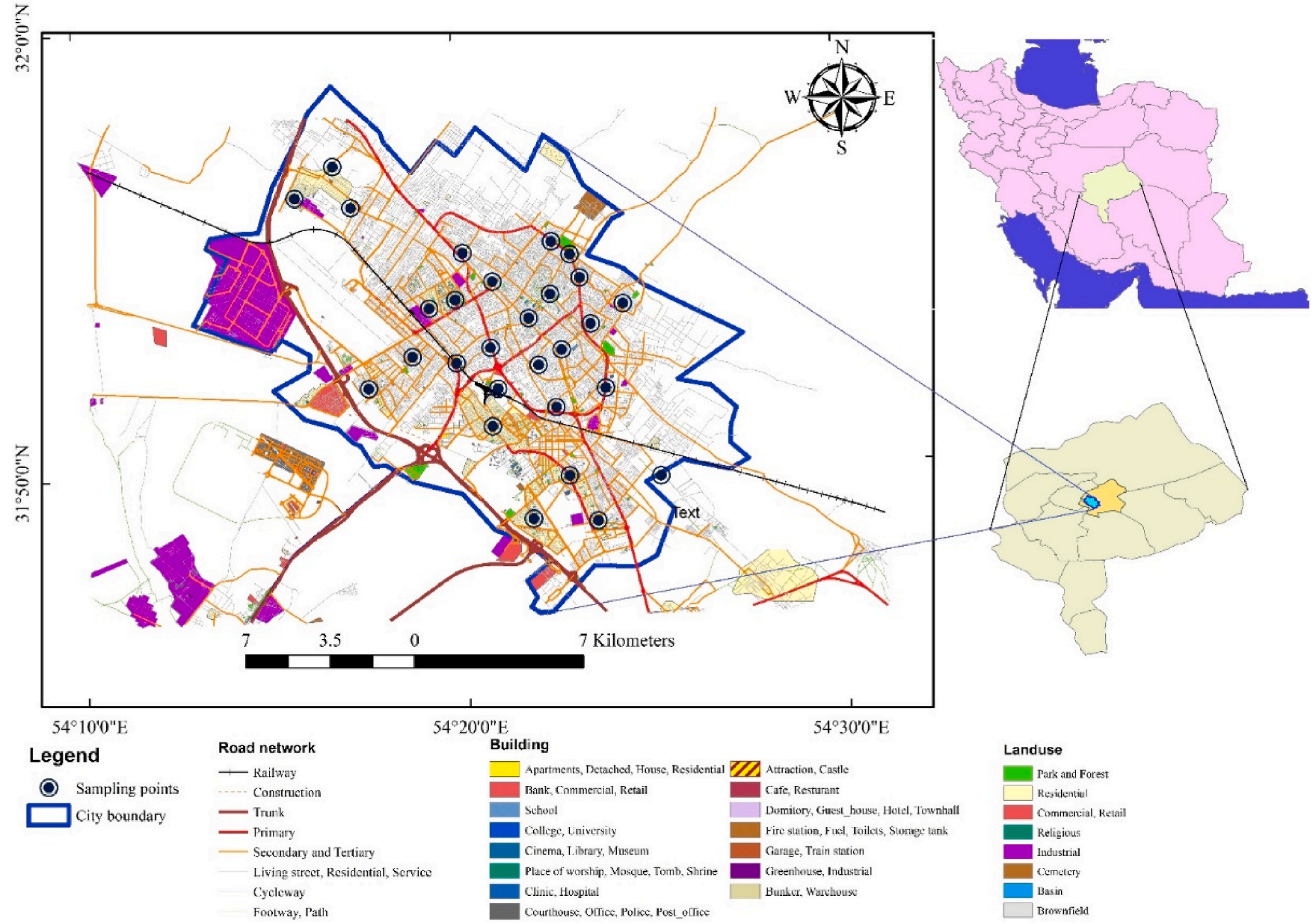

Fig. 1. Geographical location of the study area in Iran, the distribution of heavy metal sampling points, and the spatial arrangement of road networks, building types, and land use categories within the city boundary.

longitudes with an elevation of 1216 m, known for its unique geography and climate as well as its industrial importance. Due to its position in a desert region, Yazd city is characterized by its arid environment and surrounding mountains and plains, making it an important geographic hub. It has a warm, arid climate with an average annual temperature of about 19.1 °C. Summers can be extremely hot, often exceeding 40 °C, while winters are milder but can still be quite cold. The average relative humidity in Yazd is about 31 %, reflecting the arid conditions typical of desert regions. The city receives very little rainfall, averaging only 60.8 mm per year, most of which falls in late fall and winter. Dominant wind patterns include northwesterly winds in spring and summer, southeasterly winds from November to February, and westerly winds from March to October. Yazd's climatic and geographic characteristics significantly influence the patterns and intensity of air pollution and heavy metal contamination. The dry desert climate of the city experiences minimal rainfall, averaging about 60.8 mm per year, and low humidity at around 31 %. These conditions lead to dry soil and surfaces that are prone to generating dust. Dust particles can adsorb or transport heavy metals from natural sources, such as soil minerals, as well as from human activities. Yazd is known for its industrial activities, especially in the ceramic, tile, and steel industries. The Yazd Industrial Zone and the Iran Alloy Steel Company are located near the city, about 10 km and 30 km away, respectively. These industries contribute significantly to the local economy, but also pose environmental health challenges, particularly regarding heavy metal contamination of soil from industrial processes. Yazd's industrial sector contributes significantly to heavy metal emissions by discharging particulate matter containing metals like lead (Pb), cadmium (Cd), zinc (Zn), and nickel (Ni) into the atmosphere. Industrial combustion, metallurgical processing, and construction activities emit metal-laden dust and fumes that mix with naturally sourced dust particles. Yazd's basin-like topography, which is surrounded by mountains and plains, allows atmospheric pollutants to accumulate, leading to increased local exposure and environmental risks. The interaction between Yazd's natural desert environment and industrial activities creates a complex issue of heavy metal contamination in airborne particulates, highlighting the need for site-specific predictive modeling to effectively understand and mitigate pollution impacts. The interaction between Yazd's natural desert environment and industrial activities creates a complex issue of heavy metal contamination in airborne particulates, highlighting the need for site-specific predictive modeling to effectively understand and mitigate pollution impacts.

As observed in Fig. 2, extensive areas of barelands and saltlands are present in the eastern, northern, northeastern, southern, and southwestern parts of the city, which are considered sources of dust generation. Due to the geographical location of Yazd city and the surrounding barren landscapes, the frequency of dust storms around and inside Yazd City is relatively high. The expansive barelands and salt flats surrounding Yazd serve as major natural dust reservoirs. Frequent and strong winds, such as the northwesterly winds in spring and summer and the southeasterly winds in the colder months, contribute to the occurrence of recurring dust storms in and around the city. Dust storms significantly increase airborne particulate matter, facilitating the

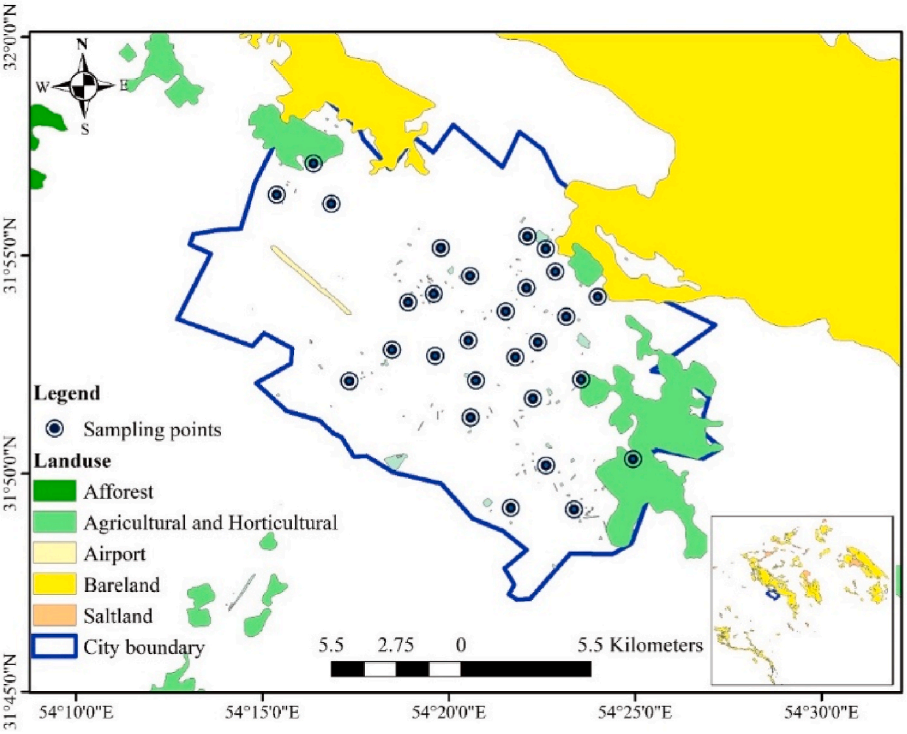

Fig. 2. The distribution of barelands and saltlands within and around the Yazd urban boundary.

transport of heavy metals and spreading them over extensive areas.

Also, based on Fig. 1, numerous roads, buildings, and industrial facilities suggest increased anthropogenic activities such as traffic, construction, and industrial operations, which contribute significantly to the heavy metal content in dust.

By comparing Yazd to similar regions, we strengthen the manuscript's argument for focusing on Yazd. Like Yazd, Al Hillah has an arid climate with frequent dust storms from the surrounding desert landscapes. Industrial activities such as steel manufacturing and cement production significantly contribute to heavy metal air pollution. In contrast to Yazd, pollution dispersion in Al Hillah is affected by varying wind patterns and is often exacerbated by environmental degradation linked to regional conflicts (Chabuk et al., 2021). The Atacama Desert region in Chile is among the driest places on Earth, where mining activities release heavy metals such as arsenic, lead, and copper into the air. Although both Yazd and the Atacama experience dust-related transport of metals, Yazd possesses a more diverse industrial base, including ceramics, steel, and tiles, which adds complexity to the sources of pollution. The Atacama Desert's extreme dryness and low population stand in stark contrast to the more urbanized environment of Yazd (González-Rojas et al., 2021). The Southwestern United States, including parts of Arizona and Nevada, faces industrial emissions alongside natural dust storms. Heavy metals from mining and smelting are prevalent pollutants. Topographical basin effects, like those in Yazd, can trap pollutants and lead to higher local concentrations. The regulatory frameworks and mitigation measures in these U.S. regions are generally more advanced, providing potential models for Yazd's environmental management (Sorooshian et al., 2024). In conclusion, Yazd shares key climatic and industrial pollution factors with similar regions; however, its unique geography, diverse industries, and specific meteorological patterns create distinct challenges and opportunities for pollution assessment and control. This highlights the need for research in Yazd to create reliable predictive models suited to its environmental context.

2.2. Data preparation

2.2.1. Dust and soil sampling

A marble dust collector was used to sample the atmospheric dust. Dust sampling was conducted during four seasons: fall and winter of 2018, and spring and summer of 2019, in the city of Yazd at a height of 3 m above ground level (on the roofs of single-story houses). To address the representativeness and limitations of the roof-based sampling method, we selected single-story rooftops at a uniform height of 3 m to minimize local variability and human interference. Although rooftop sampling may not fully capture near-surface dust dynamics, it offers a practical and consistent platform for long-term comparative analysis throughout the seasons. Samples were carefully collected at the end of each season and transported to the laboratory, where they were washed with water and reused after each sampling. Efforts were made to ensure that dust samples were not affected by rainfall predicted by the meteorological organization during the year of sediment sampling. Therefore, using dual sediment traps and coordinating with meteorological data enhances the reliability and representativeness of the collected samples. In addition, surface soil was collected once from the vicinity of the dust sampling areas.

2.2.2. Heavy metal concentration measurement

The four-acid method was used to digest the dust and soil samples from the study area, and the measurements were performed using an ICP-MS instrument (PerkinElmer model) from the USA (Amr et al., 2016). The total concentrations of the elements arsenic (As), cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), iron (Fe), manganese (Mn), nickel (Ni), lead (Pb), zinc (Zn), titanium (Ti), and zirconium (Zi) were measured. The spatial distributions of heavy metals in air dust across Yazd city during the autumn, winter, spring, and summer seasons are given in Fig. S1–S4, respectively. Overall, the concentrations of most heavy metals (As, Cd, Cr, Cu, Fe, Mn, Ni, Pb, Zn) tend to be highest during autumn and winter, with reduced levels in spring and summer seasons. The spatial patterns of heavy metals such as Pb, Zn, and Cu strongly align with urban and industrial areas, which indicates a significant influence from human-made sources. These metals are typically associated with urban and industrial activities. Seasonal changes also affect the distribution of these metals, with greater dispersion observed during summer and spring seasons and deposition occurring during autumn and winter seasons. Metals like Fe and Mn exhibit relatively stable spatial patterns across all seasons that suggest those sources are more consistent and likely influenced by both natural and human factors.

2.2.3. *Meteorological data*

Related meteorological data including maximum wind speed (Max WS), maximum wind direction (Max WD), mean wind speed (Mean WS), maximum temperature (Max T), minimum temperature (Min T), mean temperature (Mean T), mean vapor pressure (Mean VP), precipitation (P), maximum relative humidity (Max RH), minimum relative humidity (Min RH), mean relative humidity (Mean RH), total sunshine hours (Total SH), evaporation (Evp), mean dew point temperature (Mean DPT), and mean daily vapor pressure (Mean DVP) were collected from Iran Meteorological Organization for neighboring synoptic stations and Yazd city (Fig. 3). To obtain the meteorological data values at the sampling points, ArcMap software was utilized along with the Extract tools feature, which was applied based on the zoning created using the inverse distance weighting (IDW) interpolation method. The IDW method, a commonly employed spatial interpolation technique, predicts values for locations where data has not been collected by assigning weights to surrounding data points inversely proportional to their distance from the target location. In this case, IDW was used to create a continuous meteorological data surface from available synoptic stations. Subsequently, the Extract tools in ArcMap were employed to extract precise meteorological data values corresponding to the specific sampling points within the Yazd city.

2.2.4. *Distance to pollution source data*

To calculate the variables such as the nearest distance to railways, roads, industrial areas, and residential zones (as D-Roads, D-Railway, D-Industrial, and D-Residential), ArcMap software was utilized by employing the Near tool. This tool is a spatial analysis function that determines the shortest straight-line distance from sampling points to the closest feature, including railway, road, industrial area, or residential zone. Also, the average value of potential dust sources (P.D. Sources) for each sampling point is calculated using Eq. (1) for density and by considering the radius of the buffer area around each sampling point equal to 1000 m:

$$Density = \frac{SUM\_Area\ or\ Length}{Total\ buffer\ area} \times 100 \tag{1}$$

2.2.5. *Assessment of pollution indices*

For the total assessment of air pollution due to heavy metals, two indices of pollution load index (PLI) and Nemerow pollution index (PINemerow) are calculated in this study. The indices are calculated as follows (Kowalska et al., 2018):

$$PLI = \sqrt[n]{PI_1 \times PI_2 \times PI_3 \times \ldots \times PI_n} \tag{2}$$

$$PI = \frac{C_n}{GB} \tag{3}$$

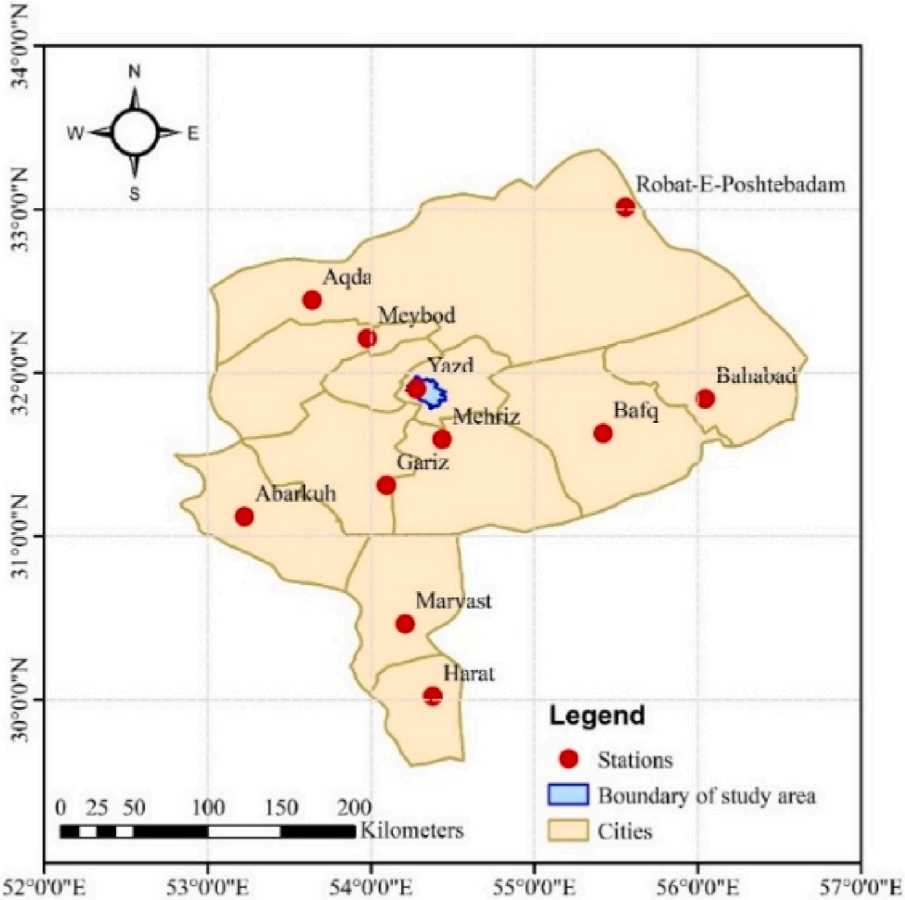

Fig. 3. Location of weather stations used to extract meteorological data for sampling points.

$$PI_{Nemerow} = \sqrt{\frac{\left(\frac{1}{n}\sum_{i=1}^{n} PI\right)^{2} + PI_{max}^{2}}{n}} \tag{4}$$

where C_n is the heavy metal concentration in air, *GB* is the geochemical background concentration, PI_{max} is the maximum value of PI among all heavy metals, and *n* is the number of heavy metals.

Based on the Kabata-Pendias (2011), the GB values for As, Cd, Cr, Cu, Mn, Ni, Pb, and Zn are considered equal to 0.67, 0.41, 59.5, 38.9, 488, 29, 27, and 70 mg kg^{-1}, respectively.

2.3. Multicollinearity test on the variables

Since strong interconnections (collinearity) among predictors reduce the effectiveness of predictions for the target variable, a collinearity test on datasets is applied before selecting variables for modeling. Collinearity can be identified using different methods. Among them, the Variance Inflation Factor (VIF) method has been extensively utilized to assess the relationship between independent variables and their influence on outcomes (Gholami et al., 2020; Ebrahimi-Khusfi et al., 2021). The VIF is defined as follows:

$$VIF = \left[\frac{1}{1 - R^{2}J}\right] \tag{5}$$

where, R^2J is the regression coefficient of determination of variable J. According to Bui et al. (2012), a VIF value greater than 10 signals a potential issue with collinearity.

2.4. Feature selection

The process of selecting relevant features is a fundamental step in the implementation of machine learning algorithms (Kursa and Rudnicki, 2010) because it directly impacts both the performance and the interpretability of models. By selecting only the most relevant features, the model complexity is reduced, it focuses on meaningful patterns, and leads to improved models' performance and interpretability. A reduction in the number of features creates a simpler model that is inherently more interpretable than a model relying on a high-dimensional feature space. It allows humans to better understand why a particular prediction or decision was made (García and Aznarte, 2020). In addition, using irrelevant or redundant features can lead to overfitting, causing the model to behave unpredictably on new data, and reducing confidence and interpretability of predictions. For local predictions, selecting features based on domain knowledge ensures the model aligns with real-world understanding of the problem. Researchers have employed different approaches to select the most important features for modeling environmental problems. This study applied three approaches of BFSA, SHAP, and wavelet coherence to select important variables controlling air pollution due to heavy metals.

The BFSA is a strong algorithm that evaluates feature importance by comparing them to randomly shuffled shadow features, ensuring that only statistically significant variables are kept for predictive modeling (Kursa and Rudnicki, 2010). The SHAP framework provides a clear interpretation of how each variable impacts model predictions, aiding in understanding the factors influencing PLI (Lundberg and Lee, 2017; Zhou et al., 2021). Zhou et al. (2021) compared traditional methods, such as Principal Component Analysis (PCA), with modern approaches like SHAP, concluding that SHAP provides superior interpretability and performance in identifying key variables. SHAP offers model-agnostic feature importance scores that are interpretable, even when dealing with nonlinear interactions. Park et al. (2022) pointed out the limitations of Tree-based Feature Importance (Tree-FI) and recommended using SHAP to address variable correlation issues. BFSA and SHAP take into consideration feature dependencies and the specific contexts of the model (Mahesswari and Maheswari, 2024). Moreover, wavelet coherence analysis greatly improves traditional feature selection by providing a dynamic and detailed approach to examining relationships between variables. Unlike traditional methods that often rely on static correlations or statistical metrics, wavelet coherence captures phase-synchronized relationships between pollution indices and predictor variables, tracking their evolution across both time and frequency domains. This dual perspective is especially effective for time-series data, where relationships can fluctuate over time or appear differently across various scales, including short-term diurnal cycles and longer-term seasonal trends. Wavelet coherence provides insights into the frequency domain, interpreted as the investment horizon, and reveals the persistence of temporal relationships and coherence patterns (Szczygielski et al., 2024).

The choice of BFSA, SHAP, and Wavelet coherence was made due to their effectiveness in managing nonlinear relationships, interpretability, and robust assessment of feature importance, which are essential for predicting air quality. These methods were applied separately, and their selected features were compared to ensure that the most robust subset was retained for modeling.

2.4.1. BFSA method

To separate influential factors from less important ones, the BFSA algorithm works based on selection criteria. The algorithm removes variables step by step, focusing on those statistically shown to be less relevant than random probes (Gholami et al., 2021). The approach enhances system reliability by minimizing the impact of correlations and unpredictable fluctuations through the application of additional randomness (Ebrahimi-Khusfi et al., 2021). The algorithm consists of eight essential steps including (1) adding multiple copies of all variables to extend the information system, (2) shuffling the newly included attributes to eliminate their correlations with PLI, (3) implementing a random forest algorithm to compute Z-score, (4) identifying the highest Z-score from the shadow variables and keeping all variables that outperform it, (5) performing an equality analysis for variables whose importance remained undetermined in the previous step, (6) removing variables with lower importance and keeping variables with higher Z-scores, (7) removing shadow variables, and (8) repeating the evaluation steps until the importance of every attribute is finalized (Gholami et al., 2021).

2.4.2. SHAP method

The SHAP uses cooperative game theory and produces Shapley values to ensure fair distribution of feature importance (Aldrees et al., 2024). Multiple steps are to be performed for implementing SHAP analysis. (1) A reference distribution is first created to analyze the model's behavior, which is commonly generated using training data samples. (2) By taking into account every potential feature combination, different sets of feature coalitions are constructed. (3) Shapley values are calculated by analyzing the model's predictions across all possible feature combinations to quantify the average contribution of each feature to the overall prediction. SHAP values indicate the importance of each feature, with positive values denoting a favorable impact and negative values reflecting an adverse influence. (4) SHAP values are visualized by drawing a plot to recognize the most important variables.

2.4.3. Wavelet coherence method

Wavelet coherence analysis is a method used to determine the time-frequency regions where two signals exhibit the strongest relationship or coherence. When applied for feature selection, it highlights the features that exhibit a strong correlation with PLI, across various temporal scales. Significant coherence values in wavelet coherence plots often imply a notable relationship between the feature and the target variable (Seifi et al., 2021). In signal processing, coherence between two time series *x (t)* and *y(t)* is defined as (Boya and Ardila-Rey, 2020):

$$C_{xy}(f)=\frac{\left|W_f^{xy}\right|^2}{W_f^x W_f^y} \tag{6}$$

$$W_f^{xy}=X(f).Y^*(t) \tag{7}$$

where f represents the frequency, W_f^x and W_f^y are the power spectral densities (PSDs) of *x(t)* and *y(t)*, respectively, *X(f)* and *Y (t)* are the Fast Fourier Transform (FFT) of *x*(*t*) and *y*(*t*), respectively. The asterisk (*) symbol is utilized to indicate the complex conjugate operator.

The continuous wavelet transform (CWT) is a mathematical tool characterized by a zero-mean function that is localized across both the time and frequency domains. The process involves breaking down a time series into a series of translated and expanded iterations of itself using a convolution operation with a mother wavelet function. This approach enables the analysis of the series at various scales while retaining its temporal dynamics. The Morlet wavelet is a popular choice as the mother wavelet in CWT applications (Boya and Ardila-Rey, 2020).

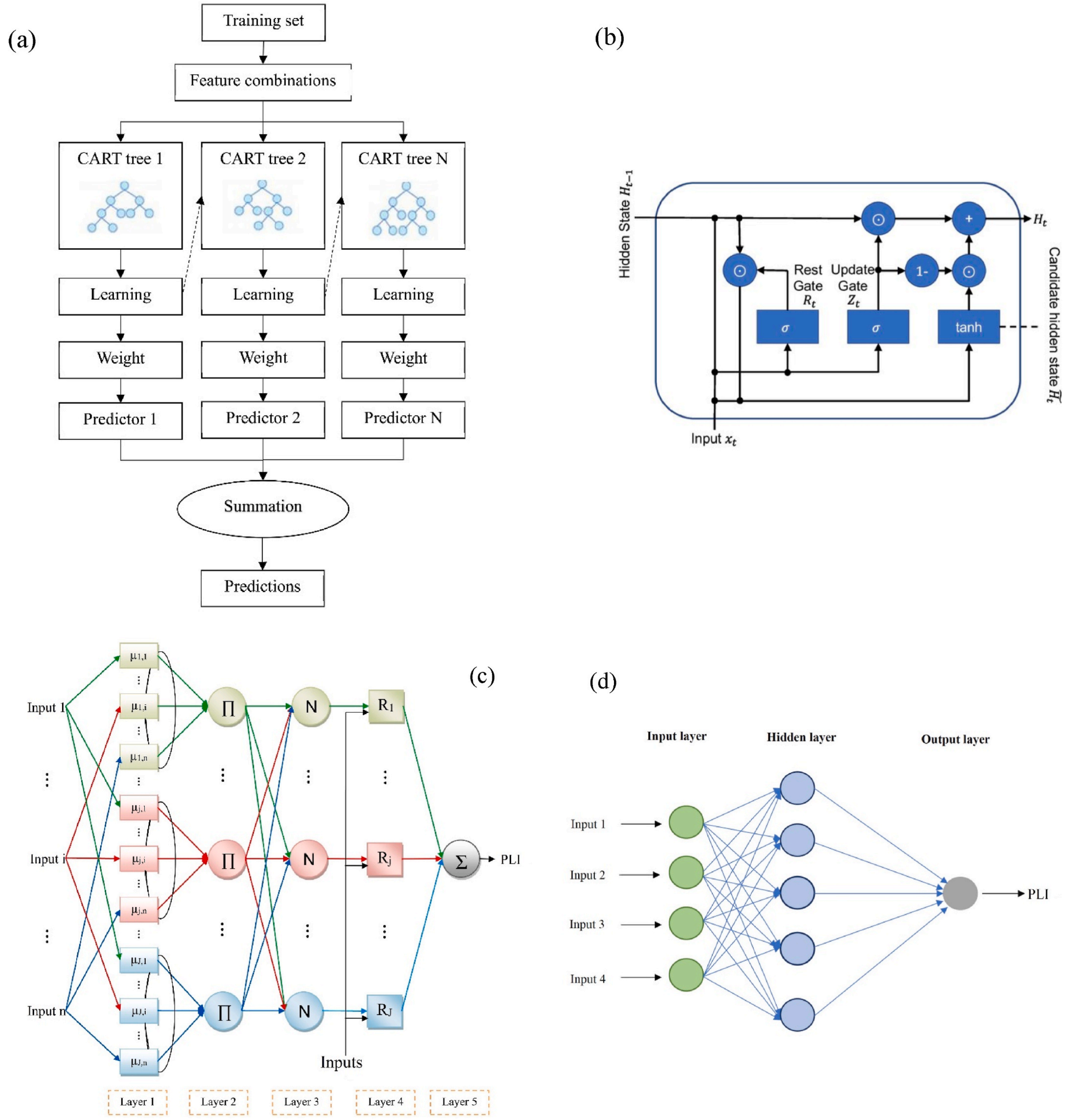

Fig. 4. The schematic diagram of (a) XGBoost, (b) GRU (Chen et al., 2024), (c) ANFIS (Dehghani et al., 2019), and MLP models.

2.5. Machine learning models

The XGBoost algorithm is an optimized and upgraded version of the gradient boosting decision tree (GBDT). This approach utilizes a series of decision trees, progressively refining the model by incorporating a new regression tree at each step. The new tree is trained to address the errors between the predicted and actual values of PLI to enhance the model's predictive accuracy. Since the primary aim of prediction is to ensure that the estimated value of PLI closely aligns with those observed, two parameters of generalization and regularization should be optimized. The regularization factor measures tree complexity and enhances the model's stability through simplification (Li et al., 2022). The schematic diagram of Fig. 4a illustrates the structure of the XGBoost model. The flowchart demonstrates that the input data is processed to build a series of decision trees that form the model's structure. Trees are iteratively trained to minimize errors from prior iterations, and their predictions are combined to produce the final output. By optimizing both the generalization and regularization parameters, XGBoost effectively enhances predictive accuracy.

The GRU architecture is based on LSTM and an optimized network structure while delivering similar performance outcomes. In comparison to the LSTM architecture, the GRU framework employs just two gate mechanisms, namely the update and reset gates. Based on Fig. 4b, the input at the current time step is provided to the GRU cell, which interacts with the previous hidden state of H_{t-1} to compute the updated hidden state of H_t. After obtaining input variables, the reset gate determines how much of the information from the previous state should be ignored (Li et al., 2021). This mechanism enables the GRU to adjust its memory dynamically and respond to evolving patterns in the input sequence. The reset gate involves the hidden state, a critical element in retaining significant information from earlier time steps while discarding irrelevant data (Chen et al., 2022). In the next step, the update gate determines what past information is incorporated into the present moment and must be retained. This gate enhances the GRU model's ability to identify and understand complex patterns in data. The next step focuses on deriving the candidate hidden state, which integrates new information into the memory. Finally, the GRU model calculates the current hidden state that is considered as the output (Seifi et al., 2024).

The ANFIS is a hybrid model that combines artificial neural networks (ANN) with fuzzy logic to address linguistic uncertainty. ANFIS employs the structure of a fuzzy inference system with three key components: (1) "if-then" fuzzy rules, (2) a database for membership functions, and (3) an inference mechanism. In this study, the Sugeno-type ANFIS model is implemented due to its efficiency. ANFIS operates through a five-layer network structure where inputs pass through fuzzification, rule application, normalization, and defuzzification layers to produce an output. Each rule is represented with a Gaussian membership function and parameters optimized to improve the prediction of the system by finding optimal membership parameters (Dehghani et al., 2019) (Fig. 4c).

The MLP is a type of ANN that operates using interconnected layers of neurons. These layers include an input layer, one or more hidden layers, and an output layer. Each neuron processes incoming signals using weighted connections and applies an activation function to produce output signals. In the study, the MLP uses a ReLU (Rectified Linear Unit) activation function in the hidden layers and a linear activation function in the output layer to process input data. Training the MLP involves backpropagation to adjust weights and biases, minimizing the error between predicted and actual outputs (Fig. 4d). However, standard backpropagation can converge slowly or become stuck in local optima, so an optimization algorithm is integrated to enhance performance.

The Optuna algorithm is an efficient and flexible hyperparameter optimization framework used to enhance the performance of applied models.

The models were developed and implemented in Python, utilizing its comprehensive libraries and tools for computational modeling and analysis.

2.6. Managing non-stationary data by MVMD

In this study, the Augmented Dickey-Fuller (ADF) test is used to assess the stationarity of the time series data. The test helps identify whether the data exhibits unit roots, which are indicative of non-stationary behavior. If non-stationarity is detected, the data is preprocessed or decomposed using the MVMD method to extract Intrinsic Mode Functions (IMFs), which represent the underlying oscillatory modes suitable for further analysis and modeling. The MVMD extends the VMD version to handle multichannel data for concurrent decomposition. The MVMD is helpful to solve the problem of adaptive selection of mode parameters using scale segmentation and offers mode separability to avoid any predefined wavelet filter bank boundaries. The MVMD process focuses on progressively optimizing each IMF by estimating its core frequencies and corresponding bandwidths. To apply the MVMD method for a set of input variables, the extracted modes must accurately reconstruct the original signal while minimizing their total bandwidth (Seifi et al., 2024). These conditions are defined through specific equations (Wang et al., 2023a,b):

$$X(t) = \sum_{k=1}^{K} U_k(t), t = 1, 2.., n \tag{8}$$

$$\left[\begin{array}{c} \underset{\{u_{k,d}\},\{w_k\}}{min} \left\{ \sum_{k=1}^{K} \sum_{d=1}^{D} \left\| \partial_t \left[u_{+}^{k,d}(t) e^{-jw_k t} \right] \right\|_2^2 \right\} \\ subject(to) \sum_{k=1}^{K} u_{k,d}(t) = x_d(t) \end{array}\right] \tag{9}$$

where $u_{+}^{k,d}(t)$ represents the extracted modes, K denotes the total number of modes, x corresponds to the variable associated with different channels, w indicates the frequency component, and ∂_t signifies the partial derivative with respect to time.

By employing the Lagrange function, the constraints are embedded into Eq. (9) and then the alternate direction method of multipliers (ADMM) as an optimization algorithm is applied for solving Lagrange function. After that, the modes and frequencies are computed as follows:

$$\widehat{u}_{k,d}^{n+1}(w) = \frac{\widehat{x}_d(w) - \sum_{i \neq k} \widehat{u}_{i,d}(w) + \frac{\overset{\frown n}{\lambda_d w}}{2}}{1 + 2\alpha(w - w_k)^2} \tag{10}$$

$$w_k^{n+1} = \frac{\sum_d \int_0^\infty w \left| \widehat{u}_{k,d}^{n+1}(w) \right|^2 dw}{\sum_d \int_0^\infty \left| \widehat{u}_{k,d}^{n+1}(w) \right|^2 dw} \tag{11}$$

where $\widehat{x}_d(w)$, $\widehat{\lambda}_d(w)$, $\widehat{u}_{k,d}^{n+1}(w)$, $\widehat{u}_{i,d}(w)$ are the Fourier transforms,

After generating IMFs, they use as inputs into different models of GRU, XGBoost, ANFIS, and MLP.

2.7. Models evaluation criteria and diagrams

The performance of different models evaluated using several statistical evaluation criteria including root mean square error (RMSE), the coefficient of determination (R^2), mean square error (MSE), and mean absolute error (MAE). The equations of these criteria are given as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(PLI_i^{ob} - PLI_i^{pr} \right)^2} \tag{12}$$

$$R^2 = \left(\frac{\sum_{i=1}^{N} \left(PLI_i^{ob} - \overline{PLI}\right)\left(PLI_i^{pr} - \overline{PLI}\right)}{\sqrt{\sum_{i=1}^{N} \left(PLI_i^{ob} - \overline{PLI}\right)^2} \sqrt{\sum_{i=1}^{N} \left(PLI_i^{pr} - \overline{PLI}\right)^2}} \right)^2 \tag{13}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| \left(PLI_i^{pr} - PLI_i^{ob}\right) \right| \tag{14}$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(PLI_i^{pr} - PLI_i^{ob}\right)^2 \tag{15}$$

Taylor diagram combines three critical metrics of standard deviation (SD), centered root mean square difference (RMSD), and R^2 into a single visualization. It offers a graphical assessment of how well model predictions align with observed data.

The performance and uncertainty of the best model in predicting PLI in different seasons is evaluated by Monte Carlo uncertainty method. The uncertainty is quantified using the 95PPU (95 % prediction uncertainty) bounds, defined by the 97.5 % and 2.5 % percentiles, and the degree of uncertainty ($\overline{d}_x$) is used to assess the goodness of fit and robustness of the models (Seifi et al., 2020).

$$\overline{d}_x = \frac{1}{k} \sum_{i=1}^{k} \left(X_U - X_L\right)_i \tag{16}$$

$$d - factor = \frac{\overline{d}_x}{\sigma_x} \tag{17}$$

$$95PPU(\%) = \frac{Count(Q \backslash X_L \leq Q \leq X_U)}{n} \times 100 \tag{18}$$

where k represents the total number of samples, σ_x denotes the standard deviation of the PLI, and $\overline{d}_x$ signifies the mean distance between the upper and lower bounds.

2.8. Model overview and development

The methodology of this study integrates MVMD, a strong ML model (XGBoost), Optuna optimization, and feature selection techniques (BFSA, SHAP, and Wavelet Coherence) into a cohesive framework to predict the spatiotemporal distribution of PLI. Each component has a specific role, and their combined use greatly enhances the accuracy, efficiency, and cost-effectiveness of predictions. Feature selection involves identifying the most relevant variables for predicting PLI. This approach decreases model complexity and data collection costs while improving both interpretability and performance. MVMD preprocesses selected non-stationary variables through feature selection to improve model performance by decomposing them into stationary IMFs. The decomposed input variables are fed into the XGBoost, GRU, MLP, and ANFIS models to predict PLI, utilizing their capability to model complex, non-linear relationships and their computational efficiency. Optuna optimizes hyperparameters for ML/DL models to improve their predictive performance and efficiency.

The model framework comprises several fundamental steps outlined below.

Step 1. Data preparation: The dataset, including heavy metal concentration and meteorological variables, is collected. Also, datasets of nearest distance to railways, roads, industrial areas, and residential zones, and potential dust sources are calculated.

The Pearson correlation coefficient heatmap and hierarchical clustering are used to determine how strongly two variables are related in a linear way (Fig. 5). It helps to determine strong relationships between variables and find out which features are too similar. From Fig. 5, meteorological variables like temperature (Mean T, Max T, Min T), relative humidity (Mean RH, Max RH, Min RH), and wind speed (Mean WS) tend to group together because they're closely linked. Their high positive correlations indicate that these factors are often influenced by seasonal shifts and atmospheric conditions. This cluster of weather-

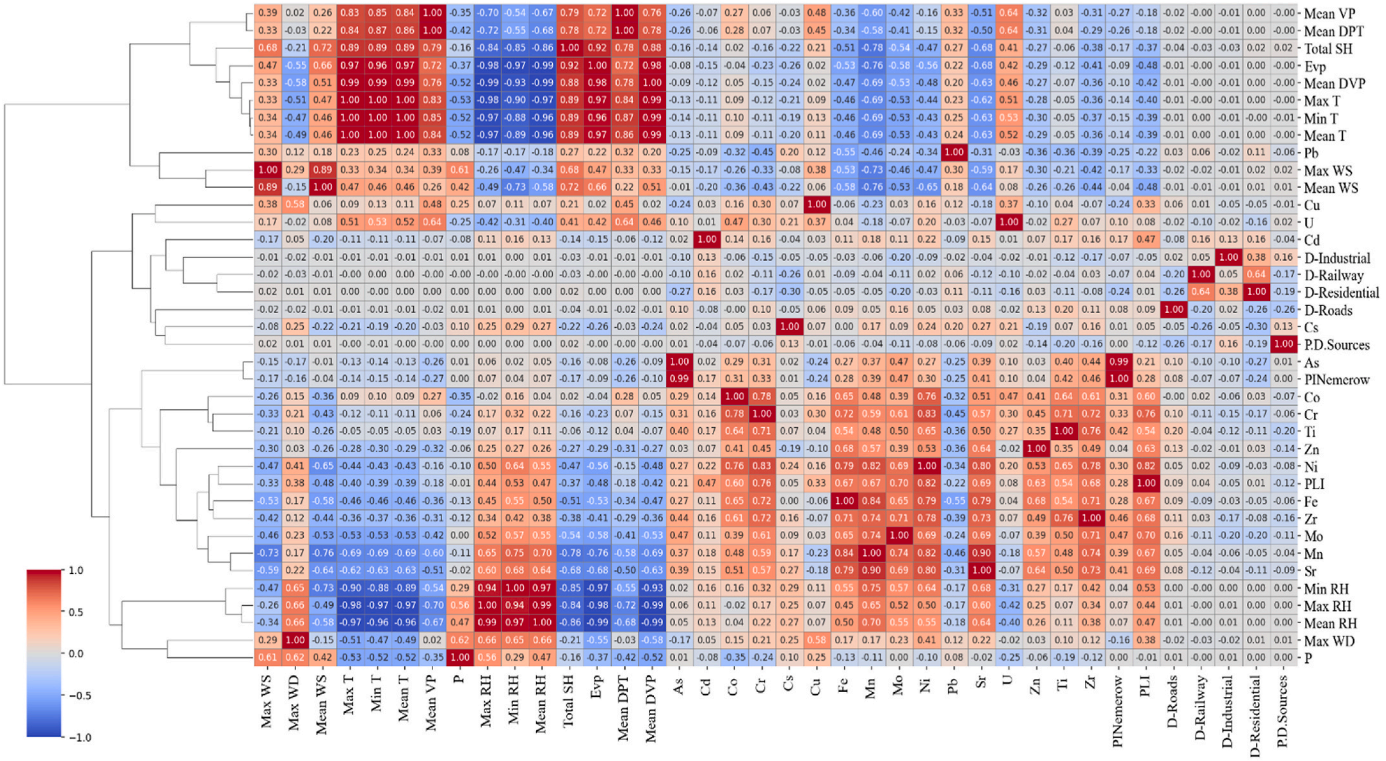

Fig. 5. The correlation heatmap combined with hierarchical clustering.

related variables might also show negative correlations with some pollutants. For instance, higher wind speeds are often associated with lower pollutant levels that reduce air pollution due to dispersion. Pollutants like heavy metals, including Pb, Cd, Zn, and Cu, tend to cluster together because they often come from similar sources or behave alike in the environment. Metals like Pb, Zn, and Cu frequently show strong positive correlations, which suggests they likely originate from similar sources such as industrial emissions, vehicle traffic, or waste burning. Variables of D-industrial, D-residential, and factors like D-road and D-railways, often cluster together. When these factors show positive correlations with pollutants, it suggests that specific types of land use contribute more to pollution. For instance, a strong link between D-industrial and heavy metals like Zn and Cu would indicate that industrial activities are a major source of these pollutants in the area. Moderate correlations between specific pollutants and land-use variables, or weaker links between pollutants and meteorological factors, suggest that these relationships aren't straightforward. They're likely influenced by a mix of factors working together or changing under different conditions that a simple correlation analysis might not fully capture. In the presented correlation matrix, it is evident that PLI has stronger correlations with other variables compared to $PI_{Nemerow}$. This indicates that PLI provides more information than $PI_{Nemerow}$ about the impact of environmental and climatic factors or the concentration of heavy elements on air dust pollution. Hence, the PLI is used as the target value of predicting models.

Step 2. Multicollinearity test: VIF test is applied on datasets to remove variables with strong collinearity. Collinearity relationships between the independent variables are represented in Fig. 6. The findings indicated that almost all of the variables exhibited a moderate level of collinearity. This is because the VIF values fell within the range of greater than 1 and up to 5 ($1 < \text{VIF} \leq 5$). These results suggest that while some correlation exists between the variables, it is not excessively strong. Then, all variables are used for the feature selection process. The variables of D-Residential and Ni had the lowest and highest VIF values, respectively. Although all VIF values are less than 5, there is a high correlation between many variables, such as Mean WS and Total SH variables. This suggests a strong interrelationship between variables that may influence the model performance. These findings necessitate careful consideration of these predictors during feature selection to avoid redundancy and enhance model interpretability.

Step 3. Feature selection: Three procedures of BFSA, SHAP, and wavelet coherence are applied to select the most important variables for predicting PLI.

There are 36 features that could produce 2^{36}-1 input vector (combinations), which significantly complicates the modeling process. Thus, the use of the feature selection method was crucial to reduce the complexity and identify the most informative features. Table 1 illustrates the optimal combination of variables utilized for predicting PLI, identified through BFSA, SHAP, and wavelet coherence analyses. According to BFSA and by applying all variables as inputs, the combination of Min RH, Cd, Cr, Cs, Ni, Pb, Zn, Ti, Zr variables has been confirmed to predict PLI. For this input vector, the Z-score of Zn is calculated equal to 2.39 that had the highest value among other variables. Therefore, the Zn variable is the most important parameter for predicting PLI. In this section, we applied the XGBoost model to determine the accuracy of the selected input vector for predicting PLI. The model was optimized using the Optuna framework, a powerful and flexible tool designed for hyperparameter optimization to enhance model performance. As seen from Table 1, the first selected input vector using BFSA (Min RH, Cd, Cr, Cs, Ni, Pb, Zn, Ti, Zr) showed high accuracy with R^2 and RMSE equal to 0.91 and 0.07, respectively, in the testing phase. The evaluated criteria in the training phase are provided in Table S1. Since one of the primary objectives of this study is to select an input vector using the minimum necessary variables to have a cost-effective characteristic for collecting, some variables were removed from the BFSA input, and alternative input combinations were evaluated. The input vector, including Max WD, Min RH, Cd, and Zn variables, created the best performance ($R^2 = 0.90$, RMSE = 0.08, MAE = 0.06) for predicting PLI. The findings indicate that Max WD demonstrated a stronger explanatory power for PLI variations compared to other meteorological variables, as observed in all input vectors.

The SHAP method was performed on the data, and the results, SHAP values (Fig. S5), showed that the SHAP values of Zn were calculated in the range of −0.21 to 0.16. For each variable, the average absolute SHAP value reflects its overall importance in the model. The average SHAP

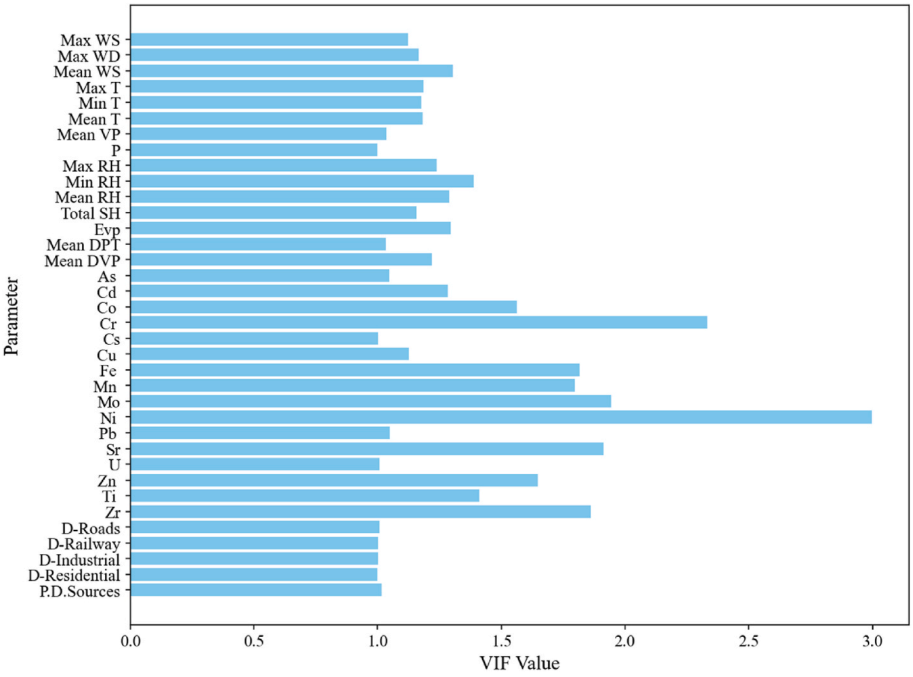

Fig. 6. VIF values to determine multicollinearity among independent variables in the dataset.

Table 1
The best feature combinations of variables based on BFSA, SHAP, and wavelet coherence techniques.

Method	Input variables	Selected variables	R^2	RMSE	MSE	MAE
BFSA	All meteorological, heavy metals, and distance parameters	**Min RH, Cd, Cr, Cs, Ni, Pb, Zn, Ti, Zr**	**0.91**	**0.07**	**0.05**	**0.006**
	All meteorological, Cd, and distance parameters	Max WD, Min T, Min RH, Cd	0.72	0.13	0.09	0.02
	All meteorological, Zn, Ni, and distance parameters	Max WD, Min RH, Zn, Ni	0.71	0.14	0.09	0.02
	All meteorological, Cd, Zn, and distance parameters	**Max WD, Min RH, Cd, Zn**	**0.90**	**0.08**	**0.06**	**0.006**
	All meteorological, Cd, Pb, and distance parameters	Max WD, Min T, Min RH, Cd	0.75	0.13	0.09	0.02
	All meteorological, Cd, Cr, and distance parameters	Max WD, Min T, Cd, Cr	0.84	0.10	0.08	0.01
	All meteorological, Cd, Cs, and distance parameters	Max WD, Min RH, Cd, Cs	0.79	0.12	0.08	0.01
	All meteorological, Cd, Ni, and distance parameters	Max WD, Min T, Cd, Ni	0.84	0.10	0.03	0.001
	All meteorological, Cd, Ti, and distance parameters	Max WD, Min T, Min RH, Cd, Ti	0.83	0.11	0.08	0.01
	All meteorological, Cd, Zr, and distance parameters	Max WD, Min T, Min RH, Cd, Zr	0.84	0.10	0.08	0.01
	All meteorological, Cd, Zr, Ni, and distance parameters	Max WD, Min T, Precipitation, Cd, Zr, Ni	0.86	0.09	0.07	0.01
	All meteorological, Cd, Zn, Ni, and distance parameters	Max WD, Cd, Zn, Ni	0.88	0.09	0.06	0.008
SHAP	All meteorological, heavy metals, and distance parameters	Zn	0.51	0.18	0.12	0.03
		Zn, Ni	0.64	0.15	0.10	0.02
		Zn, Ni, Ti	0.79	0.11	0.09	0.01
		Zn, Ni, Ti, Fe	0.84	0.10	0.08	0.01
		Max WD, Zn, Ni, Ti, Fe	0.86	0.09	0.07	0.01
Wavelet coherence analysis	All meteorological, heavy metals, and distance parameters	Total SH, Fe, Ti, D-Roads, D-Railway, D-Industrial	0.75	0.13	0.10	0.02
		Total SH, Mean RH, Evaporation, Fe, Ti, D-Roads, D-Railway, D-Industrial	0.88	0.08	0.07	0.007

values for Zn, Ni, Ti, Fe, and Max WD were calculated as 0.11, 0.027, 0.025, 0.021, and 0.02, respectively. To select the optimal input vector, different combinations of variables were tested using SHAP. The combination of Max WD, Zn, Ni, Ti, and Fe was determined to be the best (Table 1). This implies that these variables, when used together, resulted in the best performance of the model, as assessed by SHAP. The highest accuracy of the SHAP input combinations was achieved using Max WD, Zn, Ni, Ti, Fe, which is lower than the accuracy of the combination selected by the BFSA method.

Based on the color intensity (with warmer colors indicating higher coherence) in the wavelet coherence plots (Fig. S6), many variables, including Total SH, Fe, D-Roads, D-Railway, and D-Industrial, showed high coherence with PLI over significant time intervals and frequency ranges that make them strong candidates for prediction. Also, variables such as Mean RH, Evaporation, and Ti exhibit moderate coherence that suggests some importance for predicting PLI. In spite of BFSA and SHAP, the wavelet coherence method indicates the importance of spatial and industrial factors. Among wavelet coherence input combinations, Total SH, Mean RH, Evaporation, Fe, Ti, D-Roads, D-Railway, and D-Industrial, achieved the highest accuracy, although it was still lower than the accuracy obtained with the best combination selected by the BFSA method.

Overall, the BFSA method showed its strength in systematically eliminating irrelevant variables and led to the selection of the optimal feature set. While SHAP offered valuable insights into feature importance, its selected combinations did not achieve the same level of performance with BFSA. Wavelet coherence analysis, while effective in identifying relationships across time and frequency domains, yielded lower overall accuracy metrics for this dataset. Therefore, after comparing the results, the combination of Max WD, Min RH, Cd, Zn was found to be the best feature set with minimum required variables. This selection was justified by its superior performance compared to other combinations. Subsequent evaluations were conducted using this optimal feature set.

Step 4. XGBoost model training and testing: In this step, the XGBoost model optimized using Optuna was used to compare the results of different variable combinations. The datasets are split randomly into 70 % for the training and the remaining 30 % for testing. The best combination was chosen in this step.
Step 5. Develop comparing models: GRU, ANFIS, and MLP models were developed for model training, involving the Optuna optimization of hyperparameters for the best variable combination selected from step 4. The optimal values of the models' parameters are given in Table 2.

Step 6. Evaluation of models: The results of the XGBoost model for the best variable combination selected from step 4 were compared with GRU, ANFIS, and MLP models from step 5 using goodness-of-fit key criteria such as MAE, MSE, RMSE, and R^2, as well as error box plot, Taylor diagram, and heatmap plot. The best model for predicting PLI was selected from this step.
Step 7. Managing non-stationary data: Variables in the best input combination are examined for stationarity, and if necessary, processed using the MVMD method to produce IMFs. The non-stationary variables are recognized using ADF test.

Table 2
Optimal values of the models' parameters.

Model	Parameter values
GRU	Learning rate:0.0074, Number of layers:4, Number of neurons of the first layer: 444, Number of neurons of the second layer: 17, number of epochs:200
MLP	Learning rate:0.0018, Number of hidden layers:2, Number of neurons of the first layer:497, Number of neurons of the second layer:403, Batch size:32
XGBoost	Learning rate:0.0251, Max depth:3, Subsample:0.5437, Colsample_bytree:0.9031, L1 regularization:1.4342E-07, L2 regularization:4.01626E-08
ANFIS	Learning rate:4.35E-05, alpha:0.0115

Step 8. Incorporating MVMD with the best model: The best model chosen from step 6 was then proceeded using MVMD by applying non-stationary data, and two models were compared.

Step 9. Creating spatial maps: The best model for predicting PLI was chosen in the testing phase of step 8 applied for creating spatial observed and predicted PLI maps in different seasons. In addition, the temporal uncertainty analysis is done for predicting in different seasons.

The conceptual flowchart illustrating the PLI prediction model is depicted in Fig. 7.

3. Results and discussion

3.1. Evaluate the performance of other models against XGBoost for the best combination

The results of the proposed models for predicting PLI for the best input combination (Max WD, Min RH, Cd, Zn) are given in Table 3. The models were optimized using the Optuna algorithm for hyperparameter tuning. XGBoost model demonstrated the best performance on the testing set by achieving the lowest RMSE (0.08), MSE (0.06), and MAE (0.006). The GRU model followed with a strong performance ($R^2 = 0.77$, RMSE = 0.12). In contrast, the ANFIS model showed weaker performance for predicting PLI with higher errors of RMSE = 0.17 and MAE = 0.13. The MLP model outperformed the ANFIS model and achieved a good performance with an R^2 value of 0.74 during the testing phase.

The boxplots of the absolute error distributions for four models (GRU, ANFIS, XGBoost, and MLP) during the training phase and the testing phase are given in Fig. 8. XGBoost consistently exhibited the lowest absolute error in both training and testing phases, which makes it the most accurate and robust model in this comparison. GRU and MLP showed moderate performance, while ANFIS demonstrated the widest error spread and the most outliers, especially during the testing phase. The Taylor diagram for training and testing phases (Fig. 9) illustrates the comparative performance of GRU, ANFIS, XGBoost, and MLP based on their correlation, standard deviation, and RMSD relative to the observed data. The observed points are best approximated by the XGBoost model, demonstrating its superior capacity to understand and predict complex data trends. The GRU and MLP models are positioned closely together. The XGBoost is located between R lines 0.99 and 1, STD lines of 0.30 and 0.35, and under RMSD line of 0.04 in the training phase, and between R lines 0.95 and 0.96, STD lines of 0.20 and 0.30, and RMSD lines of 0.04 and 0.08 in the testing phase. Fig. 10 illustrates the correlation between

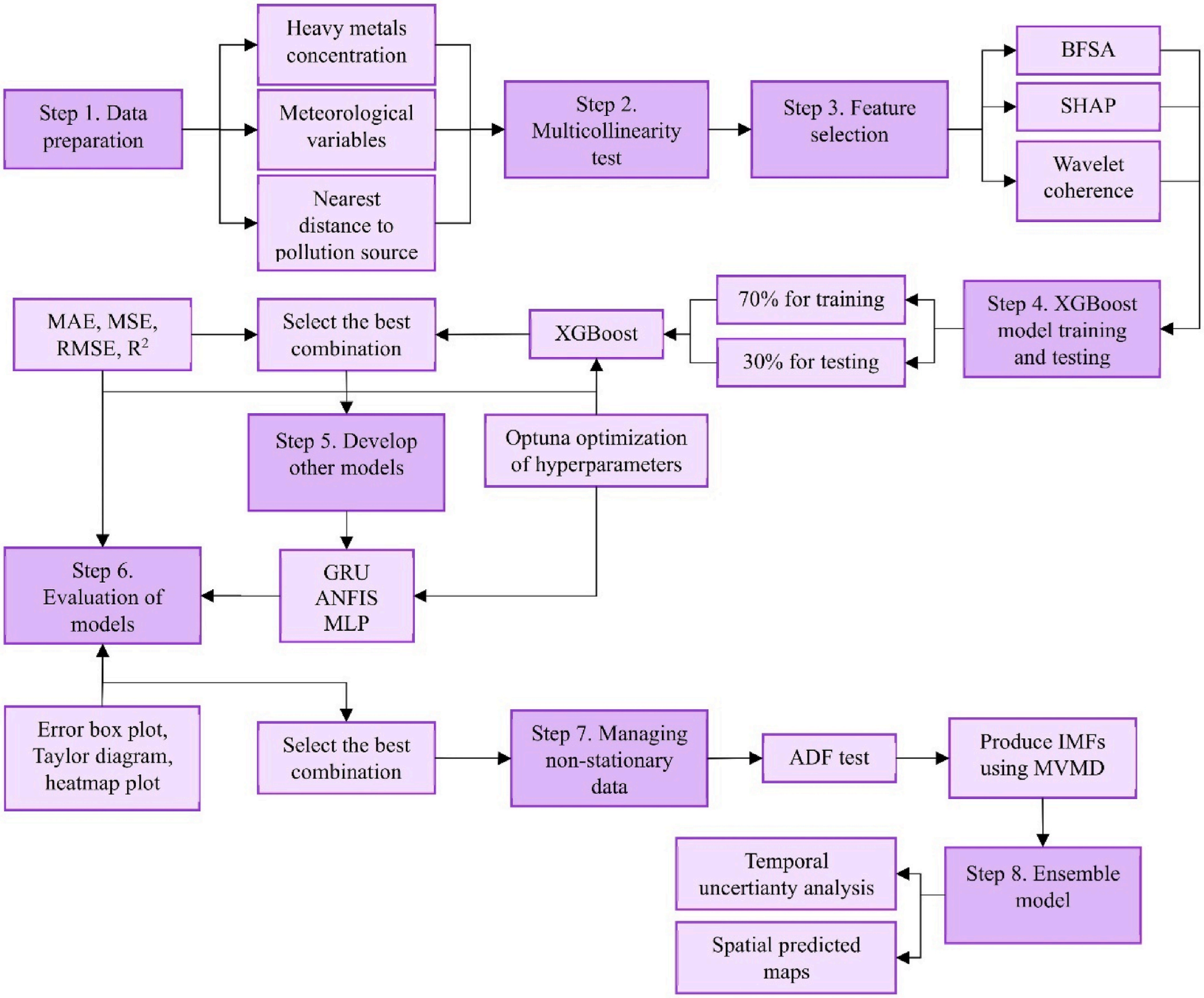
Step 1. Data preparation
Heavy metals concentration
Meteorological variables
Nearest distance to pollution source
Step 2. Multicollinearity test
Step 3. Feature selection
BFSA
SHAP
Wavelet coherence
Step 4. XGBoost model training and testing
70% for training
30% for testing
XGBoost
Select the best combination
MAE, MSE, RMSE, R²
Step 5. Develop other models
Optuna optimization of hyperparameters
GRU
ANFIS
MLP
Step 6. Evaluation of models
Error box plot, Taylor diagram, heatmap plot
Select the best combination
Step 7. Managing non-stationary data
ADF test
Produce IMFs using MVMD
Step 8. Ensemble model
Temporal uncertainty analysis
Spatial predicted maps

Fig. 7. Flowchart of modeling process for predicting PLI.

Table 3
Performance of the models in predicting the spatiotemporal pattern of PLI.

Models	Train				Test			
	R^2	RMSE	MAE	MSE	R^2	RMSE	MAE	MSE
XGBoost	0.98	0.03	0.02	0.001	0.90	0.08	0.06	0.006
GRU	0.89	0.09	0.06	0.01	0.77	0.12	0.09	0.01
ANFIS	0.70	0.14	0.11	0.02	0.51	0.18	0.13	0.03
MLP	0.92	0.07	0.06	0.01	0.74	0.13	0.08	0.02

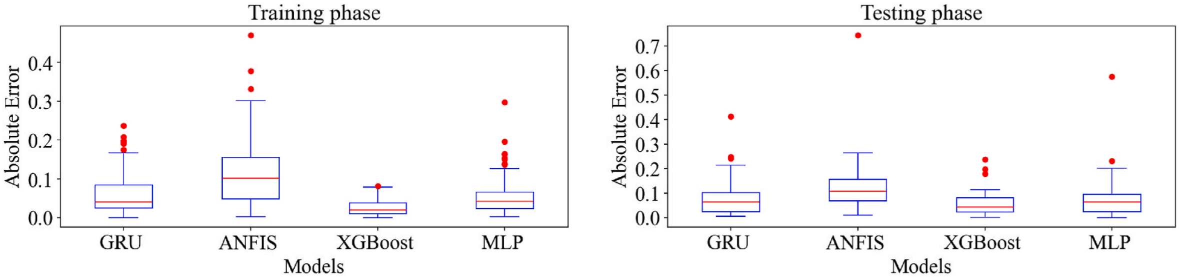

Fig. 8. Distribution of models' prediction error.

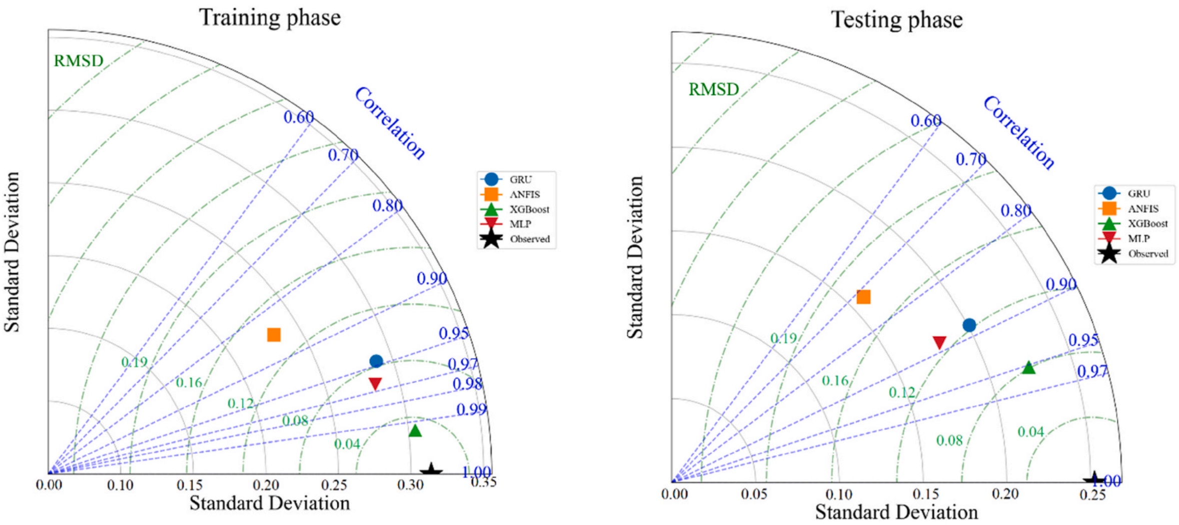

Fig. 9. The position of models' prediction against observed value using Taylor diagram.

observed and predicted PLI values across various models during the testing phase. Furthermore, normalized density values are used to color the scatter plot. A high degree of correlation between observed and predicted PLI in the XGBoost framework. For this model, the 'hot spots' are concentrated between 1.5 and 1.9 of PLI. The color spectrum for GRU and MLP models suggests slightly overpredicting and underpredicting the PLI. The high values of PLI are accurately correlated and correctly predicted by XGBoost, GRU, and MLP models. There are big 'hot spot' in ANFIS heatmap.

Overall, the XGBoost model emerges as the most effective model for predicting the spatiotemporal pattern of PLI, particularly when applied to a well-chosen input feature set. This superior performance can be attributed to several inherent characteristics and advantages of the XGBoost algorithm, which make it particularly well-suited for this application. The relationship between air dust pollutants and meteorological or environmental features (e.g., wind direction, humidity, and heavy metal concentrations) is highly non-linear and complex (Figs. 5 and 6). XGBoost is a gradient boosting framework that builds decision trees sequentially. It surpasses at modeling non-linear interactions between features and PLI by creating highly flexible decision boundaries. This allows it to capture the intricate dependencies in urban PLI data more effectively than simpler models like ANFIS or MLP. Li et al. (2022) represented that the XGBoost algorithm has a superior ability to capture the spatial and temporal variations in $PM_{2.5}$ and O_3 pollutant concentrations, outperforming the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem) model and other statistical algorithms like support vector regression (SVR), linear regression (LR), decision tree regression (DTR), and random forests (RF) particularly in urban areas.

Also, modeling PLI requires not only accurate predictions on training data but also the ability to generalize to unseen testing data, as the spatiotemporal patterns of PLI can vary significantly across regions and time. XGBoost includes built-in regularization techniques that reduce overfitting and improve generalization performance. This ensures that it performs well on both training and testing datasets. In this study, the XGBoost model demonstrated a higher processing speed compared to other models. XGBoost is highly optimized for speed and scalability. It uses parallel computing and efficient memory utilization, enabling it to process large datasets much faster than traditional models like ANFIS or MLP.

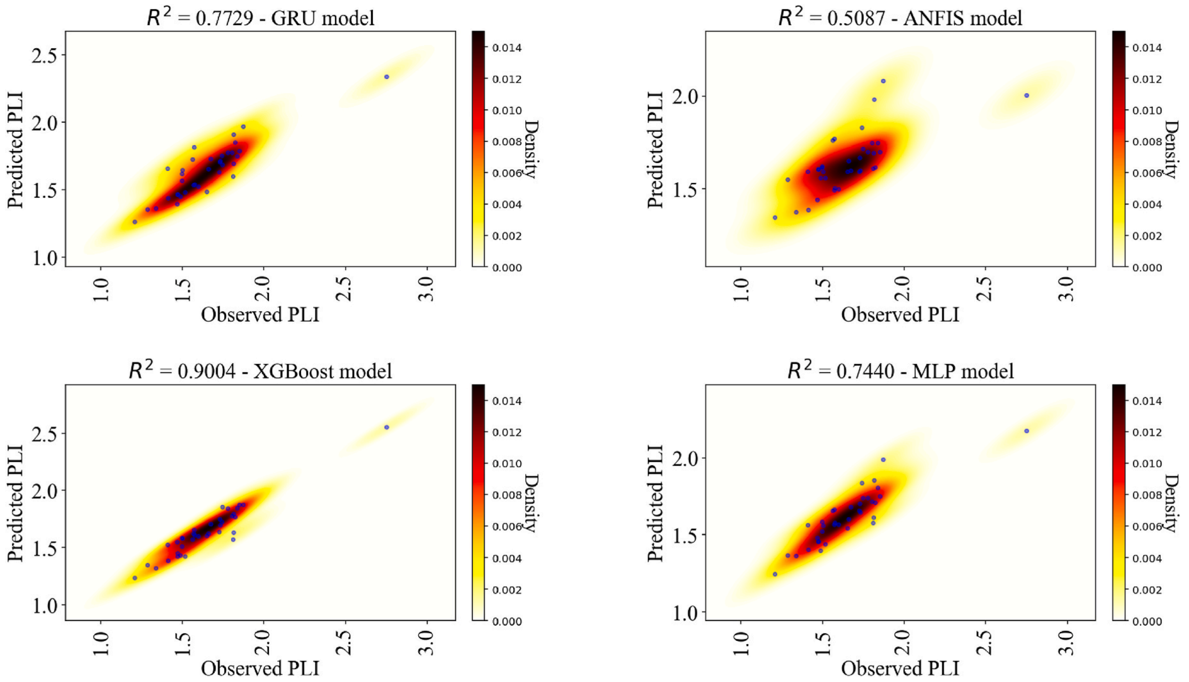

Fig. 10. The relationship between observed PLI and model predictions examined using density scatter diagrams and linear regression.

The GRU model consistently showed lower performance than XGBoost, which indicates their limited ability to handle the complexities of spatial PLI data. This discrepancy arises due to the fundamental differences in how the two models handle data, their limitations, and the specific characteristics of PLI prediction. PLI data often contains non-stationary patterns and noise due to environmental fluctuations, measurement errors, and the interaction of multiple factors. GRUs are designed for sequential modeling and assume temporal dependencies. They struggle with non-stationary signals and are less effective at capturing noise. XGBoost, on the other hand, is a tree-based model that can handle non-stationary and noisy data better, as it isolates feature interactions and splits data hierarchically. Song et al. (2023) proposed a novel hyperparameter optimization XGBoost model for accurately predicting the spatial variability of $PM_{2.5}$ concentrations. The research used Himawari-8 satellite-derived aerosol optical depth (AOD) data to map $PM_{2.5}$. The tree-structured Parzen estimator-XGBoost (TPE-XGBoost) approach outperformed grid search (GS) and random grid search (RGS) optimization algorithms and shows high prediction accuracy (R^2 values of 89.37 % in January to 83.68 % in April 2020).

In addition, with a limited number of data points and high-dimensional features, models can easily overfit during training. GRUs are prone to overfitting, especially when dealing with noisy or small datasets. Despite regularization techniques, their recurrent structure may memorize noise in the data rather than generalizing patterns effectively. XGBoost incorporates strong regularization techniques, column subsampling, and tree pruning, which help prevent overfitting and ensure better generalization to testing data. Also, GRUs require more time and computational resources for training, especially when handling large datasets or long sequences. This often results in less frequent hyperparameter tuning and suboptimal model configurations. The slower training process of GRUs can limit their ability to achieve optimal performance, particularly for spatiotemporal data. Tao et al. (2019) applied a deep learning-based model combining convolutional neural networks (1D ConvNets) and bidirectional GRU neural networks to forecast short-term air pollution by focusing on $PM_{2.5}$. The proposed model demonstrates improved prediction accuracy and lower error rates compared to BGRU, GRU, LSTM, and SVR. Chang et al. (2023) used combined W-BiLSTM(PSO)-GRU and XGBoost to enhance accuracy, stability, and robustness in predicting air pollutants ($PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, and O_3). Wavelet decomposition was employed to distinguish the low-frequency components from the high-frequency components within the time series of air quality indicators. The proposed model achieved R^2 values exceeding 0.94 and low error rates (MAE <0.02, RMSE <0.03).

3.2. Transforming non-stationary data into stationary data using MVMD

The stationarity of input variables for the best combination of Max WD, Min RH, Cd, Zn was assessed using the ADF test, as shown in Table 4. The ADF test evaluates the null hypothesis that data is non-stationary against the alternative hypothesis of stationarity. The critical values at 1 %, 5 %, and 10 % significance levels are provided, and the stationarity of each variable is determined based on whether the ADF test statistic is less than the critical value at a chosen significance level. For variables such as Max WD and Min RH, the ADF statistics do not exceed the critical value at any significance level. Additionally, their p-values are above 0.05, indicating the presence of non-stationary behavior. These variables require transformation to achieve stationarity before they can be effectively used in modeling. Features such as Zn with stationary behavior demonstrate stability over time, which makes them directly useable in predictive models. The MVMD process was implemented for extracting 3, 4, 5, and 6 IMFs of each variable. The results of MVMD-XGBoost models' accuracy against XGBoost in the testing phase are presented in Fig. 11. As seen, the ensemble model of MVMD-XGBoost accuracy with three IMFs was more than XGBoost, which exhibits the benefits of integrating MVMD preprocessing with XGBoost. MVMD enhances prediction accuracy by transforming raw

Table 4
The ADF test results for investigating stationarity in input data.

Variable	ADF Statistic	p-value	Stationarity
Max WD	−1.45	0.55	×
Min RH	−0.43	0.90	×
Cd	−7.36	~0	✓
Zn	−5.08	1.5×10^{-5}	✓
Critical value: (1 %) = −3.49, (5 %) = −2.88, (10 %) = −2.58			

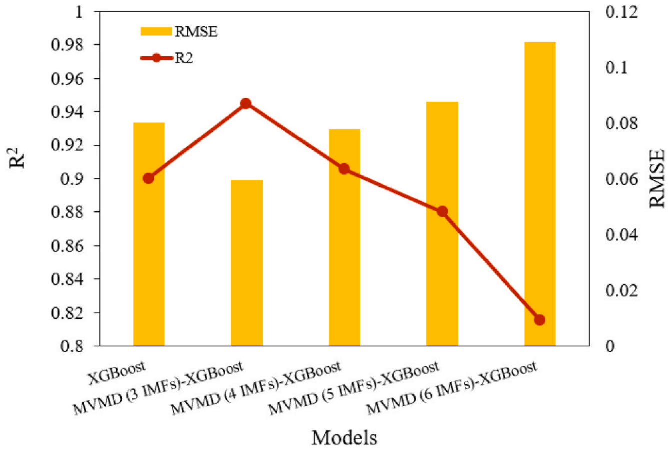

Fig. 11. Performance of MVMD-XGBoost models against XGBoost in the testing phase.

data into well-structured components, filtering noise, and isolating meaningful features to achieve better performance and more reliable predictions in complex and spatiotemporal datasets such as PLI of air dust. The RMSE values of models with more IMFs generally increased due to complexity and high dimensionality. The decomposing IMFs for non-stationary variables are given in Fig. 12.

The prediction accuracy of three models of GRU, MLP, and ANFIS based on decomposed signals of input variables, i.e., MVMD_GRU, MVMD_MLP, and MVMD_ANFIS, by using three IMFs in comparison with MVMD_XGBoost, is depicted in Table 5. Notably, the R^2 value for the testing set of all three MVMD_GRU, MVMD_MLP, and MVMD_ANFIS remains lower than MVMD_XGBoost. The consistent outperformance of MVMD_XGBoost across all metrics (R^2, RMSE, MAE, MSE) suggests this model can effectively handle nonlinear relationships between IMF components and PLI.

3.3. Predicting seasonal PLI using MVMD-XGBoost

As mentioned in the previous section, the ensemble model of MVMD-XGBoost demonstrated the best performance compared to other models for predicting PLI; therefore, this model was used to generate spatial maps for different seasons (Fig. 13). Overall, the predicted maps demonstrate a strong alignment with the observed maps to capture the general patterns and spatial variability in all seasons. In spring and summer, the predicted maps slightly overestimate the high PLI values. In autumn, while the model captures the general spatial trend, it underestimates the peak values observed in the central regions. For winter, the predicted map shows the closest agreement with the observed map and accurately represents the spatial distribution and range of values. These results highlight the model's effectiveness in reproducing seasonal spatial patterns, with minor limitations in capturing extremes.

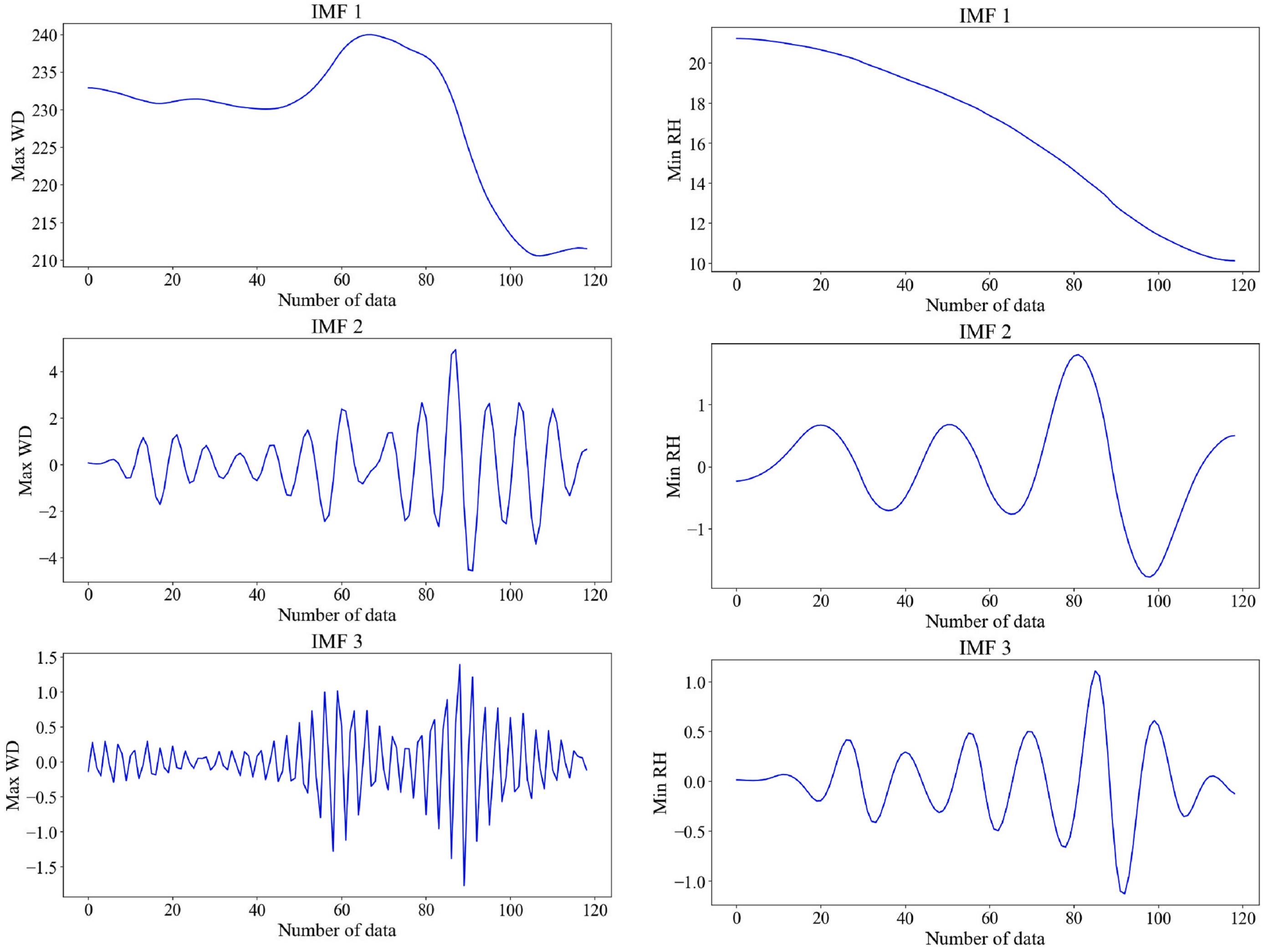

Fig. 12. Extracting IMFs using MVMD.

Table 5
Performance of the models based on the decomposed input variables in the prediction of spatiotemporal patterns of PLI.

Models	Train				Test			
	R^2	RMSE	MAE	MSE	R^2	RMSE	MAE	MSE
MVMD_XGBoost	0.99	0.001	0.001	0.00	0.94	0.06	0.05	0.00
MVMD_GRU	0.90	0.08	0.06	0.01	0.66	0.16	0.11	0.02
MVMD_ANFIS	0.75	0.13	0.10	0.02	0.53	0.17	0.13	0.03
MVMD_MLP	0.94	0.06	0.05	0.004	0.75	0.12	0.09	0.01

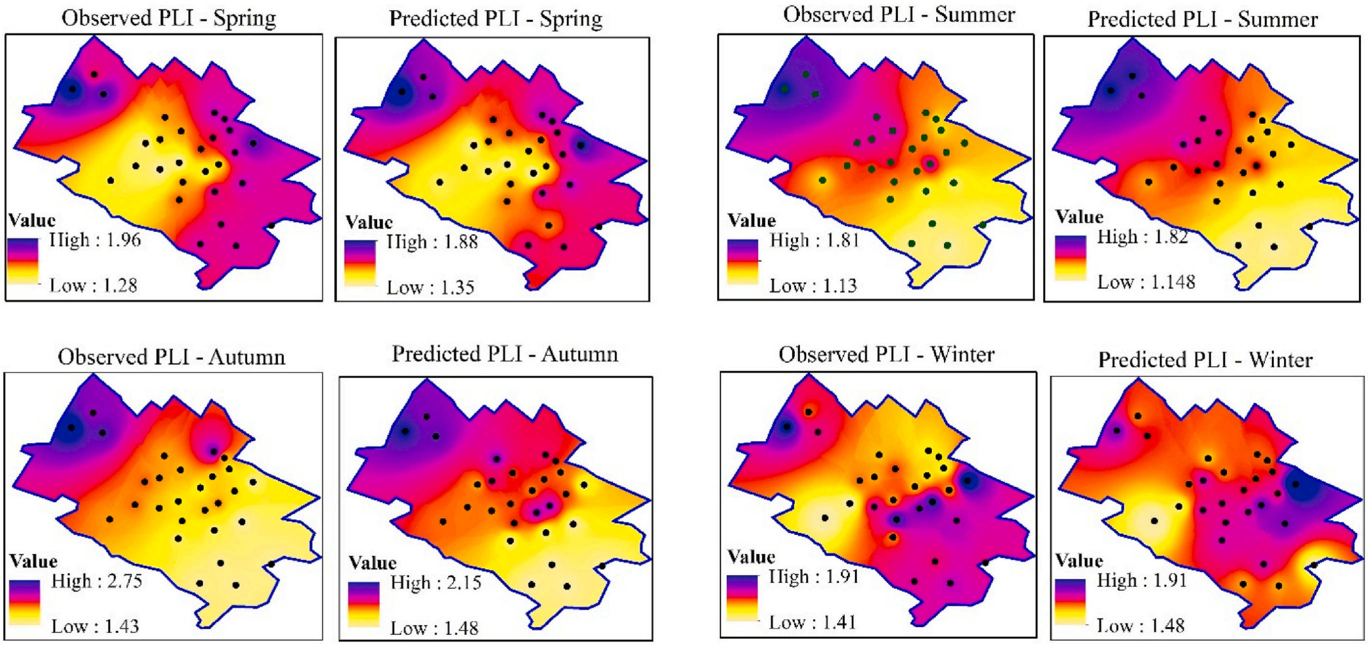

Fig. 13. Spatial map of predicted PLI using MVMD-XGBoost in different seasons.

To assess the level of pollution based on the PLI, a classification was recorded as denotes perfection (PLI <1), only baseline levels of pollution (PLI $= 1$), and deterioration of air quality (PLI >1) (Kowalska et al., 2018). As seen from observed maps, the PLI values across all seasons are consistently above one (PLI >1), which indicates a deterioration in air quality in the study area. The highest pollution levels are observed in autumn, with a maximum PLI value of 2.75, particularly concentrated in the central regions. Spring and summer exhibit moderate pollution levels, with slightly lower PLI ranges compared to autumn. Winter maps show a more uniform distribution of PLI values, with no extreme peaks but consistent air quality deterioration throughout the region. Therefore, decision-makers can focus on reducing pollution during autumn, where the highest PLI values are observed, and implement additional monitoring during autumn and winter. Especially, they can perform their decisions in hotspot regions identified in the maps (e.g., central and northern regions in autumn) to reduce localized pollution.

To gain better insights into the model's predictions, a temporal uncertainty analysis was also conducted (Fig. 14). The temporal uncertainty analysis for spring reveals that the predicted PLI values fall within the 95PPU for 83.33 % of the observed data points, with a d-factor of 1.57. The predicted trend closely follows the observed values, although minor deviations are evident in some peaks and troughs. This indicates that the model performs well in capturing the overall variability of PLI during spring, with acceptable uncertainty levels and good coverage probability. In summer, the predictive model exhibits the highest coverage probability of 96.67 %, with a d-factor of 1.14. The predicted PLI values show minimal deviation throughout the sample index. The narrower uncertainty bounds in this season demonstrate the model's strong predictive performance and reliability for summer PLI estimation. The autumn analysis shows the lowest coverage probability of 63.33 %, with a d-factor of 1.01, reflecting relatively higher deviation between observed and predicted values. Although the predicted PLI generally follows the trend of the observed data, several observed points lie outside the uncertainty bounds, particularly for higher values. This suggests that the model underestimates extreme values of PLI in autumn, indicating a need for further refinement to improve performance in this season. For winter, the model demonstrates a coverage probability of 68.97 % with the highest d-factor of 2.81. While the predicted values capture the general trend of observed PLI, the wider uncertainty bounds highlight greater variability in predictions. Peaks and troughs in the observed data are less accurately captured that suggest challenges for predicting air dust pollution in winter due to complex environmental or anthropogenic factors influencing PLI.

4. Research limitations and prospects

A key limitation is the dependence on data solely from Yazd city, which may limit the wider relevance of the conclusions. However, Yazd's specific environmental context, characterized by its arid climate, diverse industrial activities, and unique geography, makes it an ideal case study for developing and testing predictive models suited to these conditions. Expanding this research to additional regions is a valuable next step; however, it requires access to extensive, high-resolution, and continuous datasets.

The roof-based sampling method used in this study has limitations. One major concern is its inability to fully capture near-surface dust dynamics, which are essential for evaluating human exposure to air pollutants at ground level. Its limitations indicate that it should be

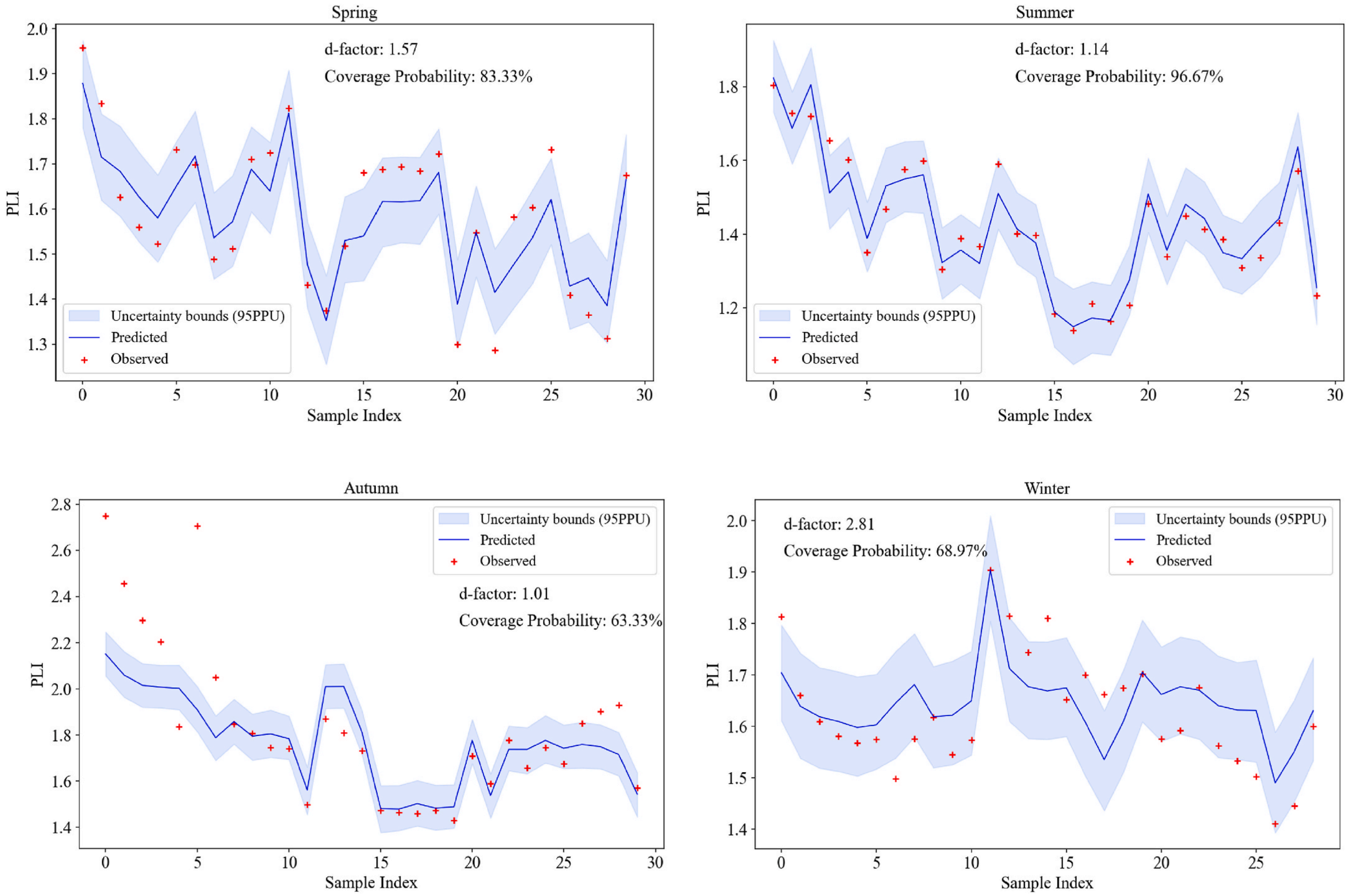

Fig. 14. Temporal uncertainty analysis of predicted PLI using MVMD-XGBoost.

combined with additional sampling strategies in future research to achieve a more thorough understanding of atmospheric dust pollution.

In addition, expanding the temporal scope of datasets and integrating satellite-derived pollution data, such as aerosol optical depth from MODIS, could improve the model's spatial-temporal resolution, especially in areas with limited ground monitoring networks.

While the current ML framework shows promise, exploring alternative ML/DL techniques and advanced parameter optimization methods could further improve predictive accuracy.

By addressing these limitations, future studies can improve the reliability and scalability of airborne heavy metal index predictions, ultimately supporting precision environmental monitoring, data-driven policy formulation, and sustainable land-use planning in dust-prone regions. Future work could investigate how the model adjusts to climate change scenarios, such as shifting wind patterns and increasing temperatures. This would enable predictions of PLI under future environmental conditions and support long-term urban resilience planning.

5. Conclusion

This study introduces a novel machine-learning framework to predict the spatiotemporal distribution of the PLI using cost-effective and readily accessible data. Three categories of datasets, including meteorological variables, heavy metal concentrations of roof dust, and distance to pollution sources, were used. By integrating advanced feature selection techniques such as Boruta, SHAP, and wavelet coherence with models like GRU, ANFIS, MLP, and XGBoost, the research highlights the effectiveness of BFSA and XGBoost for capturing complex relationships in air dust pollution. The optimal input combination, including Max WD, Min RH, Cd, and Zn, achieved superior predictive accuracy ($R^2 = 0.90$, RMSE = 0.06) without imposing heavy computational demands. In addition, the incorporation of MVMD for handling non-stationary variables further enhanced model robustness and was used for temporal uncertainty analysis across seasons. The findings indicate notable seasonal fluctuations in PLI, with the highest pollution levels observed during autumn and winter, and show uniform distribution. It was evident that specific actions are required during high pollution periods, with particular attention to hotspots like the central parts of Yazd.

The temporal uncertainty analysis further validates the reliability of the proposed MVMD-XGBoost model, particularly for spring and summer predictions. Decision-makers can use the findings of this study to implement cost-efficient monitoring strategies and develop practical measures to mitigate air pollution. This research not only bridges the gap in using low-cost meteorological and environmental variables for air dust heavy metal pollution prediction but also sets the stage for future studies to explore scalable solutions for urban air quality management in resource-limited settings. The proposed methodology provides a replicable framework for enhancing air pollution prediction and advancing environmental health strategies.

CRediT authorship contribution statement

Akram Seifi: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Somayeh Soltani-Gerdefaramarzi:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Conceptualization. **Mumtaz Ali:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apr.2025.102654.

References

Aldrees, A., Khan, M., Taha, A.T.B., Ali, M., 2024. Evaluation of water quality indexes with novel machine learning and SHapley Additive ExPlanation (SHAP) approaches. J. Water Proc. Eng. 58, 104789.

Amr, M.A., Helal, A.F.I., Al-Kinani, A.T., Balakrishnan, P., 2016. Ultra-trace determination of 90Sr, 137Cs, 238Pu, 239Pu, and 240Pu by triple quadruple collision/reaction cell-ICP-MS/MS: establishing a baseline for global fallout in Qatar soil and sediments. J. Environ. Radioact. 153, 73–87.

Boya, C., Ardila-Rey, J., 2020. A method for weather station selection based on wavelet squared coherence for electric load forecasting. IEEE Access 8, 197431–197438.

Bui, D.T., Pradhan, B., Lofman, O., Revhaug, I., Dick, O.B., 2012. Spatial prediction of landslide hazards in Hoa Binh province (Vietnam): a comparative assessment of the efficacy of evidential belief functions and fuzzy logic models. Catena 96, 28–40.

Chabuk, A., Hammood, Z.A., Abed, S.A., Kadhim, M.M., Hashim, K., Al-Ansari, N., Laue, J., 2021. Noise level in textile industries: case study Al-Hillah textile factory-company for textile industries, Al-Hillah-Babylon-Iraq. In: IOP Conference Series: Earth and Environmental Science, vol. 790. IOP Publishing, 012048, 1.

Chang, W., Chen, X., He, Z., Zhou, S., 2023. A prediction hybrid framework for air quality integrated with W-BiLSTM (PSO)-GRU and XGBoost methods. Sustainability 15 (22), 16064.

Chen, B., Stein, A.F., Castell, N., Gonzalez-Castanedo, Y., De La Campa, A.S., De La Rosa, J.D., 2016. Modeling and evaluation of urban pollution events of atmospheric heavy metals from a large Cu-smelter. Sci. Total Environ. 539, 17–25.

Chen, P.H., Chang, P.Z., Hu, Y.C., Luo, T.L., Tsai, C.Y., Li, W.C., 2024. On the robustness and generalization of thermal error models for CNC machine tools. Int. J. Adv. Des. Manuf. Technol. 130 (3), 1635–1651.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Chen, Z., Zhao, H., Zhang, Y., Shen, S., Shen, J., Liu, Y., 2022. State of health estimation for lithium-ion batteries based on temperature prediction and gated recurrent unit neural network. J. Power Sources 521, 230892.

De Nevers, N., 2010. Air Pollution Control Engineering. Waveland press.

Dehghani, M., Seifi, A., Riahi-Madvar, H., 2019. Novel forecasting models for immediate-short-term to long-term influent flow prediction by combining ANFIS and grey wolf optimization. J. Hydrol. 576, 698–725.

Ebrahimi-Khusfi, Z., Nafarzadegan, A.R., Dargahian, F., 2021. Predicting the number of dusty days around the desert wetlands in southeastern Iran using feature selection and machine learning techniques. Ecol. Indic. 125, 107499.

García, M.V., Aznarte, J.L., 2020. Shapley additive explanations for NO2 forecasting. Ecol. Inform. 56, 101039.

Gholami, H., Mohamadifar, A., Collins, A.L., 2020. Spatial mapping of the provenance of storm dust: application of data mining and ensemble modelling. Atmos. Res. 233, 104716.

Gholami, H., Mohammadifar, A., Golzari, S., Kaskaoutis, D.G., Collins, A.L., 2021. Using the Boruta algorithm and deep learning models for mapping land susceptibility to atmospheric dust emissions in Iran. Aeolian Research 50, 100682.

González-Rojas, C.H., Leiva-Guzmán, M., Manzano, C.A., Araya, R.T., 2021. Short-term air pollution events in the Atacama desert, Chile. J. S. Am. Earth Sci. 105, 103010.

He, Z., Liu, P., Zhao, X., He, X., Liu, J., Mu, Y., 2022. Responses of surface O3 and PM2. 5 trends to changes of anthropogenic emissions in summer over Beijing during 2014–2019: a study based on multiple linear regression and WRF-Chem. Sci. Total Environ. 807, 150792.

Hu, J., Chen, Y., Wang, W., Zhang, S., Cui, C., Ding, W., Fang, Y., 2023. An optimized hybrid deep learning model for PM2. 5 and O3 concentration prediction. Air Qual. Atmos. Health 16 (4), 857–871.

Kabata-Pendias, A., 2011. Trace Elements of Soils and Plants, fourth ed. CRC press, Boca Raton, pp. 28–534. Taylor &Francis Group.

Khanal, R., Furumai, H., Nakajima, F., 2015. Characterization of toxicants in urban road dust by toxicity identification evaluation using ostracod Heterocypris incongruens direct contact test. Sci. Total Environ. 530, 96–102.

Kow, P.Y., Wang, Y.S., Zhou, Y., Kao, I.F., Issermann, M., Chang, L.C., Chang, F.J., 2020. Seamless integration of convolutional and back-propagation neural networks for regional multi-step-ahead PM2. 5 forecasting. J. Clean. Prod. 261, 121285.

Kowalska, J.B., Mazurek, R., Gąsiorek, M., Zaleski, T., 2018. Pollution indices as useful tools for the comprehensive evaluation of the degree of soil contamination–A review. Environ. Geochem. Health 40, 2395–2420.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the boruta package. J. Stat. Software 36 (11), 1–13.

Leng, X., Qian, X., Yang, M., Wang, C., Li, H., Wang, J., 2018. Leaf magnetic properties as a method for predicting heavy metal concentrations in PM2. 5 using support vector machine: a case study in Nanjing, China. Environ. Pollut. 242, 922–930.

Li, C., Wang, Z., Li, B., Peng, Z.R., Fu, Q., 2019a. Investigating the relationship between air pollution variation and urban form. Build. Environ. 147, 559–568.

Li, J., An, X., Li, Q., Wang, C., Yu, H., Zhou, X., Geng, Y.A., 2022. Application of XGBoost algorithm in the optimization of pollutant concentration. Atmos. Res. 276, 106238.

Li, R., Wang, Z., Cui, L., Fu, H., Zhang, L., Kong, L., et al., 2019b. Air pollution characteristics in China during 2015–2016: spatiotemporal variations and key meteorological factors. Sci. Total Environ. 648, 902–915.

Li, W., Wu, H., Zhu, N., Jiang, Y., Tan, J., Guo, Y., 2021. Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). Information Processing in Agriculture 8 (1), 185–193.

Ling, N., Wang, Y., Song, S., Liu, C., Yang, F., Qi, X., et al., 2024. Experimentally validated screening strategy for alloys as anode in Mg-air battery with multi-target machine learning predictions. Chem. Eng. J. 496, 153824.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Mahesswari, G.U., Maheswari, P.U., 2024. SmartScanPCOS: a feature-driven approach to cutting-edge prediction of polycystic ovary syndrome using machine learning and explainable artificial intelligence. Heliyon 11, e39205.

McCartor, A., Becker, D., 2010. Top Six Toxic Threats, World's Worst Pollution Problems Report 2010. Blacksmith Institute, New York.

Olawoyin, R., Schweitzer, L., Zhang, K., Okareh, O., Slates, K., 2018. Index analysis and human health risk model application for evaluating ambient air-heavy metal contamination in chemical valley Sarnia. Ecotoxicol. Environ. Saf. 148, 72–81.

Park, J., Lee, W.H., Kim, K.T., Park, C.Y., Lee, S., Heo, T.Y., 2022. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Sci. Total Environ. 832, 155070.

Seifi, A., Ehteram, M., Soroush, F., 2020. Uncertainties of instantaneous influent flow predictions by intelligence models hybridized with multi-objective shark smell optimization algorithm. J. Hydrol. 587, 124977.

Seifi, A., Ehteram, M., Nayebloei, F., Soroush, F., Gharabaghi, B., Torabi Haghighi, A., 2021. GLUE uncertainty analysis of hybrid models for predicting hourly soil temperature and application wavelet coherence analysis for correlation with meteorological variables. Soft Comput. 25, 10723–10748.

Seifi, A., Pourebrahim, S., Ehteram, M., Shabanian, H., 2024. A robust multi-model framework for groundwater level prediction: the BFSA-MVMD-GRU-RVM model. Results Eng., 103250

Song, Y., Zhang, C., Jin, X., Zhao, X., Huang, W., Sun, X., Yang, Z., Wang, S., 2023. Spatial prediction of PM2. 5 concentration using hyper-parameter optimization XGBoost model in China. Environ. Technol. Innov. 32, 103272.

Sorooshian, A., Arellano, A.F., Fraser, M.P., Herckes, P., Betito, G., Betterton, E.A., Braun, R.A., Guo, Y., Mirrezaei, M.A., Roychoudhury, C., 2024. Ozone in the desert southwest of the United States: a synthesis of past work and steps ahead. ACS ES&T Air 1 (2), 62–79.

Szczygielski, J.J., Charteris, A., Obojska, L., Brzeszczyński, J., 2024. Capturing the timing of crisis evolution: a machine learning and directional wavelet coherence approach to isolating event-specific uncertainty using Google searches with an application to COVID-19. Technol. Forecast. Soc. Change 205, 123319.

Tao, C., Jia, M., Wang, G., Zhang, Y., Zhang, Q., Wang, X., et al., 2024. Time-sensitive prediction of NO2 concentration in China using an ensemble machine learning model from multi-source data. J. Environ. Sci. 137, 30–40.

Tao, H., Jawad, A.H., Shather, A.H., Al-Khafaji, Z., Rashid, T.A., Ali, M., et al., 2023. Machine learning algorithms for high-resolution prediction of spatiotemporal distribution of air pollution from meteorological and soil parameters. Environ. Int. 175, 107931.

Tao, Q., Liu, F., Li, Y., Sidorov, D., 2019. Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. IEEE Access 7, 76690–76698.

Tian, G., Qiao, Z., Xu, X., 2014. Characteristics of particulate matter (PM10) and its relationship with meteorological factors during 2001–2012 in Beijing. Environ. Pollut. 192, 266–274.

Wang, S., McGibbon, J., Zhang, Y., 2024. Predicting high-resolution air quality using machine learning: integration of large eddy simulation and urban morphology data. Environ. Pollut. 344, 123371.

Wang, S., Ren, Y., Xia, B., Liu, K., Li, H., 2023a. Prediction of atmospheric pollutants in urban environment based on coupled deep learning model and sensitivity analysis. Chemosphere 331, 138830.

Wang, Z., Chen, L., Chen, H., ur Rehman, N., 2023b. Monthly ship price forecasting based on multivariate variational mode decomposition. Eng. Appl. Artif. Intell. 125, 106698.

Wu, X., Yang, D., Gu, J., Wen, Y., Zhang, S., Wu, R., et al., 2021. High-resolution mapping of regional traffic emissions by using land-use machine learning models. Atmos. Chem. Phys. Discuss. 2021, 1–20.

Yang, J., Shi, B., Shi, Y., Marvin, S., Zheng, Y., Xia, G., 2020. Air pollution dispersal in high density urban areas: research on the triadic relation of wind, air pollution, and urban form. Sustain. Cities Soc. 54, 101941.

Yang, Y., Zhou, W., Wang, Z., Jiskani, I.M., Yang, Y., 2024. Accurate long-term dust concentration prediction in open-pit mines: A novel machine learning approach integrating meteorological conditions and mine production intensity. J. Clean. Prod. 436, 140411.

Zhou, H., Zhang, J., Zhou, Y., Guo, X., Ma, Y., 2021. A feature selection algorithm of decision tree based on feature weight. Expert Syst. Appl. 164, 113842.