International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 12 Issue 3 ǁ March 2024 ǁ PP. 355-364

A review of application status of machine learning in the environmental field

Zhenxiang Ji[1*], Wenxin Li[1], Shiguang Su[1]

[1] *School of Environment and Architecture, University of Shanghai for Science and Technology, Shanghai 200093, China*

Abstract

Environmental pollution has become a major problem that needs to be solved urgently, attracting the attention of scholars from all over the world. In order to better predict and alleviate this problem, scholars have explored more accurate methods for predicting pollutants in the environmental field through the practice of various methods. In recent years, machine learning has developed rapidly in the environmental field and has become one of the most popular methods. It is widely used in the monitoring and prediction of environmental media such as water, soil, and atmosphere. This article first introduces the advantages, disadvantages and principles of commonly used machine learning models. Then an overview of machine learning technology is summarized from different environmental media application fields. In addition, this article also summarizes some model input factors selected from the literature for pollutant prediction and explains the limitations that still exist in current machine learning. The research results can provide reference and have reference significance for relevant scholars in this field.

Keywords: *Environmental pollution, machine learning, input factors.*

Date of Submission: 12-03-2024 Date of acceptance: 27-03-2024

I. INTRODUCTION

Human beings, animals and plants cannot survive without the environment, and ecological and environmental governance is also an important part of national governance. The ecological environment includes the water environment, the soil environment and the atmospheric environment. Damage to any one of these environments can lead to a disruption of the ecological balance, which is bad news for plants and animals. According to the National Monthly Report on Surface Water Quality in November 2023 published by Ministry of Ecology and Environment of the People's Republic of China, the main pollutants in China's rivers and lakes are chemical oxygen demand, permanganate index and total phosphorus. The proportion of polluted surface water of category IV and above is 10.5%. Meanwhile, the main pollutants in China's soil environment are heavy metals, of which cadmium is the primary pollutant. According to the World Health Organization (WHO), as of 2019, the living environment of 99% of the world's population still does not meet the WHO's air quality standards. Therefore, it is imperative to improve and protect the ecosystem. In order to better monitor the hazards of the environment, researchers are using meteorological stations, satellite space stations and other measures to monitor the environment. At the same time, they are also combining advanced science and technology to predict pollutants in advance and take measures to prevent them, for example, by combining computers and artificial intelligence. Digital technologies represented by big data and artificial intelligence have contributed to ecological and environmental governance and green development [1].

As technology continues to advance, there is a huge amount of accessible and usable data available through the Internet and in a variety of domains. Computer applications have shifted from single data processing to artificial intelligence and machine learning, which was first introduced by Alan Turing in the 1950s along with human attempts to apply computers to the field of intelligence [2]. From the 1960s to the 1980s, artificial intelligence developed rapidly, and scientists began to engage in research on machine learning, neural networks, and other technologies. In 1985, the neural network model was proposed, and machine learning models began to be gradually applied to major fields.

Machine learning is a multi-disciplinary cross-discipline, involving computer science, probability theory knowledge, statistical knowledge, approximation theory knowledge and knowledge of complex algorithms. Its core is based on existing data and certain algorithmic rules to simulate the learning process of the human mind, and through continuous data "learning" to improve performance and make intelligent decision-making behavior [3]. Machine learning is the automation of human assistance by training algorithms on relevant data. The three main components of machine learning are supervised learning, unsupervised learning and deep learning [4]. Machine learning can identify trends in data that humans miss, replacing humans with continuous monitoring of data issues. It becomes more accurate with more training, and it is adaptive and self-learning.

This paper discusses the application of machine learning in the environmental domains of water environment, soil environment and atmospheric environment. For example, prediction of heavy metal concentrations in different environmental media, prediction of $PM_{2.5}$ concentrations in the atmosphere, prediction of physicochemical indicators in water and soil. This paper provides an overview of the various types of methods that have been applied in the environmental domain using machine learning. The paper provides an overview and analysis of commonly used models, applications in various environmental domains and future perspectives.

This paper classifies some of the existing commonly used machine learning models and describes the machine learning methods currently used for applications in the environmental domain. At the same time, this paper introduces the input and output factors that are often used by machine learning in the environmental field, which provides a better reference for subsequent research and can help researchers in this field to filter out the appropriate influencing factors. Finally, it outlines some of the current problems of machine learning in the environmental field and proposes the future development direction of this field.

II. Common machine learning models and evaluation metrics

Due to the continuous advancement of artificial intelligence, the current machine learning-based approaches to predicting air pollution fall into two main categories: non-deep learning and deep learning [5]. The non-deep learning is mainly divided into deterministic and statistical methods. Deterministic methods mainly use the principles of physical meteorology and statistical methods. It is based on atmospheric physical and chemical methods and uses dynamic data from monitoring stations to simulate the process of pollutant emission, transfer and dispersion, as well as the removal process [6]. Statistical methods use computers to construct probabilistic statistical models based on data and apply the models to predict and analyze the data. Deep learning is a powerful type of machine learning method suitable for big data processing, and has been widely used in image and speech understanding, natural language processing, and predictive [7]. Table 1 shows some of the commonly used machine learning algorithms presented in this paper.

Table 1 Advantages and disadvantages of common algorithms

Category	Model	Advantages	Disadvantages	References
Deterministic methods	CMAQ	Predictions are accurate and multiple air pollutants can be modelled at the same time	Time-consuming training	[8]
	Wrf-Chem	Predictions are accurate. Meteorological and chemical transport models are fully coupled in time and spatial resolution	Time-consuming	[9]
Statistical methods	SVM	Advantageous in solving small samples, non-linear and high-dimensional pattern recognition; simple algorithms with good robustness	Difficulty in solving multi-classification problems and sensitivity to missing data	[10-13]
	RF	It can handle high dimensional data, is fast to train, is simpler to implement, and is more capable of overfitting.	For small or low-dimensional data (data with fewer features), it may not produce a good classification.	[14]
	BPNN	It has strong nonlinear mapping ability, self-learning and self-adaptation ability, and strong generalization ability and fault tolerance ability.	The algorithm learning process is slow to converge and requires a long training time; it is highly dependent on the learning samples	[15, 16]
	LSTM	The problems of gradient vanishing and gradient explosion during long sequence training are solved.	There is a disadvantage in parallel processing and it is computationally time-consuming.	[17]

2.1 CMAQ and Wrf-Chem

Community Multiscale Air Quality (CMAQ) is one of the most widely used models for regional and urban air quality forecasting due to its multiscale and highly flexible forecasting capabilities [8]. CMAQ was first proposed by the US National Environmental Protection Agency in 1998, and can be used to forecast a wide range of pollutants according to the requirements of its own model.

The Weather Research and Forecasting model coupled with Chemistry (WRF-Chem) is a regional air quality model developed by the NOAA Prediction Systems Laboratory in the United States that couples meteorological and chemical models. It is currently used to input emission inventories into the model and combine them with meteorological conditions to simulate pollutant scenarios such as aerosolized pollutants [9] 。

2.2 SVM

Support vector machines were initially used only to solve binary classification problems. It is a machine learning method for dealing with nonlinear problems with supervised learning [18, 19]. Its main steps are as follows

Setting training samples $X_i,\ y_i$, where $X_i \in$ Rn, $y_i\ \in$ R, R is the real number field. The samples mapped to the multidimensional space are denoted by $\phi(X_i)$, and the corresponding nonlinear mapping is denoted by ϕ, which yields the following equation

$$y_i = w \cdot \phi(X_i) + b\#(1)$$

Where w is a variable weight, b is a bias value, and w and $\phi(X_i)$ are n-dimensional vectors. ξ, ξ^* are introduced as relaxation variables, and the constraint equation is

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{m} (\xi_i + \xi_i^*)$$

$$s.t \begin{cases} y_i - w^T\phi(X_i) + b \leqslant c + \xi_i^* \#(2) \\ w^T\phi(X_i) + b - y_i \leqslant \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geqslant 0 \end{cases}$$

$$(i = 1,2,3,\cdots,m)$$

where C is the penalty function, used to adjust the penalty beyond the relaxation variables. C is the penalty function, which is used to adjust the degree of penalty for exceeding the relaxation variables. Then the Lagrange multiplier method is applied to equation (3), and the equation is obtained as

$$L(w,b,\xi,\xi^*,a,a^*,r,r^*) =$$

$$\frac{1}{2} w^T w + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) - \sum_{i=1}^{m} a_i^* (y_i - w^T \cdot \phi(X_i) -$$

$$b - \varepsilon - {\xi_i}^*) - \sum_{i=1}^{m} a_i (w^T \cdot \phi(X_i) + b - y_i - \varepsilon - \xi_i) - ^{\#(3)}$$

$$\sum_{i=1}^{m} r_i - \sum_{i=1}^{m} r_i^*$$

Where a_i, a_i^*, r_i and r_i^* are Lagrange multipliers. Equation (3) is obtained by taking the partial derivatives of w and b and setting them to zero, and substituting them back into Equation (3). The dual of the above problem can be obtained by using quadratic programming. Using the quadratic programming optimisation algorithm, the optimal multipliers a_i^{new} corresponding to the parameters a_i and ${a_i}^*$ are calculated, and the prediction equation is constructed as follows.

$$f(X) = \sum_{i=1}^{m} (a_i^* - a_i) \phi(X_i)^T\ \phi(X_i) + b\#(4)$$

Where a_i^*, a_i is the Laplace operator, b is the bias value, and $\phi(X_i)\phi(X_i)^T$ is the kernel equation.

Support vector machine is currently used to predict solar and wind resources at different locations [20]. The SVM first maps the input data to a high-dimensional feature space using a kernel function and then performs linear regression in the feature space. SVM with simple structure has the advantages of good global optimality search, generalisation and adaptability [21, 22]. However, it is difficult to implement for large-scale training samples and there are difficulties in solving multi-classification problems. The optimal model is usually selected based on the fact that the correlation coefficient (R) between the output data and the observed data is high and the error generated during modelling is relatively small.

2.3 RF

The Random Forest algorithm was first developed by Breiman and is an integrated classifier based on decision trees. The algorithm uses multiple decision trees to train and predict samples. The model is created by randomly selecting training data from the initial data and the unselected data is called out of bag (OOB) [23, 24]. The Random Forest algorithm can handle a large number of input factors, while its learning process is fast. RF works by generating multiple decision tree models, each learning and making predictions independently, and finally combining them into a single prediction [25]. The construction of the model by setting the internal parameters (max_depth, min_samples_split and n_estimators, etc., for details see Table 2) in each study is now applied in various fields.

Table 2 Hyperparameters within the immediate forest algorithm section

Parameters	Definition
max_depth	Maximum depth of the tree
min_samples_split	Minimum number of samples required to split internal nodes
n_estimators	Number of decision trees
min_samples_leaf	Minimum number of samples required at leaf nodes
max_leaf_nodes	Maximum number of leaf nodes, as an integer

2.4 ANN

Artificial neural network (ANN) is a hybrid of supervised, unsupervised and reinforcement learning techniques. It is a data-driven model used to simulate the behavior of biological neural networks in the human brain [26, 27]. Currently, backpropagation neural network is one of the most widely used artificial neural network models, which is characterised by forward transmission of signals and backward transmission of errors [15]. After assigning the input and output data to the model, a special mapping is automatically created between the input and output data, allowing for further predictions.

Artificial neural network generally consists of three separate layers, including an input layer, an output layer and a hidden layer. The role of the input layer is to accept all the inputs. The role of the output layer is to present the final result. The role of the hidden layer is to define the layer between the input and output layers. In the hidden and output layers, each neuron sums the inputs and then applies a specific activation function to compute its output as detailed in Figure 1.

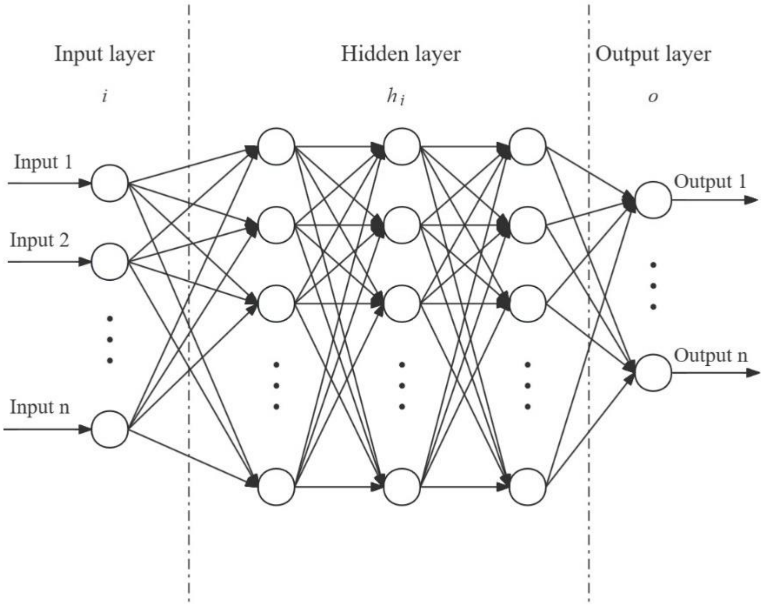

Figure 1 ANN model structure

2.5 LSTM

Currently Long-Short Term Memory neural networks consist of three main parts, which are input gates, output gates, forgetting gates [28]. Due to the nature of the Sigmoid function mapping the input to the (0, 1) interval, these three gates are used to determine the degree of embedding of the input cells, the degree of presentation of the short-term memory stream, and the degree of retention of the long-term memory stream, respectively. When the value of the sigmoid output is 1, it means that the information is completely retained. And when the output value is 0, it means that the information is completely discarded. When the value is between 0 and 1, it means that the information is preserved and deleted. When faced with time-series sensitive problems and tasks, LSTM is usually more appropriate [29]。LSTM has some advantages in sequence modelling problem, with long time memory function. LSTM is simple to implement, and solves the problem of gradient disappearance and gradient explosion in the long sequence training process. However, it has the disadvantage of parallel processing and is time-consuming to compute.

2.6 Common evaluation metrics for machine learning models

Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Correlation Coefficient (R^2) are commonly used in machine learning research to evaluate the performance of models [11, 30]. In the study, RMSE and MAPE were used to calculate the residual error and R^2 was used to assess the model fitting performance. Where the smaller the value of RMSE indicates better model performance. The closer the value of MAPE is to 0 means the smaller the error. The closer the value of R^2 is to 1, the better the model fit is.

Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Correlation Coefficient (R^2) are solved in Equations 5-7.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \#(5)}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y_i}|}{y_i} \times 100\% \#(6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \#(7)$$

Where i is the data point, n is the total number of data, y_i represents the actual value, $\widehat{y_i}$ represents the predicted value, and $\overline{y_i}$ represents the mean value of $\widehat{y_i}$.

III. Applications of machine learning

With the continuous development of artificial intelligence, machine learning has a wide range of applications in current society. It is mainly used in the fields of machine translation, image recognition, data mining and analysis, autonomous driving and intelligent transportation, and even the latest research is beginning to consider the use of deep learning to parse complex DNA sequence information [7]. In recent years, machine learning is gradually being applied to fields such as energy management (energy prediction, resource optimisation, etc.) and environmental protection. Since the atmosphere, water environment and soil contain a large number of harmful heavy metals, such as Cu (which causes abnormal liver metabolism), Zn (which increases Alzheimer's disease), and As (which causes brain tissue damage). Long-term exposure to these heavy metals is hazardous to human health, so it is particularly important to anticipate and prevent them. Therefore, the applications of machine learning in the environmental field mainly include water environment monitoring (predicting the content of heavy metals, organic microorganisms, etc., identifying and predicting water quality indicators and water levels, etc.) [31, 32], analysing pollutant content and distribution in soil [33] and predicting the concentration of atmospheric pollutants ($PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3, etc.) [5].

3.1 Soil environment field

In recent years, our economic level, science and technology have been increasing, but this has been accompanied by a series of soil problems. The progress of our society has led to harmful effects on agricultural production and people's lives. These include deteriorating soil quality, excessive concentrations of heavy metals and other pollutants in arable soils, the indiscriminate discharge of sewage from industrial and mining enterprises, and the indiscriminate discharge of industrial and agricultural waste gases [34]. Therefore, it is especially important for the treatment of soil pollution and the treatment of polluted soil industries. For the public, it is more important to understand the changes in soil pollution around us to prevent it. At present, scholars cite machine learning models on the concentration of pollutants in the soil and changes in physical and chemical indicators.

Sunori S K et al. used the support vector machine with different kernel functions to predict and evaluate the pH change in the soil to avoid the high impact of soil physical and chemical indicators on plants and animals. The results showed that the support vector machine with quadratic kernel function can better predict the change of pH in soil [35]. Uzair M and Ma W et al. used Extreme Learning Machine (ELM) for predictive assessment of organic carbon and heavy metals in soil to better identify changes in harmful heavy metals in soil. The results confirmed the feasibility of the method in predicting the concentration of pollutants in soil [36, 37]. Shen Chenchen et al. constructed machine learning models such as RF, SVM and multi-layer perceptron (MLP) to study the soil organic carbon content of three natural mixed forest species in central China [38]. Guan et al. used multiple linear regression to predict the concentration and distribution of soil heavy metals in Jiuquan based on soil heavy metal data and multispectral data [39]. Yang HR et al. constructed K-nearest neighbour, SVM and RF models to predict the spatial distribution of heavy metals in soil by using soil adsorption data and soil characteristics as input factors, and the results showed the feasibility of machine learning in this direction of data analysis [40]. Sun Y et al. constructed RF, BPNN and SVM models to predict Ni concentration in soil using hyperspectral eigenvalues and Ni concentration as input factors, and the results showed that the dimensionality reduction algorithm can improve the sensitivity of spectra [41]. Liu Jingyu et al. used multiple linear regression, RF and SVM to select soil physical properties (soil nutrients, soil acidity and alkalinity, normalised vegetation, soil weight, water content, etc.) and meteorological data as the input factors to predict the content of Cd in soil. The results showed that soil nutrients and normalised vegetation were the key variables [42].

In addition to the prediction of soil pollution levels or physical and chemical indicators, in order to better predict the crop yield and thus the soil environment for remediation and protection. At present, machine learning has also been widely used in agricultural farming. Zhou Xiuli et al. constructed multiple linear regression, RF and ANN models. Different machine models used, soil physical properties and soybean yield data were used as input factors to predict soybean yield. The results show that soil factors are more important for crop yield [43]. It also confirms the feasibility of machine learning in predicting crop yields, as well as identifying changes in soil physical properties and taking early steps for soil remediation.

3.2 Water environment field

The water environment is vital for the survival and development of all species around the globe. The combination of population growth, rapid economic and technological development, and chemical and agricultural production has led to a growing problem of water pollution problems, such as the discharge of organic dyes (at present, organic dyes are widely used in the cosmetic, textile, pharmaceutical and other industries, which has led to its penetration into the environment of various waters, and which, such as methylene blue, rhodamine B, etc., are harmful to the human body) and the discharge of industrial wastewater (at present, the effluent discharge from metal smelting, mining and other industrial activities has led to heavy metal pollution in water bodies, such as Cu heavy metal pollution in Shandong, the average concentration of Hg in the upper reaches of the Huangpu River, etc., which is harmful to human beings and the ecological environment) [44, 45]. Therefore, in order to save development costs, gain high efficiency and ease of operation, machine learning is gradually being applied to various aspects of the water environment.

Chen Yasong et al. constructed Random Forest regression model for the prediction of urban initial rainfall runoff pollution in the Yangtze River Basin. Urban characteristics (topography, climate, economy), average concentration of secondary rainfall runoff (COD, SS, TN, TP), precipitation characteristics (precipitation amount, precipitation duration, average rainfall intensity, dry period before rain), and subsurface characteristics (subsurface properties, imperviousness, and slope) were used as input factors. It was found that the prediction error was no more than 15%, and that precipitation, dry period before rain and impermeability of the subsurface were the key factors influencing the pollution of initial stormwater runoff in the Yangtze River Basin, as well as providing a new idea for predicting the pollution of initial urban stormwater [46].

With the increasing research on environmental pollutants, scholars have tried to use machine learning to predict pollutants in water quality. The results also demonstrate the feasibility and accuracy of machine learning in this field. Malygin E et al. used machine learning to simulate changes in heavy metal concentrations in Crimean river water and confirmed the high accuracy of machine learning in predicting heavy metal [47]. Jia X and Akshay R et al. used K-Nearest Neighbour algorithm (KNN), Bayesian algorithm and Decision Tree for water quality prediction comparisons, and the results showed that the accuracy of the prediction results of water quality using the above mentioned methods is high [48, 49]。

In addition to water quality prediction studies, scholars have also considered rainfall and other factors to predict water level changes. Zhu S L et al. chose Feedforward Neural Networks (FFNN) and LSTM to predict the water levels of 69 lakes in Poland. The results of the study showed that both FFNN and LSTM predicted the performance of the water level in 69 lakes in Poland with good prediction results [50]. Liang C et al. chose LSTM prediction to assess the water level of Dongting Lake in China, and found that LSTM performs better than Support Vector Machine (SVM) in prediction [51]. The above results also show that there are still some differences in the scope of application and prediction performance of each machine learning model.

3.3 Atmospheric environment field

In recent years, as scientists continue to study air pollutants, the harmful effects of each pollutant on biological health have led to increased attention to their prevention and control. At present, air pollution has become one of the problems to be solved. Air pollutants are significantly correlated with human throat disorders, and even long-term exposure to air pollution particles can lead to lung and cardiovascular disease [52, 53]. According to the World Health Organization, in 2019, 99% of the world's population lived in places that did not meet the levels of the World Health Organisation's air quality guidelines.

At present, the main sources of air pollution include fuel combustion, industrial emissions and traffic emissions. When emitted into the atmosphere, it endangers the health of human beings, animals and plants, increases the rate of disease and even cell cancer, and affects the growing environment of animals and plants, the quality of feeding and leads to diseases [54]. This in order to better monitor changes in air pollution, scholars use machine learning to make predictions for each air pollutant.

In recent years, some researchers have used neural network models to predict the assessment of $PM_{2.5}$. Wang Z et al. compared LSTM, RF and Cubist algorithms and found that LSTM obtained better accuracy and it is now becoming the most popular deep learning algorithm for time series prediction [55]. And as research continues and technology advances, while many methods utilise the learning capabilities of deep learning techniques to extract temporal or spatial correlations from air pollution data, they treat temporal and spatial correlations separately, ignoring the possible links between them. This results in poor performance of the single predictive model compared to the hybrid and improved models. Shi L et al. proposed the Balanced Social Long-Short Term Memory neural networks (BS-LSTM) to predict $PM_{2.5}$ concentration considering the spatio-temporal correlation of $PM_{2.5}$ concentration in Beijing and North China as a research object [56]. Bai Y et al. extended the LSTM model to hourly $PM_{2.5}$ prediction and simulation using the Beijing area as a research object, and proposed the ensemble long short-term memory neural network (EEMD-LSTM) [57]. Liu X et al. combined the sparrow search algorithm with the LSTM model to evaluate the prediction accuracy of the model at different time intervals in Shanghai. [6]。

In addition to the prediction of $PM_{2.5}$ concentrations, Liu H et al. used Nonlinear Auto-Regressive model with Exogenous Inputs Neural Network (NARX) with external inputs and SVM to predict SO_2, NO_2, and CO concentrations. The results show the feasibility of machine learning in predicting the concentrations of these pollutants. [10]。Cabaneros S M and Sayeed A et al. also used the machine learning approach to predict atmospheric O_3 concentrations. Cabaneros S M et al. used the hybrid model combining wavelet transform with LSTM model, where the wavelet transform preprocesses the data and the LSTM predicts the O_3 concentration [58]. Sayeed A et al. used Convolutional Neural Networks (CNN) to model the prediction of O_3 concentrations in Texas over the next 24h [59]. The results show the feasibility of machine learning methods such as CNN and LSTM for O_3 concentration prediction.

In addition to the diversity of output factors, the choice of input factors has been enriched. From the influence of meteorological data only, to the inclusion of air quality data, and even to the inclusion of magnetic parameters for the prediction of heavy metal concentrations in atmospheric particulate matter. The main reason for this is the good correlation between heavy metal content and magnetic parameters [60]. Xiao H et al. study used magnetic parameters as input factors for predicting heavy metal pollution in Xianlin, Nanjing [61]. And Salazar-Rojas T et al. constructed a support vector machine model for predicting heavy metal concentrations in street dustfall using magnetic parameters [62]. With the gradual enrichment of the input factors, the prediction model for each indicator will be improved and accurate. Some of the input and output factors in the literature are shown in the Table 3.

Table 3 Input and output factors selected for the application of machine learning to atmospheric pollutants

Input factors	Output factors	Model	References
$PM_{2.5}$, SO_2, NO_2, CO	SO_2, NO_2, CO	SVM, NARX	[10]
NO_2	NO_2	MLP, LSTM	[58]
SO_2, NO_2, O_3, CO, temperature, dew point, pressure, wind direction, wind speed	$PM_{2.5}$, SO_2, NO_2, CO, O_3	LSTM	[17]
$PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3, AQI	$PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3, AQI	RF, LSTM, MLP, XGBoost	[6]
PM_{10}, NO_2, CO, O_3, SO_2, weather, temperature, humidity, pressure, wind speed and wind direction	$PM_{2.5}$	SVM, LSTM	[11]
CO, PM_{10}, NO, NO_2, NO_x, NH_3, SO_2, benzene, toluene, temperature, wind speed and direction, rain fall, humidity	$PM_{2.5}$	LSTM, SVM, RF	[63]
$PM_{2.5}$, NO_2, SO_2, O_3, CO, χ_{LF}, χ_{ARM}, SIRM, χ_{ARM}/χ, χ_{ARM}/SIRM, and SIRM/χ_{LF}	Pb, As, Cd, Mn, Ti, Zn, Fe, V, Ni, Cu, Cr, Al, Co	SVM	[12]

IV. DISCUSSION

Although machine learning techniques have made great strides in predicting various pollutants or physicochemical indicators in the environmental field, there are still some problems to overcome. This section discusses the limitations of machine learning applications in the environmental field.

Firstly, the problem of missing data. A large amount of data is needed for machine learning model training to ensure the accuracy of the prediction results. But the datasets for the input factors in each prediction process are not all sufficient. For example, when local meteorological data are selected for the prediction of atmospheric pollutants, some of the meteorological data may be insufficient or missing due to the lack of local meteorological stations or underdeveloped systems. The lack of data can also be caused by situations such as when meteorological data collection stops during maintenance.

Second, the problem of data anomalies. Some of the data may be due to environmental or machine factors that cause that data to deviate abnormally from the true value. Therefore these abnormal data may affect the final prediction results of machine learning. Therefore it is important to establish a model that can accurately identify the abnormal data in the dataset.

Then, the problem of input factor selection. In addition to the commonly used input factors for the prediction of various pollutants, it is still worthwhile to explore whether factors such as soil utilisation, car ownership, population factors, etc. will have some influence on the prediction results.

Finally, the interpretability of machine learning. For traditional machine learning models such as linear regression, the models are more interpretable. However, for models such as deep learning and neural networks, their internal complexity leads to the input of noise into the model as well but the model prediction results still have high accuracy. This leads to non-interpretability of the model. So this part of machine learning models is also known as black box. It is important to study to explain their internal structure.

V. CONCLUSION

Current research shows that machine learning is rapidly evolving and its innovative applications in environmental prediction provide accurate and effective tools for environmental pollution problems in various countries. This paper reviews the applications of machine learning in the environmental fields (water, soil and atmosphere) and the commonly used techniques. Common machine learning models are classified and reviewed in terms of their characteristics, applications and limitations. The feasibility of machine learning for predicting various pollutant indicators in the environmental fields is also presented in terms of different application areas. The differences in the input and output factors chosen for different application areas in the process are also presented. Finally, the research problems that still exist in the current study are described. In the future, hybrid machine learning models can be developed and screened for key influencing factors to obtain more accurate research results. This paper summarizes the research progress of machine learning in predicting pollutants in the environmental field to provide a theoretical basis for researchers in this field.

REFERENCES

[1] Huang M, Zhang J, Yang Y, et al. A review of artificial intelligence applications in ecological environments (1989—2022) —— Knowledge graph analysis based on Citespace [J]. Environmental Protection Science, 2023: 1-12.

[2] Cooper S B, Leeuwen J V. Computing Machinery and Intelligence [M]. Boston: Elsevier, 2013.

[3] Wang H, Ye B, Feng J, et al. Application of Machine Learning in Steel Materials: A Survey [J]. Materials China, 2023, 42(10): 806-13.

[4] Mahadevkar S V, Khemani B, Patil S, et al. A Review on Machine Learning Styles in Computer Vision—Techniques and Future Directions [J]. IEEE Access, 2022, 10: 107293-329.

[5] Zhang B, Rong Y, Yong R, et al. Deep learning for air pollutant concentration prediction: A review [J]. Atmospheric Environment, 2022, 290.

[6] Liu X, Guo H. Air quality indicators and AQI prediction coupling long-short term memory (LSTM) and sparrow search algorithm (SSA): A case study of Shanghai [J]. Atmospheric Pollution Research, 2022, 13(10).

[7] Yan R, Liao J, Yang J, et al. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering [J]. Expert Systems with Applications, 2021, 169.

[8] Lu H, Xie M, Liu X, et al. Adjusting prediction of ozone concentration based on CMAQ model and machine learning methods in Sichuan-Chongqing region, China [J]. Atmospheric Pollution Research, 2021, 12(6).

[9] Zhang J, Li S. Air quality index forecast in Beijing based on CNN-LSTM multi-model [J]. Chemosphere, 2022, 308(Pt 1): 136180.

[10] Liu H, Wu H, Lv X, et al. An intelligent hybrid model for air pollutant concentrations forecasting: Case of Beijing in China [J]. Sustainable Cities and Society, 2019, 47.

[11] Du S, Li T, Yang Y, et al. Deep Air Quality Forecasting Using Hybrid Deep Learning Framework [J]. Ieee Transactions on Knowledge and Data Engineering, 2021, 33(6): 2412-24.

[12] Xiao H, Leng X Z, Qian X, et al. Prediction of heavy metals in airborne fine particulate matter using magnetic parameters by machine learning from a metropolitan city in China [J]. Atmospheric Pollution Research, 2022a, 13(3).

[13] Zhang J, Zhou J, Li Y, et al. Computer Simulating Effluent Quality of Vertical Tube Biological Reactor using Support Vector Machine [J]. ADVANCED RESEARCH ON INFORMATION SCIENCE, AUTOMATION AND MATERIAL SYSTEM, PTS 1-6, 2011, 219-220.

[14] Guan S, Zhang X, Zhao W, et al. A similarity distance-based space-time random forest model for estimating $PM_{2.5}$ concentrations over China [J]. Atmospheric Environment, 2023, 313.

[15] Wu D, Zhang D, Liu S, et al. Prediction of polycarbonate degradation in natural atmospheric environment of China based on BP-ANN model with screened environmental factors [J]. Chemical Engineering Journal, 2020, 399.

[16] Chen W, Chen H, Dai F, et al. Effluent water quality prediction model based on artificial neural network for wastewater treatment [J]. Water & Wastewater Engineering, 2020, 56(S1): 990-4.

[17] Seng D, Zhang Q, Zhang X, et al. Spatiotemporal prediction of air quality based on LSTM neural network [J]. Alexandria Engineering Journal, 2021, 60(2): 2021-32.

[18] Pande C B, Kushwaha N L, Orimoloye I R, et al. Comparative Assessment of Improved SVM Method under Different Kernel Functions for Predicting Multi-scale Drought Index [J]. Water Resources Management, 2023, 37(3): 1367-99.

[19] Zhang X, Zhang X, Wang C. Earthquake Magnitude Prediction Model Based on Support Vector Machine Optimized by Genetic Algorithm [J]. Journal of Hebei GEO University, 2023, 46(06): 41-6.

[20] Zendehboudi A, Baseer M A, Saidur R. Application of support vector machine models for forecasting solar and wind energy resources: A review [J]. Journal of Cleaner Production, 2018, 199: 272-85.

[21] Tian H, Zhao Y, Luo M, et al. Estimating $PM_{2.5}$ from multisource data: A comparison of different machine learning models in the Pearl River Delta of China [J]. Urban Climate, 2021, 35.

[22] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-97.

[23] Naghibi S A, Ahmadi K, Daneshi A. Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping [J]. Water Resources Management, 2017, 31(9): 2761-75.

[24] Schonlau M, Zou R Y. The random forest algorithm for statistical learning [J]. Stata Journal, 2020, 20(1): 3-29.

[25] Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1): 5-32.

[26] Gupta V, Mishra V K, Singhal P, et al. An Overview of Supervised Machine Learning Algorithm [Z]. 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART). 2022: 87-92.

[27] Di Salvo C. Improving Results of Existing Groundwater Numerical Models Using Machine Learning Techniques: A Review [J]. Water, 2022, 14(15).

[28] Raheja S, Malik S. Prediction of Air Quality Using LSTM Recurrent Neural Network [J]. INTERNATIONAL JOURNAL OF SOFTWARE INNOVATION, 2022, 10(1).

[29] Qi Y, Li Q, Karimian H, et al. A hybrid model for spatiotemporal forecasting of PM(2.5) based on graph convolutional neural network and long short-term memory [J]. Sci Total Environ, 2019, 664: 1-10.

[30] Pak U, Ma J, Ryu U, et al. Deep learning-based $PM_{2.5}$ prediction considering the spatiotemporal correlations: A case study of Beijing, China [J]. Sci Total Environ, 2020, 699: 133561.

[31] Lowe M, Qin R W, Mao X W. A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring [J]. Water, 2022, 14(9).

[32] Zhu S N, Lu H F, Ptak M, et al. Lake water-level fluctuation forecasting using machine learning models: a systematic review [J]. Environmental Science and Pollution Research, 2020, 27(36): 44807-19.

[33] Zhang Y, Lei M, Li K, et al. Spatial prediction of soil contamination based on machine learning: a review [J]. Frontiers of Environmental Science & Engineering, 2023, 17(8).

[34] Jia X. Soil Environmental Protection and Pollution Control Countermeasures [J]. Journal of Agricultural Catastrophology, 2023, 13(08): 50-2.

[35] Sunori S K, Kumar S, Anandapriya B, et al. Machine Learning Based Prediction of Soil pH [J]. 2021 5th International Conference on Electronics, Communication and Aerospace Technology, 2021: 884-9.

[36] Uzair M, Tomasiello S, Loit E, et al. Predicting the soil organic carbon by recent machine learning algorithms [M]. 2022.

[37] Ma W, Tan K, Du P, et al. Predicting soil heavy metal based on Random Forest model; proceedings of the 36th IEEE International Geoscience and Remote Sensing Symposium (IGARSS), F 2016 Jul 10-15, 2016 [C]. 2016.

[38] Shen C, Xiao W, Zhu J, et al. Characterization of soil organic carbon and key influencing factors of natural forests in Central China based on machine learning algorithms [J]. Sci Silvae Sin, 2023: 1-14.

[39] Guan Q, Zhao R, Wang F, et al. Prediction of heavy metals in soils of an arid area based on multi-spectral data [J]. J Environ Manage, 2019, 243: 137-43.

[40] Yang H, Huang K, Zhang K, et al. Predicting Heavy Metal Adsorption on Soil with Machine Learning and Mapping Global Distribution of Soil Adsorption Capacities [J]. ENVIRONMENTAL SCIENCE & TECHNOLOGY, 2021, 55(20): 14316-28.

[41] Sun Y, Chen S, Dai X, et al. Coupled retrieval of heavy metal nickel concentration in agricultural soil from spaceborne hyperspectral imagery [J]. JOURNAL OF HAZARDOUS MATERIALS, 2023, 446.

[42] Liu J, Li R, Liang Y, et al. Soil Cadmium Prediction and Health Risk Assessment of an Oasis on the Eastern Edge of the Tarim Basin Based on Feature Optimization and Machine Learning [J]. Environmental Science, 2023: 1-15.

[43] Zhang Q. Countermeasures on the Influence of Soil Fertilizers on the Quality of Agricultural Products [J]. Agricultural Mechanization Using & Maintenance, 2023, (09): 105-7.

[44] Zhu Y, Lyu M, Tian X. Research Progress on Detection Methods of Water Environmental Pollutants [J]. China Resources Comprehensive Utilization, 2022, 40(02): 127-9.

[45] Chi L. Present Situation of heavy metal pollution in water environment and discussion on detection technology [J]. World Nonferrous Metals, 2020, (19): 197-8.

[46] Chen Y, Hou X, Zhao Y, et al. Characteristics and Predictions of Initial Rainwater Runoff Pollution in the Yangtze River Basin Based on Machine Learning [J]. EnEng, 2023: 1-11.

[47] Malygin E, Lychagin M. Machine learning approach for simulation of heavy metal concentration in river water: the Crimean peninsula case study [J]. E3S Web of Conferences, 2020, 163: 06009

[48] Jia X. Detecting Water Quality Using KNN, Bayesian and Decision Tree [J]. 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), 2022: 323-7.

[49] Akshay R, Tarun G, Kiran P U, et al. Water-Quality-Analysis using Machine Learning [M]. 2022.

[50] Zhu S L, Hrnjica B, Ptak M, et al. Forecasting of water level in multiple temperate lakes using machine learning models [J]. Journal of Hydrology, 2020, 585.

[51] Liang C, Li H Q, Lei M J, et al. Dongting Lake Water Level Forecast and Its Relationship with the Three Gorges Dam Based on a Long Short-Term Memory Network [J]. Water, 2018, 10(10).

[52] Wang M, Wang X, Qu G, et al. Research progress on the correlation between air pollution and throat diseases [J]. Chin J Otorhinolaryngology-Skull Base Surg, 2020, 26(05): 599-602.

[53] Liu C, Chen R, Sera F, et al. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities [J]. New England Journal of Medicine, 2019, 381(8): 705-15.

[54] Wang R. Hazards and Treatment of Air Pollution in Environmental Engineering [J]. Journal of Agricultural Catastrophology, 2023, 13(02): 160-2.

[55] Wang Z, Zhou Y, Zhao R, et al. High-resolution prediction of the spatial distribution of $PM_{2.5}$ concentrations in China using a long short-term memory model [J]. Journal of Cleaner Production, 2021, 297.

[56] Shi L, Zhang H, Xu X, et al. A balanced social LSTM for PM(2.5) concentration prediction based on local spatiotemporal correlation [J]. Chemosphere, 2022, 291(Pt 3): 133124.

[57] Bai Y, Zeng B, Li C, et al. An ensemble long short-term memory neural network for hourly PM(2.5) concentration forecasting [J]. Chemosphere, 2019, 222: 286-94.

[58] Cabaneros S M, Calautit J K, Hughes B. Spatial estimation of outdoor NO_2 levels in Central London using deep neural networks and a wavelet decomposition technique [J]. Ecological Modelling, 2020, 424.

[59] Sayeed A, Choi Y, Eslami E, et al. Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance [J]. Neural Networks, 2020, 121: 396-408.

[60] Chen Y, Wang G, Chen J, et al. Magnetic Reponse of Heavy Metals Pollution in Urban Topsoil of Yangpu District, Shanghai City [J]. Bulletin of Soil and Water Conservation, 2017, 37(03): 28-34.

[61] Xiao H, Xu Y, Qian X, et al. Magnetic diagnosis of heavy metal pollution in atmospheric particulate matter (PM_1) from Nanjing city [J]. Acta Scientiae Circumstantiae, 2022b, 42(5): 74-82.

[62] Salazar-Rojas T, Cejudo-Ruiz F R, Calvo-Brenes G. Comparison between machine linear regression (MLR) and support vector machine (SVM) as model generators for heavy metal assessment captured in biomonitors and road dust [J]. Environmental Pollution, 2022, 314.

[63] Masood A, Ahmad K. Data-driven predictive modeling of PM(2.5) concentrations using machine learning and deep learning techniques: a case study of Delhi, India [J]. Environ Monit Assess, 2022, 195(1): 60.