

Shapley Additive Explanations for Microplastic Contamination Modelling in the Penang River: A First Implementation of a PCA-Guided Evidential Ensemble Approach for Microplastic Prediction in Water Systems

Abdullahi Garba Usman^{1, *}, Mohd Saiful Samsudin², and Dilber Uzun Ozsahin³

¹ Irfan Suat Gonsel Operational Research Institute, Near East University, Nicosia/TRNC, 99138, Türkiye; abdullahigusman@gmail.com

² Environmental Technology Division, School of Industrial Technology, Universiti Sains Malaysia, 11800, Penang, Malaysia; saifulsamsudin@usm.my

³ Department of Medical Diagnostic Imaging, College of Health Sciences, University of Sharjah, Sharjah, P.O. Box 27272, Postal Code 211 United Arab Emirates; dozsahin@sharjah.ac.ae

* Correspondence author: abdullahigusman@gmail.com

Abstract: Microplastic (MP) pollution has become a major environmental issue with major impacts on ecological and human health, especially in the tropical river systems that are vulnerable to urbanization and industrialization. This paper presents the first study in the field of evidential ensemble modelling combined with principal component analysis (PCA) and explainable artificial intelligence (XAI) to simulate the presence of microplastics in the River Penang. The objectives were two-fold (i) to establish the prevailing physicochemical gradients that determine MP variability with the help of PCA, (ii) to build and interpret advanced ensemble models that could provide both high predictive and mechanistic information. The three major components were found to explicate a greater percentage of the total variance (over 91%). PC1 was an illustration of thermo-ionic pollution gradients (with a negative correlation between temperature, conductivity, and MPs), PC2 a turbidity-pH axis (strongly negatively correlated with MPs), and PC3 gradients of sediment-dilution. A two-scenario modeling technique was then chosen, including single machine learning models and the ensemble models, namely Evidential Neural Network Ensemble (EVNNE), Robust Linear Ensemble (RLE), Boosted Tree Ensemble (BTE), and Artificial Neural Network Ensemble (ANNE). The single models show less predictive performance compared to other models, with GPR-M1 (RMSE = 4.34, MAE = 3.03) and ANN-M1 (RMSE = 4.76, MAE = 3.11) outperforming RT despite having high error performance. On the other hand, ensemble models increased predictive capability sufficiently. RLE and BTE reduced the RMSE to 2.75 and 1.71, and the MAE to 1.45 and 1.10 respectively, which are substantial errors considering standalone models. EVNNE demonstrates high generalization performance with minimal prediction error (RMSE = 0.078 and MAE = 0.02). The interpretability evaluation of the Explainable AI SHAP analysis values proved the ensemble models' clarity with noticeable insights to the underlying environmental mechanisms. Additionally, pH, electrical conductivity (CDC) and temperature have the strongest impact, while turbidity and dissolved oxygen (DO) have moderate but context-dependent impacts. This study introduced a methodological approach to the monitoring of microplastics through the integration of PCA as the environmental representation extraction method, ensemble modeling as a predictive robust approach, and XAI modeling as an interpretable approach.

Keywords: Microplastic; Evidential Neural Network; Pollution; Penang River; Environmental Technology



1. Introduction

Plastic pollution is one of the defining environmental challenges in the twenty-first century, where the annual global production of this waste exceeds 390 million tons, and a high proportion of this waste eventually enters natural systems (Hale et al., 2020). Microplastics (MPs, <5 mm) are one of the ubiquitous contaminants in soils, rivers, lakes, and oceans, where they persist and accumulate in the food chains. Freshwater rivers serve as conduits for land-based plastic leakage into marine systems, this is by having almost 80% of marine plastics originate in rivers. This is enhanced by poor waste management, excessive use of plastics, and industrial effluent in the tropical and rapidly growing urban centers such as Southeast Asia (Zeng et al., 2025). However, it poses a certain threat to the environment and human health, including MPs as carriers of the persistent organic pollutants. Such a gap limits the analysis of the risk, its management plans, and alignment with international initiatives such as the United Nations (UN), Sustainable Development Goals (SDGs 6, 12 and 14), the United States Environmental Protection Agency (EPA), and the United Nations Environmental Program on plastic pollution (Andrady, 2011).

The modeling processes of MP concentrations pose many special methodological problems, this is because their variability is determined by nonlinear interactions between physicochemical parameters, which are difficult to measure, including pH, dissolved oxygen (DO), conductivity (CDC), turbidity, and temperature (Deng et al., 2025). Conventional monitoring approaches either become expensive and spatially constrained, whereas conventional statistical approaches tend to miss nonlinear driver (Zhen et al., 2023). Another option that has become a strong alternative is the machine learning (ML) approach, where algorithms such as Gaussian Process Regression (GPR), Artificial Neural Network (ANN), and Regression Trees (RT) have shown promising ability in water quality prediction (Egbuna et al., 2025). However, these stand-alone models are not robust and are accused of being black boxes with low interpretability. To address these limitations, the use of ensemble learners, especially the Robust Linear Ensemble (RLE), Evidential Neural Network Ensemble (EVNNE), Boosted Tree Ensemble (BTE), and ANN Ensemble (ANE), ensures better prediction accuracy using the combination of hybrid prediction learners. Moreover, Explainable Artificial Intelligence (XAI), including Shapley Additive Explanations (SHAP) increases transparency by measuring the contribution of a single variable and converting ML output into mechanistically interpretable information (Kalatzis et al., 2025). Additionally, Principal Component Analysis (PCA) offers a viable pre-modeling tool to identify prevailing gradients in environmental data, which can assist in separating the thermo-ionic, turbidity-pH, and dilution-driven processes that control MPs (Kalatzis et al., 2025). Hence, to the best knowledge of the authors this is the first study in the published technical literature that demonstrates the combined integration of PCA, XAI and Evidential Neural Network Ensemble Approach in modelling Microplastic Pollution in river Penang that can be effectively and efficiently used in microplastic concentration prediction without any data leakage. Therefore, the proposed approach provides both predictive and interpretative capabilities for microplastic modelling, which can provide useful information for understanding the environmental factors associated with microplastic pollution in the studied river. Therefore, the proposed approach provides both predictive and interpretative capabilities for microplastic modelling, which can provide useful information for understanding the environmental factors associated with microplastic pollution in the studied river.

On this basis, River Penang in Malaysia has been selected as a case study due to its socio-economic and ecological significance. It is a hilly catchment river that has been under numerous man-made influences until it empties into the Malacca Strait and is therefore a typical example of a tropical river (Costache et al., 2020). This work is the first implementation of the PCA-guided evidential ensemble modeling using XAI to make predictions of MP in water system. It aimed at achieving the following specific goals: (i) conducting exploratory statistics and PCA to obtain the main physicochemical drivers of MPs, (ii) comparing standalone and ensemble ML models, (iii) interpreting the variables' contribution and threshold behaviors using XAI, and (iv) putting the findings in perspective of environmental implications and policy frameworks. This paper excels in the field by combining dimensionality reduction, predictive performance, and interpretability with implications on actual control and management of river basins to combine local objectives and production with national priorities and sustainability at the global level.

2. Materials and Methods

2.1. Experimental Method for Microplastic Determination

The determination of microplastics in the river Penang water samples was carried out according to the method reported by Sheng et al., (2021) with some modifications according to the present study (Sheng et al., 2021). The water samples were collected from the selected sampling stations at a depth of 0–20 cm using stainless-steel containers. A volume of 10 L of water was collected for each sampling point and passed through a stainless-steel mesh of 0.3 mm pore size for the separation of suspected microplastic particles. The retained materials were carefully transferred into clean Petri dishes using stainless-steel forceps. The samples were then observed under a stereomicroscope at 40× magnification, and the particles were identified based on their physical characteristics. The microplastics were classified according to shape, size and color. The identified particles were counted and the concentration was calculated as the number of microplastic particles per liter of water (particles/L).

2.2. Description of Study Area and Location

The current research was performed on a River Penang tropical river system (see Figure 1), located in Penang, Malaysia, and has an important role to play in sustaining both the ecological system and socio-economic activities. The river is formed by the hilly catchment areas of Penang Island and passes through highly populated and industrialized areas before draining into the coastal waters of the Malacca Strait. This physical shift between relatively untouched upstream areas to highly affected downstream areas qualify River Penang as an ideal case to explore the dynamics of microplastic (MP), when subjected to an interactive effect of both natural and human-generated forces. The climatic environment of the river basin can be described as humid equatorial with annual mean temperatures of 27–31 °C and a seasonal precipitation of over 2,400 mm, which is contributed by the Northeast and Southwest monsoons (Azalina et al., 2023). These climatic conditions not only control the hydrology of the river but also affect the degradation rate of plastics, in that high temperature and high rainfall events will speed up the degradation and mobilization of plastic debris. Moreover, the steep nature of the upper catchment also adds to episodic flushing of sediments, and the lower alluvial plains are also more susceptible to the accumulation of pollutants because of the lower flow velocities (Azalina et al., 2023).

The basin land use is significantly varied and includes urban centers, industrial areas, agricultural farms, and aquaculture locations. The high rate of urbanization and poor handling of waste products have led to high levels of plastic leakage into stormwater and river systems. The sources of high CDC are industrial effluents, whereas the levels of nutrients and pH are affected by domestic wastewater. Construction works and sediment transportation during the monsoons are usually linked to the turbidity peaks in the river. Such pressures, in combination with each other, provide a clear indication of the vulnerability of River Penang to the new pollutants, such as MPs that are bound to run alongside other pollutants like heavy metals and organic waste (Al-shami et al., 2010). River Penang as a study area is also justified by the fact of its regional and global significance. Incidentally, it supplies domestic, agricultural, and recreational water locally and thus becomes a direct route of MP exposure to human populations. On a regional level, the river flows into the Malacca Strait, which is among the busiest shipping routes in the world, hence serving as a channel of plastic leakage into the oceans through the land (Sultan et al., 2023). This relationship puts River Penang in the broader scope of the contribution of Southeast Asia to global oil marine pollution. Therefore, the research location is important in terms of scientific investigation as well as policy intervention (Length, 2012).



Figure 1. Map of River Penang and sampling locations

The present study was conducted using measured observations obtained from the River Penang sampling campaign. The data consist of six environmental variables, namely temperature, pH, dissolved oxygen (DO), conductivity (CDC), turbidity and microplastic concentration (MPs, Particle/L). The observations were obtained from the sampling locations identified as SP1–SP6 over three sampling occasions.

The measured microplastic concentration was used as the target variable, whereas temperature, pH, DO, CDC and turbidity were considered as the environmental input variables for the modelling process. The original measurements indicate that the microplastic concentration varied from 1.70 to 55.00 Particle/L, with an average concentration of 35.6693 Particle/L. The physicochemical observations also covered a relatively wide range, particularly for CDC and turbidity, which varied from 30.4–16,300 $\mu\text{S}/\text{cm}$ and 10.7–750 NTUs, respectively.

2.3. Standalone Modeling Techniques

Before model development, the measured data were statistically examined and the input variables were prepared for the computational modelling procedure. PCA was applied for the reduction and transformation of the input information before the machine learning stage. The resulting variables were subsequently used for the development of the individual machine learning models.

This study employs three different standalone machine learning models, namely; GPR, ANN, and RT to model the relationship between predictors and target response (Elzain et al., 2022).

The GPR is a Bayesian, non-parametric regression approach that represents the underlying input/output relationship as a distribution over functions defined by kernels function. GPR provides both predictive estimates and associated uncertainty, making it suitable for small to medium sized datasets (Van Dao et al., 2020).

ANN are data-driven models inspired by biological neural systems and consist of interconnected layers of neurons. ANNs execute a complex and nonlinear relationship through weight optimization by training algorithms such as backpropagation, feedforward etc making them effective for modeling highly nonlinear systems (Khatti et al., 2019).

Regression Trees (RT) are non-parametric models that predict continuous values by recursively partitioning the input space into homogeneous regions based on error minimization criteria. RT models are interpretable and capable of capturing nonlinear interactions, though they may be prone to overfitting when used as standalone models (Granata et al., 2017).

2.4. Ensemble Modeling Techniques

The ensemble models were developed by combining the predictions obtained from the individual machine learning models rather than relying on a single predictive model. For each ensemble model, the selected input variables were first supplied to the constituent models, and each model generated an independent prediction of microplastic concentration. The predictions from the individual models were subsequently combined to obtain the final ensemble prediction. The ensemble output was therefore generated from the collective prediction behaviour of the constituent models, with the objective of

reducing the prediction error associated with individual models and improving the generalization capability of the predictive system.

To enhance predictive robustness and generalization performance, this study employs four ensemble modeling approaches: Robust Linear Ensemble (RLE), Evidential Neural Network Ensemble (EVNNE), Boosted Tree Ensemble (BTE), and Artificial Neural Network Ensemble (ANE).

The Robust Linear Ensemble (RLE) combines predictions from multiple base learners using weighted linear aggregation. The weights are optimized to minimize overall prediction error while reducing sensitivity to noise and outliers, thereby improving model stability and reliability (Abba et al., 2019).

The Evidential Neural Network Ensemble (EVNNE) integrates multiple neural networks within an evidential framework, allowing uncertainty information from individual models to be combined using evidence theory. This approach improves predictive confidence assessment and robustness under data uncertainty (A. G. Usman et al., 2024; A. G. Usman, Jibrin, et al., 2025).

The Boosted Tree Ensemble (BTE) is an iterative ensemble technique in which multiple regression trees are trained sequentially, with each new tree focusing on correcting the errors of previous ones. This process enhances predictive accuracy and reduces bias (Elith et al., 2008).

The Artificial Neural Network Ensemble (ANE) aggregates predictions from multiple independently trained neural networks, typically using averaging or weighted fusion. ANE improves generalization performance by reducing variance and mitigating overfitting (A. G. Usman et al., 2020).

2.5. Performance objectives and validation strategy

In the current study four different objective functions were employed to determine the efficiency of both the stand-alone and ensemble techniques for the prediction of Mp using different physicochemical variables (see Equation 1 to 4). The objective functions consist of; one goodness of fits in terms of determination coefficient (R^2), and three error measuring metrics namely; mean- squared error (MSE), mean absolute error (MAE) and root mean- squared error (RMSE).

$$R^2 = 1 - \frac{\sum_{j=1}^N [(Y)_{obs,j} - (Y)_{com,j}]^2}{\sum_{j=1}^N [(Y)_{obs,j} - (\bar{Y})_{obs,j}]^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Y_{obs,i} - Y_{com,i})^2}{N}} \quad (2)$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_{obs,i} - Y_{com,i})^2 \quad (3)$$

$$MAE = \frac{1}{N} \left[\sum_{i=1}^N |Y_{obs,i} - Y_{com,i}| \right] \quad (4)$$

where N, $Y_{obs,i}$, $\bar{Y}$ and $Y_{com,i}$ are data number, observed data, average value of the observed data and computed values, respectively.

The dataset was divided into 75% training and 25% testing subsets before model development. The training dataset was used for model calibration, feature selection and optimization, whereas the testing dataset was kept completely independent for final model evaluation. A 10-fold cross-validation procedure was performed within the training dataset to assess the stability of the developed models and to reduce the possibility of overfitting during model tuning. The training data were divided into ten approximately equal subsets, where nine subsets were used for model development and the remaining subset was used for validation. The procedure was repeated ten times by changing the validation subset. Model parameters were selected based on the validation performance, after which the selected model was recalibrated using the complete training dataset. Final model performance was determined using the independent testing dataset. Model accuracy was evaluated using R^2 , RMSE, MAE, MAPE, correlation coefficient and NSE. The testing data were not used during feature selection, hyperparameter optimization or model calibration (Abba et al., 2022, 2023; Benaafi et al., 2022; Jibril et al., 2024; A. G. Usman et al., 2022, 2023; A. G. Usman, Abd, et al., 2025; J. Usman et al., 2023). The training set is used in developing the model, setting to tune hyperparameters and prevent overfitting, and the testing set to assess final model performance on unseen data. Also, to ensure robustness, k-fold cross-validation was be applied, where the dataset is partitioned into 10 equal folds. The model is trained on 10 folds and validated on the remaining fold, and the process is repeated 10 times. The average performance across all folds is then reported (Abba & Usman, 2020).

3. Results and discussions

3.1. Statistical-Based Exploratory Results

The River Penang data were analyzed descriptively to provide an overview of the input water quality variables and MPs concentrations (Table 1). The average water temperature was 29.48 °C with a reasonably small dispersion (SD = 0.87 °C), which is characteristic of a stable thermal regime of tropical rivers. The pH showed moderate change (mean = 6.63; SD = 0.73), which indicates slightly acidic to neutral conditions. DO was at an average of 3.37 mg/L, which is on the lower section of the ecological health standards, indicating that there might be organic loading or microbial activities in the river. The parameters that showed the most variability were the CDC and turbidity (TURB). The range of CDC was 30.4 to 16,300 µS/cm with a very high variance (11,752,579), indicating that the ionic compositions were heterogeneous, probably due to the point (urban/industrial effluents) and non-point (runoff) sources. The turbidity was between 10.7 to 750 (mean=110.51 NTU), which further supported the fact that the water was not always in a stable state of clarity, and that this could be due to episodic rainfall and resuspensions of the sediments. Mean concentrations of MPs were 35.67/L, and the standard deviation was 12.66. The extent was found between limits of 1.7 particles/L to 55.0 particles/L, and indicates hotspots of contamination of MP in some of the areas sampled. The values of skew and kurtosis add more distributional characteristics: The skew (-1.666) is negative, which indicates the tail that is oriented towards lower values, and the kurtosis (2.250) is moderate peak in contrast to the normal distribution.

In terms of data analytics, the large kurtosis of TURB (13.41) and CDC (4.32) means that they are distributed in the form of heavy tails, and these tails contain many extreme values, which can be very influential to the model learning process (see Table 1). Equally, pH and TURB skewness values above 1.5 will be indicative of non-normal distributions, which may need ensemble models and non-linear learners to effectively model variability. These searching fashions explain the shift of the individual models (GPR, ANN, RT) to sophisticated collections (RLE, EVNNE, BTE, ANNE) to reduce noise sensitivity and enhance predictive generalization. The descriptive findings are all based on the complex and dynamic nature of the water quality of River Penang, with the presence of physicochemical stressors, including high conductivity, turbidity, and low DO, and the high loads of MPs. These conditions highlight the importance of strong predictive models that incorporate quantification of uncertainty, including evidential ensembles, which do not just represent the mean trends but also state confidence of predictions that is necessary in environmental management and policymaking.

Table 1. Descriptive statistical summary of physicochemical parameters and microplastic concentrations in River Penang

	Temp.	pH	DO	CDC	TURB.	MPs
Mean	29.483	6.628	3.374	2584.258	110.506	35.669
Stand. Dev.	0.869	0.730	0.713	3428.204	147.724	12.656
Sample Variance	0.755	0.532	0.508	11752579.297	21822.497	160.181
Kurtosis	-0.916	3.335	-0.998	4.323	13.413	2.250
Skewness	0.126	1.619	-0.070	1.971	3.661	-1.666
Minimum	27.830	5.680	2.090	30.400	10.700	1.700
Maximum	31.200	9.130	4.710	16300.000	750.000	55.000

3.2. Dependency Feature Selection Results

Figure 2 shows the results of the dependency feature selection, demonstrating the correlation structure of physicochemical parameters and MP in River Penang. The dependency analysis reveals that there are some significant relations among physicochemical parameters and MP concentrations in River Penang. The positive (but not strong, $r = 0.452$) relationship between temperature and MPs indicated that thermal regimes affected the processes involved in polymer degradation and fracture. High water temperature could increase the degradation of larger plastics into smaller ones, and at the same time, decrease DO as the strong negative correlation between temperature and DO ($r = -0.827$) supported this claim. Conductivity was also positively correlated with MPs ($r = 0.464$), which indicates that the increased enrichment of the river due to the release of anthropogenic discharges (including domestic effluents and industrial wastewaters) may be associated with the increase in microplastic contamination. The obtained results suggest that warmer and more polluted waters provide an environment where MPs can accumulate and persist. In comparison, the turbidity has a weak negative correlation with MPs ($r = -0.268$), which means that suspended solids are also available in the system, but MPs do not necessarily have to co-migrate with the particulate matter, and might be regulated by different kinds of transport and deposition interactions.

Nevertheless, pH showed the most significant negative correlation with MPs ($r = -0.690$), which demonstrates a significant chemical control mechanism. Acidic environments can increase polymer degradation and surface weathering of plastics, resulting in more detectable fragments in water. Anthropogenic effects may also be reflected in this relationship, in which inputs that decrease pH (e.g., acidic effluents, leachates) are also the inputs that introduce microplastic contaminants. It is supported by the moderate negative correlation between DO and MPs ($r = -.221$), indicating that regions with low oxygen, which is frequently an indicator of organic pollution and stagnant flow, are also associated with an increased MP load (see Figure 2). Physically, this is an indication of a multi-stressor environment in which ionic strength, acidity, and oxygen depletion meet to amplify microplastic pollution. The fact that temperature, conductivity, and pH are highly correlated with MPs indicates that they should be modeled as important predictors, and collinearity (e.g. between temperature and DO) can be effectively taken care of with ensemble learning. These interdependencies not only offer a statistical basis on which to choose the model, but also offer an ecologically sensible explanation of how a combination of physicochemical stressors can help explain the distribution of MPs in tropical river systems.

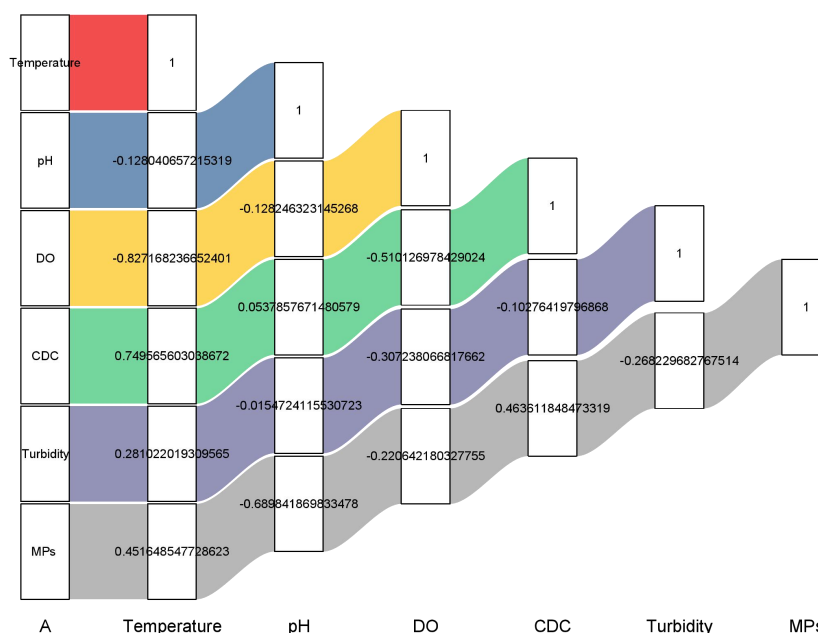


Figure 2. Heatmap of the dependency feature selection results showing the correlation structure among physicochemical parameters (Temperature, pH, DO, CDC, Turbidity) and microplastic concentrations (MPs) in River Penang.

3.3. Principal Component Analysis (PCA) Results

The PCA findings (Table 2) indicate that the first three principal components (PCs) have a cumulative variance of over 91 percent, where PC1 (45.2 percent), PC2 (27.8 percent) and PC3 (18.2 percent) represent the variances. This indicates that a lower-dimensional space with three components could be used to capture most variability of the dataset. PC1 has a good positive loading of temperature (0.579) and conductivity (0.494) and a negative loading of DO (-0.483). This implies that PC1 is largely a thermo-ionic gradient, with warmer and more conductive waters being linked to the low concentration of oxygen. Notably, MPs (loading = 0.396) also load positively on PC1, which supports the fact that MPs co-vary with thermal and chemical stressors.

Accordingly, PC1 may be seen as a complex pollution gradient with physicochemical stress and enrichment of microplastics. PC2 and PC3 also narrow down the variability structure. PC2 has high positive loading on turbidity (0.428) and pH (0.522); however, it has a strong negative contribution of MPs (-0.556). This means that less turbid and more alkaline waters are less connected to loads of microplastics, implying that MPs might not just co-migrate with suspended solids, but have different transport dynamics as a result of ionic strength and fragmentation processes. The negative loading of PC3 (0.747) indicates a sediment dilution effect with medium positive values of CDC and DO, which means that the large suspended solids darken or reallocate MPs in the water column. Collectively, these points indicate the multifactorial regulations of MPs, such as thermal and ionic mechanisms (PC1), the chemical environment (PC2), and the particle dynamics (PC3).

In the modeling framework, such findings provide the rationality of using ensemble learners, especially evidential ensembles, which have the potential of incorporating numerous nonlinear dependencies and uncertainties. Reducing the redundancy and emphasizing the structures of the dominant variance, PCA validates that conductivity, temperature, and pH are the most informative predictors of MPs. Meanwhile, the secondary but complicated role is taken by turbidity; however, it should be interpreted carefully. Figure 3 demonstrates that PC1 indicates a sign of thermo-ionic pollution gradient, where the conductivity and temperature level are higher than the DO and MPs. PC2 measures an inversely related turbidity pH axis to MPs, meaning that alkalinity and suspended solids were not the direct cause of microplastic variability. Combined, the first two components have an explanation of 73% of the total variance in the data.

Table 2. PCA of physicochemical parameters and microplastic concentrations in River Penang

Variable	PC1	PC2	PC3
CDC	0.493795	0.033191	0.267636
DO	-0.482627	-0.345123	0.203753
pH	-0.15158	0.521312	0.424928
Temperature	0.579435	0.153491	-0.330244
Turbidity	0.091457	0.42779	-0.747144
MPs	0.39516	-0.555397	-0.229447

Note: The first three principal components (PC1–PC3) explain over 91% of the total variance in the dataset. PC1 represents a thermo-ionic gradient dominated by temperature, conductivity, and MPs inversely related to DO. PC2 is characterized by turbidity and pH with a negative contribution from MPs, while PC3 reflects sediment–dilution dynamics. These components highlight the dominant environmental gradients that influence the occurrence of microplastics in the river system.

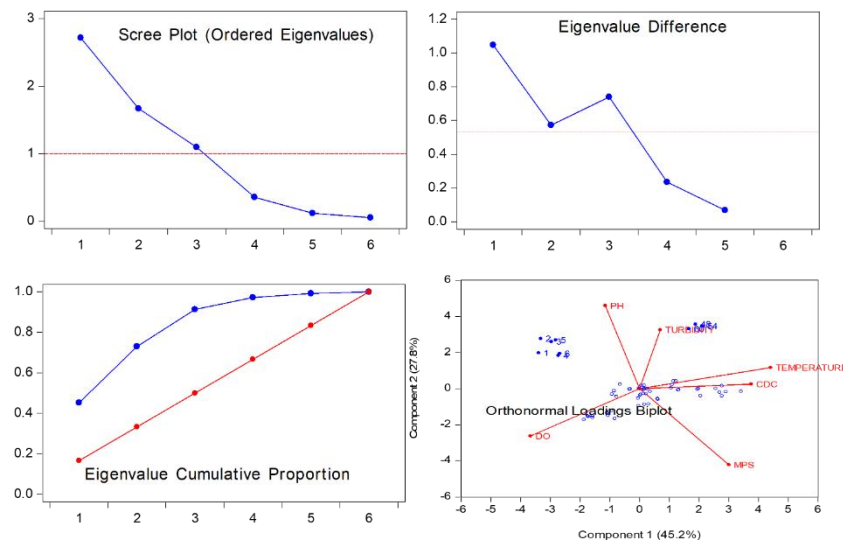


Figure 3. A biplot of the first two principal components (PC1 = 45.2%, PC2 = 27.8%)

3.4. Standalone Model Performance for Microplastic Prediction

Table 3 shows the performance of the standalone models in forecasting the MP concentrations. GPR-M1 turned out to be the strongest and most consistent in its performance, as it exhibited an outstanding performance in training ($R^2 = 0.95$, RMSE = 2.66, MAE = 2.12) and the high generalization capability in testing ($R^2 = 0.89$, RMSE = 4.34, MAE = 3.03). This shows that GPR was highly successful in recording the nonlinear relationships between MPs and water quality variables without serious overfitting. The comparatively low error measures indicate that GPR models will be suitable to deal with the complex distributions that are observed in River Penang, especially when physicochemical stressors, including conductivity, pH, and temperature, work together to modify MPs. The other variants of GPR (M2 and M3) also did fairly well, achieving R^2 values of around 0.71 to 0.75 in testing and validating the reliability of the kernel-based learning in predicting MPs.

Conversely, the Regression Tree (RT) showed unreliable and overall poor predictive power. In spite of the moderate training accuracy of RT-M2 ($R^2 = 0.73$, RMSE = 6.90), the level of testing performance decreased to $R^2 = 0.64$, and RT-M3 was almost incapable of generalizing ($R^2 = 0.11$, RMSE = 12.15).

This emphasizes that RT models are sensitive to data partitioning and that they overfit noisy models or very variable models, e.g., data caused by episodic spikes in turbidity and conductivity. ANNs had mixed results. ANN-M1 was equally competitive with GPR at the expense of having high training accuracy ($R^2 = 0.96$, RMSE = 2.58) and low testing accuracy ($R^2 = 0.87$, RMSE = 4.76). ANN-M2 and ANN-M3 were however susceptible to overfitting and underperforming (testing $R^2 = 0.53$ and 0.44 , respectively), perhaps because of poor architecture or convergence (see Figure 4).

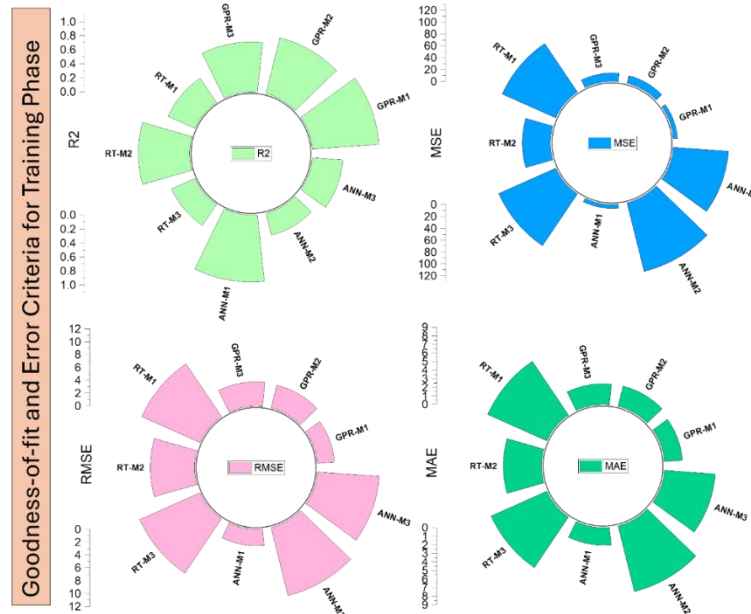


Figure 4. Radial plots of standalone machine learning models for microplastic concentration prediction in training phase

Table 3. Performance evaluation of standalone machine learning models for microplastic concentration prediction in River Penang

	Training Phase			
	R^2	MSE	RMSE	MAE
GPR-M1	0.95	7.10	2.66	2.12
GPR-M2	0.82	13.47	3.67	2.44
GPR-M3	0.71	14.37	3.79	2.44
RT-M1	0.42	99.09	9.95	7.59
RT-M2	0.73	47.62	6.90	4.49
RT-M3	0.36	105.65	10.28	7.15
ANN-M1	0.96	6.63	2.58	2.06
ANN-M2	0.33	119.83	10.95	7.92
ANN-M3	0.44	93.37	9.66	5.98
	Testing Phase			
	R^2	MSE	RMSE	MAE
GPR-M1	0.89	18.87	4.34	3.03
GPR-M2	0.71	23.39	4.84	3.57
GPR-M3	0.75	24.10	4.91	3.52
RT-M1	0.71	45.21	6.72	5.33
RT-M2	0.64	43.27	6.58	4.65
RT-M3	0.11	147.51	12.15	9.19
ANN-M1	0.87	22.64	4.76	3.11
ANN-M2	0.53	56.76	7.53	5.45
ANN-M3	0.44	93.37	9.66	5.98

Moreover, the results of the standalone model indicate that both the GPR and ANN can be used to make reasonable predictions, but are more stable under different model specifications, with GPR normally displaying more consistent results during training and testing groups. Regression Trees turned out to be underperformers; however, they highlight their shortcomings when dealing with high-variance multicollinear data. These results explain why ensemble methods (RLE, EVNNE, BTE, ANNE) are required to address the drawbacks associated with each particular model through combining

complementary strengths, and thus enhance the robustness and quantification of uncertainty in MP prediction. In the testing stage, the patterns of comparative error of standalone models indicate that large discrepancies exist in their predictive reliability. GPR-M1 had the smallest error rates, RMSE = 4.34 and MAE = 3.03, which is a relative deviation of approximately 12-15% out of the measured MP mean (35.67 particles/L). This proves that GPR was able to record the underlying trends with a little bias and variation. GPR-M2 and GPR-M3 were slightly lower in error (RMSE 4.84.9; MAE 3.5), and with a skewness of about 14-17% this is an acceptable error when dealing with environmental data, where there is high noise in such data. The values underscore the strength of the GPR models, particularly in the presence of nonlinear interaction of parameters, and in the case of uneven distributions of MPs (Figure 5).

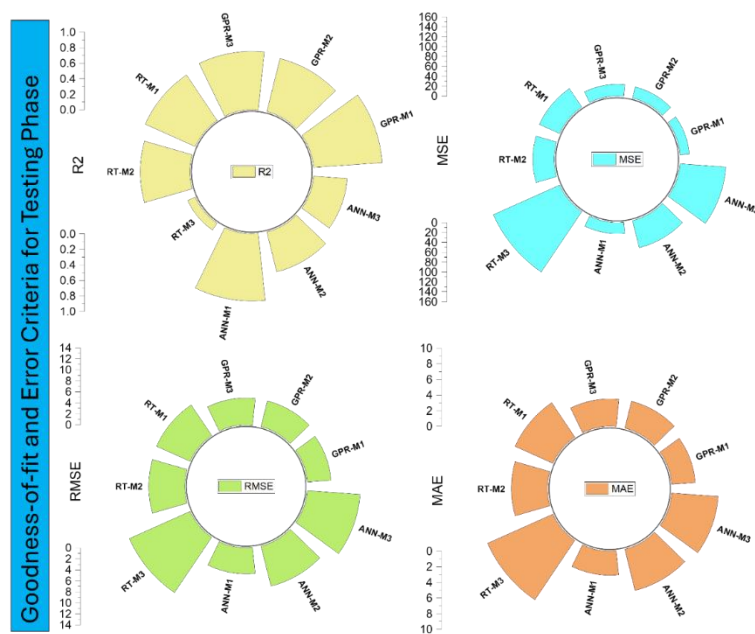


Figure 5. Radial plots of standalone machine learning models for microplastic concentration prediction in testing phase

The ANN models had a more diverse error profile. ANN-M1 performed well in generalization with RMSE = 4.76 and MAE = 3.11, which is equivalent to the error of 13-16% but similar to GPR. Nevertheless, both ANN-M2 and ANN-M3 had much bigger errors (RMSE = 7.53 and 9.66; MAE = 5.45 and 5.98), equivalent to 21-27%. This decrease implies that the network training is not stable and it could be overfit due to training noise, which decreases its predictive capacity to conditions that are not observed. In comparison, Regression Trees did the worst: RT-M1 and RT-M2 had an average RMSE of 6.6, 6.7 (18-20% error), RT-M3 was extremely inaccurate (RMSE = 12.15; MAE = 9.19), which is a >30% error. This again confirms the vulnerability of tree-based learners to variance inflation in heterogeneous data. All percentages of error analysis show that GPR-M1 and ANN-M1 are the best independent predictors, and RT and poorly tuned ANN variants have difficulties with generalization when modeling MPs.

3.5. Ensemble Model Performance for Microplastic Prediction

The ensemble models perform better than the independent methods, showing the benefits of model aggregation in the treatment of the nonlinear and complicated interactions that define the dynamics of MP in River Penang. The EVNNE was found to be the most successful of all investigated models, with a training R^2 of 0.999 and testing R^2 of 0.999. The error values associated with it were also very low (testing RMSE = 0.078; MAE = 0.02), which represents less than 0.5 per cent relative error against the mean MP concentration (Table 4). Such exceptional accuracy implies not only the predictive power of the ensemble architecture but also the ability of the evidential methods to measure the uncertainty and give strong confidence estimates, which are a significant attribute in environmental decision-making. The stability of the performance of the training and testing phases also shows that EVNNE is not prone to overfitting and hence a solid system in predicting the MP (Figure 6). Additionally, other ensemble methods performed well compared to individual models. The BTE was shown to be highly accurate (training R^2 = 0.982, testing R^2 = 0.979), with testing RMSE = 1.71 and MAE = 1.10, or an error of about 5%. This highlights the power of the ensemble in correcting weak tree learners and be able to capture nonlinearity in high-variance datasets. Likewise, the RLE preserved testing results of R^2 = 0.956 and

RMSE = 2.75 and MAE = 1.45 (8 % error) between simplicity and stability. The Artificial Neural Network Ensemble (ANNE), in comparison, was slightly weaker than the other ensembles, the testing R^2 of the ANNE being 0.892, and the RMSE of 3.86 (11% error), but again much better than the weaker standalone ANN and RT versions. The results reveal that ensembles, especially EVNNE and BTE, have a better generalization power, high accuracy, and noise resistance, making them the most appropriate instruments to be used in the modeling of MPs in complex river systems.

Table 4. Predictive results between the observed and computed

Training Phase				
	R ²	MSE	RMSE	MAE
RLE	0.968	5.543	2.353	1.364
EVNNE	0.999	0.231	0.053	0.01
BTE	0.982	3.163	1.779	0.537
ANNE	0.953	8.214	2.866	1.186
Testing Phase				
RLE	0.957	7.574	2.752	1.448
EVNNE	0.999	0.312	0.078	0.02
BTE	0.979	2.931	1.712	1.097
ANNE	0.892	14.866	3.856	2.598

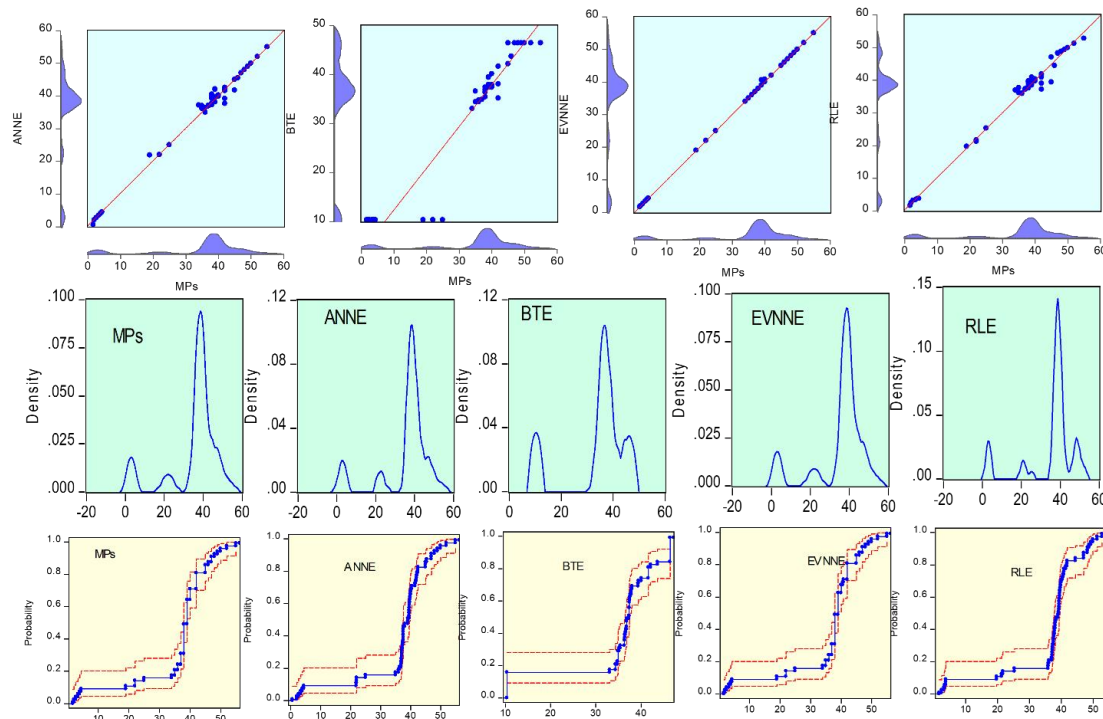


Figure 6. Ensemble model performance diagnostics for microplastic prediction in River Penang

3.6. Comparison of Standalone and Ensemble Models

When direct comparison is made between standalone and ensemble methods, it is noticeable that the performance of the model integration is substantially improved. The optimal standalone models (GPR-M1 and ANN-M1) had testing R^2 of 0.89 and 0.87, with relative errors of RMSE of 4.34 and 4.76, respectively. Conversely, the ensembles greatly minimized errors; RLE and BTE gave R^2 above 0.95 and RMSE of between 1.71 and 2.75 (5-8% error), whereas EVNNE gave predictions with a near-perfect fit with $R^2 = 0.999$ and RMSE = 0.078 (0.1 error). ANNE still improved most individual settings, indicating $R^2 = 0.892$ and RMSE=3.86, despite being the weakest ensemble. These gains show that ensembles apply the strengths of individual learners in counterbalancing their weaknesses through a combination of their complementary strengths. Standalone GPR and ANN would capture nonlinear dependencies, but were unable to be sufficiently stable in different scenarios. Ensembles, especially EVNNE and BTE, were more robust and generalized, and better fitted to more complicated noisy environmental data. This apparent performance disparity explains the significance of ensemble learning to further develop a

reliable microplastic concentration model and supports the application of XAI-based ensemble methods in environmental technology.

Table 5 shows the performance of standalone machine learning (GPR, ANN, RT) and ensemble machine learning (RLE, EVNNE, BTE, ANNE) when predicting MP at the testing stage. Of the independent methods, the GPR-M1 had the most satisfactory performance ($R^2 = 0.89$, RMSE = 4.34, MAE = 3.03), followed by ANN-M1, with an R^2 of 0.87. The regression tree (RT-M1) was performing significantly behind the optimal range ($R^2 = 0.71$, RMSE = 6.72), which shows that the model is not very effective in capturing the nonlinear variability of MPs. In general, standalone models had moderate accuracy, with errors within 3-5 MPs/liter, which can still be improved in practice. The ensemble approaches provided large accuracy improvements. The RLE enhanced the accuracy of the testing to $R^2 = 0.956$, with 87.5% lower RMSE and 52.1% lower MAE when compared to GPR-M1. BTE was even better with $R^2 = 0.979$ and 60.6% (RMSE) and 63.7% (MAE) errors reduction, which is an affirmation of boosting in correcting the shortcomings of the single trees. The best outcome was obtained with the EVNNE, which had a perfect predictive power ($R^2 = 0.999$) and a negligible error (RMSE = 0.078, MAE = 0.02), which is 98 to 99% better than the highest standalone model. The weaker ANNE also performed better than the standalone ANN and RT, with slight error differences of 11-14%. Combined, the findings are a clear indication that ensemble learning is much better at predicting MPs than individual models, and EVNNE can optimally generalize, while BTE/RLE can provide strong and steady alternatives. This underscores the importance of ensemble-based methods in environmental modeling (especially with nonlinear and noisy data), including microplastic concentrations in river systems.

Table 5. Comparison of standalone and ensemble models for MP prediction (testing phase)

Model	R^2	RMSE	MAE	RMSE Improvement (%)	MAE Improvement (%)
Standalone Models					
GPR-M1	0.89	4.34	3.03	—	—
ANN-M1	0.87	4.76	3.11	—	—
RT-M1	0.71	6.72	5.32	—	—
Ensemble Models					
RLE	0.956	2.75	1.45	36.6%	52.1%
EVNNE	0.999	0.078	0.02	98.2%	99.3%
BTE	0.979	1.71	1.10	60.6%	63.7%
ANNE	0.892	3.86	2.60	11.1%	14.2%

3.7. Explainable AI (XAI) Results

To enhance transparency in ensemble predictions, an explainable machine learning technique was applied using two complementary frameworks (SHAP and AIX). Both methods are based on game-theoretic Shapley values that measure the marginal contribution of each aspect to model predictions. The use of such methods ensured the robustness and interpretability of our results on independent platforms. The global (Python) SHAP analysis showed that pH, turbidity, and DO have the greatest impact, closely followed by CDC and temperature. In particular, the negative influence of pH was high, and greater alkalinity decreased the predictions of MP, which is in line with the field evidence that showed greater release of microplastics under acidic environments. CDC and temperature also had a positive effect, indicating ionic enrichment and thermally induced polymer degradation. Turbidity expressed a subtle effect, with the highest values of turbidity diluting MPs in the water column. DO showed low but negative impacts, which is in line with the fact that hypoxic waters are associated with organic and pollutant-rich environments. Simultaneously, MATLAB AIX offered local Shapley descriptions at the query-point level. As an illustration at the measurement point, Temperature = 28.07 °C, pH = 9.13, DO = 4.04 mg/L, CDC = 43.33 uS/cm, and Turbidity = 70.93 NTU, the model provided the highest negative value to pH (-30) Shapley value (see Figure 7). This implies that alkalinity was high, which significantly lowered the estimated MPs in this area. The contribution of CDC and turbidity was positive, although with minor magnitudes, and temperature and DO have no effect in this particular case. This localized understanding is in line with the global SHAP results, which verify that pH is the major suppressor of MPs, whereas CDC and turbidity are the secondary enhancers. This study has shown that the convergence of tools was improved by combining both global (Python SHAP) and local (MATLAB AIX) interpretability frameworks. The two validations of the environmental interpretation enhance the trust in the readings: MPs in River Penang are regulated by a combination of three chemical (pH, CDC), thermal (temperature), and hydrodynamic (turbidity) indicators, which are controlled by the dynamics of oxygen. This multi-level explainability guarantees not only predictive but also ecological interpretability, which renders the ensemble models reliable in using them to make decisions in environmental monitoring and management.

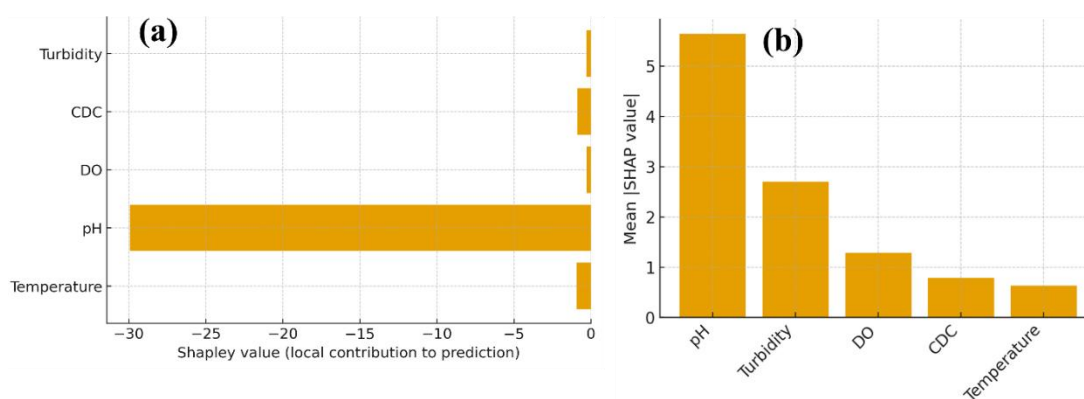


Figure 7. Explainable AI results for MP prediction using SHAP. (a) Local Shapley explanation for a single query point. (b) Global SHAP feature importance indicates that pH is the dominant predictor, followed by turbidity, DO, CDC, and temperature.

3.8. Environmental Implications

This research study has significant implications in terms of comprehending MP pollution in tropical river systems and especially in highly urbanized River Penang catchments. Physically, the explainable AI analysis showed that pH, conductivity, temperature, and turbidity are dominant factors that cause MP concentration. Through these associations, the interrelationships between hydrological, chemical, and thermal processes hold importance in the regulation of the transport of pollutants. An example is that pH and MPs have a strong negative correlation, indicating that acidic water hastens the degradation of polymers and breaks down plastics into smaller particles, which are easily transported in the river. High conductivity indicates increases in urban runoffs, industrial effluents, as well as sewage discharges, which not only indicate anthropogenic pressures but also MPs co-transport. Similarly, the effect of temperature highlights the fact that plastic degradation is sensitive to tropical weather conditions, in which an increase in thermal stress accelerates the rate of degradation. Turbidity, which was also presumed to have a positive correlation with MPs, showed a complicated relationship with the extreme loads of sediment diluting the MPs, most probably by covering or burying them in the sediment. Taken together, these observations highlight those interactions between multiple stressors that control MPs in rivers and are not a one-way pollutant pathway. From an ecological and human health perspective, the existence of MPs on the River Penang is of great concern. Microplastics may serve as carriers of persistent organic pollutants (POPs), heavy metals, and pathogenic microorganisms, and increase risks to aquatic biota and communities that depend on riverine water. The prediction was made even when the turbidity was reasonably moderate, and this means that communities that irrigate with river water, fish, or informally use river water in domestic activities could be chronically exposed to MPs. This is in line with international evidence that MPs are capable of bioaccumulation in aquatic life and with possible entry into the human food chain via fish, shellfish, and crops water irrigated with contaminated water. These threats make it apparent that the river basins with the active process of urbanization and industrial development require constant monitoring and regulatory measures.

These findings show the reflection of global and national policies that are focused on the containment of the emerging pollutants at the policy level. The United Nations Sustainable Development Goals (SDGs) are directly considered: SDG 6 (Clean Water and Sanitation) through the monitoring of water quality improvement, SDG 12 (Responsible Consumption and Production) through the strategies of minimizing the leaching of plastics, and SDG 14 (Life Below Water) through the reduction of the land-based microplastic inputs into coastal and marine systems. At the national level, the U.S. Environmental Protection Agency (EPA) and the European Chemicals Agency (ECHA), among others, have defined microplastics as emergent contaminants requiring standard monitoring protocols and predictive modeling software. The World Health Organization (WHO) has also raised concerns about MPs in drinking water as a risk to human health, but the level at which regulation is needed is still in scientific controversy. Our paper offers evidence-based modeling solutions that can be used as an addition to these initiatives, forecasting the conditions of hotspots and influencing risk evaluation.

Rivers are considered lifelines of urban population, agriculture and industry in the Malaysian and Southeast Asian setting. River Penang is not an exception, as it sustains ecological habitats, as well as social-economic livelihoods. The present study adds to the evidence-based management of river basins by proving that AI-enhanced ensemble models could be reliable predictors of MPs and offered interpretable drivers of pollution. Such predictive frameworks when incorporated in the national water

quality programs under the Department of Environment Malaysia and in accordance with the regional plastic pollution strategies developed by the ASEAN would enhance adaptive management. Moreover, explainable AI will guarantee transparency and the stakeholders and policymakers may believe the predictions and act on them, whether it is plastic waste minimization campaigns, upgrades of effluent treatment, or catchment-based interventions. Finally, the environmental consequences of this work do not only apply to the River Penang case study. With the explanation of the AI combined with ensemble modeling, we offer a versatile structure that could be extended to other rivers in Southeast Asia, Africa, and the rest of the world, where pollution pressures are increasing, but the monitoring infrastructure is possibly insufficient. These frameworks allow the proactive decision-making process to be in place, to have the local management activities aligned with the international sustainability agendas, and to have the gap between the scientific evidence and the environmental governance narrowed.

As shown in Fig. 8, the total microplastic abundance varied considerably among the sampling stations and sampling periods. At SP ST1, the abundance increased from 38 particles during SWM (10 June 2023) to 59, 81 and 81 particles during the subsequent sampling periods. SP ST3 increased from 54 to 64, 86 and 86 particles, while SP ST4 increased from 82 to 90, 92 and 92 particles. Similarly, SP ST5 increased from 82 particles during SWM to 72, 94 and 94 particles, whereas SP ST2 increased from 61 to 71, 79 and 79 particles. In contrast, SP ST6 decreased from 63 particles to 59 and subsequently 57 particles. The observed differences indicate that microplastic abundance was not uniform across the sampling stations and sampling periods. The increase observed at most stations after the June 2023 sampling may be related to temporal changes in monsoon-associated hydrological conditions. Rainfall and surface runoff can facilitate the mobilization of plastic debris from surrounding areas into river channels, while changes in flow conditions can influence the transport, deposition and resuspension of microplastic particles. Temperature can also influence plastic weathering and fragmentation, although such an effect cannot be established directly from the present results.

Importantly, the figure provides microplastic abundance according to sampling time and station, but does not contain corresponding rainfall, river discharge or detailed temperature records for each sampling event. Therefore, the observed pattern is interpreted as temporal and spatial variation occurring under the sampled monsoon conditions, rather than as direct evidence that rainfall, temperature or climate change caused the observed changes.

The findings remain relevant to climate change because changes in rainfall intensity, runoff and hydrological variability may alter the transport pathways and retention of microplastics in river systems. The spatial differences among SP ST1–SP ST6 also indicate that the response of microplastic abundance to changing environmental conditions may differ between sampling locations. Further long-term monitoring incorporating rainfall, temperature, river discharge and other hydrological parameters would be required to establish the specific contribution of climate-related factors to microplastic transport and distribution.

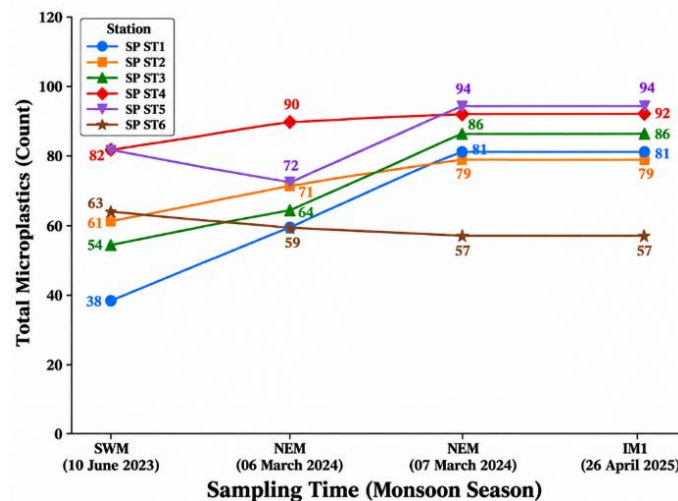


Figure 8. Temporal variation of the microplastics at each station

4. Conclusion

The paper is the first study to use evidential ensemble learning with XAI to predict the concentration of MP in the River Penang. The research showed the strength of more sophisticated models of data and ensemble models in modelling the complex, nonlinear dynamics of MP pollution in tropical rivers. The

standalone models signify moderate prediction, where GPR-M1 and ANN-M1 were found to be with higher performances. Nevertheless, the ensemble models had a significant improvement over all the individual learners. Specifically, the EVNNE had a relatively perfect predictive power ($R^2 = 0.999$, RMSE = 0.078, MAE = 0.02) in comparison to optimal standalone models. The explanation of the AI analysis revealed that the predictive framework presented clear information about the driving forces of the environment on MPs. Shapley explanations had consistent findings to pH, conductivity, and temperature as the most significant predictors, and turbidity and dissolved oxygen were considered to be context-specific. These findings are well-grounded physically, since acidic waters promote the degradation of polymers, high conductivity is indicative of pollution caused by effluents, and tropical temperatures promote the weathering of plastics. Context-specific interpretation was necessitated by the subtle contribution of turbidity, and extreme sediment loads used to dilute MPs. Combined, the results not only contribute to predictive science but also enhance ecological knowledge of the interaction of chemical and hydrodynamic factors to control the dynamics of MPs. The implications of these findings on the environment are far-reaching. Multi-stressor interactions of water chemistry, hydrology, and anthropogenic discharges influence MPs in River Penang. The study aims to indicate the urgency of incorporating predictive models in water quality management systems by aligning it with the Sustainable Development Goals (SDGs), especially SDG 6 (Clean Water and Sanitation), SDG 12 (Responsible Consumption and Production), and SDG 14 (Life Below Water). Moreover, the proven interpretability of the ensemble models guarantees that the predictions can be acted upon and will be reliable to support the country-level policies and the global policies of international organizations like the EPA, UNEP, and WHO. In summary, this study defined a new methodological standard in the field of microplastic monitoring and prediction of the riverine systems. Combining evidential ensembles with XAI makes it possible to achieve extraordinary predictive performance and easy interpretability in the transition between data science and effective environmental governance. Although the present study was restricted to River Penang, the model is portable and can be extended to other rivers that have the same issue of plastic pollution. The directions of future research should include increasing datasets in terms of time and space, remote sensing and in-situ monitoring integration, and the connection of ensemble forecasting and hydrodynamic transport. These innovations will not just perfect the predictive accuracy but will also add to an international approach to the reduction of plastic pollution and the protection of freshwater sources.

Author Contributions

Conceptualization, AGU. and MSS.; methodology, AGU. and MSS.; software, AGU.; validation, AGU.; formal analysis, AGU., DUO and MSS.; investigation, AGU., DUO and MSS.; resources, AGU., and MSS.; data curation AGU., and MSS.; writing— AGU., DUO and MSS., writing—review and editing, AGU., DUO and MSS.; visualization, AGU; supervision, DUO and MSS.; project administration, AGU. and MSS; funding acquisition, AGU. and MSS All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by both Near East University, TRNC and Universiti Sains Malaysia Grant Scheme (Grant Number: R502-KR-RUT002-0000001110-K134).

Conflict of Interest Statement

None to be declared

Data Availability Statement

Data will be provided on reasonable request by the corresponding author.

Acknowledgment

This work was supported by the Universiti Sains Malaysia, Research University Transdisciplinary (RUTrans) Grant Scheme (Grant Number: R502-KR-RUT002-0000001110-K134).

References

- Abba, S. I., Benaafi, M., Usman, A. G., & Aljundi, I. H. (2022). Sandstone groundwater salinization modelling using physicochemical variables in Southern Saudi Arabia: Application of novel data intelligent algorithms. *Ain Shams Engineering Journal*, xxx, 101894. <https://doi.org/10.1016/j.asej.2022.101894>
- Abba, S. I., Benaafi, M., Usman, A. G., Ozsahin, D. U., Tawabini, B., & Aljundi, I. H. (2023). Mapping of groundwater salinization and modelling using meta-heuristic algorithms for the coastal aquifer of eastern Saudi Arabia. *Science of The Total Environment*, 858(July 2022), 159697. <https://doi.org/10.1016/j.scitotenv.2022.159697>
- Abba, S. I., Nourani, V., & Elkiran, G. (2019). Multi-parametric modeling of water treatment plant using AI-based

- non-linear ensemble. *Journal of Water Supply: Research and Technology - AQUA*, 68(7), 547–561. <https://doi.org/10.2166/aqua.2019.078>
- Abba, S. I., & Usman, A. G. (2020). *Modeling of Water Treatment Plant Performance using Artificial Neural Network: Case Study Tamburawa Kano-Nigeria*. September.
- Al-shami, S. A., Hassanahmad, A., Azizah, S., & Nor, M. (2010). *je s c . je s c .* 22(11), 1718–1727.
- Andrady, A. L. (2011). Microplastics in the marine environment. *Marine Pollution Bulletin*, 62(8), 1596–1605. <https://doi.org/10.1016/j.marpolbul.2011.05.030>
- Azalina, N., Feisal, S., Haziqah, N., Firdaus, M., Sani, A., Khairulnisa, D., Ahmad, A., Azwan, M., Faiza, N., Razak, A., Nilam, T., & Tengku, B. (2023). Anthropogenic disturbance of aquatic biodiversity and water quality of an urban river in Penang , Malaysia. *Water Science and Engineering*, 16(3), 234–242. <https://doi.org/10.1016/j.wse.2023.01.003>
- Benaafi, M., Yassin, M. A., Usman, A. G., & Abba, S. I. (2022). *Neurocomputing Modelling of Hydrochemical and Physical Properties of Groundwater Coupled with Spatial Clustering , GIS , and Statistical Techniques*.
- Costache, R., Pham, Q. B., Sharifi, E., Linh, N. T. T., Abba, S. I., Vojtek, M., Vojteková, J., Nhi, P. T. T., & Khoi, D. N. (2020). Flash-flood susceptibility assessment using multi-criteria decision making and machine learning supported by remote sensing and GIS techniques. *Remote Sensing*, 12(1). <https://doi.org/10.3390/RS12010106>
- Deng, L., Liu, K., Fan, Y., Qian, X., Ke, T., Liu, T., Li, M., Xu, X., Yang, D., & Li, H. (2025). Interpretable machine learning models reveal the partnership of microplastics and perfluoroalkyl substances in sediments at a century scale. *Journal of Hazardous Materials*, 486(November 2024), 137018. <https://doi.org/10.1016/j.jhazmat.2024.137018>
- Egbuna, I. K., Saidu, M., Ahmad, K. H., & Ogeah, P. U. (2025). *Advancing environmental sustainability through emerging AI-based monitoring and mitigation strategies for microplastic pollution in aquatic ecosystems*.
- Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Elzain, H. E., Chung, S. Y., Senapathi, V., Sekar, S., Lee, S. Y., Roy, P. D., Hassan, A., & Sabarathinam, C. (2022). Comparative study of machine learning models for evaluating groundwater vulnerability to nitrate contamination. *Ecotoxicology and Environmental Safety*, 229, 113061. <https://doi.org/10.1016/j.ecoenv.2021.113061>
- Granata, F., Papirio, S., Esposito, G., Gargano, R., & de Marinis, G. (2017). Machine learning algorithms for the forecasting of wastewater quality indicators. *Water (Switzerland)*, 9(2), 1–12. <https://doi.org/10.3390/w9020105>
- Hale, R. C., Seeley, M. E., La Guardia, M. J., Mai, L., & Zeng, E. Y. (2020). A Global Perspective on Microplastics. *Journal of Geophysical Research: Oceans*, 125(1), 1–40. <https://doi.org/10.1029/2018JC014719>
- Jibril, M. M., Idris, S., Hauwa, M., Jibrin, B., Jibrin, U., Adamu, M., Usman, D. A. G., Dilber, A. G. I., Ozsahin, U., & Karami, Z. (2024). New random intelligent chemometric techniques for sustainable geopolymer concrete : low - energy and carbon - footprint initiatives. *Asian Journal of Civil Engineering*, 25(2), 2287–2305. <https://doi.org/10.1007/s42107-023-00908-7>
- Kalatzis, D., Katsafadou, A. I., Katsarou, E. I., Chatzopoulos, D. C., & Kiouvrakis, Y. (2025). *Explainable Artificial Intelligence for the Rapid Identification and Characterization of Ocean Microplastics*. Figure 1, 1–30.
- Khatti, T., Naderi-Manesh, H., & Kalantar, S. M. (2019). Application of ANN and RSM techniques for modeling electrospinning process of polycaprolactone. *Neural Computing and Applications*, 31(1), 239–248. <https://doi.org/10.1007/s00521-017-2996-6>
- Length, F. (2012). *Concentration of selected heavy metals in water of the Juru River , Penang , Malaysia*. 11(33), 8234–8240. <https://doi.org/10.5897/AJB11.3840>
- Sheng, W., Hadibarata, T., Yuniarto, A., Ho, K., Tang, D., Abdullah, F., Syafrudin, M., Al, D. A., & Al-mohaimeed, A. M. (2021). Characterization of microplastics in the water and sediment of Baram River estuary, Borneo Island. *Marine Pollution Bulletin*, 172(August), 112880. <https://doi.org/10.1016/j.marpolbul.2021.112880>
- Sultan, U., Abidin, Z., Lumpur, U. K., Simalungun, U., & Sumatra, N. (2023). *ANALYSIS OF THE FLASH FLOOD EVENT AND RAINFALL*. 21(1), 58–71.
- Usman, A. G., Abd, M. A., Hashem, A. F., Hussaini, M., Anwar, T., Koramar, M., & Lawal, B. (2025). *Journal of Radiation Research and Applied Sciences Intelligent techniques for modelling the effects of methanol leaves extract of Combretum hypopilinum diels (Combretaceae) on hepatic enzymes as indicators of radiation-induced liver injury*. 18(July).
- Usman, A. G., Işik, S., & Abba, S. I. (2020). A Novel Multi-model Data-Driven Ensemble Technique for the Prediction of Retention Factor in HPLC Method Development. *Chromatographia*, 83(8), 933–945. <https://doi.org/10.1007/s10337-020-03912-0>
- Usman, A. G., IŞIK, S., & Abba, S. I. (2022). Qualitative prediction of Thymoquinone in the high-performance liquid chromatography optimization method development using artificial intelligence models coupled with ensemble machine learning. *Separation Science Plus*, June, 1–9. <https://doi.org/10.1002/sscp.202200071>
- Usman, A. G., Jibrin, A. M., Mati, S., & Abba, S. I. (2025). Evidential-bio-inspired algorithms for modeling groundwater total hardness: A pioneering implementation of evidential neural network for feature selection in water resources management. *Environmental Chemistry and Ecotoxicology*, 7, 494–505. <https://doi.org/10.1016/j.enceco.2025.02.012>
- Usman, A. G., Mati, S., Jibril, M. M., Usman, J., Shah, S. M. H., Abba, S. I., & Naganna, S. R. (2024). Evidential Neural Network Model for Groundwater Salinization Simulation: A First Application in Hydro-Environmental Engineering. *Water (Switzerland)*, 16(20). <https://doi.org/10.3390/w16202873>
- Usman, A. G., Usanase, N., Abba, S. I., Ozsahin, I., Uzun, B., Yassin, M. A., Rahman, S. M., & Ozsahin, D. U. (2023). Environmental modelling of CO concentration using AI-based approach supported with filters feature extraction_ A direct and inverse chemometrics-based simulation. *Sustainable Chemistry for the Environment*,

- 2(February), 100011. <https://doi.org/10.1016/j.scenv.2023.100011>
- Usman, J., Abba, S. I., Ishola, N. B., El-Badawy, T., Adamu, H., Gbadamosi, A., Salami, B. A., Usman, A. G., Benaafi, M., Othman, M. H. D., & Aljundi, I. H. (2023). Genetic neuro-computing model for insights on membrane performance in oily wastewater treatment: An integrated experimental approach. *Chemical Engineering Research and Design*, 199, 33–48. <https://doi.org/10.1016/j.cherd.2023.09.027>
- Van Dao, D., Adeli, H., Ly, H. B., Le, L. M., Le, V. M., Le, T. T., & Pham, B. T. (2020). A sensitivity and robustness analysis of GPR and ANN for high-performance concrete compressive strength prediction using a monte carlo simulation. *Sustainability (Switzerland)*, 12(3). <https://doi.org/10.3390/su12030830>
- Zeng, H., Cui, Y., Shao, R., Li, T., Zhao, Y., Liu, R., Dai, Z., Liu, Y., Wang, J., Xie, Q., & Zhang, X. (2025). Microplastics in animal-derived products and their potential risks to human health. *TrAC Trends in Analytical Chemistry*, 185, 118187. <https://doi.org/10.1016/j.trac.2025.118187>
- Zhen, Y., Wang, L., Sun, H., & Liu, C. (2023). Prediction of microplastic abundance in surface water of the ocean and influencing factors based on ensemble learning ☆. *Environmental Pollution*, 331(P2), 121834. <https://doi.org/10.1016/j.envpol.2023.121834>