PLANT SCIENCE TODAY
Vol 12(sp1): 01–17
https://doi.org/10.14719/pst.8514

eISSN 2348-1900

REVIEW ARTICLE

Assessing long-term environmental risks of treated paper mill effluent via machine learning models: A review

Deepan Kumar Krishnamoorthy[1], Sivasubramanian K[1], Dhevagi P[1], Kumaraperumal R[2], Sumathi C S[3] & Kshatriya T T[4]

[1]Department of Environmental Science, Tamil Nadu Agricultural University, Coimbatore 643 001, Tamil Nadu, India
[2]Department of Remote Sensing and Geographic Information System, Tamil Nadu Agricultural University, Coimbatore 643 001, Tamil Nadu, India
[3]Department of Physical Sciences and Information Technology, Agricultural Engineering College and Research Institute, Tamil Nadu Agricultural University, Coimbatore 643 001, Tamil Nadu, India
[4]Department of Soil Science & Agricultural Chemistry, Tamil Nadu Agricultural University, Coimbatore 643 001, Tamil Nadu, India

*Correspondence email - deepan.pgens2023@tnau.ac.in

Received: 25 March 2025; Accepted: 01 July 2025; Available online: Version 1.0: 10 October 2025

Cite this article: Deepan KK, Sivasubramanian K, Dhevagi P, Kumaraperumal R, Sumathi CS, Kshatriya TT. Assessing long-term environmental risks of treated paper mill effluent via machine learning models: A review. Plant Science Today. 2025;12(sp1):01–17. https://doi.org/10.14719/pst.8514

Abstract

With growing water scarcity, treated paper mill effluent (TPME) is increasingly reused in agriculture for its water and nutrient value. However, long-term use raises concerns about pollutant buildup in soil and water, potentially harming ecosystems and human health. This review explores how machine learning (ML) models can help predict and manage the environmental impacts of TPME over time. The objective is to assess the effectiveness of various ML techniques such as decision trees, neural networks and ensemble models in forecasting changes in soil and water quality due to TPME irrigation. We reviewed recent studies, datasets and real-world applications to evaluate the performance and limitations of these models. Findings show that ML offers clear advantages over traditional models by handling complex, non-linear data patterns and improving prediction accuracy. However, challenges remain, including data availability, quality and the complexity of model interpretation. This review highlights the potential of ML as a powerful decision-support tool for sustainable wastewater management. It also emphasizes the need for better data practices and collaboration between environmental scientists, policymakers and technologists. By integrating ML into regulatory frameworks, the paper industry can move toward safer, more sustainable effluent reuse. In bridging technology with environmental science, this study supports the adoption of ML driven solutions to enhance long term environmental monitoring and promote greener practices in industrial water management.

Keywords: impact; machine learning; prediction model; treated effluent

Introduction

It is stated that about 7000 Bgal of wastewater is being generated in paper and pulp industries which contains a lot of nutrients and pollutants together. As the world is facing severe water crisis, utilizing the huge quantity of wastewater with proper treatment helps to minimize the water shortages. While utilizing the treated wastewater for irrigation it is essential to assess the impact it causes on the environment. Current methods for assessing these consequences are static systems that may not adequately capture the dynamic and diverse nature of environmental systems. However, the advent of ML presents a one-time chance to improve forecasting and guide better adaptation efforts. A novel ML-CEEMDAN-LSTM hybrid model has shown consistent outperformance in predicting reclaimed water volumes across different seasons, crucial for urban water management (1). ML applications extend to various aquatic environments, including groundwater, sewage and drinking water, enabling automated error detection and water quality evaluation. The integration of ML in environmental monitoring provides nuanced insights and forecasts about water quality trends, potentially revolutionizing environmental policy-making and resource management (2)

An Effluent Treatment Plant (ETP) treats raw paper industry effluent to create treated effluent. After undergoing physicochemical analysis, this is either utilized for industrial reuse or irrigation. Input data on physicochemical characteristics and environmental impacts are produced by environmental monitoring over time. In order to conduct predictive analysis for sustainable effluent management, these datasets are processed using ML algorithms and artificial intelligence (AI). The workflow for assessing and predicting the environmental impact of treated effluents using ML models is illustrated in Fig. 1.

The more challenging use of ML algorithms in models by sophisticated problems in wastewater engineering for energy loss prediction and lateral outflow forecasting by the Regression Tree M5P, Bagging and Random Forest (RF) algorithm (3). For example, ML models are more precise in terms of hydrological process estimation compared to linear regression; however, they are never used for making conclusions, as they both show inconsistent results for different algorithms (4). Still, the potential application of ML algorithms for spring discharge forecasting looks rather promising, e.g. M5P, RF and SVR algorithms; moreover, for short-term predictions, M5P is the most appropriate algorithm and for

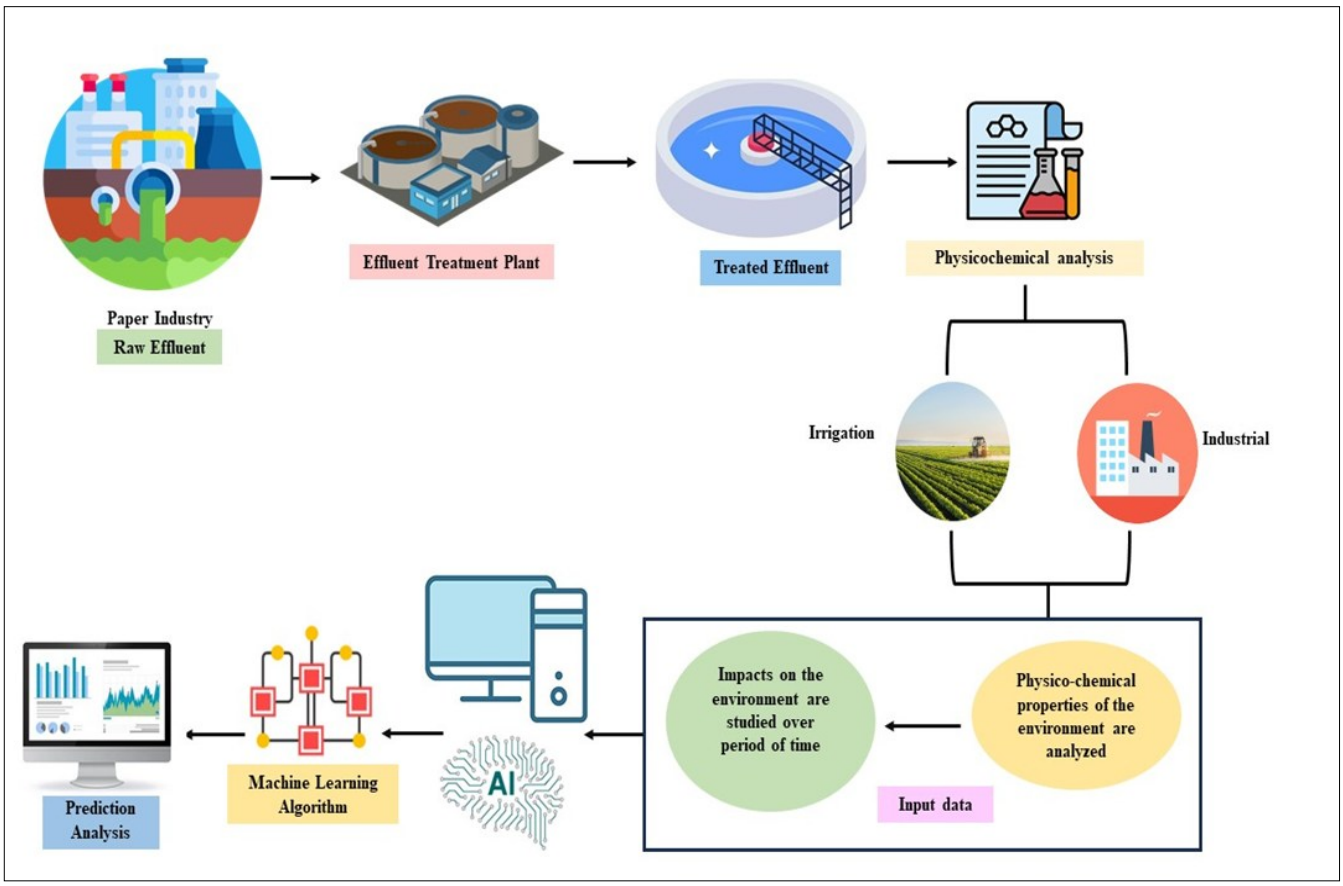

Fig. 1. Graphical workflow of effluent treatment and prediction of environmental impacts with AI/ML models.

medium-term predictions, RF algorithm can be used (3). Also, recommendations are made to apply a mixture of several ML algorithms by using purely data-driven to a more theory-based ML algorithm for hydrological studies depending on the specific situation and available data. This paper has attempted to investigate whether ML models can accurately forecast the effect of paper effluent discharge on the environment in the long run. To achieve this, it plans to combine literature as well as primary studies and datasets with current ML methods to build on what is currently known about causal connections between activities associated with paper production and their environmental consequences. After analysing several types of forecasting approaches from regression to classification and clustering models, this paper can provide the following recommendations to companies that focus on paper, governments and organizations that are responsible for the environment.

Numerous studies are emphasizing the significance of sustainability in particular about paper effluent discharge in the 21st century. The principle of sustainable development plays a critical role when selecting appropriate wastewater treatment techniques. This is because the chosen methods should not only be effective in removing pollutants but also minimize environmental impacts, reduce energy consumption and ensure long-term ecological balance. In this context, sustainable development serves as a guiding framework that emphasizes using eco-friendly, cost-effective and resource-efficient approaches in wastewater management decisions (5). The importance of sustainable development along with the need for sustainable outcomes and the role of the research community for this change has been highlighted. From these studies, it can be deduced that there must be more responsible disposal of wastewaters which comprise of paper materials (6).

Numerous studies highlighted below recognize the need for an integrated intradisciplinary and transdisciplinary approach to respond to the challenges of industry-environment-technology. A previous study emphasized the importance of ML and explained that the application of ML techniques can contribute to the environmental gains of steel production methods (7) while another study focused on how ML can be used within the framework of eco-innovation research. These studies collectively underscore the necessity of adopting a research technique that is both multi- and transdisciplinary as well as science-based to respond to the question of how to achieve balance between environmental and industrial development (8).

Bibliometric analysis

A bibliometric analysis of the research work was performed using the following public resources such as Google Scholar, Research Gate, GNKI, Wiley, PubMed and Springer. These sources were used to gather the literature related to ML. Open-source packages like biblioshiny and bibliometrix were used to conduct bibliometric analysis using VOS viewer software (9) and the results indicated that there are 317 literatures on the prediction of long-term impact of treated paper mill effluent using ML models. The keyword co-occurrence recorded that the terms “Machine learning”, “Effluent”, “Models”, “Wastewater treatment”, “Algorithm”, “Prediction” and “Deep learning” were used frequently (Fig. 2). Most of the literature was from the countries like China, India, USA, Saudi Arabia, South Korea. The search terms used included: “machine learning,” “treated effluent,” “paper mill,” “environmental impact,” “prediction models,” “wastewater treatment” and “deep learning.” The inclusion criteria consisted of peer-reviewed articles, conference papers and review papers published in English that directly addressed the use of ML for predicting the long-term effects of treated industrial effluent. Exclusion criteria included non

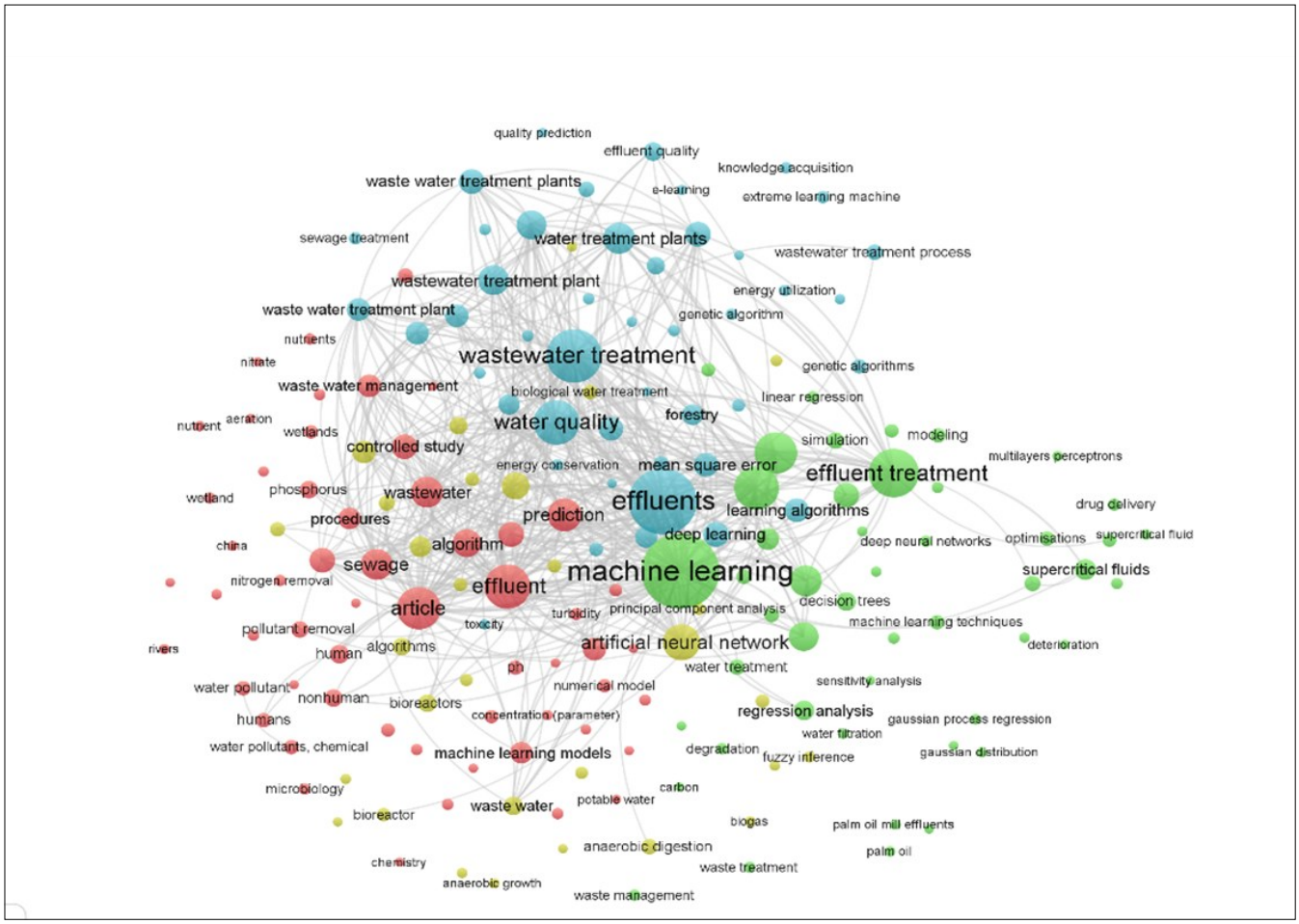

Fig. 2. Bibliometric analysis.

-English publications, unrelated industrial sectors and studies lacking ML methodology. The time frame for the literature search spanned from 2000 to 2024, ensuring the inclusion of both foundational and recent advancements.

Paper effluent - An overview

The pulp and paper industry is a major source of environmental pollution, with effluent containing high concentrations of chemicals and pollutants. Effluent is particularly problematic during the pulping process, where it can wreak havoc on the surrounding ecosystem. Effluent treatment processes such as aerobic biological treatment and anaerobic digestion are crucial for minimizing the environmental impact of paper production.

The pulp and paper industry utilizes a wide range of chemicals in its processing, including chlorine-based bleaching agents, dyes, sizing agents and additives to control pH and viscosity. These chemicals can enhance properties such as brightness, strength and printability, but also contribute to the complexity and potential environmental impact of the effluent. The use of chlorine and chlorine compounds in bleaching can result in the formation and release of chlorinated organic compounds, which can pose risks to the environment and food chain. Therefore, the industry must consider the environmental impact of these chemicals and explore alternative, more sustainable options.

The treatment and disposal of paper effluent present significant challenges due to its diverse composition, which includes organic matter, nutrients and chemical additives. Effluent treatment facilities employ various techniques such as sedimentation, filtration, biological treatment and chemical oxidation to remove or neutralize contaminants before discharge into the environment (10). The discharge of untreated or poorly treated paper effluent can have adverse effects on surrounding ecosystems and water bodies, including oxygen depletion and eutrophication. Advanced oxidation technologies such as ozonation and Fenton processes are effective in removing recalcitrant contaminants from paper effluent (10). Biological treatment methods including the use of microorganisms have also been explored for the removal of halogenated organic materials and decolorization of bleached kraft effluents.

The management of paper effluent, a complex mixture of organic and chemical substances is crucial for minimizing environmental impacts and ensuring industry sustainability (11). To address this, stringent regulatory requirements and advancements, in papermaking technology and wastewater treatment processes are being implemented (11). These efforts are part of a broader trend in wastewater treatment, which involves the removal of undesirable contaminants through physical, chemical and biological processes (11). However, the presence of emerging pollutants in treated effluents remains a challenge, necessitating the evaluation of innovative strategies for their removal (12). Similarly, the removal of metals during wastewater treatment is a key consideration, particularly in meeting more stringent discharge consents. In the context of silage effluent, which poses its own set of challenges, a multi-faceted approach is recommended, including the management of crop moisture content, infrastructure for effluent treatment and disposal and the use of natural treatment systems (13).

Sources of paper effluent in the industry

The paper industry's impact on the environment, particularly through effluent generation is a significant concern. Studies have shown that paper mill effluents contain high levels of pollutants, including total suspended solids (TSS), electrical conductivity (EC), chemical oxygen demand (COD), biochemical oxygen demand (BOD_5), phenols, Potassium (K) and NO_3-N (nitrate-nitrogen), which can harm aquatic life and the environment (14) (Table 1). The papermaking process also contributes to excessive greenhouse gas emissions and wastewater discharge. Efforts to address these issues include the development of green technologies and the promotion of a green economy in the printing industry (15).

The processing of wood pulp, a key component of paper manufacture is a complex and environmentally impactful process. The environmental challenges associated with this process, particularly in terms of effluent treatment and pollution prevention emphasizes the need for advanced treatment technologies to address the environmental impact of pulp-mill waste. while A recent study (14) underscores the importance of efficient treatment for non-wood fibre pulping effluents with alternative woody sources for pulp and paper processing to mitigate the environmental impact of wood depletion. Together, these studies collectively highlight the need for sustainable and environmentally responsible wood pulp processing in the paper industry. .

The environmental impact of traditional chlorine-based bleaching methods in the paper industry has prompted the exploration of alternative techniques. Ozone bleaching, followed by peroxide bleaching has been proposed as a more sustainable method for producing high-quality deinked pulp (16). Microbial xylanases have also been identified as a cost-effective way to reduce chlorine usage in the bleaching process while increasing brightness and reducing the discharge of chlorinated organic compounds (17). Control and optimization of bleaching reactions are crucial for ensuring product quality and minimizing operating costs. These studies collectively highlight the potential for more sustainable bleaching methods in the paper industry.

The papermaking process is a significant contributor to effluent generation due to its water-intensive operations and use of chemical additives. Effluent is laden with suspended solids and organic compounds, as well as chemical constituents from additives. The increasing concentrations of dissolved and colloidal substances in process waters pose challenges for managing wet-end chemistry. The industry's environmental impact is further exacerbated by high water consumption, solid waste generation and air emissions. To address these issues, strategies for sustainable water use, effluent reduction and pollution prevention are crucial.

The surface coating industry is undergoing significant changes to meet environmental regulations, with a focus on reducing emissions of volatile organic compounds (VOCs). Efforts to control the release of toxic effluents from the paper industry into the environment include chemical, biological and mechanical treatments, as well as bioremediation methods. The presence of emerging pollutants in reclaimed water from wastewater treatment plants (WWWTPs) is a major concern and innovative strategies for their removal are being explored (12). The textile dyeing industry is a significant source of surface water pollution, particularly due to dye-contaminated wastewater.

Benefits of predictive modelling in managing effluent impacts

Effluent management in WWTPs is another important area where use of predictive modelling offers several benefits that affect the plant's performance, operation, cost and its environmental implications. By use of artificial intelligence and internet of things, the functioning of treatment plants as well as the decisions that are made can be improved. The key benefits of predictive modelling in managing effluent impacts include several significant advantages in addressing the challenges associated with effluent treatment. Data analytics involves the use of large dataset to forecast equipment failures, chemical addition and treatment procedures. This way, the specific treatment plants can work to improve any weak links and thereby achieve better and optimum overall operations, which result in reduction of costs and general improvement of performance (18).

Cost reduction

In particular, through the use of predictive treatment plants will be able to follow, preventive maintenance measures, which in turn will limit cases of equipment breakdowns. Preventive actions have a great impact on repair costs, on distribution of resources and on increasing the high lifetime of the plant equipment; therefore could decrease the overall plant expenses.

Table 1. Typical concentration of key pollutants in paper mill effluent(15, 16)

Pollutant	Description	Typical concentration range	Environmental impact
BOD_5	Measures biodegradable organic matter	150 - 1500 mg/L	Depletes oxygen in water, harmful to aquatic life
COD	Measures total organic matter (biodegradable + non-biodegradable)	500 - 3000 mg/L	Indicates high organic load, potential toxic effects
TSS)	Undissolved solids in effluent	200 - 2000 mg/L	Reduces light penetration, affects photosynthesis
Phenols	Toxic organic compounds	0.5 - 5 mg/L	Toxic to aquatic organisms even at low concentrations
NO_3-N	Nitrogenous compound from breakdown of organic matter	1 - 10 mg/L	Contributes to eutrophication and algal blooms
K	Inorganic salt commonly used in pulping	10 - 50 mg/L	High concentrations can disrupt aquatic nutrient balance
Heavy metals (e.g., Cr, Cu, Zn)	May come from additives, inks, dyes	0.01 - 1 mg/L (varies by metal)	Bioaccumulative and toxic to aquatic life
EC	Indicator of dissolved ionic substances	500 - 3,000 µS/cm	Affects salinity and aquatic ecosystems

Environmental protection

Another application is to forecast the outcome rate of the treatment plants and thus achieve the quality of effluent that is allowed by the environment laws before the water is released. As a result, plants can avoid contamination, establish the balance of ecosystems and promote the use of water resources in activities for sustainable development.

Risk mitigation

Conducting a predictive analysis for designing the scenario and prognostics allow the treatment plants able to have a clear picture about the likely hazards including the impact of floods or cyclones, breakdown of the machines etc. Such a strategy helps plants to plan for contingencies, minimize the scenarios that they would violate the legal requirements and protect public health as well as the environment.

Data-driven decision-making

Predictive modelling puts in the hands of the treatment plant operator's information and intelligence that they can use with immediacy from the analysis of the data. AI algorithms along with IoT devices will help the operators to make the right choice, minimize time as well as effort and enhance the performance of the system with the help of predictive and trend analyses.

Overview of machine learning models

Machine learning is vital in today's approach to deal with the overwhelming data, making computer systems able to work out the reaction without being coded. Such models are trained with the help of machine learning algorithms with supervised, unsupervised or with both kinds of data in order to make them more effective in their predictions. This process entails inputting data into it and optimally fine-tuning the model's parameters for certain tasks until the development of a ML model. There are different categories of the ML models depending on the goals that are set such as classification model or prediction model. Supervised learning is the simplest type of learning and consists of establishing algorithms which receive labelled data in order to be able to determine how new data should be labelled or predicted. In actuality, unsupervised learning employs data sets that are not labelled such that it can search for patterns within the sets and this is especially helpful in uncovering patterns and relations of these models are Support Vector Machines (SVMs), Decision Trees, RFs and Neural Networks among others (19). SVMs are particularly efficient in data classification since it focuses on determining the best margins; on the other hand, Decision Trees offer an organized manner of estimating responses from the characteristics of the data. RFs is a much more sophisticated method and Neural Networks try to imitate a human brain and its nodes, where inputs are transformed to the wanted outputs (20) (Fig. 3).

When selecting the appropriate machine learning model some consideration that include the following aspects may be used; speed, accuracy, model complexity and interpretability based on the business or project being done. The picking of models may at times involve guess work as to which would be most appropriate for a specific task (Table 2). ML models can be classified into two main categories: the classification models, in which responses are members of a fixed set and the regression models, in which responses are observed as real numbers. ML models are significant in many different industries to transform the business reality and introduce advanced innovations to such branches as finance, marketing and retail chains. It is crucial for anyone who is interested in ML to understand the differences of the models and algorithms because this knowledge will help them greatly when they are working as a data scientist or as a professional using data in their daily jobs (21).

The role of machine learning in environmental prediction

Environmental prediction employs the utilization of the ML algorithm with new approaches being adopted to minimize error margins (19). These methods are especially useful in satellite data analysis, climatic modelling and evaluation of the environmental data and design. They have also been used in the contexts of air quality modelling, with an emphasis on the antecedent predictors to enhance the accuracy of forecasts (22). Nonetheless, the approach described above of applying ML in environmental pollution research has some challenges, such as the choice of model and interpretability, model selection and data accessibility (23). Nowadays, linear statistical analysis, time series analysis and deep learning models are applied to extract important information from environmental data. Sub-type of the ML field is the so-called deep learning, which is relevant in the case of environmental data due to the possibility of using big integrated environmental supervision.

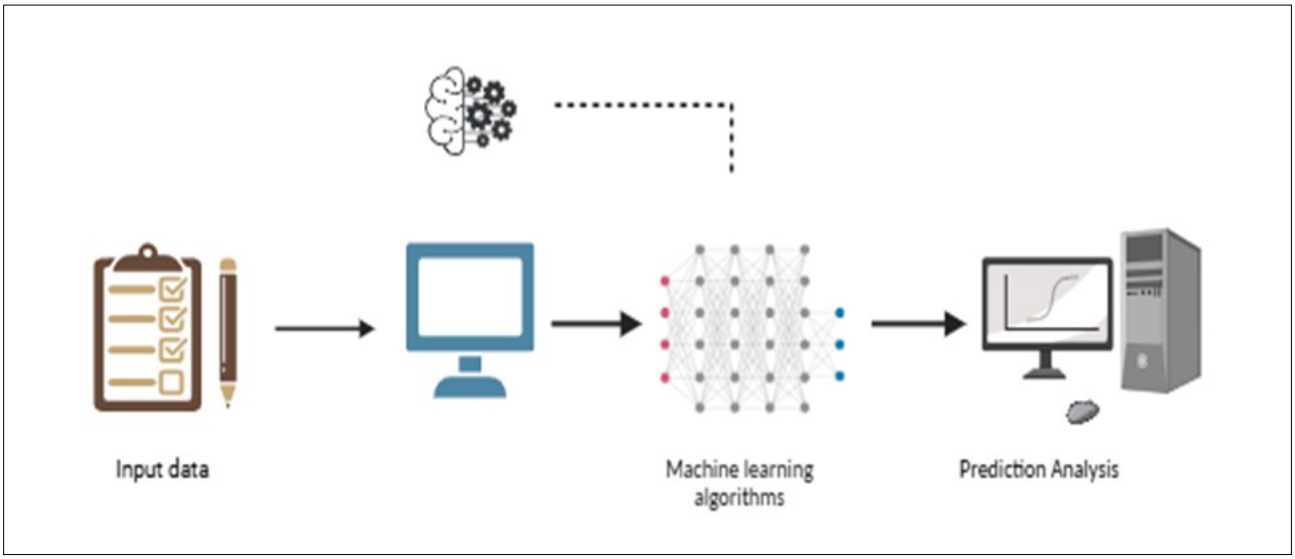

Fig. 3. Predictive ML model.

Table 2. Key data types and their uses in ML models (21, 44)

Data type	Description	Importance in ML modelling
User Navigation Data	Tracks user behaviour on websites to generate statistical data	High for user interaction analysis
Wastewater Inflow Volume	Monitors the volume of wastewater entering treatment plants	Critical for predicting treatment needs
Ambient Temperature	Measures surrounding temperature at treatment sites	Influences microbial activity and treatment efficiency

Artificial intelligence (AI), especially in the fields of toxicology is applied in increasing degrees for detection, prognosis and exploitation of environmental threats. Thus, there are several issues related to the organization of its work that needs to be addressed in order to enhance its efficiency (23). Some of the issues and recommendations that Zhu *et al*., raises pertain the general areas of concern in environmental datasets and ML models; issues such as inadequate preprocessing of data and lack of standard guidelines in the creation of the models (24). An approach to enhance the reliability and explainability of ML models for estimating the life cycle of chemicals has been described (25). A study highlighted recent applications and emerging trends in predictive toxicology, demonstrating that SVMs, RF and decision trees are among the most effective techniques (26).

A vast number of studies have been carried out to explore the use of both AI and ML techniques for environmental prediction especially in air pollution and sensor data. The artificial neural networks (ANNs) can also be used, out of which LSTM (Long Short-Term Memory) has been proven to provide a better solution for the air quality index case and the conclusion of employing the LSTM model, as it has high accuracy in predicting environmental parameters (27). A more enhanced spatial temporal LSTM model has been discussed to prove that the given approach is capable of modelling the complex dynamics of the environment with high accuracy (28). In sum, these works show how AI and ML use their resources towards environmental prediction, particularly where air pollution is concerned and with particular focus on sensors.

The use of ML is not only restricted to the research field in environmental prediction. It is also being used in practice to failure and bounded rationality in decision-making regarding climate change and environmental protection. Additionally, ML is enhancing many different systems and processes that influence the environment, ranging from using ML to estimate the better route for the supply delivery trucks that release little carbon to applying ML to find out how water can be used efficiently in farming (29-31). In conclusion, the role of ML is rising in environmental prediction because of the possibilities to improve the assessment of environmental factors. The purpose of these models is as wide as environmental toxicology, renewable energy, environmental monitoring and protection. But there are still issues to be solved, for example, lack of knowledge and technical difficulties in the sphere of data quality and model explanation. Nevertheless, the application of ML for environmental predictions can go a long way in assisting climate and environmental conservation.

Data sources and collection methods for model training

Data sources

When it comes to models decisions about data sources are crucial for models' success. Some common data sources include: Effluent discharge records, water quality monitoring data, meteorological datasets, soil and sediment quality data, biodiversity and ecotoxicological data, industrial process data and socioeconomic data.

Application data formats - for example, they can work with such data types as persons' medical records, bank transactions and registers of various connected objects such as IoT gadgets.

The data used should reflect the application as accurately as possible and should include all aspects of the issue at hand. Thus, enriching the data with such modalities as text, images, audio and video will help to cover the gap and make a model more robust (32).

Limitations of data sources - Inaccurate forecasts may result from the noise, biases and inaccuracies present in data sources. The idea of "dataset multiplicity" illustrates how test-time predictions are impacted by training data uncertainties, potentially having varying effects on various demographic groups (33). Although multimodal datasets are essential for creating adaptable AI systems, they present particular difficulties for evaluating quality and frequently don't fully integrate various modalities. Many cases can be resolved with a single modality, exposing shortcomings in existing datasets, according to a suggested two-step strategy for assessing multimodal datasets. To reduce biases and enhance model performance, these problems highlight the necessity of more thorough dataset documentation and analysis throughout production and use, combining qualitative and quantitative methods (34).

Data collection methods

As for the data collection technique, it greatly depends on the particular source of data as well as its intended application. Some common approaches include:

Web scraping and crawling (31) - Repeated processes of gathering information through websites which however, involves dealing with large amounts of somewhat unstructured data.

APIs (Application Programming Interfaces) - reads the data from the online services and databases, but it is constrained by APIs.

Crowdsourcing (35) - the phenomenon of outsourcing tasks that can be solved with the help of human intellect to a large number of people, but here the quality issue is very sensitive.

Sensors and IoT devices (36) as sources of raw data such as images, audio, location and time series data but includes additional difficulties in sensor noise and missing values.

Methods like Generative Adversarial Networks (GANs) to create synthetic data useful for augmentation and simulation.

Data preprocessing and cleaning

Depending on the type of data required, suitable data mining techniques should be applied to collect the necessary data for preprocessing and cleaning before feeding it into the model. Key steps in this process include:

I. Special attention must be given to handling missing values and outliers using feasible and realistic approaches to ensure data quality.

II. Data cleansing involves removing inconsistencies, correcting errors and eliminating irrelevant data.

III. Normalizing and scaling features to ensure that the data is on a comparable scale, which helps improve model performance.

IV. Encoding categorical variables into numerical formats so they can be effectively processed by machine learning algorithms.

In the beginning, huge datasets are grouped into training data set, validation data set and the test data set. The processes can somewhat differ depending on the nature of data being considered and depending on the design of a certain model type. For instance, if we have text data, it will require tokenization, padding or even truncation while on the other hand the image will be required to be resized, cropped or even augmented among other things.

The selection of the right data sources and methods of data collection is one of the most important activities when training ML models.. Thus, collected data should be relevant, accurate and sufficient to address the degree and impact of the issue under consideration. Data cleaning and preparation is also significant in order to create a model that learns from clean and well formatted data. This enables data scientists to develop robust generalization models that effectively apply to real-world scenarios (37).

Deep learning models for predicting long-term effluent effects

A range of studies have demonstrated the effectiveness of LSTM networks in predicting effluent parameters and key features in WWWTPs. The LSTM models, when compared to RNNs (Recurrent Neural Networks), achieved more precise predictions with lower RMSE (Root Mean Square Error) values. Another study reported that LSTM models outperformed other algorithms in predicting coagulant dosage and dissolved oxygen concentration respectively. A recent study further supported these findings, showing that LSTM-based models, particularly the exponentially-smoothed LSTM, were effective in forecasting key features of WWTPs. These studies collectively highlight the potential of LSTM networks in improving the accuracy and efficiency of effluent prediction in WWTPs (38).

A particular study employed simple RNN and LSTM structures were utilized to build models for predicting WWTP effluent parameters and the models' performance was evaluated systematically with regard to different training data situations and model structures. In the case of LSTM model, it was found that when the epoch is set to 50 and batch size to 100 gave the minimum training time and RMSE (39).

LSTM models were applied to predict the trend of COVID-19 infections in the United States, demonstrating the ability of LSTM to predict the long-term impact of epidemics (40, 41). Also, LSTM models have been used in the following areas such as concrete compressive strength for sustainable construction and energy load forecasting (42).

Other existing methods include CNN (Convolutional Neural Network), RNN and LSTM in a progressive learning higher order features of the raw inputs using traffic prediction. These models have been viewed to provide better performances than the other usual ML paradigms in handling the current complexity of transport systems (43).

The LSTM models can be used well for the long-term effluent effects from WWTPs due to the capability to identify the dependency patterns between the short-term and long-term ranges within the data set. Other factors such as tuning of the LSTM model's hyperparameters such as epoch and batch size can also enhance its results (38).

Machine learning algorithms for predicting treated effluent toxicity

Supervised learning algorithms have been employed significantly in the prognosis of the toxicity of treated industrial wastewater effluents. Such algorithms can assist in evaluating the possible effects exhibited by effluents on the environment and in strategies to control toxicity. Some of the most commonly used ML algorithms for toxicity prediction include:

SVM: Originally developed for pattern recognition, SVMs have since become one of the more commonly used algorithms in toxicity prediction employing toxicity endpoints such as hepatotoxicity, cardiotoxicity and carcinogenicity (44).

RF: The other common algorithm in the text classification is RF which has also been proven to be quite effective in toxicity prediction (44).

K-Nearest Neighbors (KNN): KNN has been used in the classification of drug induced immune thrombocytopenia toxicity with an area under a curve of 76. 9 % and the overall accuracy of 75. 6 % on an external validation dataset while it is discovered that the classifier using the binary words only scored 6 % better on an external validation set (31).

Ensemble Learning: Bagging among the best ensemble methods is the one that performs excellently well based on toxicity prediction's accuracy, correlation coefficient and other error measurements (45).

Deep Neural Networks (DNN): Thus, deep learning algorithms such as DNN have been employed to predict toxicity with reference to the fact that they are capable of learning complex patterns from the available data sets (44).

These algorithms' performance invariably hinges on considerations like the nature and quantity of the underlain dataset, the type of toxicity endpoint to be predicted and input molecular representation It has been identified that feature selection and model optimization are important in terms of accurate and reliable independency predictions or models.

Likewise, the efficient and effective analysis of the toxicity levels of industrial wastewater effluents has become relatively easy by way of the use of ML algorithms, which enhance the rate of control to enhance environmental protection (46).

Feature engineering for paper effluent prediction models

Feature selection activities are very important activities that ensure the generation of well accurate effluent prediction models for WWTPs. Thus, the indicated models enable efficient forecasting of effluent quality parameters, which is critical for managing plant operations and resources.

On the basis of influent water quality and some process control parameters, predictive models were established for effluent TN (Total Nitrogen) concentrations and total energy consumption in WWTPs (18). To improve the predictive accuracy of the F&A model (Feed and Aeration Model), advancement in the

ML techniques that include Bayesian optimization and ensemble bringing in to use RF (RF), SVM and Multilayer Perceptron (MLP) models (47).

Innovative approaches using ML models have been applied to predict wastewater quality parameters in both simulated and real-world WWTP scenarios (47). These models including RF, SVM and MLP have shown promising results in predicting TN levels and other effluent quality parameters, emphasizing the importance of high-quality data collection and understanding changes in WWTP operations for model accuracy.

Original studies implementing ML models for the accurate prediction of WCS (Water Quality Scenarios) in both randomly generated and existing served real-world WWTPs. Specifically, the RF, SVM and MLP models stated positive outcomes in forecasting TN levels and other parameters of the effluent quality underlining the necessity of qualitative data collection and identification of the changes in WWTP for enhancing the models' efficacy . The Permutation Importance (PI) and Partial Dependence Plot (PDP) methods have been applied to assess and interpret the relative influence of different features on the accuracy and effectiveness of predicting effluent quality parameters. From these analyses, it can be seen that there are high correlation coefficients between influent parameters that are followed by nitrogen contents and the consequent prediction results to suggest the importance of the feature aspect in model creating (18).

Also, the literature indicates that the application of the deep learning time series forecasting (DLTSF) with LSTM models is effective in predicting the efficiencies of the TSS, COD, BOD, ammonia and sulphide in WWTPs. These models such as deep cascade-forward backpropagation (DCB) networks have indicated high gains in accuracy for forecasting of effluent quality parameters, which are very useful in evaluating or predicting the WWTP performance (48).

The analysis of feature engineering is crucial when establishing effective models to predict effluents in WWTPs. These models can detect even the subtle changes in input variables and thus predict the effluent quality parameters effectively, which in turn can assist the WWTP in increasing its efficiency and serving the environmental conservation course.

In order to improve predictive models for WWTP effluent quality, feature engineering is essential. By improving input components, methods like factor analysis (FA) can improve the accuracy of ML models. Numerous models have demonstrated promise in forecasting effluent characteristics such as total nitrogen, nitrate nitrogen and COD, including RF, SVMs and LSTM (49). For certain parameters, these models' R2 values surpass 0.97, indicating their great accuracy. TSS and other influent parameters have a considerable impact on effluent projections, according to feature importance analysis. Furthermore, methods for reducing dimensionality and creating interaction terms can enhance model performance even further. These feature engineering techniques combined with ML techniques provide insightful information for improving WWTP management and operations (50).

The research field evaluates ML methods to forecast plant WWTP effluent quality. Different methods of selecting and engineering features have proven effective for improving model performance. The integration of ML models with advanced control strategies allowed them to enhance both operational efficiency and effluent quality results (51). A new predictive system based on Golden Jackal Optimization and a two-stage feature selection framework linked to the CNN-LSTM-TCN deep learning model for determining effluent total nitrogen has been created (52). The research demonstrates that choosing essential features along with their appropriate engineering results in better accuracy of predictive models for WWTP effluent quality assessment which enhances operational efficiency and adds to the quality of decisions.

Integrating machine learning models with environmental policies

Policy implications of predictive modelling

In different areas of work, predictive modelling has some policy conclusions of great importance. An environmental impact assessment can affect regulatory decisions by altering the perception of impacts concerning environmental affairs. It can help in the decision-making regarding health care, needed resources and cost control, especially taking into account the healthcare reforms (53, 54). However, the applicability of this model is subject to the regulation of the predictors as well as the assurance of the consequences of the model.

Several scholars have previously worked on the application of predictive modelling in environmental policy. As mentioned by a previous study, with an application of the predictive models it is possible to further improve the policy decisions and accuracy of the regulation monitoring; besides, the use of ML algorithms can increase the efficiency of the environmental law monitoring to the highest level. This is in concurrence with that furthering the issue is Task 3's future aim of clearing the outputs of these models to policy communities and citizens (49).

Across a range of industries, ML is being used more and more in policymaking because it provides notable improvements in decision-making and predictive modeling. In order to make better educated and effective policy decisions, machine learning algorithms can analyze large datasets, find trends and offer real-time insights. To minimize Clean Water Act (CWA) infractions and maximize resource distribution, ML-based systems have been utilized in environmental policy (55, 56).

But when ML is used in governance, issues of bias, transparency and data privacy arise, calling for stronger legal frameworks and better model explainability. When dealing with "prediction policy problems," which call for predictive inference as opposed to causal inference, ML approaches are especially helpful. AI and ML present promising predictive analysis solutions in environmental toxicology, facilitating more accurate and efficient environmental risk assessment along with supporting the development of evidence-based policy (57).

How ML can inform regulatory measures

ML models have displayed influence in boosting environmental regulation and policy analysis. One of the key strengths of these models lies in their ability to process large datasets and identify patterns, which can assist in pinpointing facilities that are most likely to face regulatory violations during inspections (55). In line with this, recently a study expounded further that it is equally vital to ensure that the outputs of these models are readily available and easily comprehensible by policymakers and the public. Conversely, the resource requirements of the evidence synthesis

methods might hinder the usability of ML in the decision-making process of conservation (54).

Supervised Learning Approaches for Effluent Prediction

Many researches employed supervised learning strategies in predicting and modelling the efficiency of WWTs (Wastewater Treatment Plants) particularly those dealing with the phosphorus removal. Among the implemented approaches is SVMs, which is a type of ML that is applicable to the data, which have nonlinear associations. The research focus of a study was to identify whether the use of least squares support vector machines (LS-SVMs) would enable the prediction of the total phosphorus content in the wastewaters discharged by the treatment plant and its potential to exceed a recommended threshold of 1.0 mg/L . The various kernels that were tested by the researchers included RBF (Radial Basis Function), polynomial and MLP and out of them the RBF kernel function provided the best results with 88.52 % of the classification rate. This shows that the developed SVM-based models are capable of accurately depicting the levels of compliance of the effluent quality to set standards (58-62).

Another recognized approach of the accrued technique in training models of the wastewater treatment processes is the neighborhood component analysis (NCA) that copes with the precisely supervised kind of learning. An example of an investigation that applied NCA as a tool aimed at identifying the correlations between the process factors and the efficiency of a papermaking wastewater treatment (63). NCA is a ML method that aims to acquire the distance measure that yields highest classification accuracy by nearest neighbor classifier. Used to recognize the primary process variables that define the treatment process rendering them decisive for NCA, it can shed a light on the critical factors that determine the efficiency of treatment process (60).

Although these methods give reasonable outcome, they are restricted with the requirement of labeled data, which could be difficult to get specifically for the city's complicated wastewater treatment system. In response to this, various methods of self-supervised learning have been investigated, whereby the ML models are capable of learning pertinent representations from data which is not labelled and then can be fine-tuned to provide estimations for certain feats.

In this regard, a study suggested a new hierarchical molecular graph self-supervised learning method for property prediction and the proposed method could be employed for predicting the behavior of chemical compounds present in the wastewater (61). In a similar manner, there is a self-supervised speech representation learning method known as (HuBERT) Hidden-Unit BERT (Bidirectional Encoder Representations from Transformers) that addresses the difficulties of speech data and the absence of a lexicon of input sound units during the pre-training phase. These unsupervised learning methodologies could perhaps be integrated to learn some valuable features of the WWTPs from the available data which in turn could be trained to predict certain functions (5).

The research work that also applied the self-supervised learning approach used for modelling spatial-temporal data such as crime forecasting was named the "Spatial-Temporal Self-Supervised Hypergraph Learning (ST-HSL)" framework (63). Using this hypergraph representation, the aforementioned complex relationships between different regions and time stages are incorporated. Furthermore, the learning process of the self-supervised learning framework is divided into two stages to learn the local as well as the global spatial and temporal patterns. While this framework analyzed crime occurrence, the possibility to reuse its components, viz, self-supervised learning and hypergraph representations, could potentially be beneficial for modelling WTW's spatial-temporal features.In the area of WWT, several supervised learning techniques like SVMs and NCA have been proven useful in the context of performance determination, particularly with respect to effluent quality. But due to the scarcity of labeled data, researchers have turned to self-supervised learning which could pave the way for better learning of these complex systems and enhancing the developing models for prediction (63).

Unsupervised learning methods for identifying effluent patterns

Effluent patterns are critical aspects in the case of environmental monitoring as well as wastewater management for the effectual control of industrial outfall on ecosystems. The use of unsupervised learning techniques is a much more effective approach when it comes to outlining and analyzing these patterns since they do not require the use of labeled data. Here, we explore some of the most important methods of unsupervised learning usually applied to the search for effluent patterns.

Clustering algorithms

Algorithms such as the K-means, DBSCAN and hierarchical clustering categorize data in segments with similarities within each segment in unsupervised learning (64). These algorithms are especially helpful for pattern analysis particularly in an effluent data set since they can categorize effluent data into different clusters, which reveal patterns and outliers that cannot easily be seen by simply observing the data (65). These algorithms play a significant role in data mining, especially in regards to the detection and analysis of large amounts of data at an early stage (66). However, depending on the nature of effluent data, certain algorithms might be preferable over the others, although each has its own advantages and limitations.

Principal Component Analysis (PCA)

PCA is quite useful in simplifying high-dimensional data or effluent data. Thus, moving variables to the new uncorrelated coordinate system, PCA can help to understand the data structure and relationships between variables. This makes it very useful when it comes to data analysis and in particular data visualization because one can display the relative positions of the data points in as many fewer dimensions as possible while preserving as much information as possible. However, in the case where the data distribution is not normal, then the data has to be transformed before PCA could be performed on it (68).

Self-Organizing Map (SOM)

SOMs are one of the most advantageous methods of non-hierarchical clustering of high-dimensions data used in different domains. For this purpose, SOM Toolbox is recommended especially for implementation in the MATLAB environment. In industrial engineering, SOMs are applied to process monitoring as well as modelling and have shown good results in, for example, the pulp, steel and paper industries. Through the application of SOM-ANN the water, soil and sediment quality in the petrochemical regions can be classified which then gives insight to remedial

actions (69). This body of work confirms the ability of SOMs in data organization and representation, as well as in the elucidation of spatial tendencies in the management of environments.

This paper reveals that there are numerous techniques and tools of unsupervised learning that help to identify effluent patterns in the environmental monitoring and wastewater management. Moreover, the use of clustering algorithms and dimensionality reduction techniques, anomaly detection methods, association rule mining and SOMs can help the researchers and practitioners to discover the aspects of effluent data along with the ways to use it more effectively in order to protect the environment and minimize potential threats to human health.

Comparative analysis of machine learning models

In the present world, ML has been identified as a robust technique for solving several problems in different fields. This is mainly because the use of ML techniques is still young; moreover, researchers and practitioners are always in a dilemma as to which technique is best suited for their problem. Such comparative analysis leaves many possibilities open to investigate the differences in interrelations, strengths and weaknesses of various ML models and applicability of their methods of long-term impact prediction (Table 3).

Several comparative studies have been made to analyze the effectiveness of various kinds of ML in numerous applications. The six studied ML models for landslide susceptibility modelling included extreme gradient boosting, RF, ANN, SVM, C4. 5 decision tree and naive Bayes (70). Based on the evaluations, it is revealed that the XGBoost model (XGB) was the best model in terms of accuracy and performance (71).

A more recent study summarized and compared three popular ML models, which include the Linear Model, the Forest Model and the SVM (72). This study discovered that the Linear Model is less succinct than the Forest and SVM models in terms of performance because of the former's inability to handle as many kinds of interactions as well as relationships and distinctive patterns. Forest and SVM models are relatively more complex and less explainable than the Linear Model.

However, the selection of the appropriate ML model that will yield more accurate and long-term impact prediction depends on several factors such as the type of the task to be solved, the data available as well as the level of explainability. For instance, if the relations in the problem are complex, non linear and the volume of data is large then models such as, RF or SVM might be more appropriate. On the other hand, where interpretability of model is a consideration and the problem at hand can well be solved by linear models the Linear Model could be a better option. It is imperative for long-term impacts prediction to compare the available or developed ML models in order to identify the one most suitable for the purpose. Although such algorithms as RF or SVMs may demonstrate better performance regarding the accuracy, these algorithms can be more non-transparent or intricate. This is because while the Linear Model is easy to interpret and thus more interpretable, its performance on complex problems may be lower compared to the complex models. Consequently, the selection of the most suitable model depends on the characteristics of the given problem and available resources (72).

Applications of ML in environmental science

Data analytics together with ML has been beneficial to environmental science averagely contributing more on opportunities and accessing techniques (19). They have been applied in the areas such as; climate, electricity usage and disaster management. It has also been implemented in areas like environmental and water management, where big data and processes have been used in data-aided methods (73). But, when it comes to adopt such methods in environmental science and engineering, one must focus on how to construct the models, how to interpret the results and how to evaluate the applicability of the models. An essential area of the use of the ML in Environmental Science in the production of forecast / predictions of weather conditions. The refined algorithms open up the possibilities of ML models with respect to enhance predictability based on intricate weather data. It is important when there is unexpected calamities and general management of resources in the context of evolving climate. Also, ML can be used for energy sustainability by using smart methods for energy consumption, finding patterns of energy use and integrating renewable energy into the system.

In this aspect, ML has proven to be a useful tool especially in the area of pollution control in the environment. ML can be employed for pollution change prediction and tracking, for water quality prognostication and for contaminant recognition (74, 75). A former study provides another example of how ML works to identify other high-risk facilities that should be targeted for an inspection to yield more value for the effort being expended (49). The proper and effective assessment of data and modelling of water quality is inevitable and ML can widely contribute to this step (70).

Hence, in the case of marine plastics and microplastics in particular, ML has proven to be rather useful, offering the prospect of improving identification and evaluating the environmental risks (75). Regarding the application of ML in maritime transportation, the environmental impact in ports can be reduced when it comes to emissions and energy (76). However, some of the challenges associated with ML in environmental pollution research, including air quality and post-consumer plastic recycling are interpretability of the model and data sharing (76). In the ocean studies also, ML was used in field studies like in the prediction of ocean weather and climate, modelling of habitats and identification of oil spills and pollution. These studies collectively show the capabilities of ML with an aim to solve some of the problems facing the

Table 3. Comparison of ML models for effluent prediction (79, 80)

Model	Correlation Coefficient (CC)	RMSE	Best use case	Computational cost	Interpretability
GPR	0.964 - 0.975	Low	Best overall performance	High	Moderate (complex kernel functions)
RF	0.932 - 0.943	Medium	Robust in varied data settings	Medium	High (feature importance available)
XGB	0.916 - 0.954	High	Fast processing times	Low	Medium (requires tools like SHAP)
LightGBM	0.883 - 0.890	High	Handling large data sets	Very Low	Low to Medium (requires post-hoc tools)

environment like the problem of marine plastics. Considering data analysis, the use of ML and Deep Learning is more extensive, as they allow discovering important information from big data sets. These findings can be used in decision-making in changes that have to be made concerning the environment, hence helping in resource optimization and policy implementation. With this ML's help, advancing information-centric investigation in ES, theoretical notions are supported by practice and bring visualization of the complex environmental problems, making the approaches to their solving more global (Table 4).

Limitations and challenges of ML in paper effluent prediction

ML has become an essential technique useful in many fields among them being environmental science, which can be used to predict results such as paper effluent characteristics. Nevertheless, as with any other field, there are limitations and complexities that are inherent in the use of ML in this aspect. In this discussion, the different challenges and limitations that are encountered during the use of ML in the prediction of paper effluent will be elaborated (Table 5).

The prediction of paper effluent properties using ML presents both technological and ethical issues. Training data bias can produce biased results, especially for underrepresented communities. Since only well-funded organizations may have access to advanced technologies, it is imperative to ensure equity in the adoption of ML. Data quality problems are one example of a technical barrier; effluent data is frequently inconsistent or lacking (77).

Concerns about model interpretability persist, particularly when intricate algorithms are used as "black boxes". Inadequate validation of models across many circumstances might lead to overfitting and generalization issues. Researchers provide solutions to these problems, including fairness-aware algorithms, openness policies and thorough data analysis to identify critical variables and minimal data needed for precise forecasting (78).

Data quality and quantity

One of the main issues in applying ML for paper effluent prediction is data: its deficiency and quality. It is an evident fact that the performance of ML models strongly depends on the amount and quality of data for training. However in paper effluent prediction, one of the big challenges is to collect detailed and representative data sets containing a wide range of effluent attributes. Restricted or skewed information can cause deviations in model performance and inhibit the learning capacity of ML algorithms (79, 80).

Feature selection and engineering

Another important aspect or component of ML is feature selection or feature engineering. Deciding which features would have the greatest impact on effluent characteristics contained in the paper and incorporating these into quantifiable inputs for the model is a difficult process. When dealing with paper effluent predictions, it is important to select the appropriate features that define and describe the effluent characteristics and tendencies. Lack of proper feature selection leads to poor performance of models, lower rate and accuracy of predictions (81).

Model interpretability

Another general issue of the ML models is the interpretability of the models. The predictions are often made in demanding decision-making areas, such as in an environment where everything needs to be clear and understandable. Some of the macro models used in machine learning, especially deep learning models are known to be 'black boxes'; hence, explaining the outputs of a model is not easy (82). In the case of paper effluent prediction, the models developed must give results that can be easily understood by the stakeholders especially to undertake necessary actions (83).

Overfitting and generalization

Numerous ML problems are affected by overfitting, including predicting characteristics of paper effluents (84). Different approaches have been suggested in order to reduce overfitting such as early stopping, pruning, augmentation and regularization (84). These strategies try to strike a balance between the model's complexity and its ability to generalize well across different scenarios as well as datasets, thus enhancing accurate prediction of effluent. Model simplification, reliable model assessment and data variety have been stated as significant strategies to beat overfitting in immunological applications (85). In another work, this importance has been elucidated by explaining that there must be a comprehension about the generalization equations in ML procedures for effectively combating overfitting (82).

Computational resources and scalability

ML algorithms and in general, taking large models like DNN that are computationally intensive to compute on can have a huge amounts of power consumption (23). Consideration should also be given to computational efficiency and scalability of models in certain cases such as paper effluent prediction where real or near-real time predictions may be needed (86, 87). Limited computational resources hinder the application of ML models to predict paper effluent generated in dotted settings due to computer science

Table 4. Summary of ML applications in environmental predictions (71)

Application area	ML technique	Outcome
Urban Air Quality	Regression Trees, Neural Networks	Accurate air pollution forecasting
River Water Quality	SVM	Enhanced pollution detection and remediation
Wildlife Population Dynamics	Satellite Image Analysis	Informed conservation strategies
Agricultural Impact	Predictive Modelling	Optimized farming practices for reduced runoff
Industrial Energy & Emissions	Pattern Analysis	Reduced energy consumption and emissions

Table 5. Summary of challenges in predicting environmental impacts (37)

Challenge category	Specific issues
Data Quality	Bias, inaccuracy, non-representativeness
System Complexity	Difficulty in modelling dynamic, complex environmental systems
Interpretability	"Black box" models, lack of transparency
Resource Intensity	High computational and energy costs
Ethical and Regulatory	Privacy concerns, bias, job displacement, regulatory compliance
Scalability and Generalization	Poor performance in new or varied environmental contexts

limitations (83). Considering all the above discussion, it can be concluded that while ML has great potential for prediction of effluent characteristics from papermaking; however large number of limitations and challenges need to be taken care in order to achieve fuller potentials. Addressing questions of data quality, feature selection, model interpretability and overfitting, as well as computational resources is a significant step toward creating ML models in predicting paper effluents that are strong and trustworthy (88).

Future directions for research and development

Emerging ML technologies

Current innovations in ML and DL (Deep Learning) have given a sign of hope in improving processes in water treatment. These technologies have been incorporated in different fields, some of which are WWTS in which they have been used in modelling processes, estimating technology performance and working conditions. Nevertheless, the application of ML in predictive control has made operations stable and decreased aeration in wastewater treatment processes (89). Moreover, combined with DL models like LSTM networks the prediction of major characteristics of WWTPs has presented relevant evolutions; which the implications in system availability and operational costs are highly significant (40). However, there are still issues like the lack of good data management and explainability of the models that need to be solved to unlock the full potential of these technologies (90) (Table 6).

Researchers currently emphasize the necessity of adding Explainable Artificial Intelligence (XAI) technology to ML systems that monitor and control WWWTP operations. The implementation of XAI techniques enables better understanding of complex models by enabling targeted process improvements which results in more efficient wastewater treatment (91). XAI implementation in compliance and regulatory models provides the necessary audit capabilities with traceability features needed to establish trust in AI systems. The improvement in XAI techniques becomes essential in environmental monitoring systems because it enables better understanding of model decision-making.

The importance of SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) methods as part of explainable AI (XAI) increases because they develop ML model interpretability in different application fields. Through these methods companies gain better understanding of model decisions which enables both bias detection and fairness enhancement. The interpretation of results becomes complex due to both selector choice and feature correlation with the selected model while following both processing techniques. The implementation of Explanation as a Service (XaaS) tools SHAP and LIME provides both model transparency and operator-AI system collaboration which can boost operational reliability and efficiency in different markets (91).

Potential for cross-disciplinary research

The incorporation of ML along with interdisciplinary techniques is the need of the hour to combat global issues like global warming and the depletion of resources. It can benefit climate analysis as well as teleconnections elucidation and upgraded weather alerts (91). It also has a critical application in prognosis or creating models based on data collected for assessment of the impact of climate change on the environment (92). A change from the linear model of scientific research and innovations to the sophisticated socio-technical systems of knowledge co-production, the “Triple Helix” and the “Quadruple Helix” require cross-boundary collaborations across disciplines (93-97). These models enable the societal frameworks to provide constant cooperation between academia, industry and government to promote long-term environmental innovation.

The next frontiers in effluent impact prediction

The science of using ML and AI to predict and counter the effects of effluents is still advancing and can grow. The models containing both process-based and ML-based synchronous ones seem to be the most promising since they can help increase the accuracy of estimations and computation speed, taking into account the data that is still insufficient to comprehend (98). They are also being progressively applied to environmental governance such as the scheme for prediction of air quality, typification of solid waste, mapping of the distribution of pollutants and controlling of the water environment (95). In the case of industrial effluent treatment, the use of ML and AI has been reported in constructed wetlands and Activated Sludge WWT processes and they appear to be in good effect. Another trend in the development and application of emissions monitoring systems based on mathematical models at industrial facilities is the introduction of the system of predictive emissions monitoring (96).

Model development and evaluation

The general processes go through the creation of models, the assessment and validation of the models that are used in a broad range of fields. It encompasses formulating, developing and iterating a model to reflect reality, the system or phenomenon under analysis. In model development, the first steps are usually to identify the problem that a model is proposed to solve, the goals of the model and to collect data (97). Scholars next decide which methodology is suitable for recreating the process, depending on its nature and the available information, i.e., statistical, mathematical or simulation (98).The model is then developed and it’s parameters are estimated utilizing procedures such as regression or optimization.

Another important and related process is the evaluation of the model and its performance. The final step in the validation of a model is the cross validation, where the predictions of the model are checked with respect to factual data or facts (99). Selection of certain categories and models is generally used in testing a model

Table 6. Future trends in ML and ES (7, 37)

Trend	Description	Potential Impact
AI and ML Technology Advances	Development of advanced ML models like LSTM for WWTPs	Enhances accuracy and operational efficiency
Smart Technology Integration	Merging ML with smart technologies for water management	Reduces costs and improves sustainability
Cross-Disciplinary Research	Combining ML with various scientific disciplines	Fosters innovation and addresses global challenges
Hybrid Modelling Techniques	Developing models that blend process-based and ML approaches	Improves prediction accuracy and computational performance

and measuring its performance and these include; goodness-of-fit measures, error rates and statistical significance. The model's output may show poor performance and hence, the researchers may be required to improve the quality of the model, perhaps by modifying its characteristics, optimizing or reducing the number of variables used or employing other modelling methods (97, 99).

Another factor, perhaps more significant when assessing a model is the question of the drawbacks or the relative error of the model For one, models are abstractions and often, they do not fully encompass the details of the real world system that is being modeled (98). Such limitations should be pointed out appropriately by the researchers and compute the level of uncertainty present in the predictions (99).

The model development and model evaluation are two parallel cycles of a model development process, during which one has to design, test and improve the models for the sake of their accuracy and effectiveness. Considering the common difficulties of ML, the researchers can obtain the necessary models that reflect real-life systems and can be applied in different spheres for decision-making (Table 7).

Predicting effluent impact on water quality parameters

Effluents from various industries can significantly impact water quality parameters, leading to environmental and health concerns. Several studies have highlighted the detrimental effect of industrial effluents on water bodies. Factory wastes especially the textile, rubber and other industries emit high levels of heavy metals, orgainic compounds and other toxins.For example, research on the effects of effluents on water resources have demonstrated that they change the parameters including pH, EC, alkalinity, chloride, fluoride, TDS, hardness, DO (Dissolved Oxygen), BOD and COD above their authorised limits consequently profounding water quality (96).

A case of industrial effluent pollution affecting water quality was reported in the Oken River, Nigeria, where elevated levels of salinity, turbidity, total dissolved solids (TDS), dissolved oxygen (DO), chemical oxygen demand (COD) and heavy metals were detected downstream of the effluent discharge point, rendering the water unsuitable for human consumption. Similarly, a study on the ecological impact of effluent discharge from a fish farm into the Odo-Owa stream in Nigeria revealed significant deterioration in water quality, highlighting the urgent need for monitoring and protection of water bodies from untreated effluents (94).

The pollution level of these effluents has a significant effect on the water quality characteristics as well as the quality of water available for consumption and use, hence there is need for enhancement of standards for effluent discharge, treatment and control to protect water and human health. The necessity of controlling and regulating industrial discharge is paramount as it aims to reduce the impact of effluents to water quality.

Forecasting changes in DO levels

DO, which often is used to evaluate water quality and the conditions of water inhabitants presence. Essential information must be gathered and analyzed in order to forecast the changes in DO levels with precision to allow efficient water management. Some prior research has suggested newly developed deep-learning models and functional forecasting methodologies to forecast DO levels.

The study offered a deep learning model based on a graph, which was named Graph Neural Network Sample and Aggregate (GNN-SAGE) and it proved to be the most efficient method for predicting DO levels in the Credit River Watershed, Ontario, Canada. It further presented an AUC-ROC (Area Under the Receiver Operating Characteristic Curve)of 0.97 and the RMSE of 0.34 mg/L; the results were superior to all benchmarking models. This way, the model received spatiotemporal information from the neighboring monitoring stations and subsequently, temperature was determined as one of the drivers affecting future DO levels in the water body (95).

Another method that can also be used when developing the forecast of the vertical profiles of DO percent saturation in lakes is the functional data analysis (FDA). FDA models were constructed with 2 hr interval DO measurement data and the overall DO profile was predicted for the 2 to 24 hr into the future. The FDA method performed better than the other models and extending the list of functional variables to include pH, temperature and conductivity enhanced longer-range predictions (97).

In hydroponic systems, the DO level was greatly affected by the variation in the value of environmental temperature. Autopot had a positive effect of DO and temperature while Smart Waterer Unpad 1 and Smart Waterer Unpad 2 had negative effects. Indeed, the independent Smart Watering Unpad 02 installation as mentioned earlier herein this paper was able to attain optimal mean DO levels as compared to the other used watering systems.

These studies prove the possibility of using such state-of-the-art methods as graph neural networks and functional data analysis for effective DO level prediction in different water bodies and systems. Extending the system with spatial and temporal data of the source and DO, as well as adjustment of the model to the changes in the seasons can enhance the accuracy in the DO forecast.

Futures assessment of effluent on water quality characteristics BOD, COD, pH and effects of temperature fluctuation are important approaches that must be taken to protect the actual water resources. From the available literature, knowledge on the impact of industrial wastewater effluent on surface and groundwater qualities is provided, onsetting the significance of monitoring and managing industrial wastewater effluent. From the available literatures, knowledge on the impact of industrial

Table 7. Summary of training and testing techniques (100)

Technique/Tool	Description	Used for
Cross-Validation	Assess model generalizability on independent dataset	Model Validation
Bagging	Aggregates predictions of multiple models	Reducing Variance
Boosting	Sequentially corrects errors of previous models	Reducing Bias
Accuracy	Measures correct predictions	Model Performance
Precision and Recall	Measures correctness and completeness of positive predictions	Classification Accuracy
F1 Score	Weighted average of Precision and Recall	Balance between Precision and Recall
AUC-ROC	Measures separability	Classification Problem Thresholding

wastewater effluent on surface and ground water qualities is provided; onset the significance of monitoring and managing of industrial wastewater effluent.

Impact of industrial wastewater on surface water quality

Another research conducted on An Giang Province in Vietnam showed that the AWWQA (American Water Works Quality Association), TSS, BOD, settleable solids, nitrate, ammonia and coliform had polluted the surface water through the discharge of industrial wastewaters (96). This implies that industrialization poses a great influence on the pollution of surface water and therefore needs constant treatment to avoid the use of contaminated water.

Impact of dyeing industrial effluent on groundwater quality

A previous study, which was aimed at identifying the effects of dyeing industrial effluent on groundwater quality, reported the textile industries as the major polluters in Bhilwara city in Rajasthan. Finally, the analysis also uncovered that the effluent from these industries affected the physico-chemical quality of the groundwater and thus, called for appropriate treatment and disposal of industrial effluent (100).

Assessing water quality using ANNs

An ANN and FA based on the Nemerow pollution index (NPI) were employed in the study on the treatability influence of municipal sewage effluent on surface water quality evaluation (98). The model's effectiveness in evaluating the quality of water was demonstrated by the results, with chloride being a crucial parameter. This method can be helpful in forecasting how effluent will affect various aspects of water quality.

Predicting BOD and COD: Predicting BOD and COD necessitates taking into account a number of variables, including the kind and volume of industrial effluent, the treatment methods used and the inherent characteristics of the receiving water body (99). ANNs and FA can be helpful tool for predicting how effluent might impact BOD and COD.

Predicting pH and temperature variations: It is necessary to take into consideration the same elements as well as the particular industrial processes involved in order to predict variations in pH and temperature. For instance, it has been discovered that the tanning process considerably affects the concentrations of pH, BOD, TSS, COD and TDS in wastewater (100). These effects can be lessened and the contamination of surface and groundwater resources can be avoided with the proper treatment and management of industrial wastewater.

Conclusion

In summary, ML offers advancements in accuracy, efficiency and long-term forecasting; making it a viable option for estimating the environmental impact of paper effluent discharge. However, to guarantee dependable model performance, a number of important issues need to be resolved. Among these are the quantity and caliber of training data, since inadequate or skewed datasets can impede learning and produce predictions that are not reliable. To effectively depict effluent behavior and trends, effective feature engineering and selection are also necessary. Furthermore, model interpretability is still a major challenge, especially when utilizing intricate architectures like DNN, since stakeholders need to be able to comprehend and have faith in model outputs in order to make well-informed decisions. The implementation of ML in practical settings is made more difficult by technical issues including overfitting, generalization and processing needs. However, data and resource limitations can be addressed with the use of transfer learning and model pruning techniques, which improves the adaptability of previously trained models to tasks unique to the paper sector. Overall, by carefully negotiating these hurdles, ML can play a crucial role in promoting sustainable and data-driven environmental management methods in the paper production sector.

Acknowledgements

We thank the Department of Environmental Science, Tamil Nadu Agricultural University, Coimbatore for their support and for funding this research.

Authors' contributions

KDK and SK conceptualized the study, developed methodology and carried out the formal analysis. Investigation was conducted by KDK and KTT. The original draft was written by KDK, SK and KTT. SK, KR, DP and SCS contributed to writing the review and editing of the manuscript. Supervision was provided by KR, DP, SCS and KTT. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest: The authors declare no conflict of interest.

Ethical issues: None

References

1. Zhang L, Wang C, Hu W, Wang X, Wang H, Sun X, et al. Dynamic real-time forecasting technique for reclaimed water volumes in urban river environmental management. Environ Res. 2024;248:118267. https://doi.org/10.1016/j.envres.2024.118267
2. Miller T, Łobodzińska A, Kozlovska P, Lewita K, Kaczanowska O, Durlik I. Advancing water quality prediction: the role of machine learning in environmental science. Grail Sci. 2024;36:519–25. https://doi.org/10.36074/grail-of-science.16.02.2024.092
3. Granata F, de Marinis G. Machine learning methods for wastewater hydraulics. Flow Meas Instrum. 2017;57:1–9. https://doi.org/10.1016/j.flowmeasinst.2017.08.004
4. Schmidt L, Heße F, Attinger S, Kumar R. Challenges in applying machine learning models for hydrological inference: a case study for flooding events across Germany. Water Resour Res. 2020;56 (5):e2019WR025924. https://doi.org/10.1029/2019wr025924
5. Kamali M, Khodaparast Z. Review on recent developments on pulp and paper mill wastewater treatment. Ecotoxicology and environmental safety. 2015 Apr 1;114:326-42. https://doi.org/10.1016/j.ecoenv.2014.05.005
6. McManners PJ. Increasing the impact of sustainability research-a new methodology. J Sustain Res. 2019;1(1):e190008. https://doi.org/10.20900/jsr20190008
7. Colla V, Pietrosanti C, Malfa E, Peters K. Environment 4.0: How digitalization and machine learning can improve the environmental footprint of the steel production processes. Matériaux Tech. 2020;108(5–6):507. https://doi.org/10.1051/mattech/2021007
8. Ines DM, Angel PS. A review of the use of machine learning techniques in eco-innovation research. In: 4th Int Conf Business Meets Technology. Valencia: Editorial Universitat Politècnica de

València; 2022;244–54. https://doi.org/10.4995/BMT2022.2022.15550

9. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr. 2017;11(4):959–75. https://doi.org/10.1016/j.joi.2017.08.007
10. Hermosilla D, Merayo N, Gascó A, Blanco Á. The application of advanced oxidation technologies to the treatment of effluents from the pulp and paper industry: a review. Environ Sci Pollut Res. 2015;22:168–91. https://doi.org/10.1007/s11356-014-3516-1
11. Crini G, Lichtfouse E. Wastewater treatment: an overview. Green adsorbents for pollutant removal. Environmental Chemistry for a Sustainable World. 2018;18:1–21. https://doi.org/10.1007/978-3-319-92111-2_1
12. Ouasfi N, Sabbar E, Khamliche L. Enhancing the water quality of treated effluents from wastewater treatment plants by mitigating emerging pollutants. All Sci Abstr. 2023;1(5):6. https://doi.org/10.59287/as-abstracts.1356
13. Gebrehanna M, Gordon RJ, Madani A, VanderZaag AC, Wood JD. Silage effluent management: a review. J Environ Manage. 2014;143:113–22. https://doi.org/10.1016/j.jenvman.2014.04.012
14. Shakil MSZ, Mostafa MG. Characterization of paper mill effluent and its impacts on the environment. J Chem Environ. 2023;2(1):109–22.https://doi.org/10.56946/jce.v2i01.135
15. Gu W, Li Y, Zhang X. Printing industry and the environment. Adv Mater Res. 2013;663:759–62. https://doi.org/10.4028/www.scientific.net/AMR.663.759
16. Gangwar AK, Prakash NT, Prakash R. Applicability of microbial xylanases in paper pulp bleaching: a review. Bioresources. 2014;9:3733–54. https://doi.org/10.15376/biores.9.2.3733-3754
17. Bagherzadeh F, Nouri AS, Mehrani M-J, Thennadil S. Prediction of energy consumption and evaluation of affecting factors in a full-scale WWTP using a machine learning approach. Process Saf Environ Prot. 2021;154:458–66. https://doi.org/10.1016/j.psep.2021.08.040
18. Maganathan T, Senthilkumar S, Balakrishnan V. Machine learning and data analytics for environmental science: a review, prospects and challenges. IOP Conf Ser Mater Sci Eng. 2020:955 012107. https://doi.org/10.1088/1757-899x/955/1/012107
19. Pernkopf F, Roth W, Zoehrer M, Pfeifenberger L, Schindler G, Froening H, et al. Efficient and robust machine learning for real-world systems. arXiv [preprint]. 2018. https://doi.org/10.48550/arXiv.1812.02240
20. Lowe M, Qin R, Mao X. A review on machine learning, artificial intelligence and smart technology in water treatment and monitoring. Water. 2022;14(9):1384. https://doi.org/10.3390/w14091384
21. Masih A. Machine learning algorithms in air quality modeling. Glob J Environ Sci Manage. 2019;5(4):515–34. https://doi.org/10.22034/GJESM.2019.04.10
22. Liu X, Lu D, Zhang A, Liu Q, Jiang G. Data-driven machine learning in environmental pollution: gains and problems. Environ Sci Technol. 2022;56(4):2124–33.https://doi.org/10.1021/acs.est.1c06157
23. Miller TH, Gallidabino MD, MacRae JI, Hogstrand C, Bury NR, Barron LP, et al. Machine learning for environmental toxicology: a call for integration and innovation. Environ Sci Technol. 2018;52(22):12953–5. https://doi.org/10.1021/acs.est.8b05382
24. Zhu JJ, Yang M, Ren ZJ. Machine learning in environmental research: common pitfalls and best practices. Environ Sci Technol. 2023;57(46):17671–89. https://doi.org/10.1021/acs.est.3c00026
25. Sun Y, Wang X, Ren N, Liu Y, You S. Improved machine learning models by data processing for predicting life-cycle environmental impacts of chemicals. Environ Sci Technol. 2023;57(8):3434–44. https://doi.org/10.1021/acs.est.2c04945
26. Wang MWH, Goodman JM, Allen TEH. Machine learning in predictive toxicology: recent applications and future directions for classification models. Chem Res Toxicol. 2020;34(2):217–39. https://doi.org/10.1021/acs.chemrestox.0c00316
27. Wang C, Liu B, Chen J, Yu X. Air quality index prediction based on a long short-term memory artificial neural network model. J Comput. 2023;34(2):69–79. https://doi.org/10.53106/199115992023043402006
28. O'Donncha F, Hu Y, Palmes P, Burke M, Filgueira R, Grant J. A spatio-temporal LSTM model to forecast across multiple temporal and spatial scales. Ecol Inform. 2021;69:101687. https://doi.org/10.1016/j.ecoinf.2022.101687
29. Brownlee J. Deep learning for computer vision: image classification, object detection and face recognition in python. Machine Learning Mastery; 2019 Apr 4.
30. Wang B, Tan X, Guo J, Xiao T, Jiao Y, Zhao J, et al. Drug-induced immune thrombocytopenia toxicity prediction based on machine learning. Pharmaceutics. 2022;14(5):943. https://doi.org/10.3390/pharmaceutics14050943
31. Liu X. Comparison of different machine learning models: linear model, forest and SVM. Appl Comput Eng. 2024;51:225–30. https://doi.org/10.54254/2755-2721/51/20241467
32. Meyer AP, Albarghouthi A, D'Antoni L. The dataset multiplicity problem: how unreliable data impacts predictions. FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability and Transparency. 2023;193–204. https://doi.org/10.1145/3593013.3593988
33. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis) contents: a survey of dataset development and use in machine learning research. Patterns. 2021;2(11);100336. https://doi.org/10.1016/j.patter.2021.100336
34. Daniel F, Kucherbaev P, Cappiello C, Benatallah B, Allahbakhsh M. Quality control in crowdsourcing: a survey of quality attributes, assessment techniques and assurance actions. ACM Comput Surv. 2018;51(1):1–40. https://doi.org/10.1145/3148148
35. Dang C, Liu Y, Yue H, Qian J, Zhu R. Autumn crop yield prediction using data-driven approaches: support vector machines, random forest and deep neural network methods. Can J Remote Sens. 2020;47(2):162–81. https://doi.org/10.1080/07038992.2020.1833186
36. Sun Y, Wang X, Ren N, Liu Y, You S. Improved machine learning models by data processing for predicting life-cycle environmental impacts of chemicals. Environ Sci Technol. 2022;57(8):3434–44. https://doi.org/10.1021/acs.est.2c04945
37. Wongburi P, Park JK. Prediction of wastewater treatment plant effluent water quality using recurrent neural network (RNN) models. Water. 2023;15(9):3325. https://doi.org/10.3390/w15193325
38. Fang X, Zhai Z, Xiong R, Zhang L, Gao B. LSTM-based modelling for coagulant dosage prediction in wastewater treatment plant. In: AIEE '22: Proceedings of the 2022 3rd International Conference on Artificial Intelligence in Electronics Engineering. 2022;23–7. https://doi.org/10.1145/3512826.3512847
39. Cheng T, Harrou F, Kadri F, Sun Y, Leiknes T. Forecasting of wastewater treatment plant key features using deep learning-based models: a case study. Ieee Access. 2020;8:184475–85. https://doi.org/10.1109/ACCESS.2020.3030820
40. Wu F, Shu J. Predictions for COVID-19 with deep learning models of long short-term memory (LSTM). In: Biomedical and Business Applications Using Artificial Neural Networks and Machine Learning. 2022;128–53. https://doi.org/10.4018/978-1-7998-8455-2.ch005
41. Latif SD. Concrete compressive strength prediction modeling utilizing deep learning long short-term memory algorithm for a sustainable environment. Environ Sci Pollut Res. 2021;28:30294–302. https://doi.org/10.1007/s11356-021-12877-y

42. Kashyap AA, Raviraj S, Devarakonda A, Nayak KSR, Santhosh KV, Bhat SJ. Traffic flow prediction models - a review of deep learning techniques. Cogent Eng. 2021;9(1):2010510. https://doi.org/10.1080/23311916.2021.2010510

43. Guo W, Liu J, Dong F, Song M, Li Z, Khan MKH, et al. Review of machine learning and deep learning models for toxicity prediction. Exp Biol Med. 2023;248:1952–73. https://doi.org/10.1177/15353702231209421

44. Rawat D, Meenakshi, Bajaj R. Performance analysis of drug toxicity prediction using machine learning approaches. In: 2023 3rd Int Conf Innovative Sustainable Computational Technologies (CISCT). 2023;1–6. https://doi.org/10.1109/cisct57197.2023.10351253

45. Cavasotto CN, Scardino V. Machine learning toxicity prediction: latest advances by toxicity end point. ACS omega. 2022;7(51):47536–46. https://doi.org/10.1021/acsomega.2c05693

46. Ye G, Wan J, Deng Z, Wang Y, Chen J, Zhu B, et al. Prediction of effluent total nitrogen and energy consumption in wastewater treatment plants: Bayesian optimization machine learning methods. Bioresour Technol. 2024;395:130361. https://doi.org/10.1016/j.biortech.2024.130361

47. Cechinel MAP, Neves J, Fuck JVR, de Andrade RC, Spogis N, Riella HG, et al. Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: a plant-level analysis. J Water Process Eng. 2024;58:104758. https://doi.org/10.1016/j.jwpe.2023.104758

48. Lv J, Du L, Lin H, Wang B, Yin W, Song Y, et al. Enhancing effluent quality prediction in wastewater treatment plants through the integration of factor analysis and machine learning. Bioresour Technol. 2024;393:130008. https://doi.org/10.1016/j.biortech.2023.130008

49. Hino M, Benami E, Brooks N. Machine learning for environmental monitoring. Nat Sustain. 2018;1:583–8. https://doi.org/10.1038/s41893-018-0142-9

50. Chy MKH, Buadi ON. Role of machine learning in policy making and evaluation. Int J Innov Sci Res Technol. 2024;9(10);456–63.https://doi.org/10.38124/ijisrt/IJISRT24OCT687

51. Aparna K, Swarnalatha R, Changmai M. Optimizing wastewater treatment plant operational efficiency through integrating machine learning predictive models and advanced control strategies. Process Saf Environ Prot. 2024;188:995–1008. https://doi.org/10.1016/j.psep.2024.05.148

52. Liu W, Liu T, Liu Z, Luo H, Pei H. A novel deep learning ensemble model based on two-stage feature selection and intelligent optimization for water quality prediction. Environ Res. 2023;224:115560. https://doi.org/10.1016/j.envres.2023.115560

53. Isibor PO, Kayode-Edwards II, Taiwo OS. Emerging technology and future directions in environmental nanotoxicology. In: Environmental Nanotoxicology: Combatting the Minute Contaminants; Springer. 2024;325–46. https://doi.org/10.1007/978-3-031-54154-4_16

54. Asensio OI, Mi X, Dharur S. Using machine learning techniques to aid environmental policy analysis. Case Stud Environ. 2020;4(1):961302.https://doi.org/10.1525/cse.2020.961302

55. Cheng SH, Augustin C, Bethel A, Gill D, Anzaroot S, Brun J, et al. Using machine learning to advance synthesis and use of conservation and environmental evidence. Conserv Biol. 2018;32(4):762–4. https://doi.org/10.1111/cobi.13117

56. Zhang Y, Yang J, Huang M, Liu H. Neighborhood component analysis for modeling papermaking wastewater treatment processes. Bioprocess Biosyst Eng. 2021;44:2345–59. https://doi.org/10.1007/s00449-021-02608-5

57. Zang X, Zhao X, Tang B. Hierarchical molecular graph self-supervised learning for property prediction. Commun Chem. 2023;6:34. https://doi.org/10.1038/s42004-023-00825-5

58. Hsu W-N, Bolte B, Tsai Y-HH, Lakhotia K, Salakhutdinov R, Mohamed A. Hubert: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3451–60. https://doi.org/10.1109/TASLP.2021.3122291

59. Li Z, Huang C, Xia L, Xu Y, Pei J. Spatial-temporal hypergraph self-supervised learning for crime prediction. In: 2022 IEEE 38th Int Conf Data Eng (ICDE). Kuala Lumpur, Malaysia. 2022;2984–96. https://doi.org/10.1109/ICDE53745.2022.00269

60. Zou H. Clustering algorithm and its application in data mining. Wirel Pers Commun. 2019;110:21–30. https://doi.org/10.1007/s11277-019-06709-z

61. Gupta MK, Chandra P. A comprehensive survey of data mining. Int J Inf Technol. 2020;12:1243–57. https://doi.org/10.1007/s41870-020-00427-7

62. Naik P, Nelaballi S, Pusuluri VS, Kim DK. Deep learning-based code refactoring: A review of current knowledge. Journal of Computer Information Systems. 2024 Mar 3;64(2):314-28. https://doi.org/10.1080/08874417.2023.2203088

63. Lever J, Krzywinski M, Altman N. Principal component analysis. Nat Methods. 2017;14:6412. http://dx.doi.org/10.1038/nmeth.4346

64. Maadooliat M, Huang JZ, Hu J. Integrating data transformation in principal components analysis. J Comput Graph Stat. 2015;24(1):84-103. https://doi.org/10.1080/10618600.2014.891461

65. Olawoyin R, Nieto A, Grayson RL, Hardisty F, Oyewole SA. Application of artificial neural network (ANN)-self-organizing map (SOM) for the categorization of water, soil and sediment quality in petrochemical regions. Expert Syst Appl. 2013;40(9):3634–48. https://doi.org/10.1016/j.eswa.2012.12.069

66. Tian H, Zhao Y, Luo M, He Q, Han Y, Zeng Z. Estimating PM2.5 from multisource data: a comparison of different machine learning models in the Pearl River Delta of China. Urban Clim. 2021;35:100740. https://doi.org/10.1016/j.uclim.2020.100740

67. Bilgilioglu H. A comparison of different machine learning models for landslide susceptibility mapping in Rize (Türkiye). Baltica. 2023. https://doi.org/10.5200/baltica.2023.2.3

68. Liu X. Comparison of different machine learning models: linear model, forest and SVM. Appl Comput Eng. 2024;51:225–30. https://doi.org/10.54254/2755-2721/51/20241467

69. Sun AY, Scanlon BR. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications and future directions. Environ Res Lett. 2019;14(7):073001. https://doi.org/10.1088/1748-9326/ab1b7d

70. Saab C, Zéhil G-P. About machine learning techniques in water quality monitoring. In: 2023 5th Int Conf Adv Comput Tools Eng Appl (ACTEA). Zouk Mosbeh, Lebanon. 2023:115-21. https://doi.org/10.1109/actea58025.2023.10193911

71. Bai R-H, Fan R-Q, Liu Q, Liu Q, Yan C-R, Cui J-X, et al. Overview of the application of machine learning for identification and environmental risk assessment of microplastics. Huan Jing Ke Xue. 2024;45(2):1185–95. https://doi.org/10.13227/j.hjkx.202302110

72. Mansoursamaei M, Moradi M, González-Ramírez RG, Lalla-Ruiz E. Machine learning for promoting environmental sustainability in ports. J Adv Transp.2023;2144733:17. https://doi.org/10.1155/2023/2144733

73. Tsai Y-Y, Chen P-Y, Ho T-Y. Transfer learning without knowing: reprogramming black-box machine learning models with scarce data and limited resources. In: Proc 37th Int Conf Mach Learn (ICML). PMLR. 2020;119:9614–24..

74. Jui TD, Rivas P. Fairness issues, current approaches and challenges in machine learning models. Int J Mach Learn Cybern. 2024;15:3095–125. https://doi.org/10.1007/s13042-023-02083-2

75. Chung J, Teo J. Mental health prediction using machine learning: taxonomy, applications and challenges. Appl Comput Intell Soft Comput. 2022;9970363:19 p. https://doi.org/10.1155/2022/9970363

76. Jenga K, Catal C, Kar G. Machine learning in crime prediction. J Ambient Intell Humaniz Comput. 2023;14:2887–913. https://doi.org/10.1007/s12652-023-04530-y

77. Bujang SDA, Selamat A, Ibrahim R, Krejcar O, Herrera-Viedma EE, Fujita H, et al. Multiclass prediction model for student grade prediction using machine learning. IEEE Access. 2021;9:95608–21. https://doi.org/10.1109/ACCESS.2021.3093563

78. Arsić SM, Mihić M, Petrović D, Mitrović Z, Kostić SC, Mihic O. Review of measures for improving ML model interpretability: empowering decision makers with transparent insights. In: 2024 Int Conf Artif Intell Comput Data Sci Appl (ACDSA). Victoria, Seychelles. 2024;1–5. https://doi.org/10.1109/ACDSA59508.2024.10467907

79. Batzolis E, Vrochidou E, Papakostas GA. Machine learning in embedded systems: limitations, solutions and future challenges. In: 2023 IEEE 13th Annu Comput Commun Workshop Conf (CCWC). Las Vegas, NV, USA. IEEE. 2023;0345–50. https://doi.org/10.1109/CCWC57344.2023.10099348

80. Ying X. An overview of overfitting and its solutions. J Phys Conf Ser. 2019;1168(2)022022. https://doi.org/10.1088/1742-6596/1168/2/022022

81. Gygi JP, Kleinstein SH, Guan L. Predictive overfitting in immunological applications: pitfalls and solutions. Hum Vaccin Immunother. 2023;19(2). https://doi.org/10.1080/21645515.2023.2251830

82. Rao NSV. Study of overfitting by machine learning methods using generalization equations. In: 2023 26th Int Conf Inf Fusion (FUSION). Charleston, SC, USA. 2023:1–8. https://doi.org/10.23919/fusion52260.2023.10224198

83. Sinnott RO, Guan Z. Prediction of air pollution through machine learning approaches on the cloud. In: 2018 IEEE/ACM 5th Int Conf Big Data Comput Appl Technol (BDCAT). Zurich, Switzerland. 2018:51–60. https://doi.org/10.1109/bdcat.2018.00015

84. Icke O, van Es DM, de Koning MF, Wuister JJG, Ng J, Phua KM, et al. Performance improvement of wastewater treatment processes by application of machine learning. Water Sci Technol.. 2020;82 (12):2671–80. https://doi.org/10.2166/wst.2020.382

85. Huntingford C, Jeffers ES, Bonsall MB, Christensen HM, Lees T, Yang H. Machine learning and artificial intelligence to aid climate change research and preparedness. Environ Res Lett. 2019;14 124007. https://doi.org/10.1088/1748-9326/ab4e55

86. Hamdan A, Ibekwe KI, Etukudoh EA, Umoh AA, Ilojianya VI. AI and machine learning in climate change research: a review of predictive models and environmental impact. World J Adv Res Rev. 2024;21 (01):1999–2008. https://doi.org/10.30574/wjarr.2024.21.1.0257

87. Liu J, Mooney H, Hull V, Davis SJ, Gaskell J, Hertel T, et al. Systems integration for global sustainability. Science. 2015;347(6225). https://doi.org/10.1126/science.1258832

88. Gupta PK, Yadav B, Kumar A, Himanshu SK. Machine learning and artificial intelligence application in constructed wetlands for industrial effluent treatment: advances and challenges in assessment and bioremediation modeling. In: Bioremediation for Environmental Sustainability. 2021;403–414. https://doi.org/10.1016/B978-0-12-820524-2.00016-X

89. Peng N. Application of Machine learning techniques in environmental governance: a review. Adv Eng Technol Res (ICISCTA). 2023;7(1):528–37. https://doi.org/10.56028/aetr.7.1.528.2023

90. Meshalkin VP, Skobelev DO, Vocciante M, González M, Popov AY. Predicting emissions from the chemical and energy industries: progress in applying modeling approaches. Theor Found Chem Eng. 2021;55:588–93. https://doi.org/10.1134/S0040579521040278

91. Soklaridis S, Shier R, Black G, Bellissimo G, Di Giandomenico A, Gruszecki S, et al. "My words matter": perspectives on evaluation from people who access and work in recovery colleges. Mental Health and Social Inclusion. 2024;28(2):134–143. https://doi.org/10.1108/MHSI-01-2023-0002

92. Jiao H, Udomlertsakul N, Tamprasirt A. Credential control balance: a universal blockchain account model abstract from bank to Bitcoin, Ethereum external owned account and account abstraction. arXiv [preprint]. 2024. https://doi.org/10.48550/arXiv.2402.10616

93. Ankley GT, Corsi SR, Custer CM, Ekman DR, Hummel SL, Kimbrough KL, et al. Assessing contaminants of emerging concern in the Great Lakes ecosystem: a decade of method development and practical application. Environ Toxicol Chem. 2023;42(12):2506–18. https://doi.org/10.1002/etc.5740

94. Dhaker N, Mehta P. Impact of dyeing industrial effluent on physicochemical parameters of ground water quality of industrial area of Bhilwara, Rajasthan. Stud Indian Place Names. 2020;40 (68):872–8.

95. Famoofo OO, Adeniyi IF. Impact of effluent discharge from a medium-scale fish farm on the water quality of Odo-Owa stream near Ijebu-Ode, Ogun State, Southwest Nigeria. Appl Water Sci. 2020;10:68. https://doi.org/10.1007/s13201-020-1148-9

96. Rocha PAC, Santos VO, Thé JVG, Gharabaghi B. New graph-based and transformer deep learning models for river dissolved oxygen forecasting Environments. 2023;10(12):217. https://doi.org/10.3390/environments10120217

97. Durell L, Scott JT, Nychka D, Hering AS. Functional forecasting of dissolved oxygen in high-frequency vertical lake profiles. Environmetrics. 2022;34(4):e2765. https://doi.org/10.1002/env.2765

98. Giao NT. Surface water quality influenced by industrial wastewater effluent in An Giang province, Vietnam. NIPES J Sci Technol Res. 2022;4(1):51–58. https://doi.org/10.37933/nipes/4.1.2022.4

99. Mohammed R, Al-Obaidi B. Treatability influence of municipal sewage effluent on surface water quality assessment based on Nemerow pollution index using an artificial neural network. IOP Conf Ser Earth Environ Sci. 2021;877 012008. https://doi.org/10.1088/1755-1315/877/1/012008

100. Bhardwaj A, Kumar S, Singh D. Tannery effluent treatment and its environmental impact: a review of current practices and emerging technologies. Water Qual Res J. 2023;58(2):128–52. https://doi.org/10.2166/wqrj.2023.002

Additional information

Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

Publisher information: Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.