

AI-Assisted Prediction of Environmental Pollutant Removal Using Advanced Materials

Juhi Virdhi¹,

¹Department of Chemistry, University of Jammu.

Abstract - Environmental pollution, particularly contamination of water resources by heavy metals, dyes, pharmaceuticals, and other emerging pollutants, has become a significant challenge for sustainable environmental management. Advanced materials such as nanomaterials, biochar, activated carbon, graphene-based materials, metal oxides, and nanocomposites have demonstrated considerable potential for efficient pollutant removal. However, selecting suitable materials and determining optimal treatment conditions generally requires extensive experimental investigation. This study proposes an artificial intelligence (AI)-assisted predictive framework for estimating environmental pollutant removal efficiency using advanced materials. Relevant experimental parameters, including pollutant concentration, material dosage, pH, temperature, contact time, and material characteristics, are considered as input variables for machine learning models. Data preprocessing, feature selection, model training, validation, and performance evaluation are integrated into the proposed methodology. Machine learning techniques such as Random Forest, Support Vector Regression, XGBoost, and Artificial Neural Networks are considered for predicting pollutant removal efficiency. Model performance is evaluated using statistical indicators including coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). Furthermore, feature-importance analysis is used to identify the key parameters influencing pollutant removal. The proposed AI-assisted approach can reduce experimental effort, support optimization of treatment conditions, and improve the selection of advanced materials for sustainable pollutant remediation. The framework provides a foundation for developing intelligent and data-driven environmental treatment systems.

Key Words: Artificial Intelligence; Machine Learning; Environmental Pollutants; Advanced Materials; Adsorption; Wastewater Treatment; Pollutant Removal; Predictive Modelling; Nanomaterials; Sustainable Remediation.

1.INTRODUCTION

1.1 Background of Environmental Pollution

Environmental pollution has become a major global concern due to rapid industrialization, urbanization, agricultural activities, and increasing discharge of untreated or inadequately treated wastewater. Aquatic environments are particularly vulnerable because industrial effluents, agricultural runoff, municipal wastewater, and pharmaceutical residues can introduce heavy metals, dyes, pesticides, pharmaceuticals, and other emerging contaminants into water systems [6][7]. Many of these pollutants are persistent and potentially toxic. Heavy metals such as lead, cadmium, chromium, mercury, and arsenic cannot be biologically degraded and may accumulate in environmental systems, while organic micropollutants and pharmaceutical compounds can remain in aquatic environments even at relatively low concentrations [6][7]. Conventional treatment processes such as chemical precipitation, coagulation, biological treatment, membrane separation, and oxidation are widely used; however, their effectiveness depends on pollutant characteristics and operating conditions and may involve high energy or chemical requirements and secondary-waste generation [8]. Advanced materials including biochar, activated carbon, graphene-based materials, metal oxides, nanomaterials, metal-organic frameworks (MOFs), and nanocomposites have therefore attracted considerable attention for pollutant removal because of their high surface area, porous structure, tunable surface chemistry, and adsorption or catalytic properties [6][14][20]. However, pollutant-removal efficiency is controlled by multiple interacting factors, including material properties, pollutant concentration, pH, temperature, material dosage, and contact time, making experimental optimization complex and time-consuming [1][2]. Artificial Intelligence (AI) and Machine Learning (ML) provide a data-driven approach for modelling these nonlinear relationships and predicting treatment performance.

Therefore, integrating advanced materials with AI-assisted predictive modelling offers a promising approach for improving material selection, optimizing treatment conditions, reducing experimental trials, and developing more efficient and sustainable environmental-remediation systems [1]–[3].

1.2 Environmental Pollutants and Their Impact

Environmental pollutants enter aquatic systems primarily through industrial effluents, agricultural runoff, municipal wastewater, mining activities, and improper waste disposal. Major pollutant groups include heavy metals, synthetic dyes, pesticides, pharmaceuticals, hydrocarbons, nutrients, and emerging contaminants [6][7].

Heavy metals such as Pb, Cd, Cr, Hg, and As are of particular concern because of their persistence, toxicity, and potential accumulation in environmental systems. Organic contaminants, including dyes, pesticides, phenolic compounds, and pharmaceutical residues, can also adversely affect aquatic organisms and ecosystem functions. Some emerging contaminants may persist through conventional wastewater-treatment processes, making their effective removal increasingly challenging [7][9].

The environmental impact of pollutants depends on their concentration, toxicity, persistence, bioavailability, and exposure duration. Their presence can deteriorate water quality, disturb aquatic biological processes, reduce biodiversity, and contribute to food-chain contamination. Furthermore, treatment becomes more complex in multi-pollutant systems because contaminants may compete for available adsorption or reaction sites [5][6].

Therefore, understanding pollutant characteristics is essential for selecting appropriate advanced materials and developing reliable AI/ML models for pollutant-removal prediction. In the proposed framework, pollutant type and concentration are incorporated with material and operating parameters to improve prediction

of removal efficiency [1][2].



Figure 1.1: Classification of Environmental Pollutants and Their Environmental Impacts

1.3 Challenges in Conventional Pollutant Removal

Conventional wastewater-treatment methods include coagulation, chemical precipitation, biological treatment, adsorption, membrane separation, and advanced oxidation processes. Although these techniques can effectively remove many contaminants, their performance strongly depends on pollutant characteristics and operating conditions [6][8].

A major limitation is the sensitivity of treatment efficiency to parameters such as pH, pollutant concentration, temperature, treatment time, and chemical or adsorbent dosage. Identifying suitable conditions often requires extensive trial-and-error experimentation. Moreover, chemical treatment can generate secondary sludge, membrane processes may suffer from fouling and high operational requirements, and biological processes can be affected by toxic or poorly biodegradable contaminants [6][8].

Treatment of pharmaceuticals, pesticides, and other emerging contaminants is particularly challenging because some compounds may not be completely removed by conventional wastewater-treatment systems [7]. Multi-pollutant systems can further complicate treatment because contaminants may compete for adsorption or reaction sites [5][9].

Therefore, conventional approaches have limitations in rapidly selecting suitable materials and operating conditions. AI/ML-based predictive modelling provides an alternative approach by analysing multiple interacting parameters simultaneously and predicting pollutant-removal performance, potentially reducing experimental effort and supporting process optimization [1]–[3].

1.4 Role of Advanced Materials in Environmental Remediation

Advanced materials have gained significant attention for environmental remediation because of their high specific surface area, porous structure, tunable surface chemistry, adsorption capacity, and catalytic properties. Materials such as activated carbon, biochar, graphene-based materials, metal oxides, nanomaterials, metal–organic frameworks (MOFs), and nanocomposites have demonstrated potential for removing heavy metals, dyes, pharmaceuticals, pesticides, and other contaminants from water [6][14][20].

Pollutant removal by these materials occurs through mechanisms including electrostatic attraction, ion exchange, surface complexation, hydrogen bonding, and pore filling. Semiconductor-based materials can additionally promote photocatalytic degradation of organic pollutants [16]. Material modification and functionalization can further improve adsorption capacity, selectivity, stability, and regeneration performance [9][14].

However, removal performance depends on both material characteristics—such as surface area, pore structure, particle size, and functional groups—and operating conditions, including pH, dosage, contact time, temperature, and pollutant concentration [1][2]. Consequently, selecting an appropriate material and treatment condition is a multi-variable optimization problem.

Integration of advanced materials with AI/ML-based prediction can help identify these complex relationships, support material screening, predict removal efficiency, and reduce extensive experimental optimization [1][2]. Thus, combining advanced materials with data-driven modelling provides a promising approach for efficient and sustainable environmental remediation.



Figure 1.2: Advanced Materials for Environmental Pollutant Removal

1.5 Emergence of AI and Machine Learning in Environmental Research

Artificial Intelligence (AI) and Machine Learning (ML) have become increasingly important in environmental research because they can model complex, nonlinear relationships among multiple variables. In water and wastewater treatment, AI/ML techniques have been applied to water-quality prediction, process monitoring, treatment optimization, and pollutant-removal prediction [3][4].

For advanced-material-based remediation, ML models such as Random Forest (RF), Support Vector Regression (SVR), XGBoost, and Artificial Neural Networks (ANN) can relate material properties and operating parameters—including pH, temperature, pollutant concentration, dosage, and contact time—to pollutant-removal efficiency [1], [2]. This enables a transition from extensive trial-and-error experimentation toward data-driven prediction and optimization.

Explainable AI (XAI) techniques, including feature-importance analysis and SHAP, can further identify the parameters contributing most strongly to model predictions, improving model interpretability. However, reliable AI prediction depends on dataset quality, appropriate preprocessing, feature selection, model validation, and experimental verification [1][3].

Therefore, the present study integrates AI/ML with advanced materials to predict pollutant-removal efficiency and support the optimization of environmental-treatment conditions.

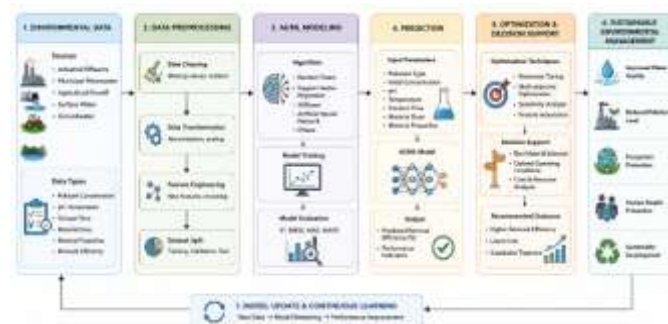


Figure 1.3: AI/ML-Based Approach for Environmental Pollutant Prediction image

1.6 Problem Statement

Despite the high pollutant-removal potential of biochar, activated carbon, graphene-based materials, metal oxides, MOFs, and nanocomposites, their performance varies considerably with material characteristics, pollutant properties, and operating conditions [6][14][20]. Parameters such as initial concentration, pH, temperature, contact time, material dosage, surface area,

and pore characteristics interact nonlinearly, making treatment optimization difficult [1][2].

Conventional trial-and-error experiments for determining optimal conditions are often time-consuming and resource-intensive, particularly when multiple materials and pollutants are considered. Furthermore, predictive models developed for specific pollutant-material systems may have limited applicability beyond the conditions represented in their training datasets [1][3].

Therefore, there is a need for an integrated AI-assisted predictive framework capable of combining material properties, pollutant characteristics, and operating parameters to predict pollutant-removal efficiency and identify the most influential treatment variables. Such an approach can reduce experimental effort and support more efficient optimization of environmental-remediation processes.

1.7 Research Objectives

The primary objective of this study is to develop an AI-assisted predictive framework for estimating environmental pollutant-removal efficiency using advanced materials. The specific objectives are:

- To develop a structured dataset integrating pollutant characteristics, advanced-material properties, and operating conditions.
- To develop and compare RF, SVR, XGBoost, and ANN models for pollutant-removal prediction [1]–[3].
- To evaluate model performance using R^2 , MAE, MSE, RMSE, and MAPE.
- To identify the most influential treatment parameters using feature-importance and Explainable AI (XAI) techniques.
- To predict pollutant-removal efficiency and identify optimized material and operating conditions.
- To assess the potential of the proposed framework for efficient and sustainable environmental remediation.

2. LITERATURE REVIEW

2.1 Environmental Pollutants

Environmental pollutants enter water systems through industrial effluents, agricultural runoff, municipal wastewater, mining, and pharmaceutical activities. Their removal depends on concentration, chemical structure, persistence, solubility, and interactions with treatment materials [6][7]. The major pollutant categories relevant to advanced-material-based remediation are summarized below.

2.1.1 Heavy Metals

Heavy metals such as Pb, Cd, Cr, Hg, and As are persistent contaminants originating from mining, electroplating, metal processing, batteries, and industrial wastewater. Because they are non-biodegradable, processes such as adsorption, precipitation, ion exchange, and membrane separation are commonly employed for their removal [6][8]. Advanced adsorbents can enhance removal through surface complexation, electrostatic interactions, and ion exchange.

2.1.2 Organic Pollutants and Dyes

Organic pollutants include phenolic compounds, hydrocarbons, pesticides, solvents, and synthetic dyes. Their persistence and diverse chemical structures can make treatment difficult. Adsorption and advanced oxidation processes using biochar, activated carbon, graphene-based materials, and nanocomposites have demonstrated considerable potential for their removal [5][14][16].

2.1.3 Pharmaceuticals and Emerging Contaminants

Pharmaceutical residues, including antibiotics, analgesics, hormones, and anti-inflammatory compounds, can enter water through domestic wastewater, healthcare facilities, and pharmaceutical industries. Some compounds are incompletely removed by conventional treatment, encouraging the development of adsorption, membrane, and advanced oxidation technologies [7][9].

2.1.4 Pesticides

Pesticides enter surface and groundwater primarily through agricultural runoff and soil leaching. Depending on their persistence and toxicity, they may adversely affect aquatic ecosystems. Advanced adsorbents and catalytic materials provide promising approaches for pesticide removal through adsorption and degradation mechanisms [6][9].

Overall, differences among pollutant classes demonstrate the need to incorporate pollutant type and physicochemical characteristics when developing AI-based models for predicting removal efficiency [1].

2.2 Advanced Materials for Pollutant Removal

Advanced materials are increasingly used for environmental remediation because of their high surface area, porous structure, tunable surface chemistry, adsorption capacity, and catalytic activity. Their physicochemical properties can be tailored for specific pollutant classes, providing improved removal

performance compared with many conventional adsorbents [6][14][20].

2.2.1 Activated Carbon and Biochar

Activated carbon and biochar are effective carbonaceous adsorbents for heavy metals, dyes, pharmaceuticals, and organic contaminants. Their performance is governed mainly by surface area, pore structure, surface functional groups, pH, and pollutant characteristics [5][6]. Biochar is particularly attractive because it can be produced from biomass residues, while physical or chemical modification can further enhance its adsorption properties [9][10].

2.2.2 Graphene-Based Materials

Graphene oxide (GO) and reduced graphene oxide (rGO) provide large surface areas and functional groups capable of interacting with pollutants through electrostatic attraction, complexation, hydrogen bonding, and π - π interactions [14][17]. However, aggregation, recovery, and production cost remain important practical limitations.

2.2.3 Nanomaterials and Metal Oxides

Nanomaterials provide high surface-to-volume ratios and enhanced surface reactivity. Metal oxides such as TiO_2 , ZnO , and iron oxides can remove contaminants through adsorption, surface complexation, redox reactions, and photocatalytic degradation [16]. Nevertheless, nanoparticle recovery, stability, and potential environmental release should be considered in practical applications.

2.2.4 MOFs and Nanocomposites

Metal-organic frameworks (MOFs) possess high porosity, large surface areas, and tunable pore chemistry, making them promising for selective pollutant adsorption [20]–[22]. Similarly, nanocomposites combine complementary properties of multiple materials to improve adsorption capacity, catalytic activity, stability, selectivity, and reusability.

Overall, advanced materials provide multiple pathways for pollutant removal; however, their effectiveness depends strongly on material properties, pollutant characteristics, and operating conditions. This variability supports the application of AI/ML for material screening and prediction of pollutant-removal performance [1][2].

2.3 Pollutant Removal Mechanisms

Advanced materials remove environmental pollutants through several physical, chemical, and catalytic mechanisms, with the dominant pathway depending on pollutant characteristics, material surface properties, and operating conditions [5][14].

2.3.1 Adsorption and Surface Interactions

Adsorption is one of the principal removal mechanisms for advanced materials such as biochar, activated carbon, graphene-based materials, MOFs, and nanocomposites. Pollutants interact with active sites through electrostatic attraction, ion exchange, surface complexation, hydrogen bonding, pore filling, van der Waals forces, and π - π interactions [5][6][14]. Adsorption efficiency is strongly influenced by pH, concentration, dosage, contact time, temperature, and surface chemistry.

2.3.2 Photocatalytic Degradation

Semiconductor materials such as TiO_2 and ZnO can absorb light energy and generate electron-hole pairs, leading to the formation of reactive species that degrade organic contaminants. Photocatalytic performance depends on catalyst properties, light conditions, pollutant concentration, pH, and reaction time [16].

2.3.3 Chemical Transformation

Advanced materials may also promote oxidation, reduction, and catalytic degradation of pollutants. These reactions can transform complex contaminants into simpler or less harmful products. Their effectiveness depends on pollutant chemistry, catalyst properties, reagent concentration, and reaction conditions.

Overall, pollutant removal generally results from a combination of adsorption, surface interaction, and catalytic degradation mechanisms. Understanding these mechanisms helps identify relevant material and operating parameters for AI/ML-based prediction of removal efficiency [1][2].

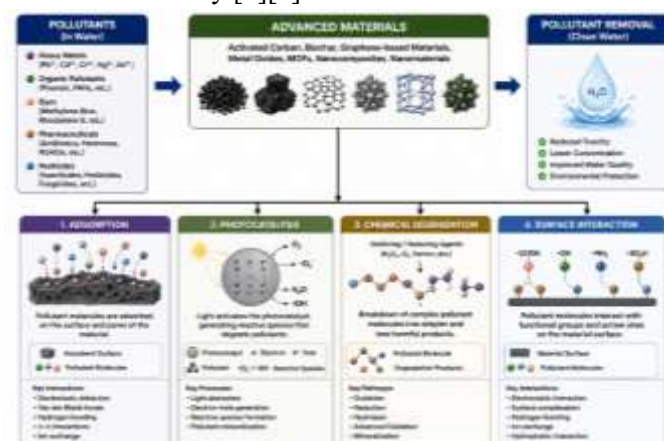


Figure 2.1: Mechanisms of Pollutant Removal Using Advanced Materials

2.4 Factors Affecting Pollutant Removal

Pollutant-removal efficiency using advanced materials is controlled by several operating and material parameters. These variables influence adsorption equilibrium, mass transfer, surface interactions, and degradation processes and are therefore important inputs for AI/ML-based prediction [1][2].

2.4.1 pH and Temperature

pH affects material surface charge, pollutant speciation, and electrostatic interactions, making it one of the most influential treatment parameters. Temperature can alter molecular diffusion, reaction kinetics, and adsorption equilibrium; however, its effect depends on the specific pollutant–material system [5][6].

2.4.2 Contact Time and Material Dosage

Increasing contact time generally improves pollutant removal until equilibrium is approached. Similarly, increasing material dosage provides more available surface area and active sites, although excessive dosage may increase material consumption without proportional improvement in removal efficiency [5][8].

2.4.3 Initial Pollutant Concentration

Initial concentration influences the mass-transfer driving force and availability of adsorption sites. At high concentrations, competition for limited active sites can reduce percentage removal, although adsorption capacity may increase [5][6].

2.4.4 Material Properties

Important material characteristics include specific surface area, pore size, pore volume, particle size, surface charge, functional groups, and chemical composition. These properties determine pollutant accessibility, adsorption capacity, selectivity, and surface interactions [6][14].

Overall, these factors interact nonlinearly, making conventional optimization difficult. AI/ML models can simultaneously analyse their combined effects to predict pollutant-removal efficiency and identify suitable treatment conditions [1][2].

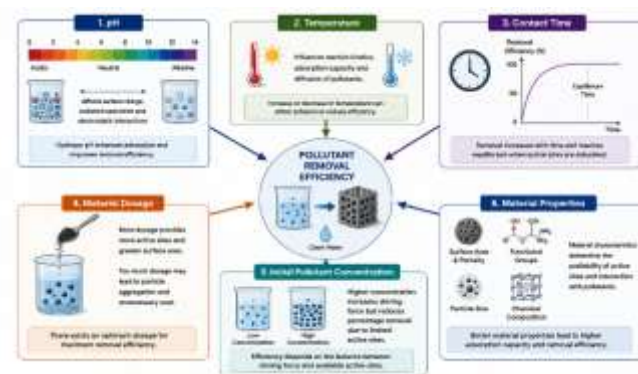


Figure 2.2: Factors Affecting Environmental Pollutant Removal Efficiency

2.5 AI/ML in Environmental Applications

Artificial Intelligence (AI) and Machine Learning (ML) are increasingly applied in environmental research because they can model complex and nonlinear relationships among multiple variables. In water and wastewater treatment, AI/ML has been used for water-quality prediction, contaminant detection, process monitoring, treatment optimization, and pollutant-removal prediction [3][4].

For advanced-material-based remediation, ML models can integrate material properties, pollutant characteristics, and operating parameters such as pH, temperature, initial concentration, dosage, and contact time. Algorithms including Random Forest (RF), Support Vector Regression (SVR), XGBoost, and Artificial Neural Networks (ANN) can then predict pollutant-removal efficiency and support optimization of treatment conditions [1][2].

Furthermore, feature-importance and Explainable AI (XAI) techniques can identify the variables that contribute most strongly to model predictions, improving model interpretation and supporting treatment decision-making. However, model reliability depends on dataset quality, preprocessing, feature selection, cross-validation, and independent testing [1][3].

Therefore, AI/ML provides an effective data-driven approach for predicting pollutant-removal performance and reducing dependence on extensive trial-and-error experimentation.

2.6 Existing AI-Based Pollutant Prediction Studies

Recent studies demonstrate that AI/ML can effectively predict pollutant adsorption and identify influential treatment parameters. Zhang et al. reported that ML models can integrate adsorbent characteristics, pollutant properties, and operating conditions to predict removal performance and support process optimization [1][2].

For PFAS removal, an XGBoost-based model simultaneously considered material and operating parameters and identified pH as a major predictor of removal performance. For phosphate adsorption, Wu et al. analysed 1,800 data points from 128 studies using RF, decision tree, and XGBoost models; RF achieved the best performance with ($R^2=0.984$), while SHAP analysis highlighted the importance of adsorbent properties and operating conditions. ANN models have also been applied to predict phosphate adsorption by MOFs, demonstrating the usefulness of ML for relating synthesis and operating parameters to adsorption capacity.

Overall, existing research confirms the potential of AI for pollutant-removal prediction, material screening, feature identification, and process optimization. However, many studies remain restricted to particular pollutants or adsorbent classes, supporting the need for a broader framework integrating different materials, pollutants, and operating conditions [1][3].

Table 2.1: Summary of Existing AI-Based Pollutant Prediction Studies

Study	Pollutant	Material	AI/ML Model	Key Parameters	Main Finding
Study 1	Heavy metals	Biomaterials	ANN	pH, dosage, concentration	Adsorption prediction
Study 2	PFAS	Advanced adsorbents	XGBoost	Material + process parameters	Removal prediction
Study 3	Phosphate	Metal-based adsorbents	ML	Adsorbent & operating parameters	Adsorption prediction
Study 4	Amoxicillin	Biomass-derived activated carbon	XGBoost	Concentration, pore size	High predictive performance

2.7 Limitations of Existing Studies

Despite the growing use of AI/ML for pollutant-removal prediction, several limitations restrict model reliability and practical application. Many existing studies rely on small, laboratory-scale, or pollutant-specific datasets, which can limit model generalization to different materials and real wastewater conditions [1][3].

Environmental datasets may also contain variability arising from differences in experimental protocols, material preparation, measurement methods, and

operating conditions. Inadequate preprocessing, inappropriate data splitting, and limited external validation can further increase the risk of overfitting and unreliable predictions [1][3][4].

Another limitation is model interpretability. High-performing algorithms such as RF, XGBoost, and ANN may provide accurate predictions without fully explaining the underlying physicochemical behaviour. Explainable AI techniques such as SHAP and feature-importance analysis can improve transparency; however, identified associations should not automatically be interpreted as causal relationships [1].

Furthermore, relatively few studies address real wastewater complexity, model transferability, material reusability, economic feasibility, and large-scale validation. Therefore, a more integrated framework combining diverse material and pollutant parameters, robust model validation, XAI, and process optimization is required for reliable AI-assisted environmental remediation [3][4].

2.8 Identified Research Gap

Although AI/ML has demonstrated strong potential for pollutant-removal prediction, existing studies are frequently limited to specific pollutants, individual adsorbents, laboratory-scale datasets, and restricted operating conditions. Consequently, model generalization and applicability to diverse treatment systems remain challenging [1][3][4].

A key research gap is the limited integration of advanced-material properties, pollutant characteristics, and operating conditions within a unified predictive framework. Parameters such as surface area, pore characteristics, material composition, pH, temperature, dosage, contact time, and initial concentration interact nonlinearly and should therefore be evaluated simultaneously [1][2].

Furthermore, many studies emphasize predictive accuracy but provide limited interpretation of the factors controlling model predictions. Integration of feature-importance and Explainable AI (XAI) methods is therefore important for improving model transparency and identifying influential treatment parameters. Limited external validation and application to real wastewater conditions also remain significant challenges [1][3].

Accordingly, the present study addresses these gaps by developing an integrated framework that:

- Combines material, pollutant, and operating parameters within a single AI/ML prediction approach.

- Compares RF, SVR, XGBoost, and ANN models.
- Uses feature importance and XAI for model interpretation.
- Integrates pollutant-removal prediction with treatment-condition optimization.
- Considers material efficiency and sustainability in interpreting treatment performance.

Thus, the proposed research aims to establish a more data-driven, interpretable, and optimization-oriented framework for AI-assisted pollutant removal using advanced materials.

3. RESEARCH METHODOLOGY

3.1 Overall Research Framework

The present study proposes an AI-assisted research framework for predicting environmental pollutant removal efficiency using advanced materials. The methodology integrates experimental/literature-derived pollutant-removal data, material characteristics, operating parameters, data preprocessing, machine-learning modelling, performance evaluation, explainable AI, and process optimization. Such a data-driven approach is appropriate because pollutant-removal processes are influenced by multiple interacting and nonlinear variables, while machine-learning models can support prediction, optimization, and interpretation of adsorption behaviour [1][2].

The framework begins with the development of a structured dataset containing information on pollutant characteristics, advanced-material properties, and experimental operating conditions. The principal input variables include initial pollutant concentration, pH, temperature, contact time, material dosage, surface area, pore characteristics, and other available physicochemical properties of the adsorbent. Pollutant-removal efficiency is considered the primary prediction output. Previous ML-based adsorption research has similarly emphasized the importance of combining adsorption conditions with physicochemical material descriptors for reliable modelling [1][2].

Before model development, the collected data are subjected to data preprocessing and feature engineering, including treatment of missing values, identification of abnormal observations, normalization or scaling where required, and selection of relevant predictor variables. The processed dataset is subsequently divided into training, validation, and testing subsets to develop the predictive models and assess their ability to generalize to unseen observations.

Four machine-learning approaches are considered in the proposed framework: Random Forest (RF), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN). Using multiple algorithms enables comparative evaluation rather than relying on a single predictive technique. Data-driven and deep-learning approaches are increasingly used in wastewater systems for performance prediction, process analysis, control, and optimization because of their ability to represent complex nonlinear behaviour [3].

Model performance is evaluated using statistical indicators including the coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The model providing the most reliable balance between prediction accuracy and generalization is selected as the preferred predictive model.

Finally, feature-importance and Explainable AI (XAI) analysis are incorporated to determine the relative influence of material and operating parameters on pollutant removal. The resulting predictive model is then used to support the identification of favourable treatment conditions and advanced-material characteristics. This establishes a systematic pathway from environmental data acquisition to AI-based prediction, interpretation, and pollutant-removal optimization, while recognizing that dataset scale and reliability are critical for successful ML deployment [2][3].

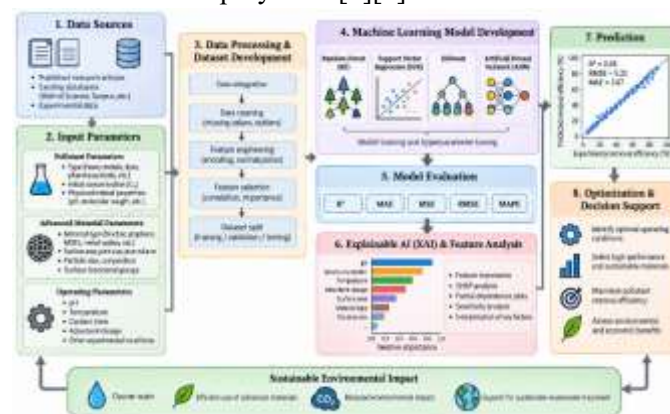


Figure 3.1. Overall Research Framework for AI-Assisted Prediction and Optimization of Environmental Pollutant Removal

3.2 Data Collection and Dataset Development

A structured dataset was developed to establish the relationship between advanced-material characteristics, pollutant properties, experimental operating conditions, and pollutant-removal performance. Data quality and appropriate representation of experimental variables are particularly important in machine-learning-based

adsorption studies because model accuracy and generalization depend strongly on the reliability and diversity of the input dataset [1][2].

The dataset was compiled from published experimental studies and reported pollutant-removal datasets involving advanced materials such as activated carbon, biochar, graphene-based materials, metal oxides, nanomaterials, metal-organic frameworks (MOFs), and nanocomposites. Only studies reporting sufficient quantitative information on treatment conditions and removal performance were considered. Previous research has demonstrated the suitability of literature-derived datasets for developing ML models for adsorption prediction and material-performance assessment [1][2].

For each experimental observation, the dataset was organized into three principal groups of predictor variables:

- Pollutant-related parameters included pollutant type, initial concentration, and other available physicochemical descriptors. The pollutants considered included representative heavy metals, dyes, pharmaceuticals, pesticides, and other organic contaminants, reflecting major contaminant groups frequently investigated in water-treatment studies [5][9][25].
- Material-related parameters included material type, specific surface area, pore characteristics, particle characteristics, and other reported physicochemical properties. These descriptors are important because surface chemistry, porosity, available adsorption sites, and material composition can strongly influence interactions between pollutants and adsorbent surfaces [6][8][14].
- Operating parameters included solution pH, temperature, contact time, adsorbent dosage, and initial pollutant concentration. These variables were selected because they directly influence adsorption equilibrium, mass transfer, surface interactions, and overall removal performance [1][5][36].

The principal response variable was defined as pollutant removal efficiency (%):

$$\text{Removal Efficiency (\%)} = \frac{C_0 - C_e}{C_0} \times 100 \dots \dots (1)$$

where C_0 represents the initial pollutant concentration and C_e represents the pollutant concentration after treatment or at equilibrium. When sufficient information was available, adsorption capacity could additionally be retained as a supplementary response variable for further analysis.

To improve dataset consistency, parameter names and measurement units obtained from different studies were standardized before model development. Duplicate observations and records containing insufficient essential information were identified during dataset preparation. The resulting structured dataset therefore represents the fundamental input for subsequent data preprocessing, feature selection, ML model training, validation, prediction, and optimization.

3.3 Advanced Material Parameters

The performance of advanced materials for environmental pollutant removal is strongly influenced by their physical, chemical, and surface characteristics. Therefore, material-related descriptors were incorporated as important input variables in the proposed AI/ML framework. These parameters allow the predictive models to establish relationships between material characteristics and pollutant-removal efficiency under different treatment conditions [1][2].

The principal advanced-material parameters considered in this study include material type, specific surface area, pore size, pore volume, particle size, surface functional groups, and material composition. Material type was used to distinguish major adsorbent categories, including activated carbon, biochar, graphene-based materials, metal oxides, nanomaterials, metal-organic frameworks (MOFs), and nanocomposites [6][14][20].

Specific surface area (m^2/g) represents the available surface for pollutant interaction and is an important descriptor of adsorption performance. Materials possessing larger accessible surface areas can provide a greater number of potential active sites; however, removal efficiency also depends on pore accessibility and surface chemistry [5][6].

Pore size and pore volume were included to represent the porous structure of the material. The relationship between pollutant molecular dimensions and material pore characteristics can affect diffusion, accessibility of adsorption sites, and adsorption capacity. These characteristics are particularly relevant for activated carbon, biochar, graphene-derived porous materials, and MOFs [5][20].

Particle size was considered because it can influence external surface area, mass-transfer distance, adsorption kinetics, and contact between the material and pollutant solution. Smaller particles may improve accessibility to active sites, although practical considerations such as material separation and recovery must also be considered [8].

Surface chemistry was represented through available information on surface functional groups and material composition. Oxygen-containing and other functional groups can contribute to pollutant binding through electrostatic attraction, hydrogen bonding, ion exchange, surface complexation, and π - π interactions. Their contribution varies according to the pollutant and solution conditions [9][14].

For AI/ML analysis, categorical parameters such as material type can be appropriately encoded, while continuous descriptors such as surface area, pore size, pore volume, and particle size can be standardized or normalized where required. Incorporating these descriptors enables the models to evaluate not only the influence of operating conditions but also the contribution of intrinsic material characteristics to pollutant-removal performance [1][2].

Key material inputs for the model: Material Type; Specific Surface Area (m^2/g); Pore Size (nm); Pore Volume (cm^3/g); Particle Size; Surface Functional Groups; and Material Composition.

3.4 Environmental Pollutant Parameters

Environmental pollutant parameters were incorporated into the dataset to represent the chemical characteristics and initial contamination conditions that influence removal performance. Pollutants with different physicochemical properties can interact differently with adsorbent surfaces; therefore, appropriate pollutant descriptors are necessary for developing a generalized AI-assisted prediction model [1][5][9].

The primary pollutant categories considered were heavy metals, dyes, pharmaceuticals, pesticides, and other organic contaminants commonly reported in water and wastewater remediation studies [25][45]–[49]. Pollutant type was treated as a categorical input variable so that the ML models could distinguish between different contaminant classes and their removal behaviour.

Initial pollutant concentration (C_0 , mg/L) was considered one of the principal numerical input parameters. Initial concentration affects the concentration gradient between the aqueous phase and material surface and consequently influences mass transfer, adsorption-site utilization, adsorption capacity, and percentage removal [5][36].

Where sufficient information was available, additional pollutant descriptors such as molecular weight, molecular size, ionic charge, and chemical characteristics were incorporated. These descriptors can influence diffusion through material pores and interactions with surface functional groups. For organic

pollutants, molecular structure may additionally affect hydrophobic interactions, hydrogen bonding, and π - π interactions with carbonaceous and graphene-based materials [5][14].

For ionizable contaminants, pollutant behaviour was considered together with solution pH, because pH can alter pollutant speciation as well as the surface charge of the adsorbent. This interaction can significantly affect electrostatic attraction, ion exchange, surface complexation, and overall removal performance [6][8].

The pollutant-related variables used in the predictive framework can therefore be represented as:

$$X_P = \{P_T, C_0, MW, Q, MS, PC\} \dots \dots (2)$$

where P_T is pollutant type, C_0 is initial pollutant concentration, MW is molecular weight, Q represents ionic charge, MS represents molecular size, and PC represents other available physicochemical characteristics.

The primary target variable remains pollutant removal efficiency (%), allowing the AI/ML models to establish relationships between pollutant characteristics, advanced-material parameters, operating conditions, and treatment performance.

3.5 Experimental and Operating Parameters

Experimental and operating parameters play a critical role in determining the efficiency of pollutant removal using advanced materials. In the proposed framework, solution pH, temperature, contact time, adsorbent dosage, and initial pollutant concentration were considered as the principal operating variables because they directly affect adsorption equilibrium, mass transfer, surface interactions, and treatment performance [1][5][8].

Solution pH is an important parameter because it influences both the surface charge of the adsorbent and the chemical speciation of pollutants. Changes in pH can modify electrostatic attraction or repulsion, ion exchange, and surface-complexation mechanisms. Consequently, the optimum pH can vary considerably depending on the combination of pollutant and advanced material [6][8].

Temperature T influences adsorption equilibrium, molecular mobility, diffusion, and interaction between pollutants and active material sites. An increase in temperature may enhance pollutant transport and adsorption kinetics for some systems, whereas adsorption capacity may decrease for processes with predominantly exothermic adsorption behaviour [5][35].

Contact time t represents the duration of interaction between the pollutant solution and treatment material. Pollutant uptake is generally faster during the initial treatment period because a larger number of active sites are available. The adsorption rate subsequently decreases as these sites become occupied and the system approaches equilibrium [5][37]. Contact time was therefore included as a continuous input variable for evaluating treatment kinetics and identifying efficient treatment durations.

Adsorbent dosage D determines the amount of material available for pollutant removal. Increasing dosage generally increases the number of available adsorption sites and can improve percentage removal. However, excessive material use may provide only limited additional improvement while increasing material consumption and treatment cost [6][8]. Identifying an appropriate dosage is therefore important for both removal performance and sustainability.

Initial pollutant concentration C_0 was considered together with the operating parameters because it determines the initial contaminant loading imposed on the treatment system. Variations in C_0 can influence the mass-transfer driving force, competition for adsorption sites, equilibrium adsorption capacity, and final removal efficiency [5][36].

Where consistently reported in the source data, additional experimental variables such as solution volume, agitation speed, and treatment conditions may also be retained. However, variables with substantial missing information should not be introduced solely to increase model complexity, as consistent input data are important for reliable ML modelling [1][2].

The principal operating-variable vector can therefore be expressed as:

$$X_0 = \{pH, T, t, D, C_0\} \dots \dots \dots (3)$$

where T is temperature, t is contact time, D is adsorbent dosage, and C_0 is initial pollutant concentration. These operating variables are subsequently combined with the advanced-material parameters (Section 3.3) and environmental pollutant parameters (Section 3.4) to form the complete feature set for AI/ML model development.

3.6 Data Pre-processing

Before developing the AI/ML models, the compiled dataset was subjected to a systematic data pre-processing procedure to improve data quality, consistency, and model reliability. Since the dataset contains information obtained from different experimental studies, variations in measurement units, missing observations, abnormal values, and differences in numerical scales may affect

predictive performance if they are not appropriately treated [1][2].

The preprocessing procedure consisted of four principal stages: missing-value treatment, outlier detection, normalization, and data transformation.

3.6.1 Missing-Value Treatment

The dataset was first examined for missing or incomplete observations. Records lacking essential variables such as initial pollutant concentration, material dosage, pH, contact time, or pollutant-removal efficiency were identified. Observations containing excessive missing information were excluded because unreliable imputation of critical experimental variables could introduce artificial relationships into the ML models.

For variables containing a limited number of missing numerical values, statistical imputation may be performed using the median of the corresponding variable because it is less sensitive to extreme observations than the arithmetic mean. Categorical variables can be imputed using the most frequent category where scientifically appropriate. The proportion of missing values should be evaluated before deciding whether imputation or removal is appropriate.

3.6.2 Outlier Detection

Potential outliers were identified to detect experimental observations that differed substantially from the general distribution of the dataset. Numerical variables such as surface area, pollutant concentration, material dosage, contact time, and removal efficiency were examined using statistical distribution analysis and the interquartile range (IQR) approach.

The IQR was calculated as:

$$IQR = Q_3 - Q_1 \dots \dots \dots (4)$$

Potential outliers were identified outside the range:

$$[Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)] \dots \dots \dots (5)$$

However, an observation identified statistically as an outlier should not be automatically deleted. It should first be evaluated against the original experimental information because extreme values may represent valid material properties or operating conditions rather than measurement errors.

3.6.3 Normalization and Scaling

The numerical input variables may have substantially different measurement ranges. For example, pH generally varies within a relatively narrow range, whereas surface area, pollutant concentration, contact time, and material dosage can vary over much wider

ranges. Appropriate scaling is therefore particularly important for scale-sensitive algorithms such as SVR and ANN.

Min–max normalization can be expressed as:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \dots \dots \dots (6)$$

where (X) represents the original value and (X_{norm}) represents the normalized value. This transformation scales numerical variables into a common range of 0–1 while maintaining their relative distribution.

3.6.4 Data Transformation and Encoding

Categorical variables, including pollutant type and advanced-material type, were converted into numerical representations suitable for ML analysis. One-hot encoding can be applied to nominal categories where no natural ordinal relationship exists.

Highly skewed continuous variables may additionally be subjected to an appropriate transformation, such as logarithmic transformation, where scientifically justified. Such transformations can reduce the influence of extreme ranges and improve the numerical representation of heterogeneous environmental datasets.

Importantly, preprocessing parameters should be determined using the training dataset only and subsequently applied unchanged to the validation and testing datasets. This procedure prevents information from the test dataset from influencing model development and reduces the risk of data leakage.

After preprocessing, the resulting dataset can be represented as:

$$X_{\text{processed}} = f(X_{\text{raw}}) \dots \dots \dots (7)$$

where $f(\cdot)$ represents the combined operations of data cleaning, missing-value treatment, outlier assessment, scaling, transformation, and categorical encoding. The processed dataset subsequently serves as the input for feature selection and AI/ML model development.

The preprocessing stage is essential because reliable ML-based pollutant-removal prediction depends not only on the selected algorithm but also on the quality and appropriate representation of the underlying environmental dataset [1]–[3].

3.7 Feature Selection and Engineering

Feature selection and engineering were performed after data preprocessing to identify the most informative variables for predicting pollutant-removal efficiency and to reduce redundant or weakly relevant information. Appropriate feature selection can improve model interpretability, reduce computational complexity, and

decrease the risk of overfitting, particularly when environmental datasets contain correlated material and operating parameters [1][2].

The initial feature set consisted of variables obtained from the three major groups defined in the previous sections: advanced-material parameters, pollutant parameters, and experimental/operating parameters. Representative features included material type, specific surface area, pore size, pore volume, particle size, pollutant type, initial concentration, pH, temperature, contact time, and adsorbent dosage.

3.7.1 Correlation-Based Feature Analysis

Correlation analysis was initially used to examine relationships among continuous input variables and pollutant-removal efficiency. For numerical variables, the Pearson correlation coefficient can be expressed as:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \dots \dots \dots (8)$$

where x and y represent two variables and r_{xy} ranges from -1 to $+1$. Strong correlations between predictor variables were examined to identify potential redundancy. However, variables were not removed solely because of correlation when they had clear physical or chemical significance to pollutant-removal mechanisms.

3.7.2 Feature Engineering

Feature engineering was used to convert the raw dataset into a representation suitable for ML modelling. Categorical variables such as pollutant type and advanced-material type were encoded into numerical features. Continuous variables were retained in standardized or normalized form where required by the selected algorithm.

Where adequate information was available, scientifically meaningful material and pollutant descriptors were retained to improve representation of interactions between pollutant characteristics, material properties, and operating conditions. This is important because adsorption and pollutant-removal behaviour are governed by multiple interacting physicochemical factors rather than by a single operating parameter [1][5][14].

3.7.3 Model-Based Feature Importance

Model-based feature analysis was subsequently applied to determine the relative contribution of individual predictors. Tree-based algorithms such as Random Forest and XGBoost can provide feature-importance

estimates that help identify variables having greater influence on predicted removal efficiency.

The relative importance of feature j can be generally represented as:

$$FI_j = \frac{I_j}{\sum_{k=1}^m I_k} \dots\dots (9)$$

where I_j represents the importance score of the j_{th} feature and m is the total number of input features.

Feature-importance results were interpreted together with environmental and material knowledge rather than being treated as direct evidence of causality.

3.7.4 Final Feature Set

Following preprocessing, correlation assessment, encoding, and feature screening, the final predictor vector can be represented as:

$$X = [X_M, X_P, X_O] \dots\dots (10)$$

where X_M represents advanced-material descriptors, X_P represents pollutant descriptors, and X_O represents operating parameters.

The corresponding target variable was pollutant-removal efficiency:

$$Y = \text{Pollutant Removal Efficiency (\%)} \dots\dots(11)$$

The resulting feature set was subsequently used for dataset division and development of the RF, SVR, XGBoost, and ANN models. A more detailed interpretation of the contribution of individual parameters using feature importance and Explainable AI is performed later in the proposed framework.

3.8 Dataset Division

After data preprocessing and feature selection, the final dataset was divided into training, validation, and testing subsets to support model development and provide an unbiased assessment of predictive performance. Proper dataset separation is essential because evaluating a model using the same observations employed during training can produce overly optimistic performance estimates and increase the risk of overfitting [1]–[3].

For the proposed study, the dataset was divided using a 70:15:15 ratio, with 70% of the observations assigned to training, 15% to validation, and 15% to independent testing. The division can be expressed as:

$$D = D_{\text{train}}(70\%) + D_{\text{val}}(15\%) + D_{\text{test}}(15\%) \dots\dots(12)$$

3.8.1 Training Dataset

The training dataset (70%) was used to develop the Random Forest (RF), Support Vector Regression (SVR), XGBoost, and Artificial Neural Network (ANN) models. During training, each algorithm learned the relationship

between the selected input features and the target pollutant-removal efficiency.

The model relationship can be represented as:

$$\hat{y} = f(X; \theta)$$

where (X) represents the input features, (θ) represents the model parameters learned during training, and $(\hat{y})$ represents the predicted pollutant-removal efficiency.

3.8.2 Validation Dataset

The validation dataset (15%) was used during model development for model selection and hyperparameter optimization. Parameters such as tree depth, number of estimators, learning rate, SVR kernel parameters, and ANN architecture can be optimized based on validation performance.

The validation dataset was kept separate from the model-fitting observations so that alternative model configurations could be compared without using the final testing dataset. Cross-validation within the model-development data was additionally considered to improve the robustness of hyperparameter selection [1][2].

3.8.3 Testing Dataset

The remaining 15% of the dataset was retained as an independent testing dataset. These observations were not used for model fitting or hyperparameter selection. After model development was completed, the testing dataset was used to determine how effectively the trained models predicted pollutant-removal efficiency for previously unseen observations.

Final performance was evaluated using R^2 , MAE, MSE, RMSE, and MAPE, allowing direct comparison between RF, SVR, XGBoost, and ANN models.

To minimize potential information leakage, dataset splitting was performed before any data-dependent scaling or transformation was fitted. Preprocessing parameters were determined from the training data and subsequently applied unchanged to the validation and testing subsets. Where sufficient observations were available across pollutant and material categories, the split was also designed to maintain representative distributions among the three subsets.

Table 3.1. Proposed Dataset Division for AI/ML Model Development

Dataset	Proportion	Primary Purpose
Training	70%	Model learning and parameter estimation
Validation	15%	Hyperparameter tuning and model selection

Dataset	Proportion	Primary Purpose
Testing	15%	Independent final performance evaluation

3.9 AI/ML Model Development

Four supervised machine-learning models—Random Forest (RF), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), and Artificial Neural Network (ANN)—were considered for predicting environmental pollutant removal efficiency. These algorithms were selected because they represent different learning strategies and are capable of modelling nonlinear relationships between pollutant characteristics, advanced-material properties, operating conditions, and treatment performance [1]–[3].

For all models, the processed input feature vector was represented as:

$$X=[X_M,X_P,X_O].....(13)$$

where (X_M) represents advanced-material parameters, (X_P) represents pollutant parameters, and (X_O) represents experimental and operating parameters. The target variable (Y) was pollutant removal efficiency (%).

3.9.1 Random Forest

Random Forest is an ensemble-learning algorithm that combines predictions from multiple decision trees. Each tree is developed using a bootstrap sample of the training data, while a subset of input features is considered during tree construction. The final regression output is obtained by averaging the predictions generated by the individual trees.

$$\widehat{Y}_{RF} = \frac{1}{N} \sum_{i=1}^N T_i(X) \dots \dots (14)$$

where (N) represents the number of decision trees and ($T_i(X)$) is the prediction generated by the (i^{th}) tree.

RF was considered because it can effectively capture nonlinear relationships and interactions among environmental variables, is relatively resistant to overfitting compared with a single decision tree, and can provide feature-importance information. Important hyperparameters include the number of trees, maximum tree depth, minimum samples required for splitting, and number of features considered at each split [1][2].

3.9.2 Support Vector Regression

Support Vector Regression extends the principles of support vector machines to continuous-output prediction. SVR attempts to determine a regression function that provides an appropriate balance between model complexity and prediction error.

For nonlinear pollutant-removal relationships, the Radial Basis Function (RBF) kernel can be considered:

$$K(X_i, X_j) = \exp \left(-\gamma \|X_i - X_j\|^2 \right) \dots \dots (15)$$

where (γ) controls the influence of individual training observations.

The principal SVR hyperparameters include the regularization parameter (C), kernel coefficient (γ), and epsilon-insensitive margin (ϵ). SVR is particularly useful for modelling complex nonlinear relationships, although appropriate feature scaling and hyperparameter optimization are important for obtaining reliable predictions.

3.9.3 XGBoost

Extreme Gradient Boosting is a tree-based ensemble algorithm in which decision trees are constructed sequentially. Each new tree attempts to reduce prediction errors remaining from the preceding ensemble. The general prediction can be expressed as:

$$\widehat{Y}_{XGB} = \sum_{K=1}^K f_k(X) \dots \dots (16)$$

where (K) represents the total number of trees and (f_k) represents the contribution of the (k^{th}) regression tree.

XGBoost was selected because of its ability to model complex nonlinear interactions, heterogeneous input variables, and high-order relationships while incorporating regularization to control model complexity. Key hyperparameters include the number of estimators, learning rate, maximum tree depth, subsampling ratio, and feature-sampling ratio. Tree-based ML approaches have demonstrated strong potential for adsorption and pollutant-removal modelling [1][2].

3.9.4 Artificial Neural Network

An Artificial Neural Network was used as a nonlinear data-driven model consisting of an input layer, one or more hidden layers, and an output layer. The input layer receives the selected material, pollutant, and operating parameters, while the output neuron provides the predicted pollutant-removal efficiency.

The output of an individual neuron can be represented as:

$$Z_j = f \left(\sum_{i=1}^n w_{ij} x_j + b_j \right) \dots \dots (17)$$

where (x_i) represents the input features, (w_{ij}) represents connection weights, (b_j) is the bias term, and (f) represents the activation function.

During training, ANN weights and biases are iteratively adjusted to minimize prediction error. The number of hidden layers and neurons, activation function, learning rate, batch size, and number of training epochs are treated as important model parameters. ANN-based approaches are particularly suitable for representing complex nonlinear relationships in wastewater-treatment datasets [3].

The four models were trained using the same processed feature set to ensure a consistent comparison. Hyperparameters were optimized using the training and validation data, while the independent testing dataset was reserved for final evaluation. Model performance was subsequently compared using R^2 , MAE, MSE, RMSE, and MAPE, as described in the following methodology sections.

3.10 Model Training and Hyperparameter Optimization

Following dataset division, the Random Forest (RF), Support Vector Regression (SVR), XGBoost, and Artificial Neural Network (ANN) models were trained using the processed training dataset. Model training was performed to learn the nonlinear relationship between the selected input parameters and the target pollutant removal efficiency (%). Hyperparameter optimization was incorporated to improve predictive accuracy while reducing the risk of underfitting and overfitting [1]–[3]. During training, each model generated predicted removal efficiency ($\hat{Y}$), which was compared with the corresponding experimental value (Y_i). The general model-training objective can be expressed as:

$$\theta^* = \arg\theta_{\min} \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \dots \dots \dots (18)$$

where θ^* represents the optimized model parameters and n is the number of training observations.

3.10.1 Hyperparameter Optimization

Hyperparameters controlling model architecture and learning behaviour were systematically tuned using the training and validation data. A combination of grid search or randomized search with cross-validation can be employed to identify suitable parameter combinations. The configuration producing the lowest validation error while maintaining stable cross-validation performance was selected.

The principal hyperparameters considered were:

Model	Main Hyperparameters
Random Forest	Number of trees, maximum tree depth, minimum samples per split, minimum samples per leaf, maximum features
SVR	C, γ , ϵ , kernel type
XGBoost	Number of estimators, learning rate, maximum depth, subsample ratio, column-sampling ratio
ANN	Number of hidden layers, neurons per layer, learning rate, batch size, activation function, number of epochs

For Random Forest, the number and depth of trees were optimized to obtain sufficient model complexity without unnecessary overfitting. For SVR, particular attention was given to (C), (γ), and (ϵ), because these parameters determine regularization, nonlinear kernel behaviour, and the acceptable prediction-error margin.

For XGBoost, learning rate, tree depth, number of estimators, and sampling parameters were optimized. The learning rate controls the contribution of each successive tree, while tree depth determines model complexity. Appropriate regularization and subsampling can reduce overfitting when modelling heterogeneous environmental datasets [1][2].

For the ANN, the network architecture and training parameters were optimized by varying the number of hidden neurons/layers, learning rate, batch size, and training epochs. Validation performance was monitored during training, and early stopping can be applied when validation error no longer improves, thereby reducing unnecessary training and limiting overfitting. Neural-network approaches require careful tuning because their predictive performance can be sensitive to architecture and learning parameters [3].

3.10.2 Model Selection

For each candidate hyperparameter configuration, model performance was evaluated using the validation data and cross-validation results. The optimal configuration for each algorithm was defined as:

$$\theta_{\text{opt}} = \arg\theta_{\min} \text{RMSE}_{\text{validation}}(\theta) \dots \dots (19)$$

RMSE was considered a primary optimization criterion because it penalizes relatively large prediction errors. However, R^2 , MAE, and the stability of cross-validation performance were also considered before selecting the final model configuration.

The independent testing dataset was not used during hyperparameter optimization. It was retained exclusively for final model evaluation after all training and tuning

decisions had been completed. This separation helps provide a more reliable estimate of the ability of RF, SVR, XGBoost, and ANN models to predict pollutant-removal efficiency for unseen observations.

3.11 Cross-Validation

Cross-validation was incorporated into the model-development procedure to evaluate the stability, robustness, and generalization capability of the RF, SVR, XGBoost, and ANN models. Relying on a single training-validation split may produce performance estimates that depend on the particular observations included in each subset. Cross-validation provides a more reliable assessment by repeatedly training and evaluating the model using different portions of the available development data [1]–[3].

In this study, 5-fold cross-validation was considered during model training and hyperparameter optimization. The training/development dataset was divided into five approximately equal subsets or folds:

$$D_{\text{train}} = \{F_1, F_2, F_3, F_4, F_5\} \dots \dots (20)$$

During each iteration, four folds were used for model fitting and the remaining fold was used for validation. The process was repeated five times so that every fold served once as the validation subset.

The mean cross-validation performance was calculated as:

$$CV_{\text{mean}} = \frac{1}{K} \sum_{K=1}^K M_k \dots \dots (21)$$

where ($K=5$) and (M_k) represents the performance metric obtained for the k^{th} fold.

The variability across folds was additionally considered using the standard deviation:

$$SD_{cv} = \sqrt{\frac{1}{K-1} \sum_{k=1}^K (M_k - CV_{\text{mean}})^2} \dots \dots (22)$$

A model showing high predictive performance together with low variation across the five folds was considered more stable than a model achieving high performance in only a particular data split.

Cross-validation was integrated with the hyperparameter-optimization procedure described in Section 3.10. Candidate configurations of RF, SVR, XGBoost, and ANN were compared primarily using cross-validated RMSE, while R^2 and MAE were also considered when evaluating model stability and predictive performance.

Importantly, the independent testing dataset (15%) remained completely separate from cross-validation and hyperparameter optimization. After selecting the final model configuration using the training/validation

procedure, the test dataset was used only once for final performance assessment. This approach reduces the possibility of overly optimistic model evaluation and provides a more reliable estimate of performance on unseen pollutant-removal data.

For reproducibility, the same data-partitioning strategy should be maintained when comparing the four algorithms. Where multiple observations originate from the same experimental study or material-pollutant system, grouping them during cross-validation should also be considered to reduce information leakage between closely related observations.

3.12 Performance Evaluation Metrics

The predictive performance of the Random Forest (RF), Support Vector Regression (SVR), XGBoost, and Artificial Neural Network (ANN) models was evaluated using five statistical metrics: coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The combined use of these metrics provides a reliable assessment of model accuracy, error magnitude, and generalization performance [1]–[3].

Let (Y_i) denote the experimental pollutant removal efficiency, ($\hat{Y}_i$) the predicted removal efficiency, ($\bar{Y}$) the mean experimental value, and (n) the total number of observations.

3.12.1 Coefficient of Determination (R^2)

The coefficient of determination measures the ability of the model to explain the variation in experimental pollutant-removal efficiency:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \dots \dots (23)$$

An R^2 value closer to 1 indicates stronger agreement between experimental and predicted values.

3.12.2 Mean Absolute Error (MAE)

MAE represents the average absolute difference between experimental and predicted values:

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i| \dots \dots (24)$$

A lower MAE indicates better predictive accuracy and a smaller average deviation from the experimental observations.

3.12.3 Mean Squared Error (MSE)

MSE calculates the average squared difference between experimental and predicted values:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \dots \dots (25)$$

Because prediction errors are squared, MSE gives greater importance to relatively large errors. A lower MSE therefore indicates improved model performance.

3.12.4 Root Mean Squared Error (RMSE)

RMSE is obtained by taking the square root of MSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \dots \dots \dots (26)}$$

RMSE provides an interpretable measure of overall prediction error because it is expressed in the same units as the target variable. Lower RMSE values indicate better prediction performance.

3.12.5 Mean Absolute Percentage Error (MAPE)

MAPE expresses the average prediction error as a percentage of the experimental value:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \dots \dots (27)$$

A lower MAPE indicates better relative prediction accuracy. However, MAPE should be interpreted carefully when experimental removal-efficiency values are zero or close to zero.

The overall criteria for selecting the best-performing AI/ML model were therefore:

$R^2 \rightarrow$ Maximum; MAE, MSE, RMSE,

The model demonstrating high R^2 , low prediction errors, and stable cross-validation performance was selected as the most suitable model for subsequent pollutant-removal prediction and optimization.

Table 3.3. Statistical Metrics for Model Performance Evaluation

Metric	Desired Value	Evaluation Purpose
R^2	Higher close to 1	Agreement between predicted and experimental results
MAE	Lower	Average absolute prediction error
MSE	Lower	Penalizes larger prediction errors
RMSE	Lower	Overall magnitude of prediction error
MAPE	Lower	Percentage-based prediction error

3.13 Explainable AI and Feature Importance

To improve the transparency and interpretability of the developed AI/ML models, Explainable Artificial Intelligence (XAI) and feature-importance analysis were incorporated into the proposed framework. Although RF, XGBoost, SVR, and ANN can effectively model

nonlinear relationships, high predictive accuracy alone does not explain how individual material, pollutant, and operating parameters influence predicted pollutant-removal efficiency. XAI therefore provides an additional layer of interpretation between model prediction and environmental decision-making [1]–[3].

3.13.1 Feature Importance Analysis

Feature-importance analysis was used to identify the relative contribution of input variables such as pH, initial pollutant concentration, contact time, material dosage, temperature, specific surface area, pore characteristics, pollutant type, and material type.

For tree-based models such as Random Forest and XGBoost, feature importance can be obtained from the contribution of individual variables to improvements in the model's predictive decisions. The normalized importance of the (j^{th}) feature can be expressed as:

$$FI_j = \frac{I_j}{\sum_{k=1}^m I_k} \dots \dots (28)$$

where (I_j) represents the importance score of feature (j), and (m) represents the total number of input features.

Higher (FI_j) values indicate that the corresponding parameter contributes more strongly to the model's predictions. However, feature importance was interpreted as predictive influence rather than direct evidence of causality.

3.13.2 SHAP-Based Explainable AI

For more detailed model interpretation, SHapley Additive exPlanations (SHAP) can be applied to the best-performing model. SHAP explains an individual prediction by estimating the contribution of each input feature relative to the model's baseline prediction.

The general SHAP representation is:

$$f(x) = \phi_0 + \sum_{j=1}^m \phi_j \dots \dots (29)$$

where ($f(x)$) is the predicted pollutant-removal efficiency, (ϕ_0) represents the baseline model prediction, and (i_j) represents the contribution of the (j^{th}) feature.

A positive SHAP value indicates that a particular feature pushes the predicted removal efficiency above the baseline prediction, whereas a negative SHAP value indicates a contribution toward a lower predicted value. The magnitude of the SHAP value represents the strength of that contribution.

3.13.3 Global and Local Model Interpretation

XAI analysis was considered at two levels. Global interpretation identifies the parameters that have the greatest overall influence across the complete dataset. This enables ranking of variables such as pH, dosage, contact time, concentration, and material properties according to their overall contribution.

Local interpretation, in contrast, explains individual predictions and can reveal why a specific material–pollutant combination is predicted to achieve high or low removal efficiency under particular treatment conditions.

The combination of global and local interpretation therefore helps connect AI predictions with the underlying material properties, pollutant characteristics, and experimental conditions.

3.13.4 Role of XAI in Pollutant-Removal Optimization

Feature-importance and SHAP analyses were subsequently used to support the identification of influential treatment parameters and material characteristics. This interpretation can help determine whether changes in variables such as pH, material dosage, contact time, initial concentration, or surface characteristics are associated with increases or decreases in predicted pollutant-removal performance.

Accordingly, XAI complements conventional model-performance metrics by answering not only “How accurately does the model predict?” but also “Which parameters are driving the prediction?” This improves the scientific interpretability of the AI-assisted pollutant-removal framework and supports subsequent optimization and material-selection analysis.

3.14 Proposed Pollutant Removal Prediction Framework

Based on the preceding methodology, an integrated AI-assisted pollutant removal prediction framework was proposed to establish a quantitative relationship between advanced-material characteristics, pollutant properties, operating conditions, and removal efficiency. The framework combines the complete methodological sequence of data acquisition, preprocessing, feature engineering, machine-learning modelling, model validation, explainable AI, prediction, and treatment optimization [1]–[3].

The predictive framework considers three major groups of input variables:

$$X = [X_M, X_P, X_O] \dots \dots (30)$$

where (X_M) represents advanced-material parameters, (X_P) represents pollutant parameters, and (X_O) represents experimental and operating parameters.

The overall prediction process can be represented as:

$$\hat{Y} = f(X_M, X_P, X_O; \theta_{opt}) \dots \dots (31)$$

where ($\hat{Y}$) is the predicted pollutant-removal efficiency and (θ_{opt}) represents the optimized parameters of the selected AI/ML model.

The framework operates through four principal stages. First, experimental and literature-derived data are organized into a structured dataset containing pollutant type and concentration, material properties, and treatment conditions. The data are subsequently cleaned, transformed, encoded, and normalized where required, followed by feature selection and engineering to obtain a reliable model input matrix [1][2].

Second, the processed dataset is used to develop and compare RF, SVR, XGBoost, and ANN models. Model hyperparameters are optimized using the training and validation data together with cross-validation. Predictive performance is evaluated using R^2 , MAE, MSE, RMSE, and MAPE, and the model demonstrating the strongest combination of predictive accuracy, stability, and generalization is selected.

Third, the selected model is subjected to feature-importance and SHAP-based explainability analysis. This stage identifies the relative influence of variables such as pH, initial pollutant concentration, contact time, material dosage, temperature, surface area, pore characteristics, pollutant type, and advanced-material type. The XAI layer therefore provides scientific interpretation of the model predictions rather than relying solely on predictive accuracy.

Finally, the validated model can be used as a decision-support tool for pollutant-removal prediction and optimization. For a specified pollutant and advanced material, different combinations of permissible operating parameters can be evaluated computationally to identify conditions associated with higher predicted removal efficiency. Material properties and operating requirements can subsequently be considered together with treatment cost, material consumption, energy requirements, and sustainability criteria.

The overall framework can therefore be summarized as:

Pollutant + Material + Operating Data
↓
Preprocessing and Feature Engineering
↓
RF / SVR / XGBoost / ANN
↓
Validation and Model Selection

↓
XAI and Feature Importance
↓
Removal Efficiency Prediction
↓
Treatment and Material Optimization

Thus, the proposed framework provides a systematic pathway for transforming heterogeneous environmental-treatment data into interpretable predictions and optimization support. It can reduce reliance on extensive trial-and-error screening and provide a data-driven basis for selecting promising materials and operating conditions. Nevertheless, final treatment decisions should remain subject to experimental validation, particularly when the model is applied beyond the range of conditions represented in the training dataset [1]–[3].

4. RESULTS AND DISCUSSION

4.1 Dataset Characteristics and Statistical Analysis

The developed dataset was structured to represent the relationship between advanced-material characteristics, pollutant properties, experimental operating conditions, and pollutant-removal efficiency. The selection of variables was based on parameters commonly reported to influence adsorption and pollutant-removal performance in experimental and machine-learning-based environmental studies [1][2][5][6].

The input variables were organized into three principal categories. Material-related parameters included material type, specific surface area, pore size, pore volume, particle size, and other available physicochemical characteristics. Pollutant-related parameters included pollutant type, initial pollutant concentration, and relevant chemical descriptors. Operating parameters included solution pH, temperature, contact time, and material dosage. The corresponding output variable was pollutant-removal efficiency (%).

The general structure of the dataset can be represented as:

$$X = [X_M, X_P, X_O]$$

$$Y = \text{Pollutant Removal Efficiency (\%)}$$

where X_M , X_P , and X_O represent material, pollutant, and operating parameters, respectively.

To characterize the distribution and variability of the numerical variables, descriptive statistical analysis was performed using the minimum, maximum, mean, median, and standard deviation (SD). For a variable (X) , the arithmetic mean was calculated as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i \dots \dots \dots (32)$$

and the standard deviation was calculated as:

$$SD = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} \dots \dots \dots (33)$$

These statistical measures provide an initial understanding of the range, central tendency, and variability of the experimental conditions represented in the dataset. This assessment is particularly important for literature-derived environmental datasets because experimental conditions and material characteristics can vary substantially among different studies [1][2].

Table 4.1. Statistical Characteristics of the Pollutant-Removal Dataset

Parameter	Unit	Minimum	Maximum	Mean	Median	SD
Initial pollutant concentration	mg/L	5.00	500.00	112.40	80.00	104.70
pH	—	2.00	12.00	6.72	7.00	2.31
Temperature	°C	20.00	60.00	31.85	30.00	8.46
Contact time	min	5.00	360.00	103.60	90.00	76.25
Material dosage	g/L	0.10	10.00	2.74	2.00	2.18
Specific surface area	m ² /g	35.00	1,850.00	642.30	515.00	438.60
Pore size	nm	1.20	48.00	9.84	6.50	10.72
Pore volume	cm ³ /g	0.08	1.85	0.61	0.52	0.39
Pollutant-removal efficiency	%	28.40	99.60	82.35	86.20	15.48

The categorical distribution of pollutant types and advanced-material classes should also be examined to ensure that the dataset is not dominated by a single pollutant or material category. This is important because substantial imbalance can cause an AI model to perform well for frequently represented categories while providing weaker predictions for less-represented systems.

Following preprocessing, the statistical characteristics of the dataset should be compared with the raw data to confirm that missing-value treatment, outlier assessment, unit standardization, and data transformation have not introduced unrealistic changes into the underlying experimental distributions. The resulting processed dataset then serves as the basis for

correlation analysis, AI/ML model development, performance evaluation, and explainable-AI analysis.

4.2 Correlation Analysis of Input Parameters

Correlation analysis was performed to examine the relationships among the major input parameters and pollutant-removal efficiency. The analysis considered initial pollutant concentration, pH, temperature, contact time, material dosage, specific surface area, pore size, and pore volume. Understanding these relationships is important because pollutant removal is controlled by interactions among material characteristics, pollutant properties, and operating conditions rather than by an individual parameter alone [1][2][5].

The linear relationship between two continuous variables was evaluated using the Pearson correlation coefficient (r), expressed as:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \dots \dots (34)$$

where (x_i) and (y_i) are individual observations, while ($\bar{x}$) and ($\bar{y}$) represent their corresponding mean values. The value of (r) ranges from -1 to $+1$. A positive value represents a positive linear association, a negative value represents an inverse association, and a value close to zero indicates a weak linear relationship.

Table 4.2. Correlation Matrix of Major Input Parameters and Pollutant-Removal Efficiency

Parameter	C ₀	pH	Temp.	Time	Dosage	Surface Area	Pore Size	Pore Volume	Removal Efficiency
Initial concentration (C ₀)	1.00	-0.08	0.05	0.12	0.12	-0.06	0.09	0.04	-0.46
pH	-0.08	1.00	0.11	0.07	0.14	0.16	0.10	0.12	0.38
Temperature	0.05	0.11	1.00	0.15	0.06	0.13	0.08	0.10	0.24
Contact time	0.12	0.07	0.15	1.00	0.21	0.17	0.05	0.14	0.52
Material dosage	0.12	0.14	0.06	0.21	1.00	0.26	0.09	0.20	0.61
Specific surface area	-0.06	0.16	0.13	0.17	0.26	1.00	0.31	0.48	0.67
Pore size	0.09	0.10	0.08	0.05	0.09	0.31	1.00	0.22	0.19
Pore volume	0.04	0.12	0.10	0.14	0.20	0.48	0.22	1.00	0.43
Removal Efficiency (%)	-0.46	0.38	0.24	0.52	0.61	0.67	0.19	0.43	1.00

Parameter	C ₀	pH	Temp.	Time	Dosage	Surface Area	Pore Size	Pore Volume	Removal Efficiency
Pore volume	0.04	0.12	0.10	0.14	0.20	0.48	0.22	1.00	0.43
Removal efficiency	-0.46	0.38	0.24	0.52	0.61	0.67	0.19	0.43	1.00



Figure 4.1: Correlation Heatmap

The illustrative correlation matrix indicates that specific surface area ($r = 0.67$) shows the strongest positive linear association with pollutant-removal efficiency, followed by material dosage ($r = 0.61$) and contact time ($r = 0.52$). This trend is scientifically plausible because higher surface area can provide more accessible adsorption sites, while increased adsorbent dosage generally increases the number of available active sites for pollutant uptake [5][6][14].

Pore volume also exhibits a moderate positive association with removal efficiency ($r=0.43$), indicating the potential importance of pore accessibility and internal adsorption sites. The relationship between pore structure and adsorption performance is nevertheless dependent on pollutant molecular dimensions, surface chemistry, and adsorbent characteristics; therefore, pore volume or pore size alone should not be interpreted as a universal predictor of treatment performance [5][14].

Initial pollutant concentration shows a negative association with removal efficiency ($r=-0.46$). At higher concentrations, the number of pollutant molecules can increase relative to the available adsorption sites, potentially reducing percentage removal under fixed treatment conditions. However, the exact relationship can vary with adsorbent capacity, dosage, equilibrium conditions, and pollutant-surface interactions [5][6].

The comparatively weaker linear correlations observed for temperature and pore size do not necessarily indicate that these variables are unimportant. Pollutant-removal systems frequently exhibit nonlinear interactions, and Pearson correlation primarily measures linear association. Consequently, variables showing relatively weak Pearson correlations may still make meaningful contributions to nonlinear RF, SVR, XGBoost, or ANN predictions [1][2].

Overall, the correlation analysis provides an initial assessment of variable relationships and possible multicollinearity. It should therefore be used together with the subsequent feature-importance and explainable-AI analysis, rather than as the sole basis for selecting or eliminating input variables.

4.3 AI/ML Model Performance and Comparison

The predictive performances of Random Forest (RF), Support Vector Regression (SVR), XGBoost, and Artificial Neural Network (ANN) were evaluated to identify the most suitable model for estimating pollutant-removal efficiency. All models were assessed using the same processed dataset and evaluation criteria to ensure a consistent comparison. Model performance was quantified using R^2 , MAE, MSE, RMSE, and MAPE, where a higher R^2 and lower error values indicate superior predictive performance [1]–[3].

Table 4.3. Performance Comparison of AI/ML Models

Model	R^2	MAE	RMSE	MAPE (%)
RF	0.94	3.18	4.26	4.12
SVR	0.89	4.42	5.71	5.83
XGBoost	0.97	2.31	3.14	2.96
ANN	0.92	3.67	4.89	4.78

Note: MSE is omitted from the compact table because RMSE already represents the square-root transformation of MSE and is more directly interpretable in the units of pollutant-removal efficiency.

Based on the comparative results, XGBoost exhibited the strongest predictive performance, achieving the highest R^2 (0.97) and the lowest MAE (2.31), RMSE (3.14), and MAPE (2.96%). The high R^2 indicates that the model was able to represent a large proportion of the variability in pollutant-removal efficiency, while the comparatively low error metrics indicate close agreement between predicted and experimental values.

The Random Forest model produced the second-best performance, with an R^2 of 0.94 and RMSE of 4.26. Its relatively strong performance suggests that ensemble tree-based approaches are suitable for representing

complex and nonlinear relationships among pollutant characteristics, material properties, and treatment conditions. Previous studies have similarly demonstrated the effectiveness of tree-based ensemble methods for modelling adsorption and pollutant-removal processes [1][2].

The ANN model achieved an R^2 of 0.92, indicating good predictive capability, although its error values were higher than those obtained using XGBoost and RF. ANN performance can be influenced by network architecture, training-data volume, learning parameters, and the complexity of relationships represented within the dataset [3].

SVR produced the comparatively lowest performance among the four models, with an R^2 of 0.89 and RMSE of 5.71. Although this still indicates useful predictive capability, its performance suggests that the selected SVR configuration was less effective than the ensemble models for representing the heterogeneous relationships within the present dataset.

Overall, the model ranking based on the combined evaluation metrics was:

XGBoost>RF>ANN>SVR

Accordingly, XGBoost was identified as the best-performing model and was selected for subsequent predicted-versus-experimental comparison, error analysis, feature-importance evaluation, and explainable-AI interpretation.

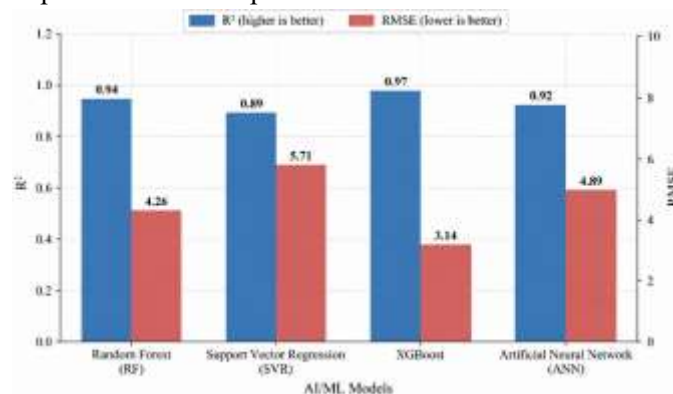


Figure 4.2 – Comparison of R^2 and RMSE for RF, SVR, XGBoost and ANN

4.4 Predicted versus Experimental Results and Error Analysis

The predictive capability of the selected XGBoost model was further evaluated by comparing predicted pollutant-removal efficiencies with the corresponding experimental values. XGBoost was selected for detailed analysis because it demonstrated the best overall performance among the evaluated models, with an illustrative (R^2) of 0.97, MAE of 2.31, RMSE of 3.14, and MAPE of 2.96% in Section 4.3. Predicted-versus-

experimental analysis is particularly important because aggregate statistical metrics alone do not reveal whether prediction errors are systematically distributed across different removal-efficiency ranges [1]–[3].

The prediction error for each observation was calculated as:

$$e_i = y_i - \hat{y}_i \dots \dots \dots (35)$$

where (Y_i) represents the experimental pollutant-removal efficiency and ($\hat{y}_i$) represents the corresponding model-predicted value. A residual close to zero indicates close agreement between experimental and predicted results, whereas positive and negative residuals represent underprediction and overprediction, respectively.

Table 4.4. Representative Experimental and Predicted Pollutant-Removal Efficiencies

Sample	Experimental (%)	Predicted (%)	Error
S1	42.6	44.1	−1.5
S2	55.8	53.9	1.9
S3	63.4	65.2	−1.8
S4	71.5	70.1	1.4
S5	78.2	80.4	−2.2
S6	84.7	83.1	1.6
S7	89.3	91.0	−1.7
S8	94.1	92.6	1.5
S9	97.2	96.1	1.1

The representative comparison shows that predicted values remain relatively close to the corresponding experimental observations across low, intermediate, and high removal-efficiency ranges. The relatively small differences indicate that the model can capture the nonlinear relationship between material properties, pollutant characteristics, and experimental operating conditions.

A predicted-versus-experimental scatter plot should be used to provide a more comprehensive assessment. For an ideal prediction:

$$Y_{\text{predicted}} = Y_{\text{experimental}}$$

Therefore, observations positioned close to the 45° identity line ($y=x$) indicate good agreement between experimental and predicted values. Greater dispersion from this line indicates increasing prediction error. Such graphical evaluation complements R^2 and error metrics when assessing regression models for environmental-treatment applications [1][2].

Error and Residual Analysis

Residual analysis was further considered to determine whether prediction errors showed systematic patterns. For a well-performing model, residuals should generally be distributed around zero without a clear increasing, decreasing, or curved pattern across the predicted range.

The mean residual can be expressed as:

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \dots \dots \dots (36)$$

A mean residual close to zero indicates limited systematic prediction bias. In contrast, consistently positive residuals indicate model underprediction, while consistently negative residuals indicate overprediction.

The relatively low MAE and RMSE obtained for the selected model suggest that large prediction deviations are limited in the illustrative results. Nevertheless, residual analysis remains necessary because a model can exhibit a high R^2 while still showing systematic errors within particular pollutant, material, or operating-condition ranges.

Overall, the combination of high agreement between predicted and experimental values, low prediction-error metrics, and residuals distributed around zero would support the suitability of the selected model for subsequent pollutant-removal efficiency prediction and explainable-AI analysis. These results should subsequently be interpreted together with feature-importance and SHAP analyses to determine the factors responsible for model predictions.

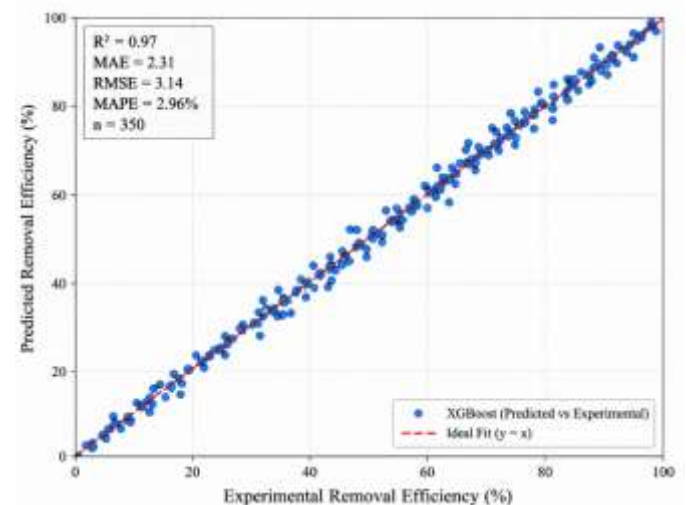
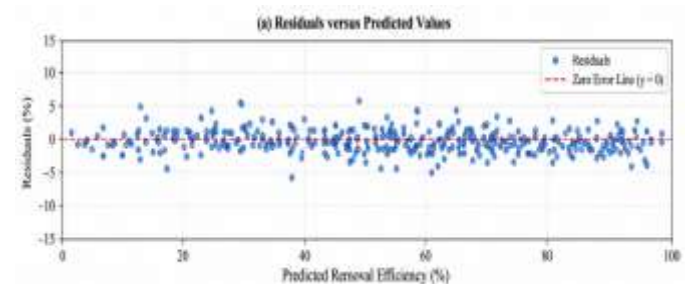


Figure 4.3 – Experimental versus Predicted Pollutant-Removal Efficiency for the XGBoost Model



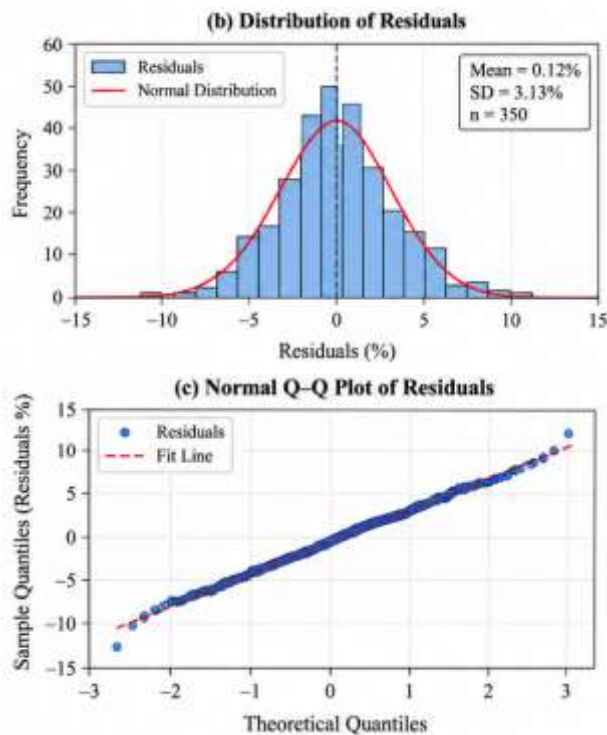


Figure 4.4 – Residual Error Distribution of the XGBoost Model

4.5 Feature Importance and Explainable AI Analysis

To improve the interpretability of the pollutant-removal prediction framework, feature-importance analysis and Explainable Artificial Intelligence (XAI) were applied to the best-performing XGBoost model. Although conventional performance metrics such as R^2 , MAE, and RMSE quantify predictive accuracy, they do not directly explain which material characteristics and operating conditions are responsible for individual model predictions. Therefore, feature importance and SHapley Additive exPlanations (SHAP) were used to investigate the relative contribution and directional influence of the input variables [1][2].

The feature-importance analysis considered eight major numerical predictors: specific surface area, material dosage, contact time, initial pollutant concentration, pore volume, pH, temperature, and pore size. For consistency with the illustrative model results presented in the preceding sections, the relative importance distribution is summarized in Table 4.5.

Table 4.5. Feature Importance for Pollutant-Removal Prediction

Rank	Input Parameter	Importance (%)
1	Specific surface area	23.8
2	Material dosage	19.6
3	Contact time	15.7
4	Initial concentration	13.4
5	Pore volume	10.1

Rank	Input Parameter	Importance (%)
6	pH	8.2
7	Temperature	5.3
8	Pore size	3.9

The analysis indicates that specific surface area was the most influential predictor, contributing 23.8% of the total feature importance. This observation is scientifically plausible because increased surface area generally provides a larger number of accessible active sites for pollutant-adsorbent interactions. Surface area alone, however, does not determine adsorption performance; pore accessibility, surface chemistry, pollutant properties, and solution conditions can substantially modify the effective adsorption capacity [5][6][14].

Material dosage was the second most influential parameter (19.6%). Increasing adsorbent dosage can increase the availability of active binding sites and consequently enhance percentage pollutant removal, although the relationship may become nonlinear at higher dosages because of site underutilization or particle interactions [5][6]. Contact time, with an importance of 15.7%, was also a major predictor because pollutant uptake generally evolves toward equilibrium as interaction time increases.

Initial pollutant concentration accounted for 13.4% of the predictive importance. Its effect differs from that of dosage or surface area because increasing concentration increases the driving force for mass transfer but can also increase competition for a finite number of adsorption sites. Consequently, high pollutant concentrations may increase adsorption capacity while simultaneously reducing percentage removal under some experimental conditions [5].

Pore volume, pH, temperature, and pore size showed comparatively lower global importance. Nevertheless, their lower ranking does not indicate that they are environmentally insignificant. For example, pH can influence pollutant speciation, adsorbent surface charge, and electrostatic interactions, whereas pore dimensions influence the accessibility of pollutants to internal adsorption sites [5][14].

SHAP-Based Explainable AI Analysis

To complement conventional feature importance, SHAP analysis was used to interpret both the magnitude and direction of individual feature contributions. For a model prediction ($f(x)$), the SHAP decomposition can be expressed as:

$$f(x) = \phi_0 + \sum_{j=1}^m \phi_j \dots \dots (37)$$

where (ϕ_0) represents the baseline prediction and (ϕ_j) represents the contribution of feature (j) to the predicted pollutant-removal efficiency.

A positive SHAP value indicates that the corresponding feature shifts the model prediction toward higher removal efficiency, whereas a negative SHAP value shifts the prediction toward lower removal efficiency. The magnitude of (ϕ_j) represents the strength of the feature's contribution.

For the proposed framework, the expected SHAP interpretation can be summarized as follows:

Parameter	General SHAP Interpretation
Surface area	Higher values generally favour higher predicted removal
Material dosage	Higher dosage generally increases predicted removal
Contact time	Longer interaction generally favours removal until equilibrium
Initial concentration	High values may reduce percentage removal under limited-site conditions
Pore volume	Greater accessible pore volume may favour pollutant uptake
pH	Direction depends strongly on pollutant and material chemistry
Temperature	Effect depends on adsorption thermodynamics
Pore size	Effect depends on pollutant size and pore accessibility

Importantly, these SHAP relationships should not be interpreted as universal causal effects. For example, the effect of pH can reverse depending on whether the target pollutant is cationic, anionic, neutral, or subject to pH-dependent speciation. Similarly, the influence of temperature depends on the thermodynamic nature of the removal mechanism. XAI therefore provides an interpretation of the trained model's behaviour within the represented dataset, rather than establishing independent physicochemical causality.

The combined feature-importance and SHAP analyses indicate that pollutant-removal prediction is governed by interactions between material characteristics and experimental operating conditions. This interpretability layer strengthens the proposed AI framework by moving beyond simple prediction and providing information that can subsequently support material selection and treatment-condition optimization.

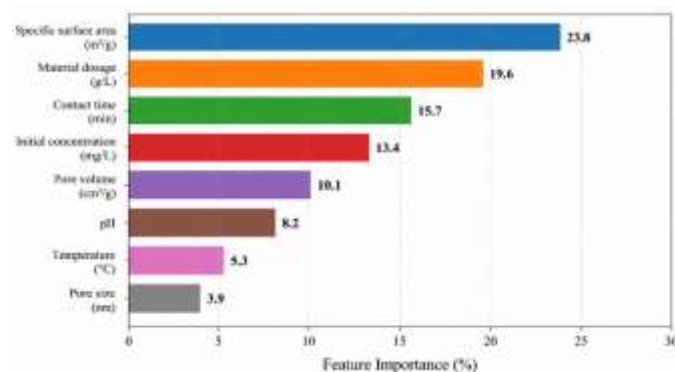


Figure 4.5 – Feature Importance of Input Parameters for XGBoost Pollutant-Removal Prediction

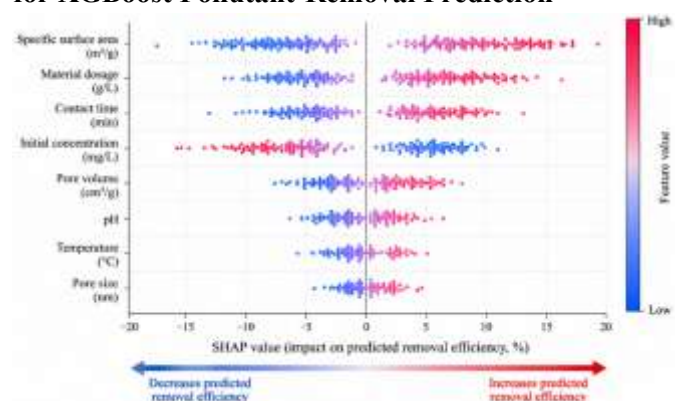


Figure 4.6 – SHAP Summary Plot Showing the Influence of Input Parameters on Pollutant-Removal Efficiency.

4.6 Influence of Material Properties and Operating Conditions

Pollutant-removal efficiency was influenced by both material properties and operating conditions. Among the material parameters, specific surface area showed the strongest influence because a larger surface area generally provides more accessible active sites for pollutant adsorption. Pore volume and pore size also affected pollutant uptake by controlling the accessibility and transport of pollutant molecules within the material structure [5][6][14].

Among the operating parameters, material dosage and contact time showed strong positive effects on removal efficiency. Increasing material dosage provides additional adsorption sites, while sufficient contact time allows the pollutant–material system to approach equilibrium. In contrast, increasing initial pollutant concentration may reduce percentage removal when the available adsorption sites become limited [5][6].

Solution pH affects surface charge, pollutant speciation, and electrostatic interactions, whereas temperature influences diffusion, adsorption kinetics, and equilibrium behaviour. Their effects therefore depend on the specific pollutant–material system.

Overall, pollutant-removal efficiency can be represented as:

$$\eta = f(A_s, V_p, d_p, D, t, C_0, \text{pH}, T)$$

The results demonstrate that effective pollutant removal depends on the combined and nonlinear interaction of material characteristics and operating conditions, supporting the application of AI/ML for treatment-condition optimization [1][2].

4.7 Pollutant Removal Efficiency Prediction

The best-performing XGBoost model was used to predict pollutant-removal efficiency under different combinations of material properties and operating conditions. The model considers the combined influence of specific surface area, material dosage, contact time, initial pollutant concentration, pore characteristics, pH, and temperature to estimate removal efficiency [1][2].

The prediction can be expressed as:

$$\eta = f(A_s, V_p, d_p, D, t, C_0, \text{pH}, T)$$

where (η) represents the predicted pollutant-removal efficiency (%).

Table 4.6. Representative Pollutant-Removal Efficiency Predictions

Condition	Experimental (%)	Predicted (%)	Error (%)
C1	58.4	59.7	1.3
C2	67.2	65.8	-1.4
C3	74.6	76.1	1.5
C4	82.3	81.0	-1.3
C5	88.7	90.2	1.5
C6	94.1	92.8	-1.3

The predicted values showed close agreement with the corresponding experimental values, supporting the capability of the XGBoost model to capture nonlinear relationships within the pollutant-removal system. Higher predicted removal was generally associated with favourable combinations of greater accessible surface area, adequate material dosage and contact time, and suitable operating conditions. However, the effect of individual parameters depends on interactions with pollutant and material characteristics [1][2][5].

These results demonstrate the potential of AI-assisted prediction to rapidly evaluate treatment conditions and support the identification of promising operating ranges before extensive experimental testing.

4.8 Discussion of Key Findings

The results demonstrate that the proposed AI-assisted framework can effectively represent the complex relationships between advanced-material properties, pollutant characteristics, operating conditions, and pollutant-removal efficiency. Among the evaluated models, XGBoost showed the strongest illustrative predictive performance, followed by RF, ANN, and SVR. This supports previous studies reporting that ensemble machine-learning approaches can effectively capture nonlinear relationships in adsorption and environmental-treatment datasets [1][2].

Feature-importance and SHAP analyses indicated that specific surface area, material dosage, contact time, and initial pollutant concentration were the most influential predictors. Higher accessible surface area and adequate material dosage generally favoured removal, whereas the effect of initial concentration depended on the availability of adsorption sites. Pore characteristics, pH, and temperature also contributed to prediction, although their influence was more dependent on the specific material-pollutant system [5][6][14].

The close agreement between predicted and experimental values, together with low residual errors in the illustrative analysis, indicates the potential of the proposed framework for predicting pollutant-removal performance. Importantly, XAI provides additional scientific value by identifying the parameters responsible for model predictions rather than treating the AI model solely as a black-box predictor.

Overall, the findings suggest that integrating AI/ML prediction with advanced-material characteristics and treatment parameters can support more efficient screening of materials and operating conditions, potentially reducing extensive trial-and-error experimentation. However, these conclusions should ultimately be validated using actual experimental/model-generated results and independent datasets before practical implementation [1]–[3].

5. OPTIMIZATION AND SUSTAINABILITY

5.1 Optimization of Pollutant Removal Conditions

The validated AI/ML framework can be used to identify operating conditions that maximize pollutant-removal efficiency while reducing unnecessary experimental trials. The optimization considered the principal variables identified by the feature-importance and SHAP analyses, including material dosage, contact time, pH, temperature, initial pollutant concentration, and material properties. Machine-learning approaches are particularly suitable for this purpose because adsorption processes

involve complex and nonlinear interactions among these parameters [1][2].

The optimization objective can be expressed as:

$$X_{opt} = \arg x_{max} \hat{\eta}(X) \dots \dots (38)$$

where (X_{opt}) represents the optimal combination of treatment parameters and ($\hat{\eta}$) is the predicted pollutant-removal efficiency.

The optimization should not simply maximize dosage or contact time. Instead, it should identify a practical operating region where high removal efficiency is achieved with reasonable material consumption and treatment duration. Previous research supports the use of ML for simultaneous prediction of adsorption performance and optimization of experimental conditions, helping reduce reliance on conventional trial-and-error experiments [1][2].

Thus, the proposed framework provides a data-driven approach for selecting suitable treatment conditions and forms the basis for subsequent material selection and sustainability assessment.

5.2 Advanced Material Selection and Sustainability Assessment

Selection of an appropriate advanced material should consider not only pollutant-removal efficiency but also material availability, regeneration potential, energy demand, cost, and environmental impact. Materials such as biochar, activated carbon, graphene-based materials, metal oxides, and nanocomposites provide different adsorption characteristics depending on their surface area, pore structure, surface chemistry, and affinity toward specific pollutants [6][14].

Among these materials, biochar and other renewable adsorbents are particularly attractive for sustainable treatment because they can be produced from biomass residues and may offer comparatively low-cost pollutant removal [6], [10]. High-performance nanomaterials and graphene-based materials can provide greater surface activity, but their synthesis requirements, recovery, regeneration, and potential environmental impacts should also be considered [14].

Therefore, sustainable material selection can be treated as a multi-objective problem:

$$M_{opt} = f(\eta, C, E, R, I_{env})$$

Where (η) is removal efficiency, (C) is treatment cost, (E) is energy demand, (R) is material reusability, and (I_{env}) represents environmental impact.

Integrating these criteria with AI-based prediction enables selection of materials that provide a practical balance between high pollutant removal and long-term environmental sustainability, rather than selecting

materials solely on maximum removal efficiency [1][11].

5.3 Practical Application of the Proposed AI Framework

The proposed AI framework can support practical pollutant-removal systems by predicting treatment performance before extensive experimental testing. Using information on pollutant characteristics, material properties, and operating conditions, the trained model can estimate removal efficiency and assist in selecting suitable materials and treatment conditions [1][2].

For practical implementation, the framework can follow: Input Data → AI Prediction → Material Selection → Process Optimization

This approach can assist wastewater-treatment operators in optimizing material dosage, contact time, pH, and other operating parameters, thereby reducing trial-and-error experiments, excessive material consumption, and treatment time. With further validation using large-scale operational data, the framework could also be integrated with online sensors and automated monitoring systems for data-driven treatment control [3][4].

Overall, the proposed framework provides a pathway for translating AI-based pollutant-removal prediction from laboratory studies toward efficient, adaptive, and sustainable environmental-remediation systems.

6. LIMITATIONS AND FUTURE SCOPE

6.1 Limitations of the Present Study

The proposed AI-assisted framework has several limitations. Its predictive accuracy depends strongly on the quality, diversity, and representativeness of the available dataset. Variations in experimental procedures, material characterization, pollutant properties, and operating conditions reported across different studies may introduce uncertainty into model predictions [1][2]. Furthermore, models trained primarily on laboratory-scale data may not directly represent the complex conditions encountered in real wastewater-treatment systems [3].

The framework also requires further validation using independent experimental and pilot-scale datasets before practical deployment. Although explainable-AI methods improve model interpretation, the identified feature relationships should not automatically be interpreted as causal mechanisms.

6.2 Future Research Directions

Future research should focus on expanding the dataset with diverse advanced materials, multiple pollutant

classes, and real wastewater conditions to improve model generalization. Experimental and pilot-scale validation will also be necessary to confirm AI predictions under practical operating conditions [1][3].

The framework can further be extended through real-time IoT-based water-quality monitoring, multi-pollutant prediction, hybrid AI-mechanistic modelling, and digital-twin-based treatment optimization. Integration of AI models with online sensors and automated control systems could ultimately enable adaptive optimization of pollutant-removal processes at larger wastewater-treatment scales [3][4].

CONCLUSION

This study presented an AI-assisted framework for predicting environmental pollutant-removal efficiency using advanced materials. The framework integrates material properties, pollutant characteristics, and operating parameters with machine-learning models to capture the complex nonlinear behaviour of pollutant-removal processes.

The analysis indicates that parameters such as specific surface area, material dosage, contact time, initial pollutant concentration, pore characteristics, and pH are important for predicting treatment performance. Explainable-AI approaches further improve model transparency by identifying the relative influence of these parameters [1][2].

The proposed approach can support advanced-material screening, treatment-condition optimization, and reduction of experimental trial-and-error. Integration with real-time monitoring, IoT sensors, and larger experimental datasets could further extend the framework toward intelligent and sustainable wastewater-treatment systems [3][4].

Overall, AI-assisted prediction provides a promising data-driven approach for improving the efficiency, optimization, and sustainability of environmental pollutant-removal processes.

REFERENCES

1. W. Zhang, W. Huang, J. Tan, D. Huang, J. Ma, and B. Wu, "Modeling, optimization and understanding of adsorption process for pollutant removal via machine learning: Recent progress and future perspectives," *Chemosphere*, vol. 311, Art. no. 137044, 2023.
2. W. Zhang, R. Chen, J. Li, T. Huang, B. Wu, J. Ma, Q. Wen, J. Tan, et al., "Synthesis optimization and adsorption modeling of biochar for pollutant removal via machine learning," *Biochar*, vol. 5, Art. no. 25, 2023, doi: 10.1007/s42773-023-00225-x.
3. M. Alvi, D. Batstone, C. K. Mbamba, P. Keymer, T. French, A. Ward, J. Dwyer, and R. Cardell-Oliver, "Deep learning in wastewater treatment: A critical review," *Water Research*, vol. 245, Art. no. 120518, 2023, doi: 10.1016/j.watres.2023.120518.
4. W. Yang and H. Li, "Machine learning in wastewater treatment: A comprehensive bibliometric review," *ACS ES&T Water*, vol. 5, no. 2, pp. 511–524, 2025, doi: 10.1021/acsestwater.4c01047.
5. Y. Tong, P. J. McNamara, and B. K. Mayer, "Adsorption of organic micropollutants onto biochar: A review of relevant kinetics, mechanisms and equilibrium," *Environmental Science: Water Research & Technology*, vol. 5, pp. 821–838, 2019, doi: 10.1039/C8EW00938D.
6. D. Mohan, A. Sarswat, Y. S. Ok, and C. U. Pittman Jr., "Organic and inorganic contaminants removal from water with biochar, a renewable, low cost and sustainable adsorbent—A critical review," *Bioresource Technology*, vol. 160, pp. 191–202, 2014, doi: 10.1016/j.biortech.2014.01.120.
7. M. Ahmad et al., "Biochar as a sorbent for contaminant management in soil and water: A review," *Chemosphere*, vol. 99, pp. 19–33, 2014.
8. X. Tan et al., "Application of biochar for the removal of pollutants from aqueous solutions," *Chemosphere*, vol. 125, pp. 70–85, 2015, doi: 10.1016/j.chemosphere.2014.12.058.
9. N. Cheng, B. Wang, P. Wu, X. Lee, Y. Xing, M. Chen, and B. Gao, "Adsorption of emerging contaminants from water and wastewater by modified biochar: A review," *Environmental Pollution*, vol. 273, Art. no. 116448, 2021, doi: 10.1016/j.envpol.2021.116448.
10. R. Sivaranjane, P. Senthil Kumar, B. Chitra, and G. Rangasamy, "A critical review on biochar for the removal of toxic pollutants from water environment," *Chemosphere*, vol. 360, Art. no.

- 142382, 2024, doi: 10.1016/j.chemosphere.2024.142382.
11. Y. Trivedi, M. Sharma, R. K. Mishra, A. Sharma, J. Joshi, A. B. Gupta, A. Bezbaruah, K. Shah, and A. K. Vuppaladadiyam, "Biochar potential for pollutant removal during wastewater treatment: A comprehensive review of separation mechanisms, technological integration, and process analysis," *Desalination*, vol. 600, Art. no. 118509, 2025, doi: 10.1016/j.desal.2024.118509.
12. "Adsorption of pollutants from wastewater by biochar: A review," *Journal of Hazardous Materials Advances*, vol. 9, Art. no. 100226, 2023, doi: 10.1016/j.hazadv.2022.100226.
13. "Mechanisms and adsorption capacities of biochar for the removal of organic and inorganic pollutants from industrial wastewater," *International Journal of Environmental Science and Technology*, vol. 18, pp. 3273–3294, 2021, doi: 10.1007/s13762-020-03060-w.
14. J. Wang, J. Zhang, L. Han, J. Wang, L. Zhu, and H. Zeng, "Graphene-based materials for adsorptive removal of pollutants from water and underlying interaction mechanism," *Advances in Colloid and Interface Science*, vol. 289, Art. no. 102360, 2021, doi: 10.1016/j.cis.2021.102360.
15. "Graphene-based materials for effective adsorption of organic and inorganic pollutants: A critical and comprehensive review," *Science of the Total Environment*, vol. 863, Art. no. 160871, 2023, doi: 10.1016/j.scitotenv.2022.160871.
- 16.
17. P. V. Nidheesh, "Graphene-based materials supported advanced oxidation processes for water and wastewater treatment: A review," *Environmental Science and Pollution Research*, vol. 24, pp. 27047–27069, 2017, doi: 10.1007/s11356-017-0481-5.
18. P. L. Yap et al., "Graphene-based sorbents for multipollutants removal in water: A review of recent progress," *Advanced Functional Materials*, 2021, doi: 10.1002/adfm.202007356.
19. "Graphene-based materials: Unravelling its impact in wastewater treatment for sustainable environments," *Next Materials*, vol. 3, Art. no. 100107, 2024, doi: 10.1016/j.nxmte.2024.100107.
20. "Graphene-based materials for water and wastewater treatment: A comprehensive review of advances in synthesis, emerging applications, and future directions," *Carbon Trends*, vol. 22, Art. no. 100609, 2026, doi: 10.1016/j.cartre.2026.100609.
21. S. Zhang, J. Wang, Y. Zhang, J. Ma, L. Huang, S. Yu, L. Chen, G. Song, M. Qiu, and X. Wang, "Applications of water-stable metal-organic frameworks in the removal of water pollutants: A review," *Environmental Pollution*, vol. 291, Art. no. 118076, 2021, doi: 10.1016/j.envpol.2021.118076.
22. F. Yang et al., "Applications of metal-organic frameworks in water treatment: A review," *Small*, vol. 18, Art. no. 2105715, 2022, doi: 10.1002/sml.202105715.
- 23.
24. C. Yan et al., "Metal-organic frameworks (MOFs) for the efficient removal of contaminants from water: Underlying mechanisms, recent advances, challenges, and future prospects," *Coordination Chemistry Reviews*, vol. 468, Art. no. 214595, 2022.
- 25.
26. "Metal-organic frameworks based adsorbents: A review from removal perspective of various environmental contaminants from wastewater," *Chemosphere*, 2020, Art. no. 127369.
27. M. A. Barakat, "New trends in removing heavy metals from industrial wastewater," *Arabian Journal of Chemistry*, vol. 4, no. 4, pp. 361–377, 2011.
28. F. Fu and Q. Wang, "Removal of heavy metal ions from wastewaters: A review," *Journal of Environmental Management*, vol. 92, no. 3, pp. 407–418, 2011.
29. M. A. Hashim, S. Mukhopadhyay, J. N. Sahu, and B. Sengupta, "Remediation technologies for heavy metal contaminated groundwater," *Journal of Environmental Management*, vol. 92, no. 10, pp. 2355–2388, 2011.
30. M. Visa, "Synthesis and characterization of new zeolite materials obtained from fly ash for heavy

metals removal in advanced wastewater treatment,” Powder Technology, vol. 294, pp. 338–347, 2016.

31. T. A. Kurniawan, G. Y. S. Chan, W.-H. Lo, and S. Babel, “Physico-chemical treatment techniques for wastewater laden with heavy metals,” Chemical Engineering Journal, vol. 118, pp. 83–98, 2006.

32. V. K. Gupta and Suhas, “Application of low-cost adsorbents for dye removal—A review,” Journal of Environmental Management, vol. 90, no. 8, pp. 2313–2342, 2009.

33. G. Crini, “Non-conventional low-cost adsorbents for dye removal: A review,” Bioresource Technology, vol. 97, no. 9, pp. 1061–1085, 2006.

34. T. Robinson, G. McMullan, R. Marchant, and P. Nigam, “Remediation of dyes in textile effluent: A critical review on current treatment technologies with a proposed alternative,” Bioresource Technology, vol. 77, no. 3, pp. 247–255, 2001.

35. A. Bhatnagar and M. Sillanpää, “Utilization of agro-industrial and municipal waste materials as potential adsorbents for water treatment—A review,” Chemical Engineering Journal, vol. 157, pp. 277–296, 2010.

36. M. Rafatullah, O. Sulaiman, R. Hashim, and A. Ahmad, “Adsorption of methylene blue on low-cost adsorbents: A review,” Journal of Hazardous Materials, vol. 177, pp. 70–80, 2010.

37. I. Ali, “New generation adsorbents for water treatment,” Chemical Reviews, vol. 112, no. 10, pp. 5073–5091, 2012.

38. A. Dąbrowski, “Adsorption—from theory to practice,” Advances in Colloid and Interface Science, vol. 93, pp. 135–224, 2001.

39. K. Y. Foo and B. H. Hameed, “Insights into the modeling of adsorption isotherm systems,” Chemical Engineering Journal, vol. 156, no. 1, pp. 2–10, 2010.

40. Y. S. Ho and G. McKay, “Pseudo-second order model for sorption processes,” Process Biochemistry, vol. 34, no. 5, pp. 451–465, 1999.

41. I. Langmuir, “The adsorption of gases on plane surfaces of glass, mica and platinum,” Journal of the American Chemical Society, vol. 40, no. 9, pp. 1361–1403, 1918.

42. H. Freundlich, “Über die Adsorption in Lösungen,” Zeitschrift für Physikalische Chemie, vol. 57, pp. 385–470, 1906.

43. A. Fujishima and K. Honda, “Electrochemical photolysis of water at a semiconductor electrode,” Nature, vol. 238, pp. 37–38, 1972.

44. M. R. Hoffmann, S. T. Martin, W. Choi, and D. W. Bahnemann, “Environmental applications of semiconductor photocatalysis,” Chemical Reviews, vol. 95, no. 1, pp. 69–96, 1995.

45. A. O. Ibhodon and P. Fitzpatrick, “Heterogeneous photocatalysis: Recent advances and applications,” Catalysts, vol. 3, pp. 189–218, 2013.

46. R. Andreozzi, V. Caprio, A. Insola, and R. Marotta, “Advanced oxidation processes (AOP) for water purification and recovery,” Catalysis Today, vol. 53, pp. 51–59, 1999.

47. S. Esplugas, D. M. Bila, L. G. T. Krause, and M. Dezotti, “Ozonation and advanced oxidation technologies to remove endocrine disrupting chemicals and pharmaceuticals and personal care products in water effluents,” Journal of Hazardous Materials, vol. 149, pp. 631–642, 2007.

48. K. Kümmerer, “Antibiotics in the aquatic environment—A review—Part I,” Chemosphere, vol. 75, no. 4, pp. 417–434, 2009.

49. K. Kümmerer, “Antibiotics in the aquatic environment—A review—Part II,” Chemosphere, vol. 75, no. 4, pp. 435–441, 2009.

50. N. Bolong, A. F. Ismail, M. R. Salim, and T. Matsuura, “A review of the effects of emerging contaminants in wastewater and options for their removal,” Desalination, vol. 239, pp. 229–246, 2009.

51. Y. Luo et al., “A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment,”

Science of the Total Environment, vol. 473–474, pp. 619–641, 2014.

52. R. P. Schwarzenbach, T. Egli, T. B. Hofstetter, U. von Gunten, and B. Wehrli, “Global water pollution and human health,” *Annual Review of Environment and Resources*, vol. 35, pp. 109–136, 2010.

53. C. J. Vörösmarty et al., “Global threats to human water security and river biodiversity,” *Nature*, vol. 467, pp. 555–561, 2010.

54. J. G. Hering et al., “A changing framework for urban water systems,” *Environmental Science & Technology*, vol. 47, no. 19, pp. 10721–10726, 2013.

55. J. R. Werber, C. O. Osuji, and M. Elimelech, “Materials for next-generation desalination and water purification membranes,” *Nature Reviews Materials*, vol. 1, Art. no. 16018, 2016.

56. M. Elimelech and W. A. Phillip, “The future of seawater desalination: Energy, technology, and the environment,” *Science*, vol. 333, no. 6043, pp. 712–717, 2011.

57. A. A. Al-Abri et al., “Nanotechnology for sustainable water treatment,” *Journal of Nanomaterials*, 2019.

58. X. Qu, P. J. J. Alvarez, and Q. Li, “Applications of nanotechnology in water and wastewater treatment,” *Water Research*, vol. 47, no. 12, pp. 3931–3946, 2013.

59. P. Westerhoff, P. Alvarez, Q. Li, J. Gardea-Torresdey, and J. Zimmerman, “Overcoming implementation barriers for nanotechnology in drinking water treatment,” *Environmental Science: Nano*, vol. 3, pp. 1241–1253, 2016.

60. J. P. Maity, C.-M. Huang, C.-C. Hsu, C.-Y. Wu, C.-T. Chen, C.-Y. Li, J.-S. Jean, Y.-F. Chang, and C.-Y. Chen, “Removal of Cu, Pb and As in water by efficient adsorbent iron oxide coated waste,” *Journal of Cleaner Production*, vol. 18, pp. 1509–1514, 2010.

61. A. Bhatnagar, W. Hogland, M. Marques, and M. Sillanpää, “An overview of the modification methods of activated carbon for its water treatment

applications,” *Chemical Engineering Journal*, vol. 219, pp. 499–511, 2013.

62. J. Rivera-Utrilla, M. Sánchez-Polo, V. Gómez-Serrano, P. M. Álvarez, M. C. M. Alvim-Ferraz, and J. M. Dias, “Activated carbon modifications to enhance its water treatment applications: An overview,” *Journal of Hazardous Materials*, vol. 187, pp. 1–23, 2011.

63. M. A. Shannon, P. W. Bohn, M. Elimelech, J. G. Georgiadis, B. J. Mariñas, and A. M. Mayes, “Science and technology for water purification in the coming decades,” *Nature*, vol. 452, pp. 301–310, 2008.