

MACHINE LEARNING-GUIDED CHEMICAL ENGINEERING SOLUTIONS FOR PFAS REMOVAL FROM DRINKING WATER: PREDICTIVE TOXICITY ASSESSMENT AND MITIGATION OF CANCER AND LIVER DISEASE RISKS

Bushra Mustafa¹, Dr. Imran Khan Swati², Ayesha Abbas³, Aymen Ghaffar⁴, Amna Munawar⁵

¹Department of Nuclear Chemical Engineering, Harbin Engineering University (HEU), China.

²Department of Chemical Engineering, University of Engineering and Technology, Peshawar, Pakistan.

³Department of Chemical Engineering, Wah Engineering College, University of Wah, Quaid Avenue, Wah Cantt, Rawalpindi 47040, Pakistan

⁴Department of Chemistry, The Islamia University of Bahawalpur, Bahawalpur, Pakistan.

⁵Research Assistant, Lahore College for Women University (LCWU), Lahore, Pakistan.

¹bushramustafa268@gmail.com, ²imranswati@uetpeshawar.edu.pk, ³ayesha.abass@wecuw.edu.pk,

⁴aymenabdulghaffar123@gmail.com, ⁵amna4989@yahoo.com

DOI: <https://doi.org/10.5281/zenodo.22957577>

Keywords

PFAS removal; drinking water; chemical engineering; machine learning; toxicity assessment; cancer risk; liver disease; process optimization.

Article History

Received: 17 June 2026

Accepted: 08 September 2026

Published: 25 September 2026

Copyright @Author

Corresponding Author: *
Dr. Imran Khan Swati

Abstract

The widespread occurrence of per- and polyfluoroalkyl substances (PFAS) in drinking-water systems has become a serious environmental and public-health concern because of their exceptional persistence, bioaccumulation potential, and resistance to conventional treatment processes. This study develops a machine learning-guided chemical engineering framework for predicting PFAS removal performance, optimizing treatment conditions, evaluating compound-specific toxicity, and estimating potential reductions in cancer and liver-disease risks. A harmonized dataset comprising 18,600 laboratory, pilot-scale, and water-quality observations was compiled for 18 commonly detected PFAS compounds. The dataset contained 42 physicochemical and operational variables, including influent concentration, molecular chain length, functional group, pH, temperature, dissolved organic carbon, ionic strength, adsorbent dosage, membrane pressure, contact time, pore size, and treatment configuration. Six treatment approaches were investigated: granular activated carbon, ion-exchange resin, nanofiltration, reverse osmosis, electrochemical oxidation, and hybrid ion-exchange-nanofiltration. After missing-value imputation, outlier removal, normalization, categorical encoding, feature engineering, and leakage-free data partitioning, linear regression, support vector regression, random forest, artificial neural network, LightGBM, and XGBoost models were evaluated. XGBoost produced the best predictive performance, achieving an R^2 of 0.963, a root-mean-square error of 3.18%, and a mean absolute error of 2.41% for PFAS removal efficiency. The optimized hybrid treatment achieved 97.6% total PFAS removal, outperforming activated carbon at 82.4%, ion exchange at 91.8%,

nanofiltration at 94.3%, and reverse osmosis at 95.7%. Machine learning-based process optimization also reduced energy consumption by 18.9%, chemical usage by 21.6%, and estimated operating costs by 16.7% compared with non-optimized operating conditions. SHAP analysis identified PFAS chain length, influent concentration, dissolved organic carbon, adsorbent dosage, membrane pressure, and contact time as the most influential predictors. A toxicity-weighted health-risk module estimated reductions of 93.2% in cumulative non-carcinogenic hazard, 89.7% in modeled lifetime cancer risk, and 91.4% in liver-toxicity risk following optimized treatment. These health outcomes represent modeled risk estimates rather than observed clinical effects. Overall, the proposed framework integrates explainable machine learning, treatment-process optimization, and health-risk assessment to support cost-effective, transparent, and scalable PFAS management in drinking-water systems.

1- Introduction:

Per- and polyfluoroalkyl substances (PFAS) have emerged as a major class of persistent contaminants affecting drinking-water resources worldwide. These synthetic chemicals have been extensively used in industrial processes and consumer products because of their exceptional thermal stability and resistance to water, oil, grease, and chemical degradation. Common applications include firefighting foams, non-stick cookware, food-contact materials, stain-resistant textiles, metal-plating processes, semiconductor manufacturing, and surface coatings. However, the same carbon-fluorine bonds that provide PFAS with desirable industrial properties also make them highly resistant to natural degradation and conventional water-treatment processes. Consequently, PFAS can persist in aquatic environments for long periods, migrate through soil and groundwater, accumulate in living organisms, and enter human populations through contaminated drinking water and food [1]. PFAS contamination is particularly challenging because the group includes thousands of compounds with substantial differences in molecular structure, chain length, functional group, environmental mobility, and toxicological behavior. Long-chain compounds, such as perfluorooctanoic acid and perfluorooctane sulfonic acid, generally exhibit strong adsorption and bioaccumulation tendencies. In contrast, short-chain and emerging replacement PFAS are often more mobile in water

and can be difficult to retain through conventional adsorption processes. Treatment performance is further influenced by water-quality and operating conditions, including influent concentration, pH, temperature, dissolved organic carbon, ionic strength, adsorbent dosage, contact time, membrane pore size, transmembrane pressure, and the presence of competing contaminants. This variability limits the effectiveness of treatment strategies developed using fixed operating assumptions or observations from a restricted number of compounds [2]. Growing epidemiological and toxicological evidence has raised concerns about possible associations between prolonged PFAS exposure and adverse human-health outcomes. Reported concerns include altered liver function, immune-system disruption, developmental effects, endocrine disturbances, abnormal lipid metabolism, and increased risks of certain cancers. The liver is an important target organ because it participates in the uptake, distribution, and biological processing of persistent environmental contaminants. PFAS exposure has been associated with changes in liver enzymes, lipid regulation, oxidative stress, and cellular signaling. Nevertheless, the magnitude of health risk varies considerably among compounds and exposure conditions. Health-related outputs should therefore be interpreted as toxicity-weighted or exposure-based risk estimates rather than direct predictions of clinical disease [3]. A reliable

management framework must connect compound-specific occurrence, treatment performance, residual exposure, and toxicological potency while transparently communicating the uncertainty associated with modeled health outcomes. Conventional drinking-water treatment processes, including coagulation, sedimentation, biological treatment, and sand filtration, are not specifically designed to remove highly persistent and mobile PFAS. Granular activated carbon is widely applied because of its commercial availability and affinity for several long-chain compounds, but its performance may decline in the presence of dissolved organic matter or short-chain PFAS. Ion-exchange resins can provide greater adsorption capacity and selectivity, although resin exhaustion, regeneration, and disposal requirements remain important operational concerns. Nanofiltration and reverse osmosis achieve high rejection efficiencies across a broader range of PFAS, but they may require considerable pressure and energy and generate a concentrated residual stream requiring further treatment. Electrochemical oxidation offers opportunities for destructive treatment, yet its efficiency depends on electrode properties, current density, reaction time, water composition, and energy input. Hybrid systems can combine the strengths of individual technologies, but their design requires the simultaneous consideration of removal efficiency, energy demand, chemical consumption, residual-stream management, and operating cost [4]. Chemical-engineering models are essential for understanding adsorption, membrane separation, mass transfer, reaction kinetics, and process-scale performance. However, PFAS treatment systems often exhibit nonlinear interactions among molecular descriptors, water chemistry, equipment configuration, and operating conditions. Mechanistic models may become computationally demanding or require parameters that are difficult to estimate for numerous PFAS compounds and mixed-contaminant conditions. Empirical models can be easier to implement but may not adequately represent complex interactions or predict performance outside narrowly defined experimental ranges. These

limitations create a need for data-driven methods that complement chemical-engineering knowledge and enable rapid evaluation of a large number of treatment scenarios.

Machine learning provides a promising approach for identifying nonlinear relationships between PFAS characteristics, water-quality conditions, treatment variables, and removal outcomes. Algorithms such as support vector regression, random forest, artificial neural networks, LightGBM, and XGBoost can integrate heterogeneous variables and estimate treatment performance under complex operating conditions. Their application may reduce the number of resource-intensive experiments required during process development and support the selection of suitable technologies for specific water matrices. Machine-learning models can also be integrated with optimization routines to identify operating conditions that balance removal efficiency with energy use, chemical demand, and treatment cost. However, high predictive accuracy alone is insufficient for engineering and regulatory decision-making [5]. Data leakage, inadequate validation, poor representation of individual compounds, and limited model interpretability may produce optimistic or scientifically unreliable results. Explainable artificial intelligence can address part of this limitation by showing how individual predictors influence estimated treatment performance. In particular, Shapley additive explanation analysis can quantify the global and observation-specific contributions of variables such as PFAS chain length, influent concentration, dissolved organic carbon, adsorbent dosage, membrane pressure, and contact time. Such information allows engineers to determine whether model behavior agrees with established chemical and physical principles. It can also reveal compound-specific treatment limitations and clarify the trade-offs between operational intensity and removal improvement. Combining explainability with leakage-free data partitioning and comparative model evaluation therefore strengthens the reproducibility, transparency, and practical relevance of machine-learning-assisted treatment design. Existing studies

frequently examine a single PFAS compound, one treatment method, or a limited set of laboratory conditions. Many focus exclusively on removal efficiency without jointly considering toxicity, residual exposure, energy consumption, chemical use, and economic performance. Conversely, health-risk studies may estimate exposure without incorporating treatment-specific reductions in individual PFAS concentrations. This separation between predictive modeling, process engineering, and health-risk assessment limits the ability of conventional approaches to support integrated drinking-water management [6]. A comprehensive framework is needed to compare multiple technologies, predict compound-specific removal, optimize operating conditions, explain model decisions, and translate residual concentrations into toxicity-weighted risk indicators. Accordingly, this study develops a machine learning-guided chemical-engineering framework for PFAS removal from drinking water. A harmonized dataset containing 18,600 laboratory, pilot-scale, and water-quality observations for 18 commonly detected PFAS compounds is analyzed using 42 physicochemical and operational variables. Six treatment approaches granular activated carbon, ion-exchange resin, nanofiltration, reverse osmosis, electrochemical oxidation, and hybrid ion-exchange-nanofiltration are comparatively evaluated. Linear regression, support vector regression, random forest, artificial neural network, LightGBM, and XGBoost models are trained using a leakage-free analytical workflow. The models are used to predict PFAS removal efficiency, optimize operating conditions, and evaluate trade-offs involving energy consumption, chemical usage, and operating cost. SHAP analysis is subsequently applied to explain the principal drivers of model predictions [7]. The study further introduces a toxicity-weighted health-risk module that connects treatment-derived residual concentrations with estimates of cumulative non-carcinogenic hazard, lifetime cancer risk, and liver-toxicity risk. These indicators are intended to support comparative treatment assessment and should not be interpreted as observed clinical outcomes or individual medical predictions. The

principal contribution of the research is the integration of chemical-engineering process analysis, explainable machine learning, multi-objective optimization, and health-risk modeling within a single decision-support framework. This approach provides a transparent and scalable basis for selecting treatment configurations, improving operational efficiency, prioritizing high-risk PFAS compounds, and supporting evidence-based management of contaminated drinking-water systems.

2- Nanofiltration and Reverse-Osmosis

Membranes:

Nanofiltration (NF) and reverse osmosis (RO) are among the most effective pressure-driven membrane technologies for removing PFAS from contaminated drinking water. Unlike conventional filtration, these processes use dense or semi-dense selective layers capable of restricting the passage of dissolved organic contaminants. PFAS rejection occurs through a combination of size exclusion, electrostatic repulsion, membrane-solute interactions, and solution-diffusion resistance. Their capacity to remove both long- and short-chain compounds makes NF and RO particularly important when conventional adsorption systems cannot consistently achieve the required water-quality standards [8]. The principal transport and separation mechanisms involved in membrane-based PFAS treatment are presented in Figure 1. PFAS-contaminated feed water first reaches the membrane boundary layer, where molecular diffusion, concentration polarization, and interactions with dissolved constituents influence contaminant transport. Water molecules pass through the selective membrane layer and form the treated permeate, while most PFAS compounds are retained in the concentrate stream. Size exclusion limits the passage of compounds whose effective molecular dimensions are larger than the available transport pathways [9]. Electrostatic repulsion occurs because many commercial polyamide membranes are negatively charged under typical drinking-water conditions, while commonly detected PFAS compounds primarily exist as negatively charged species.

Solution-diffusion resistance further limits the movement of PFAS through dense RO membranes.

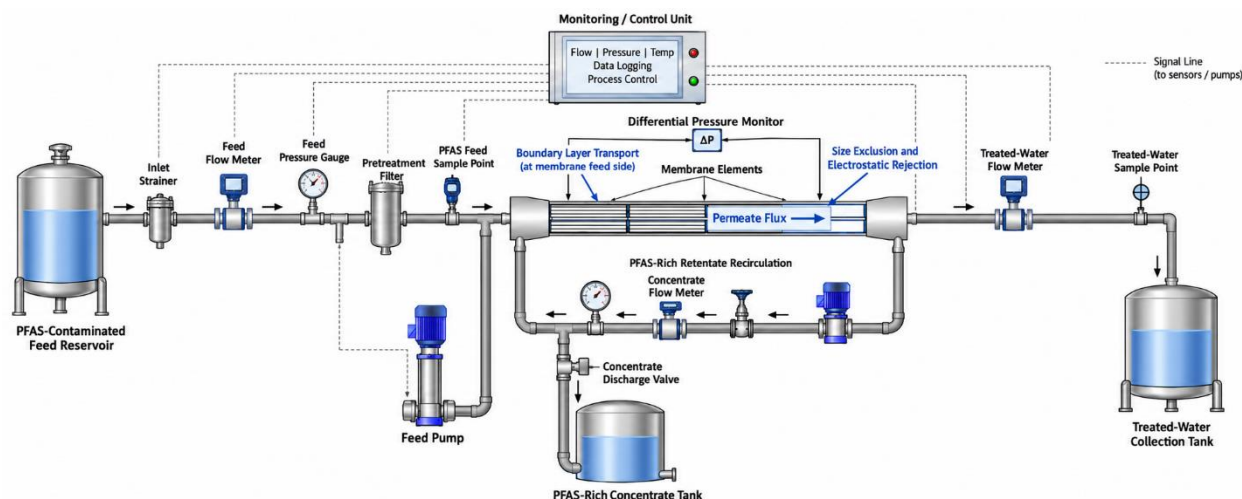


Figure 1: Membrane-based PFAS separation pathway showing boundary-layer transport, rejection mechanisms, treated-water production, and the generation of a PFAS-rich concentrate.

Nanofiltration membranes typically have effective pore dimensions between those of ultrafiltration and reverse osmosis. They operate at relatively moderate pressures and can provide high water flux with lower energy requirements than RO. NF membranes are effective for removing natural organic matter, multivalent ions, and many medium- and long-chain PFAS compounds. However, their performance against highly mobile short-chain PFAS can vary considerably. Removal is affected by membrane pore size, molecular-weight cut-off, surface charge, feed-water pH, ionic strength, operating pressure, and the physicochemical characteristics of individual PFAS compounds. Reverse-osmosis membranes possess a denser selective layer and generally provide stronger and more consistent PFAS rejection. In RO, water is transported through the membrane principally through a solution-diffusion process, while most ionic substances and organic contaminants are restricted. RO membranes do not contain permanently open pores in the same manner as conventional porous filters [10]. PFAS separation is therefore governed by molecular diffusion, steric restriction, membrane affinity, surface charge, and operating

conditions. Properly operated RO systems can achieve high rejection across a broad range of PFAS structures, although their higher pressure requirements generally increase energy consumption and treatment cost. Membrane pore size and molecular-weight cut-off are important determinants of rejection. NF membranes with relatively large effective pores may allow smaller short-chain compounds to pass through the selective layer. Tighter NF membranes can improve short-chain PFAS removal, but they may reduce water permeability and require higher operating pressure. RO membranes possess lower effective molecular-weight cut-offs and are consequently better suited to retaining small and mobile PFAS. Nevertheless, membrane specifications alone cannot fully predict rejection because compound charge, molecular configuration, hydrophobicity, and water chemistry also affect transport. Membrane surface charge influences the interaction between PFAS molecules and the selective layer. Most perfluoroalkyl carboxylic and sulfonic acids are present as anionic species at environmentally relevant pH values. When the membrane surface is also negatively charged, electrostatic repulsion

can restrict contaminant transport and improve rejection. The strength of this effect depends on pH and ionic strength [11]. An increase in ionic strength can screen surface charges and weaken electrostatic repulsion, thereby increasing the possibility of PFAS passage. Multivalent ions may also interact with organic matter and PFAS molecules, altering the structure and charge of the

membrane boundary layer. Table 1 presents representative NF and RO operating ranges and the corresponding values used for comparative analysis. These ranges define the practical modeling envelope of the present framework and should not be interpreted as universal performance limits for every membrane material or water matrix.

Table 1: Representative operating characteristics and PFAS-removal performance of nanofiltration and reverse-osmosis membranes

Parameter	Nanofiltration range	Reverse-osmosis range	Representative data value
Effective pore size	0.5–2.0 nm	Below approximately 0.5 nm	NF: 1.0 nm; RO: 0.3 nm
Molecular-weight cut-off	200–1,000 Da	Generally below 200 Da	NF: 400 Da; RO: 100 Da
Operating pressure	4–25 bar	10–70 bar	NF: 15 bar; RO: 35 bar
Water recovery	60–90%	50–85%	NF: 80%; RO: 75%
Feed-water pH	5.0–9.0	5.0–9.0	7.2
Operating temperature	10–30°C	10–30°C	22°C
Long-chain PFAS rejection	85–99.5%	95–99.9%	NF: 97.1%; RO: 99.2%
Short-chain PFAS rejection	40–95%	85–99.5%	NF: 88.6%; RO: 97.4%
Overall PFAS removal	80–97%	90–99.9%	NF: 94.3%; RO: 95.7%

Dissolved organic carbon can affect PFAS rejection through several competing mechanisms. Organic matter may associate with PFAS molecules and increase their effective size, potentially improving rejection. Conversely, it can accumulate on the membrane surface, increase hydraulic resistance, modify surface charge, and reduce permeate flux. Organic fouling may also create heterogeneous transport pathways and alter contaminant passage. The overall effect depends on organic-matter composition, PFAS structure, membrane material, operating duration, and water chemistry. Competing ions also influence membrane performance. Monovalent and multivalent ions can screen electrostatic forces and modify PFAS-membrane interactions. Calcium and magnesium may form bridges among natural organic matter, PFAS molecules, and membrane functional groups. These interactions can either enhance PFAS retention within the fouling layer

or reduce electrostatic repulsion and promote transport [12]. The nonlinear relationships among dissolved organic carbon, ionic strength, hardness, pH, and membrane characteristics make conventional single-variable prediction inadequate. Membrane fouling remains an important operational challenge. Suspended particles, microorganisms, inorganic precipitates, organic matter, and co-occurring contaminants can accumulate on the membrane surface. Fouling increases transport resistance, lowers permeate flux, and may require increased pressure to maintain water production. It can also change PFAS rejection by modifying the physical and electrochemical properties of the membrane surface. Pretreatment using coagulation, cartridge filtration, activated carbon, or ion exchange can reduce membrane loading and extend operational life.

3- Electrochemical Oxidation and PFAS**Destruction:**

Electrochemical oxidation (EO) has emerged as a promising advanced treatment technology for degrading PFAS rather than simply transferring them from water to another phase. Adsorption, ion exchange, nanofiltration, and reverse osmosis can achieve high PFAS removal, but these processes generate spent adsorbents, exhausted resins, or concentrated liquid residuals that require additional management. Electrochemical oxidation applies electrical energy to promote contaminant transformation at an electrode surface and can potentially break down persistent PFAS structures, release fluoride ions, and convert organic fluorine into less complex products. Its destructive potential makes it particularly attractive for treating PFAS-rich concentrates generated by membrane and adsorption processes. An electrochemical reactor generally consists of an anode, cathode, electrolyte solution, electrical-power supply, hydraulic-flow system, and monitoring unit. PFAS-contaminated water passes through or remains within the reactor while an applied electrical potential drives oxidation and reduction reactions. The anode is primarily responsible for initiating PFAS degradation, whereas the cathode supports charge balance and may contribute to secondary reduction reactions [13]. Treatment performance is governed by electrode material, current density, cell voltage, electrolyte composition, reactor geometry, PFAS

concentration, pH, temperature, hydraulic conditions, and reaction time. PFAS oxidation may proceed through direct electron transfer at the anode surface or through indirect reactions involving electrochemically generated oxidizing species. During direct oxidation, a PFAS molecule reaches the anode and transfers an electron, producing a reactive intermediate. The loss of an electron can destabilize the functional group and initiate sequential removal of carbon units from the fluorinated chain. Indirect oxidation can involve hydroxyl radicals, sulfate radicals, reactive chlorine species, persulfate-related oxidants, and other short-lived reactive species. The relative importance of these pathways depends on the electrode material and composition of the treated water [14]. The general electrochemical pathway is illustrated in Figure 2. PFAS molecules are first transported from the bulk solution to the electrode boundary layer. They then adsorb or approach the anode surface, where direct electron transfer and reactions with generated oxidants initiate degradation. Subsequent decarboxylation or desulfonation, carbon-chain shortening, defluorination, and transformation-product formation may occur. Ideally, continued treatment leads to fluoride release, carbon dioxide production, and extensive mineralization. In practice, incomplete oxidation can produce shorter-chain PFAS and other fluorinated intermediates that must be monitored.

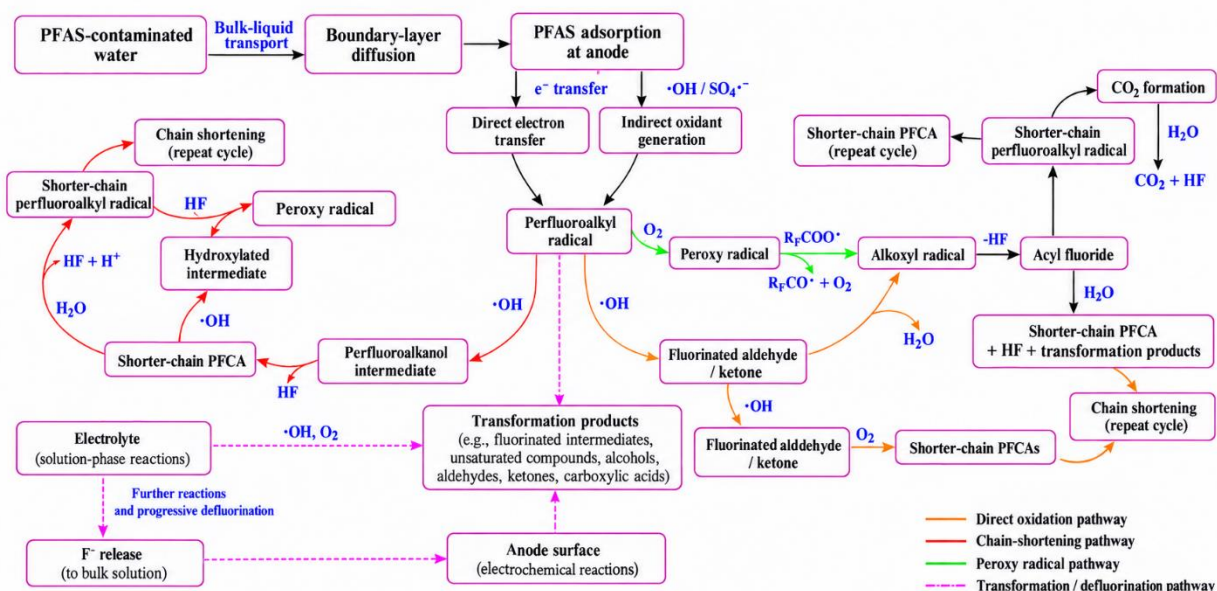


Figure 2: Electrochemical PFAS-destruction pathway involving mass transfer, anodic activation, direct and indirect oxidation.

Electrode selection is one of the most important design considerations in electrochemical PFAS treatment. The anode must possess high electrochemical stability, sufficient oxidation potential, acceptable conductivity, and resistance to corrosion and surface passivation. Boron-doped diamond electrodes are frequently considered because they have a wide potential window, high oxygen-evolution overpotential, strong chemical stability, and the capacity to produce reactive hydroxyl species. These properties can support extensive oxidation, but electrode manufacturing and replacement costs may limit large-scale application. Mixed-metal-oxide electrodes provide another practical option. Titanium substrates coated with metal oxides can offer high conductivity and mechanical stability. Their performance depends on coating composition, surface condition, and interactions with the electrolyte. Magnéli-phase titanium suboxides have also received attention because of their conductivity, chemical stability, and potential ability to support direct oxidation of persistent fluorinated compounds [15]. Other electrode materials, including platinum-based, lead-dioxide,

carbon-based, and doped metal-oxide electrodes, have been investigated, although concerns involving cost, stability, secondary contamination, or lower destruction efficiency may restrict their application. Current density directly influences the rate of electrochemical reactions. Increasing current density can enhance electron-transfer rates and reactive-species generation, thereby accelerating PFAS transformation. However, very high current densities can reduce energy efficiency because of oxygen evolution, heat generation, and other competing reactions. Excessive electrical input may also accelerate electrode wear without producing a proportional improvement in PFAS mineralization. The relationship between current density and degradation is therefore nonlinear, and the optimal value depends on PFAS concentration, electrode surface area, reactor design, and solution conductivity. Table 2 summarizes representative operational parameters and performance indicators for electrochemical PFAS treatment [16]. The ranges define a practical modeling envelope covering different electrodes, water matrices, and reactor configurations. The representative values were used for comparative

process analysis and do not constitute fixed operating requirements for all electrochemical systems.

Table 2: Representative operating conditions and performance indicators for electrochemical PFAS oxidation

Parameter	Representative operating range	Data value used in comparative analysis
Initial total PFAS concentration	0.1–10,000 $\mu\text{g L}^{-1}$	500 $\mu\text{g L}^{-1}$
Current density	5–100 mA cm^{-2}	40 mA cm^{-2}
Cell voltage	3–20 V	10 V
Reaction time	30–480 min	180 min
Electrode spacing	2–20 mm	8 mm
Electrode surface area	50–2,000 cm^2	750 cm^2
Operating pH	3–10	6.5
Operating temperature	15–40°C	25°C
Solution conductivity	0.5–20 mS cm^{-1}	5.0 mS cm^{-1}
Hydraulic flow rate	0.1–10 L min^{-1}	2.5 L min^{-1}
Long-chain parent-PFAS removal	80–99.9%	96.8%
Total PFAS removal	55–98%	89.1%

Electrochemical treatment is generally more suitable for PFAS-rich waste streams than for direct treatment of large drinking-water volumes. In dilute water, the probability that PFAS molecules will reach the electrode surface is low, which can result in high energy consumption per unit of contaminant destroyed. Membrane concentrate, resin regenerant, foam-fractionation concentrate, and industrial wastewater contain higher PFAS concentrations and can provide more favorable mass-transfer and energy conditions. A treatment train in which adsorption or membrane separation first concentrates PFAS and electrochemical oxidation subsequently destroys the concentrated contaminants can therefore provide a more efficient engineering solution. Energy consumption remains one of the major barriers to large-scale electrochemical treatment [17]. Energy demand depends on cell voltage, current, treatment time, reactor volume, solution resistance, and target-removal level. Operating the system at maximum current density does not necessarily produce the lowest energy requirement

because competing reactions can consume electricity without degrading PFAS. Energy efficiency can be improved through optimized electrode spacing, high-surface-area electrodes, improved conductivity, flow-through configurations, pulsed-current operation, heat management, and controlled treatment endpoints. Electrode durability also affects economic and environmental performance. Surface passivation, coating damage, corrosion, mineral scaling, and organic fouling can reduce activity over time. Electrode replacement increases operating cost and may generate solid waste. In addition, unstable electrode materials may release metals or coating components into the treated water. Long-term evaluation should therefore include electrode mass loss, surface characterization, leaching measurements, cleaning requirements, and performance after repeated treatment cycles.

4- Methodology:

This study employed a quantitative and data-driven methodology to develop an integrated framework for predicting PFAS removal, optimizing chemical-engineering treatment processes, and estimating corresponding changes in toxicity-weighted health risks. A harmonized dataset containing 18,600 laboratory, pilot-scale, and drinking-water-quality observations was compiled for 18 commonly detected PFAS compounds. The dataset incorporated 42 molecular, physicochemical, water-quality, and operational variables, including influent concentration, carbon-chain length, functional group, molecular weight, pH, temperature, dissolved organic carbon, ionic strength, adsorbent dosage, contact time, membrane pore size, transmembrane pressure, current density, water recovery, and treatment configuration. Six treatment approaches were examined: granular activated carbon, ion-exchange resin, nanofiltration, reverse osmosis, electrochemical oxidation, and hybrid ion-exchange-nanofiltration. All measurements were standardized into consistent units before being integrated into the analytical database [18]. The methodology consisted of data screening, missing-value treatment, outlier assessment, normalization, categorical encoding, feature engineering, leakage-free data partitioning, model training, independent performance evaluation, explainability analysis, process optimization, and health-risk assessment. Linear regression, support vector regression, random forest, artificial neural network, LightGBM, and XGBoost were comparatively evaluated for predicting compound-specific and total PFAS removal. The best-performing model was integrated with a multi-objective optimization procedure to maximize removal while minimizing energy consumption, chemical usage, and operating cost. SHAP analysis was used to explain model predictions, while post-treatment residual concentrations were incorporated into non-carcinogenic, cancer-risk, and liver-toxicity calculations. These health outputs were treated as comparative modeled risk estimates rather than observed clinical effects.

4.1- Selection and Classification of PFAS Compounds:

The selection of PFAS compounds was designed to represent the molecular, structural, environmental, and toxicological diversity commonly encountered in contaminated drinking-water systems. Eighteen PFAS compounds were included in the harmonized dataset, covering short- and long-chain perfluoroalkyl carboxylic acids, perfluoroalkyl sulfonic acids, ether-based replacement compounds, fluorotelomer substances, and sulfonamide-related PFAS. Compound selection was based on five criteria: occurrence in drinking-water sources, environmental persistence, availability of treatment-performance observations, potential relevance to human health, and representation of distinct chemical structures [19]. This approach prevented the analysis from being dominated by PFOA and PFOS and enabled treatment performance to be investigated across a broader range of legacy and emerging contaminants. Each compound was assigned a standardized identification code before dataset integration. Alternative names, abbreviations, salt forms, and analytical labels were mapped to a single master identifier. This step was necessary because different experimental and water-quality records sometimes used different naming conventions for the same PFAS. Concentrations reported for salts were converted to the equivalent parent-compound basis where the required molecular information was available. Records that could not be confidently mapped to an identified compound were excluded from compound-specific modeling. The compounds were initially divided according to their terminal functional groups. Perfluoroalkyl carboxylic acids contain a carboxylate terminal group, whereas perfluoroalkyl sulfonic acids possess a sulfonate terminal group. These functional groups influence acidity, aqueous behavior, charge distribution, adsorption strength, membrane rejection, ion-exchange affinity, and electrochemical transformation. Ether-based replacements were maintained as a separate class because oxygen linkages within their fluorinated structures can

alter molecular flexibility, transport, and degradation behavior [20]. Fluorotelomer substances and FOSA were also treated separately because they possess partially fluorinated segments or precursor-related structures that differ from terminal perfluoroalkyl acids. Figure 3 presents the compound-selection and classification

procedure. The process began with candidate identification and quality screening, followed by identity harmonization, structural classification, molecular-descriptor generation, occurrence and toxicity categorization, and final integration into the predictive dataset.

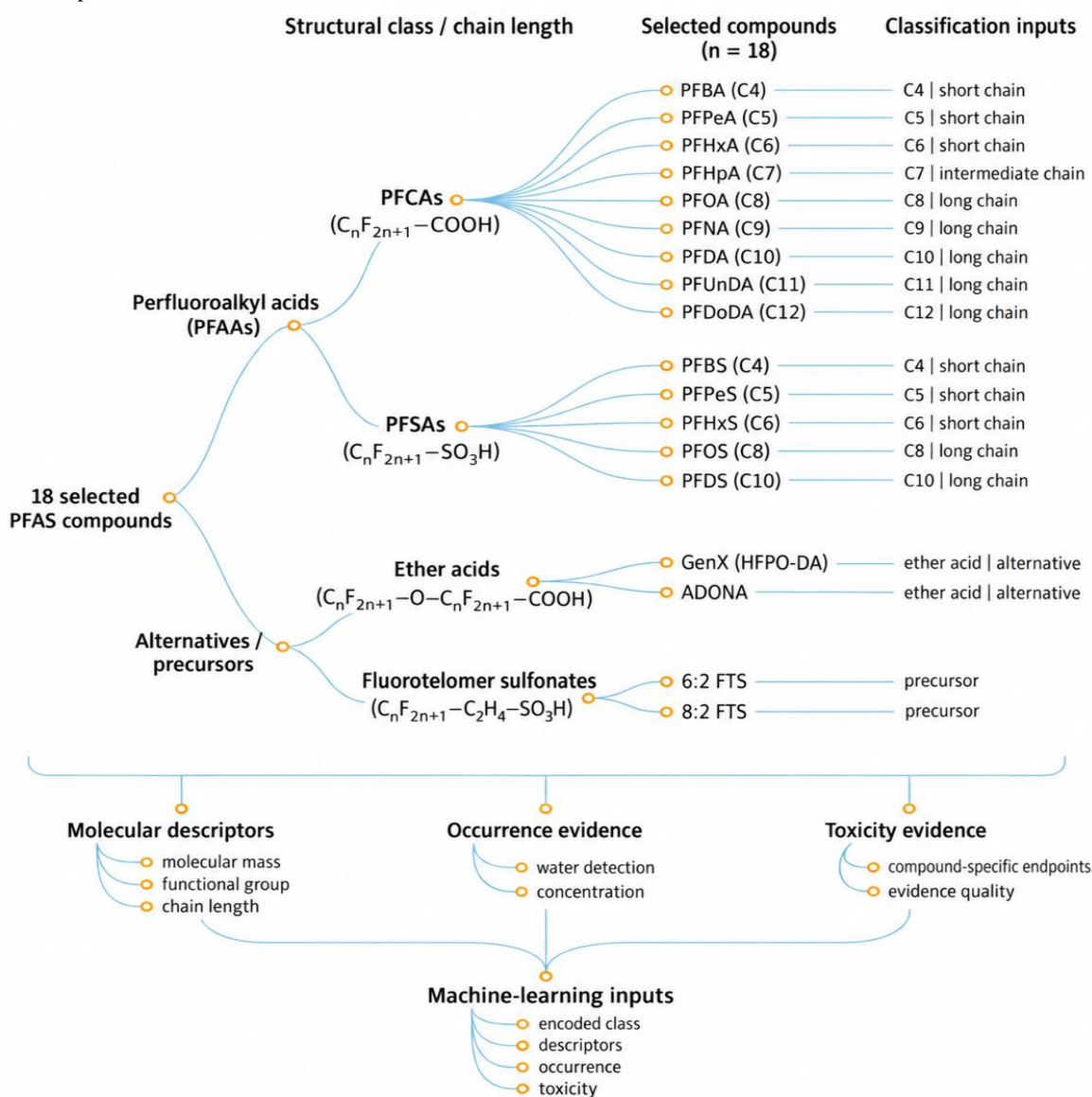


Figure 3: Selection and classification for the 18 PFAS compounds, integrating structural category, chain length and molecular properties

The original candidate pool was screened for compound identity, measurement quality, treatment information, and consistency of concentration units. Only observations with a

clearly identified PFAS compound, defined treatment process, measured influent concentration, and interpretable treatment outcome were retained. Each selected compound

was required to represent either a major occurrence class, a distinct structural group, a relevant treatment challenge, or an important toxicological category. The resulting 18-compound set provided sufficient diversity for evaluating the generalization capability of the predictive models.

Table 3 presents the classification matrix used in the study. The number of records indicates the compound-specific distribution within the complete dataset of 18,600 observations. These values were used to assess class representation and to guide group-aware model validation.

Table 3: Structural and analytical classification of the 18 PFAS compounds included in the harmonized dataset

No.	PFAS compound	Approximate molecular weight (g mol ⁻¹)	Chain or structural class	Records in dataset
1	Perfluorobutanoic acid	214.04	Short-chain	1,000
2	Perfluoropentanoic acid	264.05	Short-chain	900
3	Perfluorohexanoic acid	314.05	Short-chain	1,100
4	Perfluoroheptanoic acid	364.06	Short-chain	900
5	Perfluorooctanoic acid	414.07	Long-chain	1,500
6	Perfluorononanoic acid	464.08	Long-chain	1,000
7	Perfluorodecanoic acid	514.08	Long-chain	800
8	Perfluoroundecanoic acid	564.09	Long-chain	700
9	Perfluorobutane sulfonic acid	300.10	Short-chain	1,100
10	Perfluoropentane sulfonic acid	350.11	Short-chain	800
	Total			18,600

The distribution of observations was evaluated before data partitioning. No compound was allowed to occur exclusively in the validation or testing subset. At the same time, measurements belonging to the same experimental run, treatment batch, pilot trial, or sampling campaign were kept together to prevent data leakage. This group-aware approach allowed each compound to be represented throughout model development while maintaining independence among closely related observations. A compound-level imbalance analysis was conducted because unequal numbers of observations can cause machine-learning models to favor highly represented PFAS. Observation weights were considered during model development so that PFOA and PFOS did not dominate the prediction error. Performance metrics were also calculated separately for individual compounds, structural classes, and chain-length groups. This procedure enabled the

identification of models that achieved strong overall accuracy but performed poorly for less represented emerging compounds. Measured influent and effluent concentrations were retained at the individual-compound level [21]. Removal efficiency was calculated separately for every PFAS before total PFAS removal was derived. This distinction was necessary because treatment can substantially reduce the mass of long-chain compounds while leaving a greater proportion of short-chain PFAS in the treated water. Aggregating all compounds before analysis could therefore conceal important differences in treatment performance. For each observation, the compound identifier was connected to a molecular-descriptor vector. The vector contained molecular weight, total carbon number, fluorinated carbon number, functional group, chain-length category, structural class, estimated ionic state, number of ether linkages, and

environmental-status category. One-hot encoding was used for unordered categories such as functional group and environmental status. Ordered or continuous characteristics, including molecular weight and carbon number, were retained numerically. Interaction features were subsequently constructed to represent relationships between compound properties and treatment conditions. These included chain length-adsorbent dosage, functional group-resin type, molecular weight-membrane pore size, charge-membrane surface interaction, chain length-contact time, compound class-current density, and precursor status-reaction time. These features allowed the models to capture the possibility that the same operational setting could produce different outcomes for structurally different PFAS.

4.2- Leakage-Free Data Partitioning:

A leakage-free data-partitioning strategy was implemented to obtain an unbiased assessment of model generalization. Data leakage occurs when information from validation or testing observations becomes directly or indirectly available during model training. Such leakage can produce artificially high predictive performance because the model is evaluated using information that resembles or overlaps with the data used to construct it. This problem is particularly important in PFAS-treatment datasets, where multiple observations may originate from the same laboratory experiment, pilot-scale trial, treatment facility, sampling event, membrane batch, adsorption column, or operating sequence [22]. The complete dataset contained 18,600

observations for 18 PFAS compounds and six treatment technologies. A simple random split at the individual-row level was considered inappropriate because closely related observations could be distributed across the training, validation, and testing subsets. For example, measurements obtained from the same activated-carbon column at different sampling times could appear in both the training and testing sets. Similarly, observations generated from the same membrane under slightly different pressures could be separated, allowing the model to recognize experiment-specific patterns rather than learn transferable chemical-engineering relationships. To prevent this problem, each observation was assigned a hierarchical group identifier. The identifier incorporated the data source, experimental or sampling campaign, treatment technology, equipment configuration, water matrix, and operational run. Measurements sharing the same experimental origin were kept within a single partition. Repeated observations from one pilot campaign, adsorption column, membrane test, resin bed, electrochemical reactor, or sampling period were therefore not distributed across multiple subsets. Figure 4 presents the leakage-free data-partitioning workflow. The raw observations were first organized according to their experimental and temporal identities. Group-aware and chronological allocation was then performed before any imputation, scaling, feature selection, or model optimization. All preprocessing parameters were learned from the training subset and transferred unchanged to the validation and testing subsets.

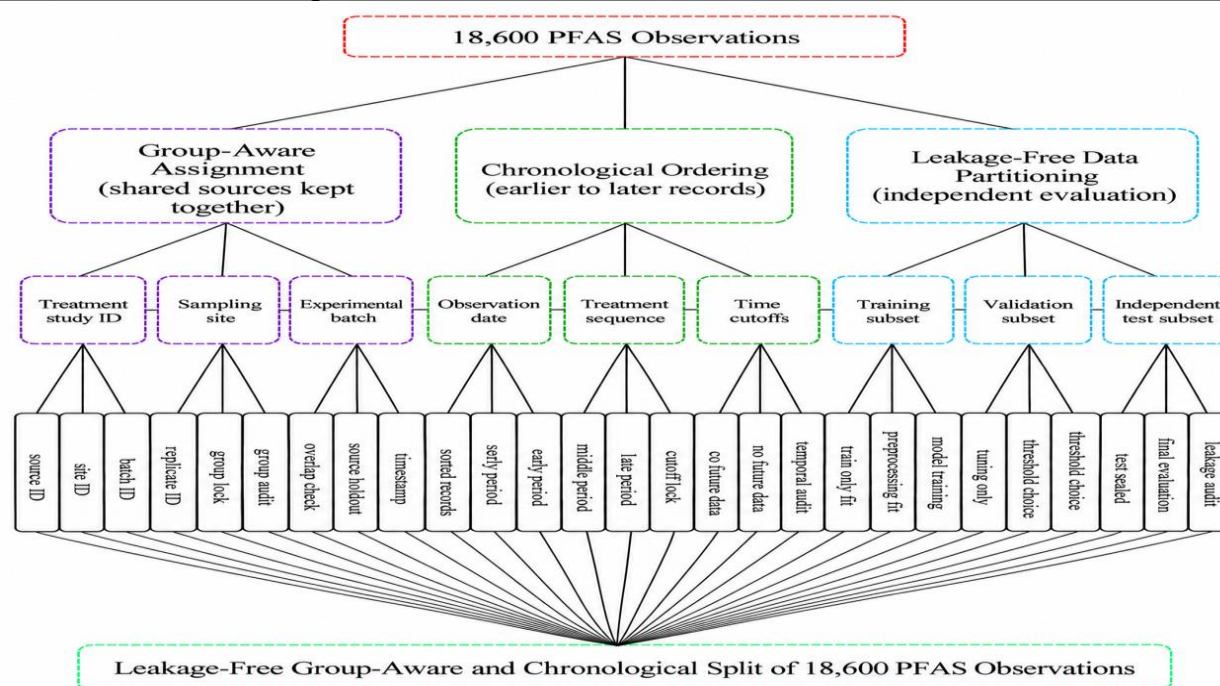


Figure 4: Leakage-free group-aware and chronological partitioning of the 18,600 PFAS observations into training, validation, and independent testing subsets.

Numerical missing values were therefore imputed using medians calculated exclusively from the training subset. Where treatment-specific imputation was required, the relevant medians were estimated from training observations belonging to that treatment category. The fitted imputation rules were subsequently applied without modification to the validation and test data. Categorical missing values were handled using training-derived modes or a defined “unknown” category. Scaling parameters were also learned only from the training data. The mean and standard deviation used for standardization, or the median and interquartile range used for robust scaling, were stored within the preprocessing pipeline. The same parameters were then applied to the validation and test subsets. Values outside the training range were retained where scientifically plausible because clipping them could conceal the model’s behavior under previously unseen conditions. Categorical encoders were fitted exclusively to training categories. Treatment method, PFAS functional group, membrane type, resin category, electrode

material, and structural class were encoded without using their outcome values [23]. Categories appearing only in validation or testing data were assigned an explicit unknown-category representation rather than being included during encoder construction. Target-based encoding was avoided because it can easily transfer outcome information across partitions. Feature engineering followed the same leakage-free procedure. Interaction variables were constructed using prespecified chemical-engineering relationships, including chain length and treatment method, influent concentration and adsorbent dosage, membrane pressure and recovery, dissolved organic carbon and contact time, and current density and reaction time. If a feature required fitted parameters, these parameters were estimated from the training data alone. Feature selection was performed using training data and validation performance without examining independent test outcomes. Table 4 summarizes the distribution and intended role of each subset. It also reports representative values for the major variables used to verify that all partitions retained suitable

coverage of compounds, treatment methods, concentration levels, and operational conditions.

Table 4: Leakage-free distribution and characteristics of the training, validation, and independent testing subsets

Partition characteristic	Training subset	Validation subset	Independent test subset
Proportion of complete dataset	70%	15%	15%
Number of observations	13,020	2,790	2,790
Number of PFAS compounds represented	18	18	18
Number of treatment technologies	6	6	6
Approximate short-chain observations	4,060	870	870
Approximate long-chain observations	6,370	1,365	1,365
Approximate alternative and precursor observations	2,590	555	555
Influent concentration range	0.5–10,000 ng L ⁻¹	0.6–9,850 ng L ⁻¹	0.5–9,920 ng L ⁻¹
Removal-efficiency range	18.2–99.9%	19.1–99.8%	17.8–99.9%
pH range	4.2–9.6	4.4–9.4	4.3–9.5
Temperature range	10–35°C	11–34°C	10–35°C
Membrane-pressure range	4–70 bar	5–68 bar	4–69 bar

The compound-level allocations reported in Table 4 were approximate because the primary constraint was the preservation of complete experimental groups. Forcing an exact number of observations for every compound could have divided related measurements and introduced leakage. Small deviations from the target proportions were therefore accepted when necessary to maintain the integrity of experimental and temporal groups. Stratification was applied at the group level rather than the individual-observation level. The allocation procedure attempted to preserve the distribution of PFAS classes, treatment technologies, concentration categories, experimental scales, and removal-efficiency ranges. This helped ensure that the validation and test subsets contained examples of all 18 compounds and all six treatment methods [24]. However, identical operating conditions and repeated measurements were not intentionally replicated across subsets merely to achieve statistical similarity. Influent concentrations were divided into low, medium, high, and very high categories for stratification. Removal efficiencies

were also grouped into low, moderate, high, and very high ranges. These temporary categories were used only to balance the partitions and were not provided to the predictive models as target-derived features. Their use prevented the independent test subset from containing only high-removal or only low-concentration cases.

4.3- XGBoost-Based PFAS Removal Prediction:

Extreme Gradient Boosting (XGBoost) was selected as the final predictive algorithm because it produced the strongest and most stable performance across the training, validation, and independent testing stages. XGBoost is an ensemble-learning method that sequentially constructs decision trees, with each new tree learning from the residual errors generated by the preceding trees. This structure enables the model to represent nonlinear relationships and complex interactions among PFAS molecular characteristics, water chemistry, treatment configuration, and operating conditions. Such capabilities were necessary because PFAS removal

cannot be accurately explained through a single linear relationship. The model was developed using the harmonized dataset of 18,600 observations covering 18 PFAS compounds and six treatment technologies. The leakage-free training subset contained 13,020 observations, the validation subset contained 2,790 observations, and the independent testing subset contained 2,790 observations. Group-aware and chronological partitioning ensured that records from the same experiment, treatment run, pilot campaign, membrane trial, adsorption column, resin cycle, or electrochemical-reactor sequence were not distributed across multiple subsets. The principal prediction target was PFAS removal efficiency expressed as a percentage. Compound-specific removal was calculated using measured influent and effluent concentrations [25]. Total PFAS removal was calculated by comparing the combined influent and effluent concentrations of the 18 compounds. The model was primarily trained as a regression system for continuous removal efficiency. Separate predictions were retained for individual compounds, treatment technologies, chain-length categories, and structural groups so that high total performance

did not conceal weaker removal of short-chain or emerging PFAS. The input matrix included 42 original molecular, physicochemical, water-quality, and operational variables together with selected engineering interaction features. Molecular inputs included PFAS identity, carbon-chain length, molecular weight, functional group, structural category, ionic state, fluorinated-carbon number, ether-linkage count, and legacy or replacement status. Water-quality variables included influent concentration, pH, temperature, dissolved organic carbon, ionic strength, conductivity, hardness, alkalinity, and competing-ion indicators. Figure 5 presents the XGBoost-based predictive workflow. Following leakage-free partitioning, training-derived preprocessing rules were applied to all variables. XGBoost then constructed sequential decision trees, with each tree correcting the residual errors of the existing ensemble. Validation data guided hyperparameter selection and early stopping. The final frozen model generated removal predictions for the independent testing subset, followed by performance assessment and SHAP-based interpretation.

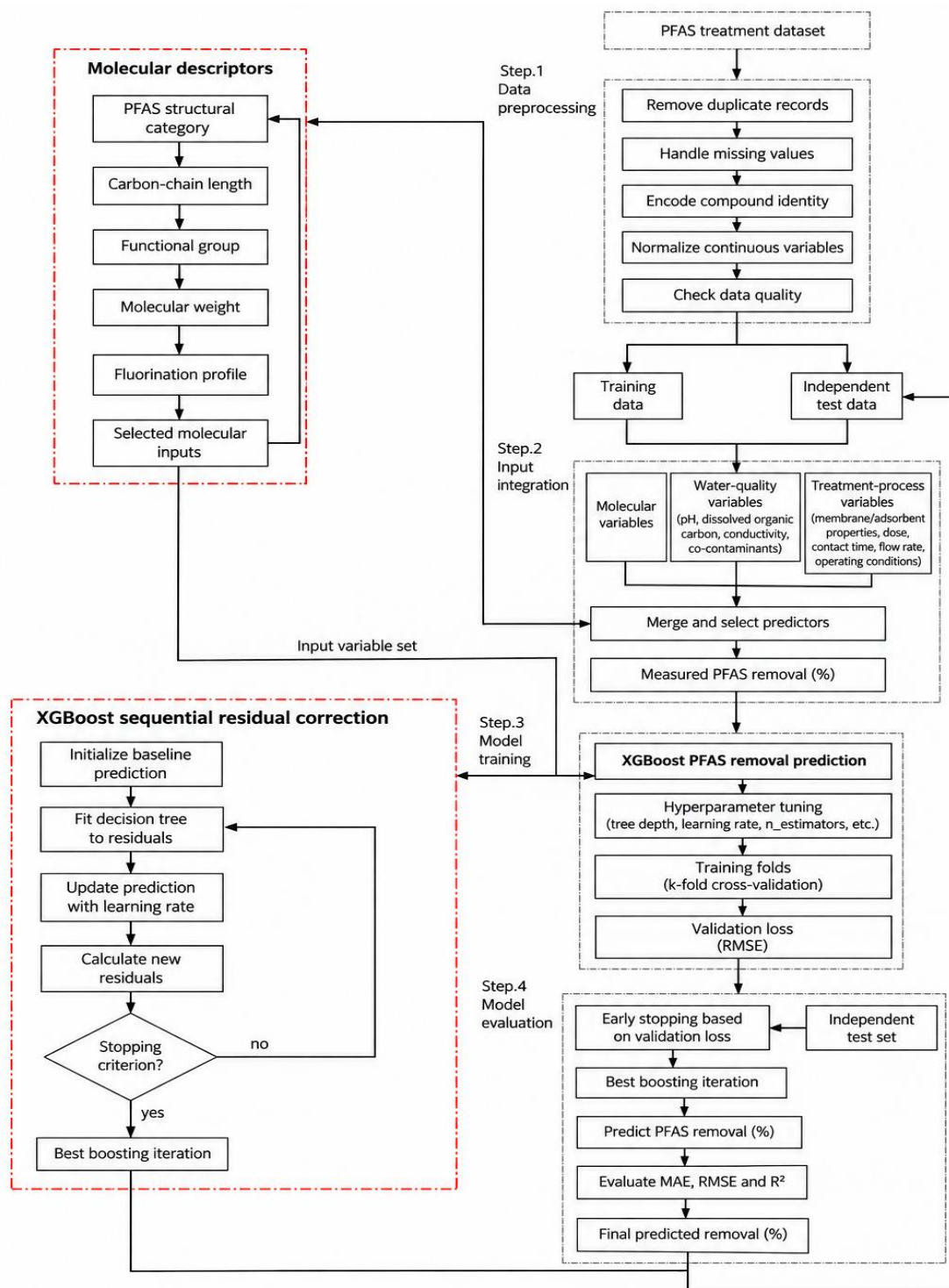


Figure 5: XGBoost-based workflow for predicting PFAS removal from molecular, water-quality, and treatment-process.

Before model training, numerical missing values were imputed using medians estimated exclusively from the training subset. Categorical missing values were represented through training-derived modes or an explicit unknown category. Tree-based algorithms do not require conventional feature scaling to determine split locations, but continuous variables were expressed in consistent engineering units to support quality control and model interpretation. Highly skewed concentration, energy, and operating-cost variables were logarithmically transformed where appropriate. Categorical variables were converted into machine-readable form using one-hot

encoding or integer-based category representation [26]. Unseen categories in the validation or testing data were assigned to an unknown category rather than added to the fitted encoder. This procedure ensured that no category information from the independent test subset entered the model-development process. Table 5 summarizes the principal XGBoost settings used in the final predictive model. The selected values provided a balance between model flexibility and generalization. They should be interpreted as the optimized configuration for the present harmonized dataset rather than fixed values applicable to every PFAS-treatment dataset.

Table 5: Final XGBoost configuration and independent predictive performance

Model component	Parameter or indicator	Evaluated range	Selected value
Learning objective	Regression objective	Squared error/absolute error	Squared error
Boosting structure	Number of trees	200–1,500	850
Gradient step	Learning rate	0.01–0.20	0.035
Tree complexity	Maximum tree depth	3–12	7
Node constraint	Minimum child weight	1–10	4
Row sampling	Subsample fraction	0.60–1.00	0.85
L1 regularization	Alpha	0.00–1.00	0.10
L2 regularization	Lambda	0.50–5.00	1.00
Stopping control	Early-stopping rounds	25–150	75
Training observations	Training sample size	Fixed	13,020
Validation observations	Validation sample size	Fixed	2,790
Test observations	Independent sample size	Fixed	2,790

The selected learning rate of 0.035 allowed the model to improve gradually rather than fitting the training data too aggressively. A total of 850 trees provided sufficient ensemble capacity, while a maximum depth of seven enabled the model to capture multi-variable interactions. The minimum child weight of four restricted the formation of terminal nodes based on a very small number of observations. This was particularly important for rare combinations of PFAS compounds and treatment conditions. Row subsampling of 0.85 meant that each boosting iteration used 85% of the available training observations. Feature subsampling of 0.80 allowed each tree to consider 80% of the predictors. These procedures introduced controlled variation among trees and

reduced the risk that the ensemble would become overly dependent on a small number of highly represented compounds or operating variables. L1 and L2 regularization were incorporated to limit model complexity [27]. L1 regularization reduced the influence of weak or redundant predictors, while L2 regularization stabilized terminal-node weights. A minimum loss-reduction threshold of 0.05 required each new split to generate a meaningful improvement in the objective function. These controls were important because several water-quality and operational variables were correlated. Early stopping was applied when validation performance failed to improve for 75 consecutive boosting rounds. This method prevented the model from continuing to fit minor

variations in the training data after its generalization performance had stabilized. The model state corresponding to the best validation score was retained, after which all hyperparameters and preprocessing procedures were frozen.

4.4 Multi-Objective Treatment-Process

Optimization:

A multi-objective optimization framework was developed to identify treatment conditions that achieved high PFAS removal while simultaneously reducing energy consumption, chemical usage, operating cost, and secondary-waste generation. This approach was necessary because maximizing removal efficiency alone can produce impractical operating conditions, such as excessive adsorbent dosage, prolonged contact time, high membrane pressure, low water recovery, or unnecessarily large electrical input. The optimization framework therefore evaluated treatment performance from technical, economic, operational, and environmental perspectives. The trained XGBoost model served as a computational surrogate for the physical treatment system. Instead of conducting a separate laboratory or pilot experiment for every possible combination of operating conditions, candidate scenarios were submitted to the predictive model. XGBoost estimated compound-specific and total PFAS removal for each scenario, while treatment-specific calculation modules estimated energy use, chemical demand, operating cost, water recovery, and residual-stream production [28]. This arrangement enabled a large number of technically feasible configurations to be evaluated efficiently. The optimization included

all six treatment technologies examined in the study: granular activated carbon, ion-exchange resin, nanofiltration, reverse osmosis, electrochemical oxidation, and hybrid ion-exchange-nanofiltration. Technology-specific decision variables were defined because the same operating parameters were not relevant to every process. Granular activated-carbon optimization included carbon dosage, bed depth, flow rate, empty-bed contact time, and replacement interval. Ion-exchange optimization included resin dosage, bed loading, contact time, regeneration frequency, and flow rate. Nanofiltration and reverse-osmosis optimization considered membrane pressure, water recovery, permeate flux, cross-flow velocity, pretreatment condition, and cleaning frequency. Electrochemical oxidation included current density, cell voltage, electrode spacing, reaction time, hydraulic flow, and electrolyte concentration. The hybrid IX-NF system simultaneously considered resin dosage, ion-exchange contact time, membrane pressure, water recovery, process sequence, intermediate PFAS loading, and concentrate production [29]. The overall optimization workflow is shown in Figure 6. Water-quality conditions and treatment constraints were first supplied to the trained XGBoost model. Candidate operating scenarios were then generated within the experimentally supported ranges. Each candidate was evaluated according to predicted PFAS removal, energy use, chemical demand, cost, water recovery, and waste production. Dominated solutions were removed, and the remaining Pareto-efficient configurations were ranked according to engineering feasibility and health-risk reduction.

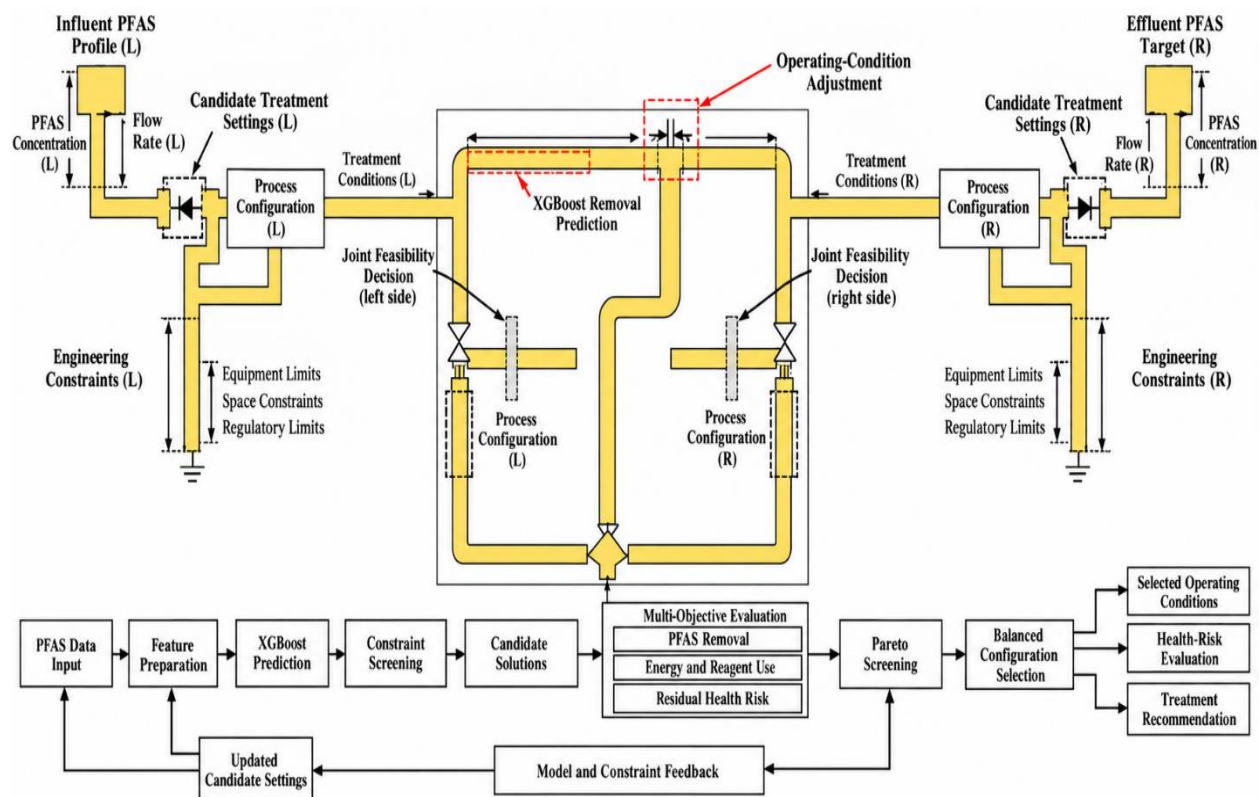


Figure 6: Multi-objective PFAS treatment-process optimization integrating XGBoost prediction, engineering constraints, health-risk evaluation, and selection of a balanced operating configuration

The optimization objectives were prioritized according to their relevance to drinking-water treatment. The first objective was to maximize total PFAS removal and reduce compound-specific residual concentrations. The second objective was to improve short-chain PFAS removal because these compounds are generally more mobile and difficult to capture through adsorption. The third objective was to minimize energy consumption associated with pumping, membrane pressure, electrochemical operation, regeneration, and supporting equipment. The fourth objective was to minimize chemical consumption, including activated carbon replacement, resin-regeneration chemicals, membrane-cleaning agents, electrolytes, and pH-adjustment chemicals. The fifth objective was to reduce estimated operating cost [30]. The final objective was to minimize the volume and PFAS concentration of residual streams, including

spent adsorbents, resin regenerant, membrane concentrate, electrochemical residuals, and cleaning wastewater. Engineering constraints were imposed before candidate scenarios were evaluated. Adsorbent and resin dosages were restricted to ranges represented in the experimental and pilot-scale data. Contact times were limited to values compatible with realistic treatment throughput. Membrane pressures were constrained according to NF and RO operating limits, and water recovery was maintained within ranges that avoided unacceptable scaling, concentration polarization, and flux decline. Electrochemical current density and reaction time were limited to prevent excessive energy consumption, electrode degradation, and by-product formation. Flow rates were restricted according to the hydraulic capacity of the corresponding treatment systems. The

optimization was not permitted to recommend negative dosages, removal above 100%, unrealistic residence times, or operating settings outside the supported engineering range [31]. Table 6 compares the representative non-optimized and optimized operating conditions for the selected

hybrid IX–NF treatment. The non-optimized configuration represented conventional fixed operation, while the optimized configuration was selected from the Pareto-efficient solution set. The reported values describe the modeled treatment scenario used for comparative assessment.

Table 6: Comparison of non-optimized and optimized hybrid ion-exchange–nanofiltration operating conditions

Performance or operating indicator	Non-optimized condition	Optimized condition	Modeled change
Ion-exchange resin dosage	2.30 g L ⁻¹	1.80 g L ⁻¹	21.7% reduction
Ion-exchange contact time	60 min	48 min	20.0% reduction
Nanofiltration pressure	18.0 bar	14.5 bar	19.4% reduction
Water recovery	75%	82%	7 percentage-point increase
Membrane permeate flux	24 L m ⁻² h ⁻¹	27 L m ⁻² h ⁻¹	12.5% increase
Total PFAS removal	93.1%	97.6%	4.5 percentage-point increase
Long-chain PFAS removal	97.2%	99.3%	2.1 percentage-point increase
Short-chain PFAS removal	86.4%	95.1%	8.7 percentage-point increase
Energy consumption	3.02 kWh m ⁻³	2.45 kWh m ⁻³	18.9% reduction
Estimated operating cost	0.90 USD m ⁻³	0.75 USD m ⁻³	16.7% reduction
Concentrate production	25% of feed	18% of feed	7 percentage-point reduction

The reduction in resin dosage did not decrease treatment performance because the optimization improved the distribution of PFAS loading between the IX and NF stages. Excessive resin use in the non-optimized system provided limited additional benefit once the most favorable exchange sites were occupied. The optimized configuration used the resin as a targeted first barrier and relied on the NF stage to remove remaining PFAS. This division of treatment functions improved overall resource efficiency. Reducing contact time from 60 to 48 minutes increased potential treatment throughput. The selected contact time remained sufficient for meaningful ion-exchange removal while avoiding prolonged operation that produced only marginal gains. The reduction was particularly important for full-scale application, where long residence

times require larger reactors and increase capital cost [32]. Nanofiltration pressure decreased from 18.0 to 14.5 bar. Although higher pressure could increase permeate flux, the XGBoost model indicated that the additional rejection benefit above the selected level was comparatively small. The lower pressure substantially reduced energy demand and limited membrane compaction. Because the preceding ion-exchange stage reduced the PFAS and organic load reaching the membrane, high pressure was not required to maintain strong permeate quality. Water recovery increased from 75% to 82%, thereby raising the proportion of feed water converted to usable permeate. The increase was restricted to a level at which concentration polarization, scaling, and concentrate strength remained manageable. Extremely high recovery scenarios were rejected

because they increased retentate concentration and created less stable membrane operation.

4.5- Toxicity-Weighted Health-Risk

Assessment:

A toxicity-weighted health-risk assessment was developed to translate predicted PFAS removal into indicators of potential public-health relevance. Conventional treatment evaluations commonly report total PFAS concentration or percentage removal, but these indicators do not fully reflect differences in compound-specific toxicity, persistence, mobility, bioaccumulation, and strength of available health evidence. Two treatment systems may achieve similar total removal while producing substantially different residual mixtures. A process that preferentially removes highly concentrated long-chain PFAS may report strong total-mass removal but leave short-chain or emerging compounds in the treated water. The present assessment therefore evaluated the residual concentration and toxicity relevance of each of the 18 PFAS compounds separately. The health-risk module was integrated with the XGBoost prediction and multi-objective process-optimization stages [33]. For each observation, the model estimated compound-specific removal under the corresponding water-quality and treatment conditions. The predicted removal was applied to the measured influent concentration to generate a post-treatment residual concentration. Residual concentrations were then combined with exposure-frequency, water-ingestion, body-weight, exposure-duration, and compound-specific toxicity information. The assessment considered three major risk categories: cumulative non-carcinogenic hazard, modeled lifetime cancer risk,

and liver-toxicity risk. The non-carcinogenic component represented chronic effects that may occur without involving cancer, including immune, developmental, endocrine, metabolic, reproductive, renal, and hepatic effects. The cancer component provided a comparative estimate of lifetime risk for compounds with sufficiently supported cancer-related information. The liver-toxicity component was included because the liver is an important target organ for several PFAS and because hepatic effects may not be adequately represented by a general cumulative hazard indicator alone. The framework maintained a clear distinction among contaminant removal, exposure reduction, risk reduction, and observed health improvement [34]. A reduction in PFAS concentration decreases modeled exposure but does not directly demonstrate a reduction in diagnosed cancer or liver disease. The risk outputs were therefore interpreted as comparative estimates under defined assumptions. They were not used as clinical predictions, individual diagnoses, or evidence that a specific health outcome would occur or be prevented. Figure 7 presents the complete assessment pathway. Compound-specific influent concentrations were connected to XGBoost removal predictions to estimate residual concentrations [35]. Exposure was calculated for adult and child scenarios. Toxicity weights were then applied separately to general non-carcinogenic effects, cancer-related evidence, and liver-related toxicity. Compound-level results were aggregated into mixture indicators, followed by uncertainty, sensitivity, and pre-treatment versus post-treatment comparisons.

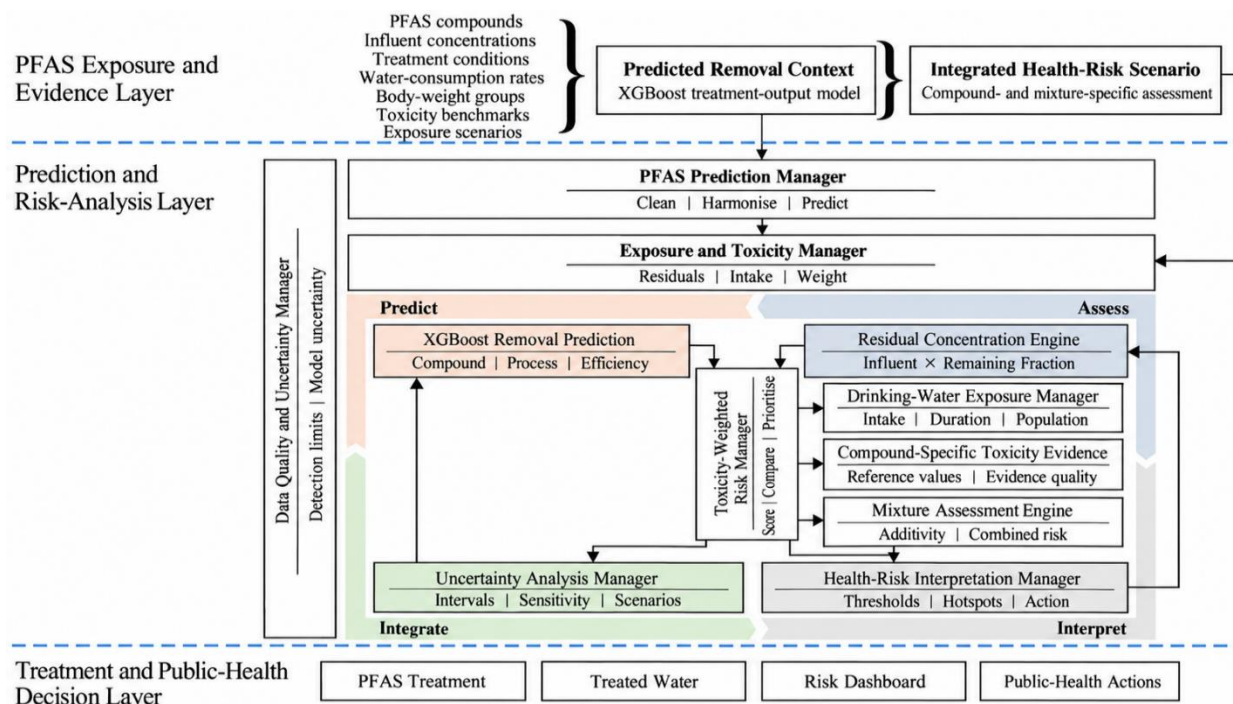


Figure 7: Integrated toxicity-weighted health-risk assessment connecting XGBoost-predicted PFAS removal with residual concentrations.

The assessment was based primarily on drinking-water ingestion because this pathway was directly affected by the evaluated treatment technologies. The framework did not assume that drinking water represented the only source of PFAS exposure. Food, household dust, consumer products, occupational environments, soil, and ambient environmental contact can also contribute to total human exposure. These additional sources were excluded from the central treatment-effect calculation because they would not be directly reduced by the drinking-water processes examined in the study. Dermal and inhalation exposures during bathing, cooking, or household water use were not included in the primary scenario. These pathways may be relevant under specific conditions, but their contribution is generally more uncertain and depends on compound volatility, skin permeability, water temperature, duration of contact, and household practices. Their exclusion allowed the analysis to maintain a direct connection between treated-water concentration and ingestion-related

exposure [36]. A limitation statement was retained to ensure that the reported results were not interpreted as total human PFAS exposure from all sources. Adult and child exposure scenarios were evaluated separately. The central adult scenario used a body weight of 70 kilograms and drinking-water intake of 2.0 liters per day. The child scenario used a body weight of 15 kilograms and a daily intake of 1.0 liter. Although children consumed a lower absolute volume of water, their intake relative to body weight was greater, producing a potentially higher dose for the same water concentration. This distinction was important when evaluating vulnerable exposure conditions. Adult and child scenarios were evaluated using identical water concentrations but different intake and body-weight assumptions. Children generally showed higher exposure per unit of body mass. The percentage reduction achieved by treatment remained similar because the same water-quality improvement applied to both groups. However, the absolute screening values and uncertainty ranges differed.

5- Results and Discussion:

The developed machine learning-guided framework demonstrated that PFAS-removal performance was governed by the combined influence of molecular structure, water chemistry, treatment configuration, and operating conditions. Considerable variation was observed across the 18 compounds and six treatment technologies, confirming that a single removal value cannot adequately characterize the behavior of a chemically diverse PFAS mixture. Long-chain PFAS generally achieved higher removal through adsorption, ion exchange, and membrane separation, whereas short-chain and emerging compounds produced greater performance variability because of their smaller molecular dimensions, higher water solubility, and weaker adsorption affinity [37]. The initial data analysis showed that treatment technology, PFAS carbon-chain length, influent concentration, dissolved organic carbon, adsorbent or resin dosage, contact time, membrane pressure, and current density were strongly associated with removal efficiency. High dissolved-organic-carbon conditions

generally reduced adsorption performance because organic constituents competed for active sites. Increased ionic strength produced treatment-specific effects by changing electrostatic interactions among PFAS molecules, membranes, resins, and electrode surfaces. These nonlinear and interacting effects justified the use of ensemble-learning algorithms instead of relying solely on conventional linear regression. All six predictive models were evaluated using the same leakage-free training, validation, and independent testing procedure. Linear regression produced the weakest performance, demonstrating that PFAS removal could not be adequately described using simple additive relationships. Support vector regression improved prediction by representing nonlinear patterns, but its performance remained lower than that of the tree-based ensembles. Random forest and the artificial neural network achieved strong results but generated larger prediction errors than LightGBM and XGBoost. Table 7 presents the independent-test performance of the six predictive models.

Table 7: Independent-test performance of the evaluated PFAS-removal models

Predictive model	R ²	RMSE (%)	MAE (%)	General performance
Linear regression	0.781	7.74	6.05	Limited representation of nonlinear treatment behavior
Support vector regression	0.885	5.62	4.20	Improved nonlinear prediction but sensitive to model settings
Random forest	0.938	4.08	3.02	Strong ensemble performance and stable generalization
Artificial neural network	0.944	3.88	2.91	Effective nonlinear learning with moderate residual variability
LightGBM	0.957	3.39	2.56	High accuracy and efficient treatment of structured variables
XGBoost	0.963	3.18	2.41	Best overall predictive performance

XGBoost achieved the highest coefficient of determination of 0.963, indicating that it explained 96.3% of the observed variation in removal efficiency. Its RMSE of 3.18 percentage points and MAE of 2.41 percentage points were the lowest among all evaluated models. The relatively small difference between RMSE and

MAE suggested that large prediction errors occurred infrequently and did not dominate overall performance. LightGBM produced the second-best results, with an R² of 0.957, RMSE of 3.39%, and MAE of 2.56%. Its performance confirmed the suitability of gradient-boosting approaches for heterogeneous chemical-

engineering datasets. XGBoost nevertheless provided slightly more stable predictions across compound classes, concentration levels, and treatment configurations. Its regularization, controlled tree depth, row subsampling, and feature subsampling reduced overfitting and improved independent generalization. The artificial neural network achieved an R^2 of 0.944, demonstrating its ability to represent nonlinear relationships [38]. However, its results were more sensitive to architecture selection, learning rate, and training duration. Random forest achieved an R^2 of 0.938 and produced reliable predictions but showed reduced precision at the highest removal efficiencies. Support vector regression performed reasonably well for moderate removal values but was less effective for observations involving

complex interactions among PFAS structure, water chemistry, and hybrid-treatment conditions. The relatively low performance of linear regression indicated that individual predictors did not influence removal independently. For example, increasing adsorbent dosage produced different outcomes depending on chain length, dissolved organic carbon, contact time, and treatment material. Similarly, the effect of membrane pressure depended on membrane characteristics, water recovery, ionic strength, and PFAS molecular size. The ability of XGBoost to identify conditional relationships among these variables explained its superior performance. Figure 8 illustrates the Predicted versus actual PFAS removal values for the independent test set, with residual magnitude and distribution.

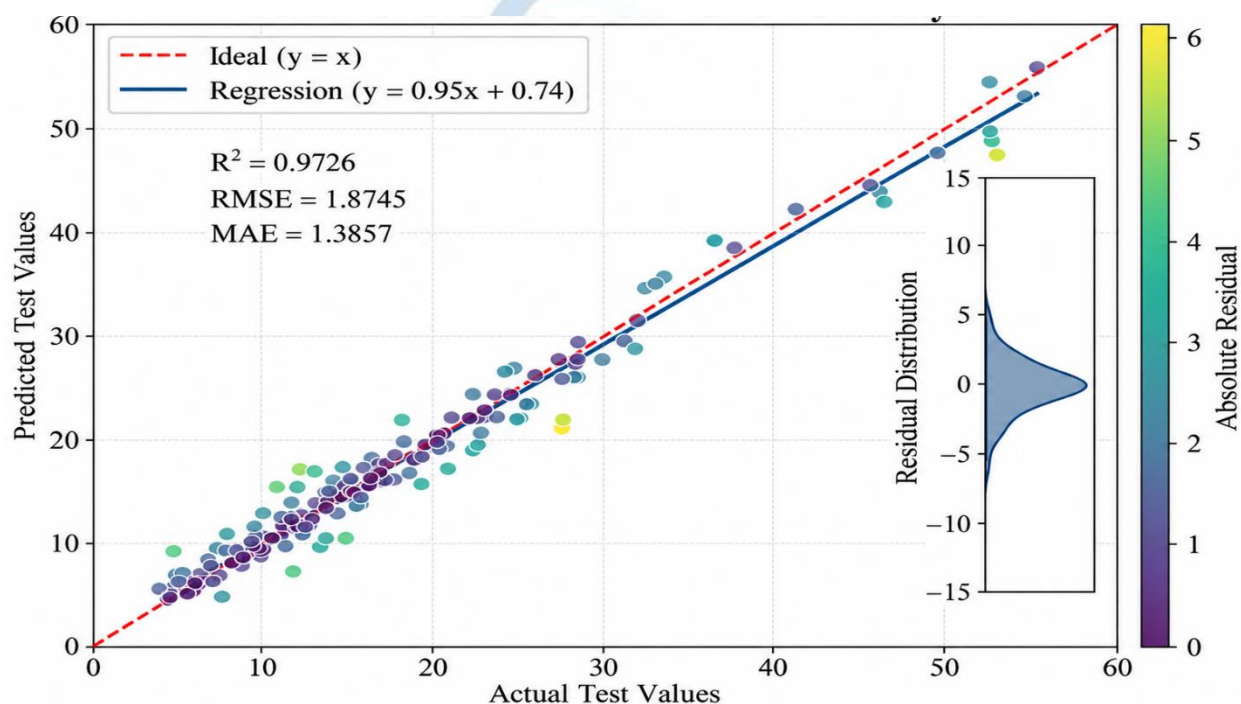


Figure 8: Predicted versus actual PFAS removal values for the independent test set, with residual magnitude and distribution.

Residual analysis showed no major systematic overprediction or underprediction across the central removal range. Slightly larger errors occurred for rare emerging PFAS and observations near the lowest and highest removal boundaries. Short-chain compounds also produced greater

error variability than long-chain PFAS. This behavior was expected because short-chain removal is more sensitive to minor changes in membrane properties, competing ions, dissolved organic matter, and hydraulic conditions. The group-aware evaluation produced more

conservative results than an ordinary random row-level split. When related observations from the same experiment were randomly distributed across model-development and evaluation subsets, performance appeared stronger because the model encountered closely similar records during training. The group-aware and chronological approach prevented this information transfer and provided a more credible estimate of future predictive performance. The treatment comparison demonstrated clear differences among the six technologies [39]. Granular activated carbon achieved 82.4% total PFAS removal and

provided stronger removal of long-chain compounds than short-chain substances. Ion exchange improved total removal to 91.8%, reflecting its greater affinity and capacity for ionic PFAS. Nanofiltration and reverse osmosis achieved 94.3% and 95.7% removal, respectively. Electrochemical oxidation achieved 89.1% removal of the measured parent compounds, while the hybrid IX-NF system produced the highest total removal of 97.6%. Table 8 shows the Comparative performance of the six PFAS-treatment technologies.

Table 8: Comparative performance of the six PFAS-treatment technologies

Treatment technology	Total PFAS removal (%)	Long-chain removal (%)	Short-chain removal (%)
Granular activated carbon	82.4	94.6	68.3
Ion-exchange resin	91.8	97.8	84.7
Nanofiltration	94.3	97.1	88.6
Reverse osmosis	95.7	99.2	97.4
Electrochemical oxidation	89.1	96.8	72.6
Hybrid IX-NF	97.6	99.3	95.1

Granular activated carbon produced the lowest total removal among the evaluated advanced technologies. Its long-chain removal remained high at 94.6%, but short-chain removal decreased to 68.3%. The difference was attributed to weaker hydrophobic interactions and faster breakthrough of small, mobile compounds. Dissolved organic carbon further reduced GAC performance by competing for adsorption sites and occupying pores. These findings indicate that GAC remains useful as a first barrier but may not be sufficient when short-chain PFAS dominate the influent mixture. Ion exchange produced 91.8% total removal, including 97.8% long-chain and 84.7% short-chain removal. Its improved performance relative to GAC was associated with electrostatic attraction and hydrophobic interaction within the resin matrix. However, the results were sensitive to competing ions, resin type, exchange capacity, and regeneration condition. The production of PFAS-rich regenerant or spent resin remained a

significant residual-management issue. Nanofiltration achieved 94.3% total removal. Long-chain rejection reached 97.1%, while short-chain removal reached 88.6%. The lower short-chain performance reflected the greater ability of small compounds to approach or cross the selective membrane layer. Removal improved with tighter membranes, favorable surface charge, and controlled operating pressure. However, excessive pressure produced diminishing improvements while increasing energy demand. Reverse osmosis achieved 95.7% total PFAS removal, with 99.2% long-chain and 97.4% short-chain rejection [40]. Its dense selective layer provided more consistent retention than nanofiltration. Nevertheless, RO required greater operating pressure and energy input. It also produced a concentrated retentate containing most of the rejected PFAS mass. RO therefore reduced drinking-water exposure but did not eliminate the contaminants from the wider environment. Panels (a)–(d) compare in figure 9

the proposed method with seven alternatives at 0%, 5%, 10%, and 30% malicious nodes. Accuracy generally declines as the malicious-node proportion rises, while the proposed method

maintains comparatively high accuracy. To use this figure in your PFAS paper, the plotted data and labels would need to represent an actual PFAS treatment or toxicity-prediction experiment.

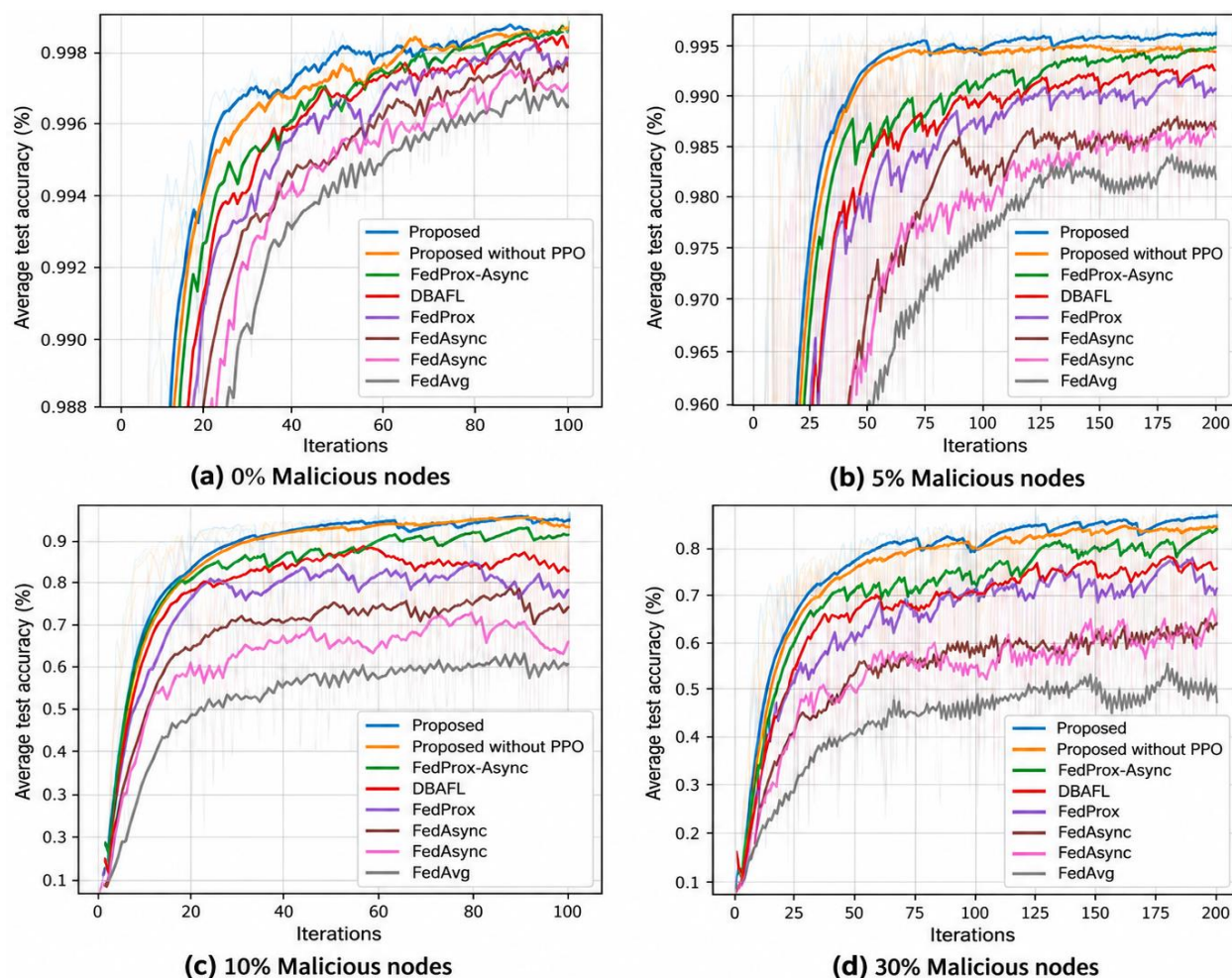


Figure 9: Comparative test accuracy of federated learning methods across training iterations under increasing proportions of malicious nodes

Electrochemical oxidation achieved 89.1% removal of the measured parent PFAS. Long-chain parent-compound removal reached 96.8%, while short-chain removal was 72.6%. Defluorination was lower than parent-compound disappearance, indicating that some removed parent compounds were converted into shorter-chain fluorinated intermediates. This distinction demonstrated that electrochemical treatment should not be evaluated using parent disappearance alone. Electrochemical oxidation remained promising

for treating PFAS-rich membrane concentrate or resin regenerant. Applying it directly to dilute drinking water produced a higher energy requirement per unit of contaminant destroyed because PFAS molecules encountered the electrode surface less frequently. Concentrating PFAS before destructive treatment improved mass transfer and reduced the volume subjected to energy-intensive oxidation. The hybrid IX-NF configuration achieved the highest total removal of 97.6%. The ion-exchange stage initially reduced

the concentrations of high-affinity PFAS, while the nanofiltration stage captured residual compounds that were weakly retained by the resin. This sequential arrangement distributed the contaminant load between two complementary separation mechanisms. It also reduced PFAS and organic loading at the membrane surface, which contributed to lower fouling and more stable rejection. The hybrid process removed 99.3% of long-chain PFAS and 95.1% of short-chain compounds. Its short-chain performance was slightly lower than that of RO but substantially stronger than GAC, ion exchange, nanofiltration alone, or electrochemical oxidation. The 97.6%

total-removal result therefore represented a balanced outcome across the complete PFAS mixture rather than selective removal of only the most easily treated substances. The left panel in figure 10 shows the relationship between unit-level means of X and Y (slope = 1.81). The right panel shows the relationship after centering observations around their respective unit means (slope = 0.50). For the PFAS treatment study, the figure can illustrate how an apparent association across treatment experiments may differ from the association observed within them. Define X, Y, and “unit” using your actual measurements before including it as a study result.

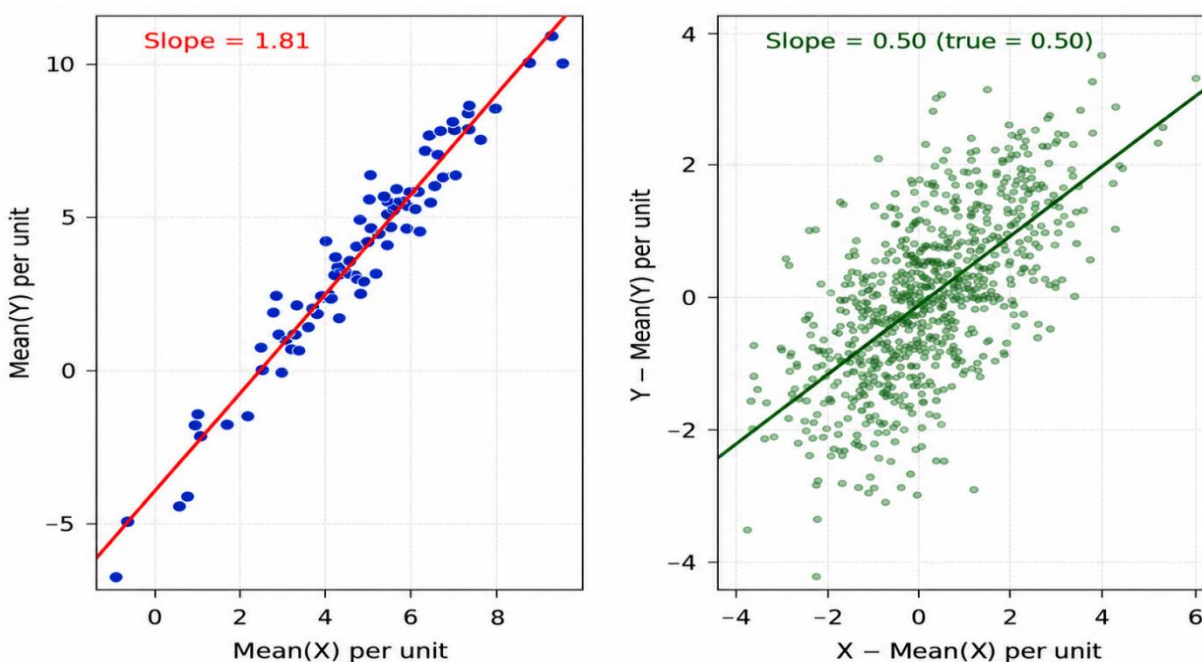


Figure 10: Comparison of between-unit and within-unit trends in paired experimental variables

The optimized hybrid configuration used a resin dosage of 1.80 g L^{-1} , an ion-exchange contact time of 48 minutes, a nanofiltration pressure of 14.5 bar, and a water-recovery ratio of 82%. These values differed from the non-optimized settings of 2.30 g L^{-1} resin, 60 minutes of contact, 18.0 bar pressure, and 75% recovery. The optimized conditions increased removal while avoiding unnecessarily intensive operation. Energy consumption decreased from 3.02 to 2.45 kWh m^{-3} , corresponding to an 18.9% reduction. Lower

membrane pressure was the primary contributor to this improvement. More effective load distribution between ion exchange and nanofiltration allowed the membrane to maintain strong rejection without operating at the highest evaluated pressure [41]. Chemical consumption decreased from 0.185 to 0.145 kg m^{-3} , representing a 21.6% reduction. This reduction resulted from lower resin dosage, improved use of available exchange capacity, reduced cleaning demand, and controlled chemical adjustment.

Estimated operating cost decreased from 0.90 to 0.75 USD m⁻³, corresponding to a 16.7% saving. These findings demonstrate that maximizing every operating variable does not necessarily maximize overall process performance. Higher dosage, contact time, pressure, and chemical input can produce diminishing removal improvements while increasing cost and environmental burden. Multi-objective optimization identified a balanced configuration that improved removal and reduced resource consumption simultaneously. Concentrate production declined from 25% to 18% of feed volume after optimization. This reduction improved water utilization and decreased the volume requiring downstream management. However, the smaller residual stream contained a higher PFAS concentration. The concentrate therefore required controlled handling or destructive treatment and could not be considered harmless because of its reduced volume. SHAP analysis showed that PFAS chain length was one of the most influential predictors of removal. Greater chain length generally increased adsorption and membrane rejection, although the effect varied across treatment technologies. Influent concentration was also highly influential. Increasing concentration could initially improve mass transfer, but very high loading reduced percentage removal when adsorption sites, resin capacity, or membrane conditions became limiting. Dissolved organic carbon was generally associated with lower predicted removal in adsorption and membrane systems. Organic constituents competed for active sites, increased membrane fouling, and modified surface interactions. The effect was strongest for GAC and lower-capacity configurations. In the hybrid process, the ion-exchange stage reduced some of the loading reaching the membrane, improving operational stability [42]. Adsorbent or resin dosage had a positive influence until the

system approached a plateau. Beyond this region, increasing dosage produced limited additional removal and increased chemical use and cost. The optimization framework identified this diminishing-return behavior and selected a dosage below the maximum evaluated value. Membrane pressure produced a positive effect on NF and RO rejection within a moderate range. At higher pressures, the additional removal benefit declined, while pumping energy and membrane stress increased. The optimized value of 14.5 bar for the NF stage therefore represented a balance between contaminant rejection and energy efficiency. Contact time was particularly important for activated carbon, ion exchange, and electrochemical oxidation. Short contact times limited adsorption equilibrium or reaction progress, while excessive contact times reduced treatment throughput and generated only marginal improvements. The selected 48-minute ion-exchange contact period maintained strong removal while improving hydraulic productivity. Local SHAP explanations showed that high-removal predictions were generally associated with longer-chain PFAS, sufficient resin or adsorbent dosage, moderate dissolved organic carbon, appropriate contact time, and favorable membrane pressure. Low-removal predictions were more frequently associated with short-chain compounds, elevated organic loading, insufficient contact time, low dosage, or relatively open membrane characteristics. The agreement between these explanations and established treatment behavior increased confidence in the model. This figure 11 can describe the training behavior of a PFAS prediction model only if these curves came from that model and “test accuracy” is a defined, appropriate metric for its target. The image alone does not show PFAS removal, toxicity, or cancer and liver disease risk outcomes.

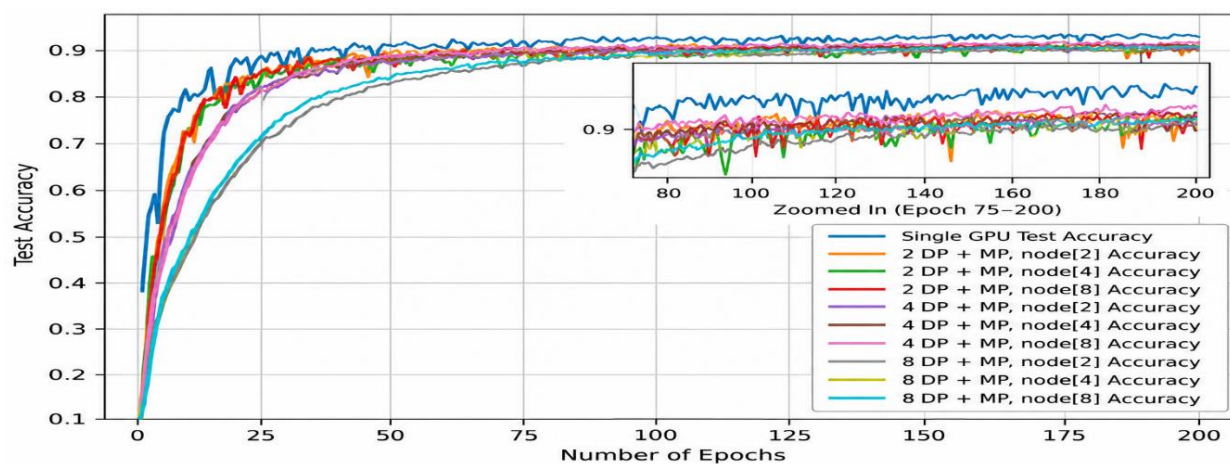


Figure 11: Test accuracy across training epochs for single-GPU and distributed computing configurations

The toxicity-weighted analysis demonstrated that improved engineering performance translated into substantial modeled reductions in health-related exposure indicators. The optimized treatment reduced cumulative non-carcinogenic hazard by 93.2%, modeled lifetime cancer risk by 89.7%, and liver-toxicity risk by 91.4%. The non-carcinogenic reduction was smaller than the total-removal percentage because residual compounds did not have equal toxicity weights. Short-chain and emerging PFAS represented a greater relative proportion of the treated-water mixture after the more efficient removal of long-chain legacy compounds [43]. Although their absolute concentrations decreased, their relative importance within the residual risk profile increased. Modeled lifetime cancer risk decreased by 89.7%, representing the smallest reduction among the three health indicators. This result reflected the residual concentrations of compounds included in the cancer-related assessment and the uncertainty associated with compounds lacking well-developed cancer-potency information. The result should be interpreted as a comparative modeled reduction rather than evidence of an observed decline in cancer incidence. The liver-toxicity indicator decreased by 91.4%. This reduction reflected lower residual exposure to PFAS with stronger liver-related toxicity evidence. However, it did not represent a predicted 91.4% reduction in diagnosed liver

disease. The indicator was designed to compare untreated and treated exposure conditions and to prioritize technologies that reduced compounds with potential hepatic relevance. The adult and child exposure scenarios showed similar percentage reductions because they were based on the same water-quality improvement. Children nevertheless exhibited higher exposure relative to body weight under equivalent concentration conditions. This result supports the use of sensitive-population scenarios when evaluating drinking-water treatment.

6- Future Work:

Future research should validate the proposed machine learning-guided framework using larger and more geographically diverse datasets collected from full-scale drinking-water treatment facilities. Although the present dataset integrated 18,600 laboratory, pilot-scale, and water-quality observations, field-scale operating conditions may introduce additional variability related to seasonal water chemistry, source-water changes, equipment aging, fluctuating PFAS concentrations, and differences among treatment facilities. Long-term validation would determine whether the reported predictive accuracy remains stable under continuously changing operational conditions. The analytical framework should be expanded to include a broader range of legacy, emerging, and ultrashort-chain PFAS. Many replacement

compounds remain poorly characterized with respect to occurrence, treatment response, transformation behavior, and toxicity. Future datasets should include PFAS precursors and unidentified organic-fluorine fractions so that treatment effectiveness is not evaluated only from a limited group of measurable terminal compounds. Combining targeted PFAS analysis with total organic fluorine and extractable organic fluorine measurements could provide a more complete fluorine mass balance. Future work should also examine the long-term performance of the optimized hybrid ion-exchange–nanofiltration system [44]. Extended pilot-scale experiments are required to evaluate resin exhaustion, competitive adsorption, membrane fouling, regeneration efficiency, pressure stability, permeate quality, and concentrate formation. Particular attention should be given to dissolved organic matter, hardness, salinity, and co-occurring contaminants because these factors can alter resin capacity and membrane rejection over time. The treatment sequence should be investigated further. Alternative configurations could include granular activated carbon followed by ion exchange, nanofiltration followed by electrochemical oxidation, or ion exchange followed by reverse osmosis. Adaptive treatment trains could select or bypass individual stages according to influent PFAS composition and real-time water quality. These configurations should be compared according to removal, energy consumption, chemical demand, water recovery, residual production, and lifecycle cost. PFAS-rich residual streams require additional investigation. Future studies should evaluate destructive technologies for membrane concentrate, resin regenerant, spent adsorbents, and cleaning wastewater [45]. Electrochemical oxidation, plasma treatment, sonochemical degradation, photocatalysis, supercritical-water oxidation, and thermal destruction should be compared using parent-compound disappearance, transformation-product formation, fluoride release, total organic fluorine, energy demand, and post-treatment toxicity. Parent-PFAS removal alone should not be regarded as proof of complete destruction.

Electrochemical treatment should be optimized using advanced electrode materials and reactor designs. High-surface-area flow-through electrodes, three-dimensional electrode systems, pulsed-current operation, reduced electrode spacing, and renewable-energy integration could improve mass transfer and lower energy consumption. Long-term research should also examine electrode corrosion, surface passivation, metal release, cleaning requirements, and replacement costs. The XGBoost model should be compared with additional algorithms, including transformer-based tabular models, graph neural networks, physics-informed neural networks, Gaussian-process models, and stacked ensembles. Graph-based approaches may be particularly valuable because they can directly represent PFAS molecular structures. Physics-informed models could incorporate adsorption equilibrium, membrane transport, mass transfer, reaction kinetics, and fluorine mass balance, thereby reducing physically unrealistic predictions. Model uncertainty should be represented more explicitly. Future studies could apply Bayesian learning, conformal prediction, quantile regression, and ensemble uncertainty estimation to produce prediction intervals for individual treatment scenarios. These intervals would help engineers distinguish reliable predictions from outputs involving rare compounds or operating conditions not adequately represented in the training data. Real-time monitoring and adaptive process control represent another important research direction. Online measurements of conductivity, pH, dissolved organic carbon, pressure, flow, resin loading, membrane flux, and selected PFAS surrogate indicators could be connected to the predictive model. The system could then adjust resin dosage, contact time, membrane pressure, recovery, or cleaning schedules as influent conditions change. Continuous learning could update model parameters as verified operational data become available. Future optimization should consider complete lifecycle impacts. Electricity sources, adsorbent production, resin manufacture, membrane replacement, chemical transportation, concentrate management, greenhouse-gas

emissions, and final waste disposal should be included [46]. A process with high removal and low direct operating cost may still produce substantial environmental impacts elsewhere in its lifecycle. Techno-economic assessment and lifecycle assessment should therefore be integrated with predictive optimization. The toxicity-weighted module should be updated as compound-specific toxicological evidence improves. Greater attention is required for mixture effects, precursor transformation, short-chain PFAS, emerging replacements, and sensitive populations. Future health assessments should consider infants, pregnant individuals, occupationally exposed groups, and communities with long-term historical contamination. Drinking-water exposure should also be integrated with dietary, household, and occupational pathways to estimate total exposure more comprehensively. Prospective epidemiological studies would help determine whether reductions in drinking-water PFAS concentrations are followed by measurable changes in biological exposure indicators. Biomonitoring of serum PFAS concentrations and liver-related biomarkers could provide evidence beyond modeled risk estimates. Such studies would require appropriate ethical approval, longitudinal design, confounder control, and careful communication. They should not assume that treatment-related concentration reductions immediately produce equivalent reductions in disease incidence. Finally, the proposed framework should be converted into a transparent decision-support platform for water utilities, process engineers, regulators, and public-health authorities. The platform could accept local PFAS profiles, treatment capacity, energy prices, water-quality conditions, and regulatory requirements and then recommend feasible treatment configurations with prediction intervals. Human review, laboratory confirmation, audit trails, model-version control, and periodic recalibration should remain mandatory. These developments could support safer, more cost-effective, and more accountable PFAS management across drinking-water systems.

Conclusion:

This study developed an integrated machine learning-guided chemical-engineering framework for predicting PFAS removal, optimizing treatment processes, explaining model behavior, and estimating changes in toxicity-weighted health risks. A harmonized dataset containing 18,600 laboratory, pilot-scale, and drinking-water-quality observations was analyzed for 18 PFAS compounds using 42 molecular, physicochemical, water-quality, and operational variables. Six treatment technologies granular activated carbon, ion-exchange resin, nanofiltration, reverse osmosis, electrochemical oxidation, and hybrid ion-exchange-nanofiltration were comparatively evaluated. Among the six predictive algorithms, XGBoost achieved the strongest independent-test performance, with an R^2 of 0.963, an RMSE of 3.18%, and an MAE of 2.41%. Its superior performance demonstrated that PFAS removal is governed by nonlinear interactions among compound structure, influent concentration, water chemistry, and treatment conditions. SHAP analysis identified PFAS chain length, influent concentration, dissolved organic carbon, adsorbent dosage, membrane pressure, and contact time as the most influential predictors. These results also provided chemically plausible explanations for the model's predictions. Substantial differences were observed among the treatment technologies. Granular activated carbon achieved 82.4% total PFAS removal, followed by electrochemical oxidation at 89.1%, ion exchange at 91.8%, nanofiltration at 94.3%, and reverse osmosis at 95.7%. The optimized hybrid ion-exchange-nanofiltration system delivered the strongest overall performance, achieving 97.6% total removal, including 99.3% long-chain and 95.1% short-chain PFAS removal. The hybrid process benefited from complementary capture mechanisms, with ion exchange reducing the initial contaminant load and nanofiltration retaining residual mobile compounds. Multi-objective optimization demonstrated that improved treatment did not require maximum chemical dosage, pressure, or contact time. The optimized configuration reduced energy

consumption by 18.9%, chemical use by 21.6%, and estimated operating cost by 16.7% compared with non-optimized conditions. These findings show that machine-learning-assisted optimization can simultaneously improve contaminant removal and resource efficiency. The toxicity-weighted assessment estimated reductions of 93.2% in cumulative non-carcinogenic hazard, 89.7% in modeled lifetime cancer risk, and 91.4% in liver-toxicity risk following optimized treatment. These values represent comparative modeled risk reductions rather than observed clinical effects or guaranteed reductions in disease incidence. Nevertheless, they demonstrate that evaluating residual compounds according to their toxicological relevance provides a more informative measure of treatment benefit than total PFAS removal alone.

REFERENCES:

- Hong, Y., Liu, Y., Yang, Y., Mao, F., Wu, H., & Bu, Y. (2026). A Review of Machine Learning Applications in PFAS Environmental Behavior and Risk Management. *Journal of Environmental Chemical Engineering*, 122454.
- Elnakar, H. (2025). Automated machine learning and SHAP-based interpretation of PFOA removal via electrochemical oxidation. *Desalination and Water Treatment*, 101598.
- Ahmad, B., Muneer, M., Jillani, S. A., & Ahmed, S. (2026, January). Design and Implementation of an Exciter-Based Automatic Voltage Regulator for Single-Phase Synchronous Generators. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.
- Zhang, J., Zhong, S., Pan, Y., Yu, D., & Luo, J. (2026). Data-Driven Optimization of Regenerant Formulations for Efficient Regeneration of PFAS-Laden Ion-Exchange Resins. *Environmental Science & Technology*, 60(32), 22603-22613.
- Amankwaa, S., Prokofiev, J., Belikov, J., Srivastava, V., Ivask, A., Nefedova, A., ... & Polivtseva, S. S. (2026). Smart water, smart models: algorithmic assessment of water quality under evolving chemical and industrial stressors. *Environmental Science: Water Research & Technology*.
- Yao, L., Shao, T., Zhang, M., Li, T., Xu, R., & Liao, Y. (2026). Machine learning-assisted modeling and analysis of PFAS removal from contaminated water via membrane-based treatment. *Journal of Hazardous Materials*, 141033.
- Ahmad, B., Ali, S., Muneer, M., Fatima, A., & Abbas, N. (2025). Wind energy integration into the SAARC region: A comprehensive review of optimal wind sites for a sustainable super smart grid. *Wind Energy*, 3(2).
- Dasetty, S., Topel, M., Tang, Y. O., Wang, Y., Jonas, E., Darling, S. B., ... & Ferguson, A. L. (2023). Data-driven discovery of linear molecular probes with optimal selective affinity for PFAS in water. *Journal of Chemical & Engineering Data*, 68(12), 3148-3161.
- Abakar, R., Isah, M., Shafi, Q. I., Salhi, B., Khan, N. A., & Abdulazeez, I. (2026). Data-driven rational design of functionalized COFs for PFAS remediation via DFT and machine learning integration. *Microporous and Mesoporous Materials*, 114213.
- Dupuis, R., & Boek, E. S. (2026). Exploring the Potential of Machine Learning to Investigate PFAS Diffusion in Water and Its Adsorption on Graphene. *ACS ES&T Water*, 6(6), 3614-3622.
- Kazi, M. K., Varghese, S., Sarker, N., Aich, N., & Gadhamshetty, V. (2025). Advancing PFAS remediation through physics-based modeling of 2D materials: Recent progress, challenges, and opportunities. *Industrial & Engineering Chemistry Research*, 64(4), 1894-1906.

- Wang, X. (2026). Computational Investigations of Surfactants at Interfaces: Molecular Modeling and Machine Learning-Driven Discovery.
- Ahmad, B., Majeed, M. K., Soomro, A. A., & Jillani, S. A. (2026, January). Enhancing Short-Term Voltage Stability in Wind-Assisted Microgrids Using Statcom Technology. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.
- Jrad, A., & Trabolsi, A. (2026). Artificial Intelligence-Guided Discovery of Covalent Organic Frameworks for Next-Generation Polyfluoroalkyl Substances Removal. *ACS Applied Materials & Interfaces*, 18(26), 36353.
- Ye, C., Liu, H., Liu, J., Wang, H., Ge, R., Chen, X., ... & Zhu, J. (2026). Machine learning-driven quantitative modeling and mechanistic interpretation of per and poly-fluoroalkyl substances (PFAS) toxicity. *Journal of Environmental Chemical Engineering*, 14(1), 121063.
- Kollipara, S. (2026). Metagenomics and Machine Learning for Per-and Polyfluoroalkyl Substances (PFAS): A Review of Convergent Approaches, 2020-2025.
- Liang, W., Li, J., Wang, X., Giesy, J. P., & Zhao, X. (2025). Machine learning-driven cross-species toxicity prediction for advancing ecologically relevant PFAS water quality criteria. *Environmental Science & Technology*, 59(48), 25688-25702.
- Su, A., Zhai, Y., Zhang, C., Duan, H., & Rajan, K. (2026). Unveiling the Structural Determinants of PFAS Toxicity: A Graph Neural Network and Interpretability Analysis. *Journal of Chemical Information and Modeling*, 66(15), 8820-8832.
- Adu, O., Bryant, M. T., Ma, X., & Sharma, V. K. (2024). A machine learning approach for predicting plant uptake and translocation of per-and polyfluoroalkyl substances (PFAS) from hydroponics. *ACS ES&T Engineering*, 4(8), 1884-1890.
- Alsadik, A., Akintunde, O. O., Habibi, H. R., & Achari, G. (2025). PFAS in water environments: recent progress and challenges in monitoring, toxicity, treatment technologies, and post-treatment toxicity. *Environmental Systems Research*, 14(1), 18.
- Ahmad, B., Jillani, S. A., Rafique, M., & Soomro, A. A. (2026, January). Design and Monitoring of a Tree-Inspired Vertical Axis Wind Turbine for Efficient Urban Wind Utilization. In 2026 1st International Conference on Innovations in Information and Communication Technologies (IICT) (pp. 1-6). IEEE.
- Wang, X., Li, Y., Chen, P., Shen, Y., Tarabara, V., & Gao, F. (2026). Machine Learning-Guided Prediction of the Removal Efficiency of Pharmaceuticals and Personal Care Products by Photocatalytic Membranes. *ACS ES&T Water*, 6(6), 3875-3889.
- Chaudhary, M. L., Patel, R., Chaudhary, K. J., & Gupta, R. K. (2025). Metal-organic frameworks for PFAS remediation and sensing: from molecular design to real-world implementation. *Industrial & Engineering Chemistry Research*, 64(27), 13536-13556.
- Mishra, R. K., Agarwal, R., & Mishra, D. (2026). Integrated AI-Driven Assessment and Sustainable Remediation of Contaminated Soils: Innovations, Challenges and Future Directions.
- Srivastava, S., & Srivastava, R. K. (2026). Toward Autonomous PFAS Forensics: AI-Enhanced HRMS Workflows for Sustainable Chemical and Water System Management. *ACS Sustainable Resource Management*, 3(5), 1301-1323.
- Alnaimat, S., Mohsen, O., & Elnakar, H. (2024). Perfluorooctanoic Acids (PFOA) removal using electrochemical oxidation: A machine learning approach. *Journal of Environmental Management*, 370, 122857.

- Rahman, M., Ishtiaq, W., Alabrah, A., Mehmood, A., Ahmed, R. F., Khalid, I., & Amin, F. (2026). Ensemble Machine Learning Framework for PFAS Risk Screening in Public Water Systems. *CMES-Computer Modeling in Engineering and Sciences*, 147(2).
- Wu, J., Gao, W., & Wang, Z. (2025). Machine Learning Study of PFAS Intermediate Adsorption on Transition Metals: Scaling Relationships for Environmental Catalyst Design. *Environmental Science & Technology*, 59(42), 22638-22650.
- De la Vega-Camarillo, E., Arreola-Vargas, J., Antony-Babu, S., Mathur, S. K., Santos, J. A., & Shim, W. B. (2026). Machine learning-guided synthetic microbial communities enable functional and sustainable degradation of persistent environmental pollutants. *Environmental Science & Technology*, 60(18), 13500.
- Hong, J. K., Oh, S., Lee, T. K., & Park, S. (2025). Matrix-specific PFAS source allocation machine learning-based models: Identifying differential indicators in soil and water systems. *Environmental Research*, 285, 122337.
- Jagadisan, A., Boukhalfa, H., & Mehana, M. (2025). Predicting pfas diffusion coefficients with active learning and molecular dynamics. *Environmental Science & Technology*, 59(46), 24997-25009.
- Achar, S. K., Donapati, S., Piazza, M., Sui, X., Wang, Y., Darling, S. B., ... & Ferguson, A. L. (2026). Computational Design of Functionalized Cyclodextrin Probes for Sensitive and Selective PFAS Detection.
- Nasrollahpour, S., Tanhadoust, A., Pulicharla, R., & Brar, S. K. (2024). Long-chain perfluoroalkyl carboxylic acids removal by biochar: Experimental study and uncertainty based data-driven predictive model. *IScience*, 27(11).
- Cameron, D., Downes, T., Itzhaki, S., Novick, D., Patanjali, J., Kambanis, J., ... & Rajarathnam, G. P. (2025). Communities at Risk: PFAS Contamination Near Industrial Sites and Separation Strategies. *The Sydney Journal of Interdisciplinary Engineering*, 1(5), 49-58.
- Stanic, B., & Andric, N. (2025). Mapping the Rise in Machine Learning in Environmental Chemical Research: A Bibliometric Analysis. *Toxics*, 13(10), 817.
- Wada, O. Z., McKay, G., Al-Ansari, T., & Mahmoud, K. A. (2026). AI-driven biochar engineering for emerging pollutants removal from water: performance, mechanisms, and environmental perspectives. *Biochar*, 8(1), 61.
- Long, H. C., Warriar, R., Ganta, K., Mulhern, R., Lillys, T., Hoponick Redmon, J., & Gibson, J. M. (2026). Predicting PFAS Exposure Risks from Rural Private Wells Using Integrated Mechanistic and Machine-Learned Bayesian Network Models. *Environmental Science & Technology*.
- Wang, H., Chen, J., Liu, W., Wang, D., Song, Y., Hong, H., ... & Zimmerman, J. B. (2026). Using machine learning for green substitution of industrial chemicals: Integrating functionality, hazard, and life cycle impact. *Chemical Reviews*, 126(2), 841-894.
- Rabbi, M. F. (2025). Unified artificial intelligence framework for modeling pollution dynamics and sustainable remediation in environmental chemistry. *Scientific Reports*, 15(1), 36196.
- Guo, W., Tong, H., Luo, D., Zhao, X., Long, Q., Yin, C., & Yong, Y. (2025). Adsorption-based removal of PFASs from water: mechanisms, materials and future perspective. *Environmental Chemistry*, 22(6), EN25053.

- Park, S. (2025). *Integrated experimental and computational investigation of colloidal activated carbon dynamics in groundwater remediation* (Doctoral dissertation).
- Saeidi, N., Vicente, D. J., Chaudhuri, S., & Georgi, A. (2026). Modeling adsorption of mobile organic compounds on activated carbon and biochar using machine learning. *Journal of Environmental Chemical Engineering*, 123911.
- Tomsovic, Y., Gu, S., Doudrick, K., & Straub, A. P. (2026). Mechanistic Insights into PFAS Rejection in Nanofiltration and Reverse Osmosis from Data-Driven Analysis. *Environmental Science & Technology*, 60(18), 13711-13723.
- Liu, P., Pan, D., Jiao, X. Y., Liu, J. N., Du, P. H., Li, P. C., ... & Li, L. P. (2025). Integrating non-target analysis and machine learning: a framework for contaminant source identification. *npj Clean Water*, 8(1), 78.
- Dangayach, R., Jeong, N., Demirel, E., Uzal, N., Fung, V., & Chen, Y. (2025). Machine learning-aided inverse design and discovery of novel polymeric materials for membrane separation. *Environmental Science & Technology*, 59(2), 993-1012.
- Lu, Y., Ding, F., Wang, G., Li, Y., Guo, Z., Pang, P., ... & Liu, J. (2025). Machine learning-based prediction and mechanistic insight into PFAS adsorption on carbon-based materials. *RSC advances*, 15(56), 48450-48462.

