Explainable AI Reveals Statistical Associations Between Industrial Activity and PFAS Contamination of Public Water Systems

Priyanshu R Gupta[1], Hogan Tyler Nance[1], Khang Nguyen[1], Mateo Srivathanakul[1], Emily Tang[1], Jaden C Deegan[1], Raman Dhiman[2] , Manish Kumar[1,2*]

[1] *Maseeh Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin,*

[2] *McKetta Department of Chemical Engineering, The University of Texas at Austin,*

Table of Contents:

Table S1: **Features used to construct watershed-level PFAS risk prediction.** The 30 predictors used in developing the model, including industrial sectors, waste management, AFFF users, and sociodemographic factors, are grouped into thematic categories by the authors. NAICS codes follow formal regulatory classifications, while the thematic feature groups were developed to aid interpretation of model outputs.

S. no.	NAICS Code	NAICS Name / Feature Name	Thematic Feature Group
1	325199	All Other Basic Organic Chemical Manufacturing	Chemical Manufacturing
2	325510	Paint and Coating Manufacturing	
3	325611	Soap and Other Detergent Manufacturing	
4	325612	Polish and Other Sanitation Good Manufacturing	
5	325998	All Other Miscellaneous Chemical Product and Preparation Manufacturing	
6	335999	Miscellaneous Electrical Equipment Manufacturing	Electronics & Electrical Equipment
7	334419	Other Electronic Component Manufacturing	
8	334413	Semiconductor and Related Device Manufacturing	
9	332812	Metal Coating, Engraving (except Jewelry and Silverware), and Allied Services to Manufacturers	Metal Treatment & Fabrication
10	332813	Electroplating, Plating, Polishing, Anodizing, and Coloring	
11	332999	Miscellaneous Fabricated Metal Product Manufacturing	
12	324191	Petroleum Lubricating Oil and Grease Manufacturing	Petroleum & Petrochemical
13	324110	Petroleum Refineries	
14	326111	Unlaminated Plastics Film and Sheet	Plastics & Polymer Products
15	325211	Plastics Material and Resin Manufacturing	
16	326112	Plastics Packaging Film and Sheet	
17	313210	Broad woven Fabric Mills	Textile, Paper & Printing
18	323111	Commercial Printing (except Screen and Books)	
19	322121	Paper (except Newsprint) Mills	
20	562213	Solid Waste Combustors and Incinerators	Waste Management & Chemical Handling
21	424690	Chemical and Allied Products Merchant Wholesalers	
22	562112	Hazardous Waste Collection	
23	562219	Hazardous Waste Treatment and Disposal	
24	562211	Solid Waste Landfill	
25	562219	Other Nonhazardous Waste Treatment and Disposal	
26	-	Military Bases	AFFF-users
27	-	AFFF-certified Airports	
28	-	Fire fighting training facilities	
29	-	Total Population	Sociodemographic factors
30	-	Neighborhood Affluence	

Table S2: Frequency of PFAS detections reported in UCMR5 (as of October 2024). Number of detections reported for individual PFAS analytes in U.S. public water systems, based on EPA's Unregulated Contaminant Monitoring Rule 5 (UCMR5). The top 8 contaminants with >500 detections were selected for model development in this study.

Name of Contaminant	Number of Detections reported
PFPeA	1891
PFBA	1752
PFHxA	1681
PFBS	1544
PFOS	1264
PFOA	1211
PFHxS	967
PFHpA	526
6:2 FTS	163
PFNA	72
PFPeS	50
HFPO-DA	45
PFDA	14
8:2 FTS	9
PFUnA	6
PFDoA	4
PFHpS	4
NFDHA	4
ADONA	3
4:2 FTS	2
PFMPA	2
NMeFOSAA	1
9Cl-PF3ONS	1
NEtFOSAA	1
PFMBA	1

Table S3. Performance of the classification algorithm using different ML architectures across 8 PFAS-specific classifiers and 1 SUMPFAS model. Mean accuracy, precision, recall, F1-score and Youden's J-Index are reported for models trained on individual PFAS compounds (e.g., PFOA, PFOS, PFBS) and for the aggregated "SUMPFAS" model. Values are reported at default threshold (0.5); optimized results (Youden's J-max) in parentheses.

XGBoost				
	Mean Accuracy	s.d.	Mean AUCROC	s.d.
SUMPFAS	69.32%	0.23%	0.76	0.0008
PFPeA	75.62%	0.14%	0.78	0.0016
PFBA	68.68%	0.15%	0.70	0.0026
PFHxA	78.29%	0.16%	0.80	0.0011
PFBS	75.75%	0.10%	0.79	0.0010
PFOS	76.94%	0.13%	0.75	0.0016
PFOA	79.90%	0.14%	0.79	0.0018
PFHxS	77.15%	0.13%	0.74	0.0026
PFHpA	85.43%	0.23%	0.75	0.0021

Random Forest				
	Mean Accuracy	s.d.	Mean AUCROC	s.d.
SUMPFAS	71.97%	0.03%	0.78	0.0007
PFPeA	77.81%	0.15%	0.81	0.0005
PFBA	72.99%	0.20%	0.75	0.0010
PFHxA	79.56%	0.18%	0.82	0.0006
PFBS	76.69%	0.39%	0.80	0.0027
PFOS	79.07%	0.23%	0.78	0.0010
PFOA	82.41%	0.08%	0.82	0.0013
PFHxS	80.33%	0.14%	0.76	0.0006
PFHpA	87.73%	0.12%	0.79	0.0026

LightGBM

	Mean Accuracy	s.d.	Mean AUCROC	s.d.
SUMPFAS	68.68%	0.09%	0.75	0.0028
PFPeA	75.69%	0.26%	0.78	0.0027
PFBA	69.63%	0.09%	0.70	0.0009
PFHxA	77.96%	0.39%	0.80	0.0018
PFBS	75.21%	0.12%	0.78	0.0011
PFOS	76.80%	0.17%	0.74	0.0022
PFOA	80.48%	0.18%	0.78	0.0017
PFHxS	77.68%	0.23%	0.72	0.0037
PFHpA	85.30%	0.08%	0.74	0.0037

Table S4: Performance of the Random Forest classification model trained for each of the 8 PFAS of interest and 1 SUMPFAS model.

Name of PFAS model	AUCROC	Accuracy	Specificity	Sensitivity	F1-Score
SUMPFAS	0.78	69.1%	0.74	0.74	0.74
PFPeA	0.81	81.6%	0.86	0.70	0.78
PFBA	0.75	75.7%	0.79	0.65	0.71
PFHxA	0.82	81.6%	0.75	0.78	0.77
PFBS	0.81	77.9%	0.73	0.85	0.79

PFOS	0.77	73.5%	0.75	0.67	0.71
PFOA	0.82	83.8%	0.86	0.79	0.83
PFHxS	0.75	75.7%	0.79	0.65	0.71
PFHpA	0.78	86%	0.83	0.72	0.77

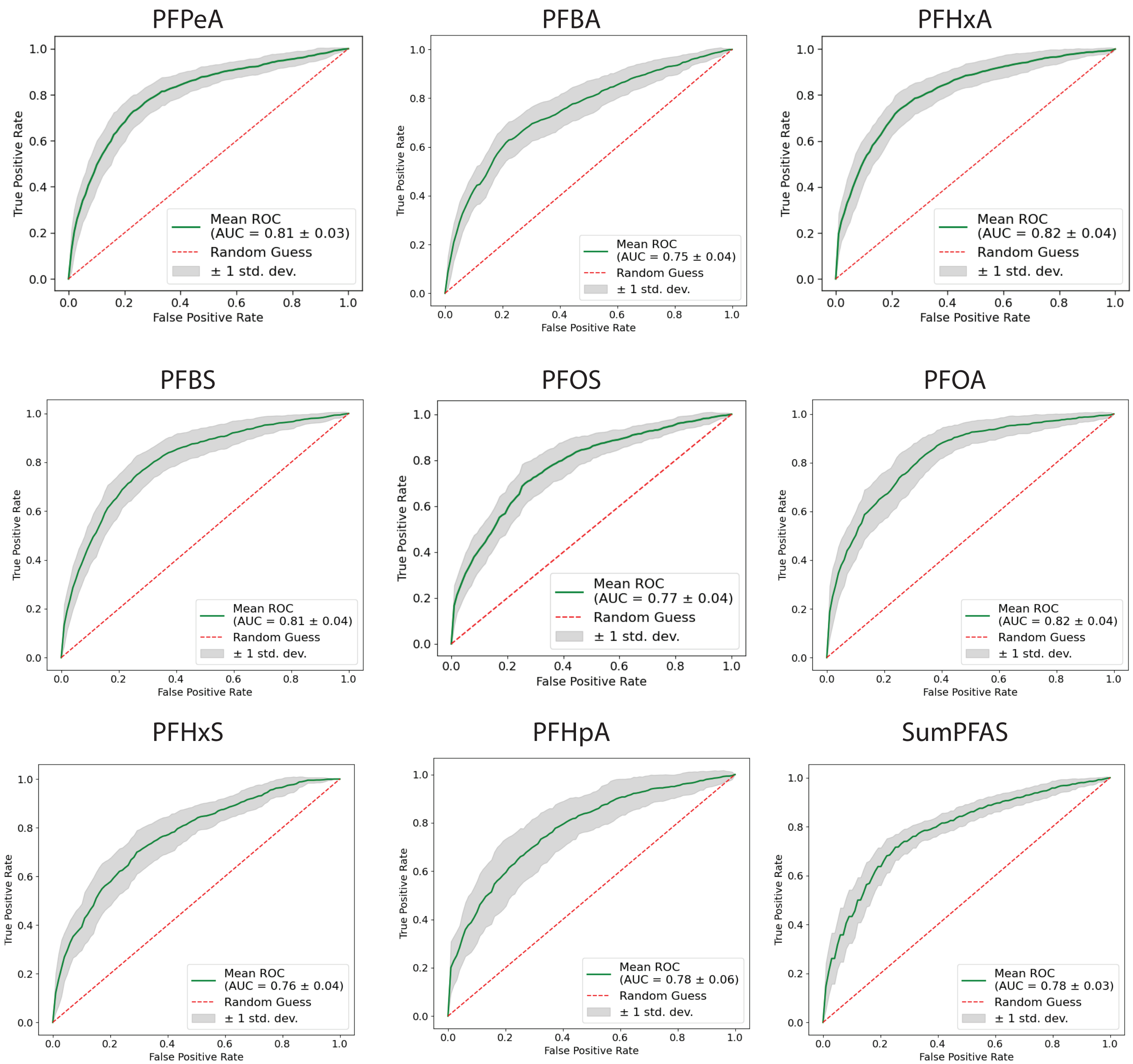

Figure S1: Cross-validated ROC curves for SUMPFAS detection model.

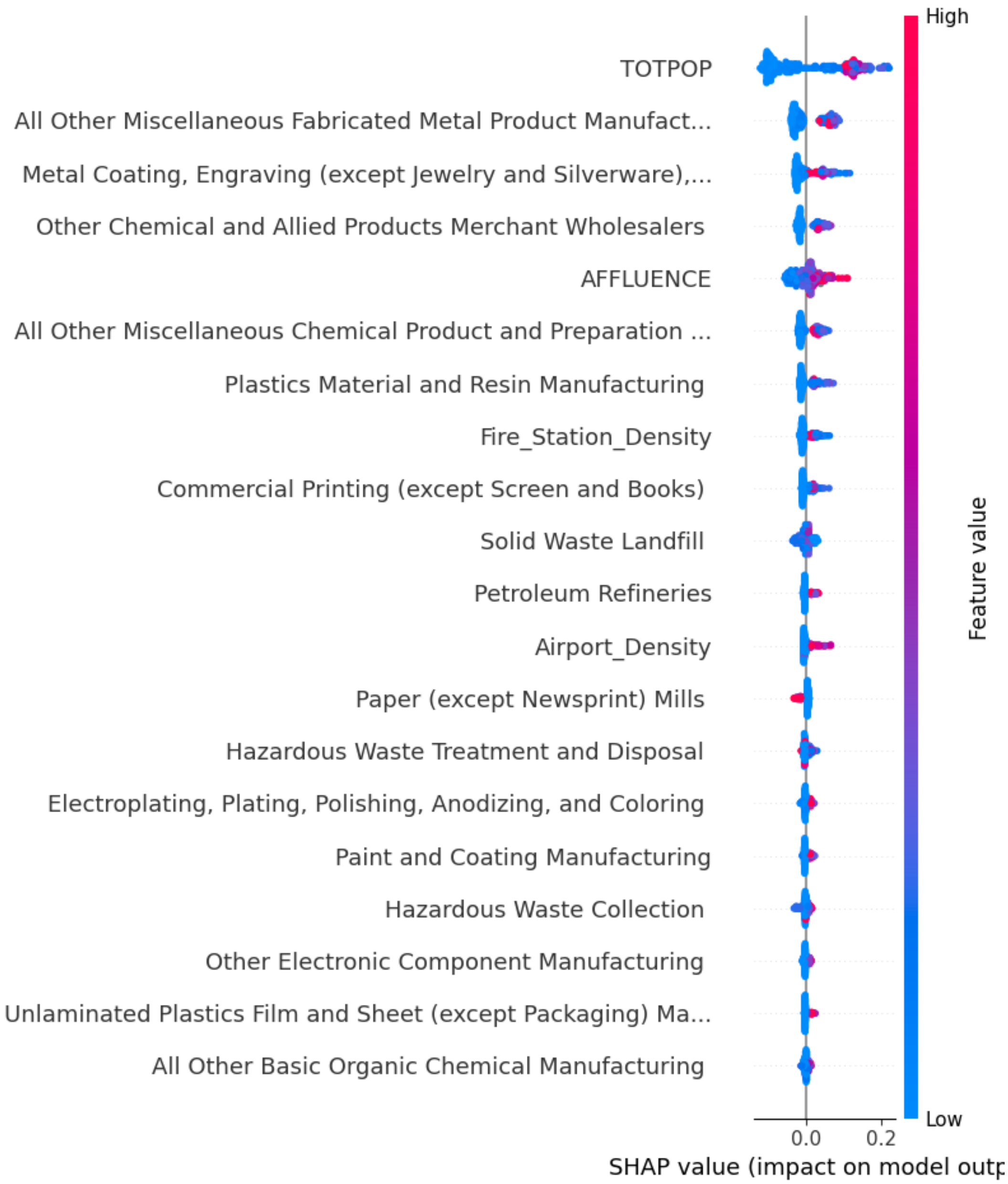

Figure S2: Full SHAP summary plot showing feature contributions across all subbasins. SHAP value distributions for all modeled features across all PWS subbasins in the national dataset. Each point represents the SHAP value for a feature in one subbasin, with color indicating the underlying feature magnitude. Positive SHAP values indicate a greater contribution to PFAS detection classification, while negative values push the prediction toward non-detection.

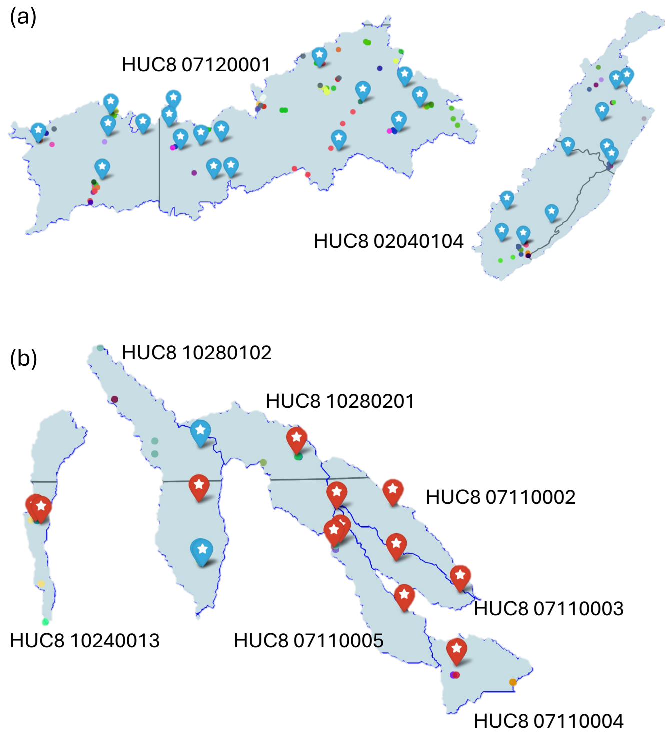

Figure S3: Illustrative examples of prediction mismatches between model output and observed PFAS detection in UCMR 5. (a) Exemplar HUC8 regions 07120001 and 02040104 with high modeled probability of PFAS detection (based on dense industrial presence), yet no PFAS detections reported in any of the PWSs (blue star marker) reported in UCMR5 measurements. The false positives from our model may reflect regions with effective treatment technologies, robust source protection by the PWSs, or under-reporting in the current sampling round. (b) HUC8 regions with minimal or no modeled industrial activity (of sectors used in this study), where model predicts non-detection, yet UCMR5 results confirm PFAS presence in all cases. These false negatives could indicate missing sources such as legacy contamination, surface runoffs, long distance transport, or non-industrial contributors like septic systems or biosolids. These examples highlight the limitations in input data in this study, pointing to the need for more comprehensive source inventories and representation of diffused contamination pathways.