APPLICATION OF THE *crAIRsis* AI-BASED FRAMEWORK FOR THE ANALYSIS OF PCB-170 IN HUMAN BREAST MILK

Timea Bezdan[1],
[0000-0001-6938-6974]
Gordana Jovanović[2],
[0000-0001-8657-423X]
Andreja Stojić[1,2],
[0000-0002-5293-9533]
Snježana Herceg Romanić[3],
[0000-0003-2382-4734]
Mirjana Perišić[1,2]*
[0000-0002-8287-4136]

[1]Singidunum University, Belgrade, Serbia

[2]Institute of Physics Belgrade, Belgrade, Serbia

[3]Institute for Medical Research and Occupational Health, Zagreb, Croatia

Abstract:

Breast milk is a reliable, non-invasive matrix for monitoring internal exposure to polychlorinated biphenyls (PCBs) and organochlorine pesticides (OCPs), particularly in vulnerable populations such as infants. Ongoing monitoring studies underscore the need for a deeper understanding of the distribution and health impacts of these persistent organic pollutants. Although artificial intelligence (AI) has been widely applied across scientific disciplines, its use in environmental exposure analysis, particularly in biological matrices like breast milk, remains limited. This study investigates the distribution of PCB-170, a highly chlorinated and toxicologically relevant PCB congener, and identifies key predictive factors using an advanced AI-based framework. The analysis was performed using the crAIRsis platform, which integrates ensemble machine learning algorithms, metaheuristic optimisation, and explainable AI methods such as Shapley additive explanations (SHAP) and Shapley additive global importance (SAGE). This approach enables the modelling of complex, non-linear relationships between variables. Breast milk samples from 186 mothers in Zadar, Croatia, were analysed for 17 PCB congeners and 7 OCPs. The most influential predictors of PCB-170 levels were PCB-180, PCB-153, and PCB-138, indicating strong co-behaviour and likely shared exposure pathways. These congeners showed relative SHAP impacts ranging from -40% to over 60%. Demographic variables, including maternal age and birth order, had minimal influence, with SHAP impacts below 10%. The study demonstrates the dominant role of higher-chlorinated PCBs in shaping internal burdens and highlights the value of explainable AI in environmental health research. The crAIRsis framework offers a robust, transferable methodology for human biomonitoring and evidence-based exposure assessment.

Keywords:

Human Biomonitoring, Polychlorinated Biphenyls, Machine Learning, Metaheuristics, Explainable Artificial Intelligence.

INTRODUCTION

Human milk is a dynamic, bioactive fluid that provides essential nutrients, immune protection, and bioactive compounds crucial for infant development and lifelong health. Its composition varies according to maternal and infant health, diet, and environmental factors, rendering it a form of personalised nutrition. Research into human milk contaminants has expanded since the 1950s, with DDT first detected in 1951 and its metabolites subsequently identified in nearly all tested samples globally. Later surveillance efforts have focused on persistent organic pollutants (POPs), including organochlorine pesticides (OCPs), polychlorinated biphenyls (PCBs), dioxins, organophosphate pesticides,

Correspondence:
Mirjana Perišić

e-mail:
mirjana.perisic@ipb.ac.rs

DOI: 10.15308/Sinteza-2025-221-227

bisphenols, and polycyclic aromatic hydrocarbons—lipophilic chemicals that accumulate in fatty tissues due to their stability and resistance to degradation. Their presence in breast milk leads to prolonged infant exposure, posing risks such as endocrine disruption, neurodevelopmental effects, and immune dysfunction. As POPs emerged as a global concern, biomonitoring initiatives broadened in scope, revealing temporal trends in legacy pollutants [1] [2] [3]; however, the understanding of pollutant co-occurrence and interrelationships remains limited.

Whilst explainable artificial intelligence (XAI) has been widely applied [4], its integration into environmental research, and its potential to enhance understanding of pollutant dynamics in environmental matrices, remains largely unexplored or frequently misinterpreted. Recently, Huang et al. [5] introduced an innovative approach to assessing chemical exposure risks in breastfeeding infants using an explainable machine learning (ML) model. By integrating ensemble resampling and advanced feature selection techniques, their framework enhances predictive accuracy in identifying high-risk chemicals such as POPs. A key innovation lies in the use of Shapley additive explanations (SHAP), which quantify the contribution of individual features—in this case, chemical properties—to the model's predictions. This approach improves the understanding of molecular factors that influence the transfer of high-risk compounds into human milk and supports more targeted risk mitigation strategies. By identifying molecular fragments linked to high-risk chemicals, the focus shifts from general chemical properties to specific molecular features, thus advancing the field of lactation toxicology. Building on our earlier research [6] [7] [8], which employed ML techniques to investigate dependencies among OCP and PCB congeners in mothers' milk, this study focuses specifically on PCB-170. This highly chlorinated and toxicologically relevant congener has gained attention due to its persistence, bioaccumulative properties, and distinct distribution patterns in environmental and biological matrices.

In this study, we employed a comprehensive AI framework developed within the crAIRsis project [9], which autonomously conducts all stages of the analysis [10] [11]. The objective was to identify the factors influencing PCB-170 distribution in breast milk and evaluate its potential as a predictive marker for broader PCB exposure assessments. The framework integrates seven ensemble regression models selected for their robust predictive power and generalisation performance. In the subsequent phase, each model is systematically evaluated using 25 metaheuristic optimisation algorithms to fine-tune the hyperparameters of the best-performing models and improve predictive accuracy. Model performance is assessed using a suite of evaluation metrics tailored to the specific problem type—classification or regression. Once the optimal model is identified, the framework proceeds to the interpretation phase by incorporating XAI techniques to quantify the contribution of each predictor, both locally and globally, thereby ensuring transparency, interpretability, and actionable insights throughout the analytical workflow. Modelling results are complemented by interactive visualisations that facilitate exploration and interpretation of model behaviour.

Although the primary aim of this paper is to characterise the distribution of PCB-170, the broader ambition of this research is to demonstrate a flexible and robust analytical framework applicable to a wide range of environmental modelling challenges. By combining ML, metaheuristics, and XAI, the proposed approach offers a transferable methodology that supports evidence-based decision-making and strengthens the effectiveness of human biomonitoring strategies.

2. METHODOLOGY

Sample collection and chemical analysis of PCBs and OCPs

Breast milk samples were collected between 2014 and 2019 from 186 healthy mothers (primiparae, secundiparae, and multiparae – third delivery), aged 19 to 41 years, residing in the Zadar region, Croatia. Participants reported no history of accidental or occupational exposure to persistent organic pollutants. Detailed sampling protocols have been described previously [6] [12]. Chemical analysis of PCBs and OCPs followed established procedures outlined in earlier studies [13] [14] [15]. The analysis focused on six indicator PCB congeners (IUPAC numbers: 28, 52, 101, 138, 153, 180), chosen due to their prevalence in technical mixtures, the environment, and biological tissues. Additionally, eleven toxicologically relevant congeners (IUPAC numbers: 60, 74, 105, 114, 118, 123, 156, 157, 167, 189, 170) were included in the analysis.

Data analysis

The data analysis, with PCB-170 as the target variable, was carried out using the crAIRsis framework, a modular and automated AI-based platform that inte-

grates advanced machine learning, metaheuristics, and explainable artificial intelligence techniques. At its core, the framework incorporates seven ensemble regression algorithms: AdaBoost, CatBoost, ExtraTrees, Gradient Boosting, Histogram Gradient Boosting, LightGBM, and XGBoost [16] [17] [18] [19]. These algorithms were selected for their proven ability to capture complex, non-linear patterns and their robustness against overfitting. Each model was evaluated using five-fold cross-validation, ensuring generalisability and minimising bias. Based on the evaluation metrics—specifically R-squared, mean absolute error, mean squared error, and others—the three best-performing models were selected for further optimisation. To enhance predictive accuracy, the hyperparameters of these top models were fine-tuned using the Sine Cosine Algorithm and Harris Hawks Optimisation metaheuristic methods [20] [21]. These approaches efficiently explore the hyperparameter search space and have demonstrated strong performance in solving complex, non-convex optimisation problems in machine learning contexts. Once the final model was selected, explainability and interpretation were undertaken using SHAP and SAGE. SHAP quantifies the influence of each input feature on individual predictions, providing a detailed understanding of model behaviour at the instance level [22], whereas SAGE evaluates feature importance by aggregating their contributions across the entire dataset, thereby identifying variables with the strongest overall impact on model performance [23]. o support interpretation, we further derived relative and normalised SHAP values and introduced a categorical framework referred to as inherent SHAP values. Relative SHAP values express the proportion of a feature's absolute SHAP value in relation to the total attribution for a given prediction, offering insight into the feature's contextual importance. Normalised SHAP values, scaled to the expected model output, simplify the interpretation of impact magnitudes. Inherent SHAP values were used to group SHAP effects into interpretable categories, whereby high negative impacts were defined as those falling below the mean of all negative SHAP values. All these methods contribute to enhanced model transparency and facilitate the interpretability of complex relationships between input variables and the target outcome. To gain a deeper understanding of feature interactions and group-level behaviour, cluster analysis was applied to the SHAP values. Dimensionality reduction was performed using Pairwise Controlled Manifold Approximation [24], followed by clustering using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [25] [26]. This stage enabled the identification of patterns, subgroups, and outliers within the dataset based on model behaviour.

The crAIRsis framework automates all key stages, including model training, evaluation, optimisation, explanation, and visualisation of both raw data and results, offering a robust and transferable methodology for environmental data modelling. All outputs, including evaluation metrics, feature contributions, and visualisations, are automatically generated and structured for subsequent interpretation and reporting.

3. RESULTS AND DISCUSSION

Monitoring of POPs in human milk in Croatia dates back to the 1970s, primarily focusing on the assessment of PCB and OCP levels and their temporal trends. Our previous studies and review [2] [12], centred mainly on monitoring efforts, revealed a consistent decline in p,p'-DDE, HCB, β-HCH, and PCBs from 1976 to 2019, largely attributed to regulatory restrictions introduced under the Stockholm Convention since the 1990s. As previously reported [2] [12], within the dataset used for this study, compounds from the DDT group—particularly p,p'-DDE—were the most prevalent, followed by PCB-153, PCB-138, PCB-180, β-HCH, PCB-118, γ-HCH, HCB, PCB-156, and PCB-170, in decreasing order. The least abundant compounds included PCB-28, PCB-105, and PCB-60. Detailed descriptive statistics and comparisons with other studies have also been documented previously.

Our pioneering research, aimed at gaining a more precise understanding of the levels, interrelations, and associations of PCBs and OCPs in human milk with maternal factors such as age and parity, employed machine learning techniques. In Jovanović et al. [6], the Guided Regularised Random Forest (GRRF) algorithm identified key factors influencing POP levels, revealing strong non-linear relationships among pollutants and the complexity of their pathways in breast milk. The model achieved prediction errors below 30% and correlation coefficients exceeding 0.90 between predicted and observed values.

Building on this, Jovanović et al. [8] applied advanced machine learning methods—XGBoost and SHAP—to investigate PCB-138 interactions with other non-dioxin-like congeners, maternal age, and parity, identifying PCB-170 and PCB-153 as key drivers of PCB-138 behaviour in milk samples and promising targets for further research.

The ExtraTrees model, optimised using the Sine Cosine Algorithm (SCA), demonstrated strong predictive performance for PCB-170, achieving an R^2 of 0.9125, RMSE of 0.7354, and MAPE of 0.3234, indicating high accuracy and a substantial proportion of explained variance in predicted concentrations. The most influential variables in predicting PCB-170 concentrations were PCB-180, PCB-153, and PCB-138, all of which exhibited significantly higher importance scores than other predictors (Figure 1). These three congeners dominated the model's output, indicating strong associations and potential co-behaviour patterns with PCB-170. While this may reflect similarities in chemical structure and possible overlap in exposure sources or metabolic pathways, such patterns do not necessarily imply a shared origin or an increased tendency for bioaccumulation. Rather, the observed co-behaviour suggests that these congeners may be influenced by related biological or environmental factors that shape their distribution in human milk [27][28]. Among the other variables, PCB-156, p,p'-DDE, and PCB-118 also contributed to the model, though with notably lower impact values. In contrast, factors such as birth order, maternal age, and certain OCPs (e.g. β-HCH, p,p'-DDT) showed minimal influence, suggesting a limited role in explaining PCB-170 variability within the analysed dataset.

Figure 2 presents SHAP dependence plots illustrating the relative impact of PCB-180, PCB-153, and PCB-138 on the prediction of PCB-170 concentrations. By applying a clustering approach to the SHAP values, we identified distinct groups of variable impacts, each representing a specific cluster that highlights how key factors influence the biological and exposure-related context affecting PCB-170 levels. These clusters are shown in different colors on the plot, and individual samples (represented as dots) are colored according to the cluster to which they belong. A more in-depth exploration of these patterns lies beyond the scope of this study. Across all plots, a clear monotonic relationship is observed between each specific congener and PCB-170, with higher concentrations of each predictor corresponding to an increased contribution to the predicted PCB-170 level. The strongest relative impacts, up to 60% for PCB-138 and PCB-153, and up to 40% for PCB-180, are associated with mid-to-high concentration ranges of PCB-153 and PCB-180, suggesting that these congeners play a dominant role in shaping the modelled PCB-170 dynamics. Marginal effects remain stable within low-concentration ranges but exhibit greater variance and influence as values increase, particularly for PCB-153. This indicates that interactions with other features are not fully disentangled, which could potentially be addressed by expanding the dataset, both in terms of sample size and inclusion of additional relevant variables. These patterns further support the notion of strong co-behaviour and likely shared exposure pathways among higher-chlorinated PCBs in the analysed population. Moreover, in the lower concentration range of the dominant predictors, negative contributions to PCB-170 prediction are also observed (with relative impacts of up to -40%), indicating that, in certain breast milk samples, the co-occurrence of these compounds at observed concentration levels is associated with reduced PCB-170 levels. This may reflect competitive metabolic pathways, differential accumulation patterns, or individual differences in exposure and elimination dynamics.

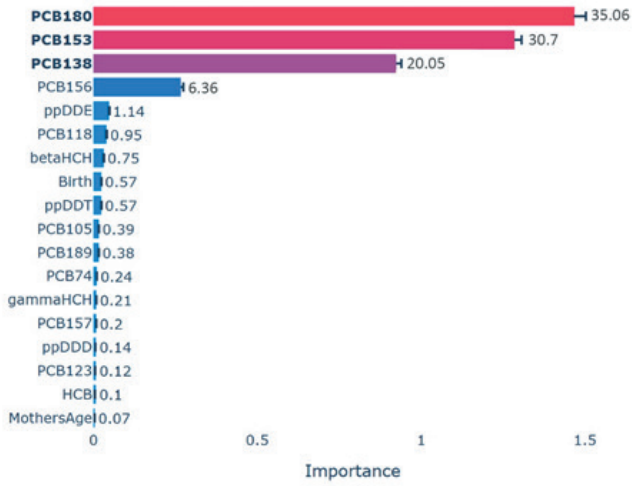

Figure 1. Global feature importance based on SAGE values. Normalized SAGE values [%] were provided as numbers within or next to the bars

The analysis indicates that maternal age has minimal influence, with SHAP values distributed closely around zero across the entire age range (Figure 3, upper panel), suggesting that, within the present dataset, age does not significantly contribute to the prediction of PCB-170 levels. In contrast, birth order exhibits slightly greater variation in its impact, particularly among mothers with three children (Figure 3, lower panel). Although the overall relative influence remains modest, there is a slight tendency towards higher predicted PCB-170 concentrations in multiparous women. Older mothers are generally subject to greater cumulative exposure to POPs, whereas parity is often associated with lower POP levels, as breastfeeding facilitates the reduction of maternal body burdens. This was demonstrated in a Norwegian study, where age was positively associated with POP concentrations, while increased parity correlated with lower levels [29]. More recent research, including a study involving over one thousand primiparous women in the Czech Republic, found that maternal diet before and during pregnancy influenced PCB levels in breast milk, whereas body weight and age at delivery had no significant effect [27]. Nevertheless, the effect size of maternal age and parity is substantially smaller than that of chemical predictors related to concentration levels, suggesting their limited apparent role in shaping internal PCB-170 burden. This likely reflects insufficient characterisation of maternal status, highlighting the need for additional variables that capture physiological and exposure-related factors more comprehensively.

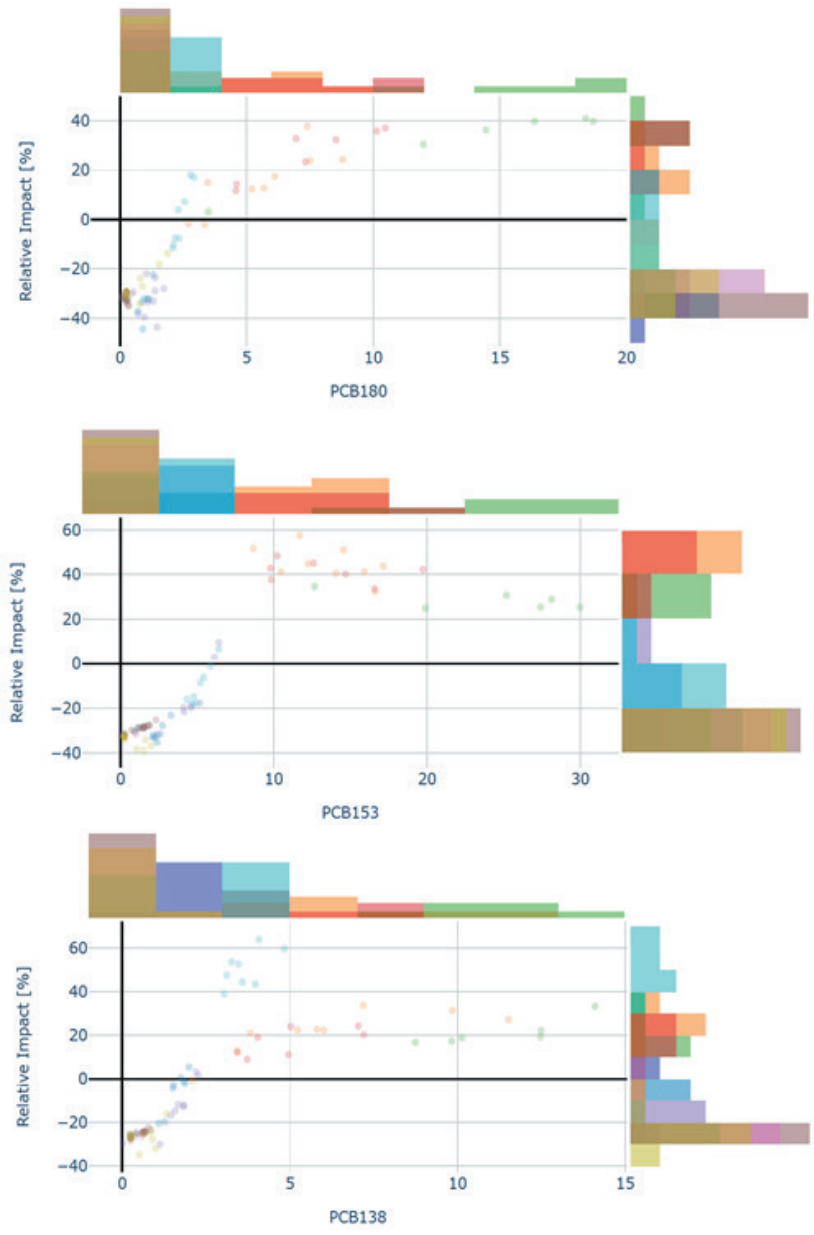

Figure 2. Relative SHAP impacts of the most important variables PCB-180, PCB-153 and PCB-138 on PCB-170 dynamic

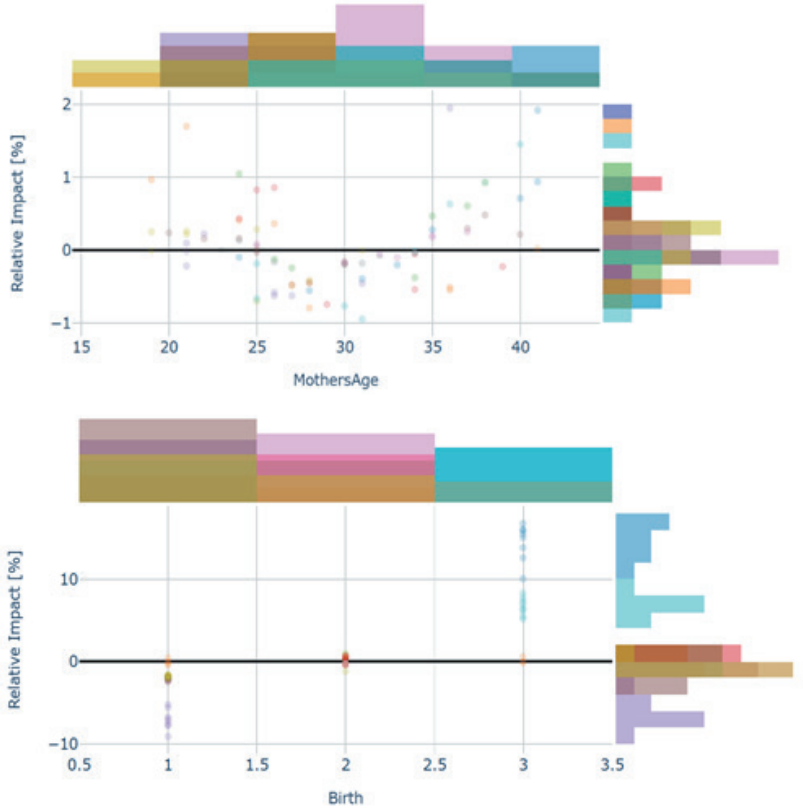

Figure 3. Relative SHAP impacts of the mothers age and birth order on PCB-170 dynamic

4. CONCLUSION

This study demonstrated the effectiveness of the crAIRsis AI-based framework in modelling and enhancing the understanding of PCB-170 distribution in human breast milk. By combining ensemble machine learning algorithms, metaheuristic optimisation, and explainable AI, the framework provided more transparent, data-driven insights into the complex, non-linear relationships between PCB-170 and its predictors. The results identified PCB-180, PCB-153, and PCB-138 as the most influential variables, indicating strong co-behaviour and likely shared exposure pathways with PCB-170. Moderate contributions were observed from other chemical compounds such as PCB-156 and p,p'-DDE, while demographic factors, maternal age and birth order, exerted minimal influence, suggesting that direct chemical exposures play a dominant role in shaping PCB-170 concentrations. Beyond supporting the role of higher-chlorinated congeners, this work illustrates the potential of an automated and interpretable framework for environmental data analysis and human biomonitoring. Future research will build on these findings by conducting a more detailed evaluation of all included variables, exploring potential interactions and non-linear effects. In addition, clustering of SHAP values will be systematically integrated into the interpretation phase to identify subpopulations with distinct exposure profiles. Finally, the methodology offers a transferable tool for evaluating exposure to persistent organic pollutants and supporting evidence-based health risk assessments.

5. ACKNOWLEDGEMENTS

The authors acknowledge funding provided by the Institute of Physics Belgrade through a grant from the Ministry of Education, Science and Technological Development of the Republic of Serbia, as well as by the Science Fund of the Republic of Serbia (Grant No. 7373, Characterizing crises-caused air pollution alterations using an artificial intelligence-based framework – crAIRsis). The analysis of PCBs and OCPs was carried out using facilities and equipment funded by the European Regional Development Fund, under the project KK.01.1.1.02.0007 "Research and Education Centre of Environmental Health and Radiation Protection – Reconstruction and Expansion of the Institute for Medical Research and Occupational Health", and co-funded by the European Union – Next Generation EU (project EnvironPollutHealth, Program Contract dated 8 December 2023, Class: 643-02/23-01/00016, Reg. No. 533-03-23-0006).

REFERENCES

[1] J. Fång, ""Spatial and temporal trends of the Stockholm Convention POPs in mothers' milk — a global review."," *Environ. Sci. Pollut. Res.*, p. 8989–9041, 2015.

[2] S. H. Romanić, ""Persistent organic pollutants in Croatian breast milk: An overview of pollutant levels and infant health risk assessment from 1976 to the present" .," *Food. Chem. Toxicol*, p. 113990, 2023.

[3] K. R. Nermo, "Trend analyses of persistent organic pollutants in human milk from first-time mothers in Norway between 2002 and 2021," *Int. J. Hyg. Environ.*, p. 114458, 2025.

[4] D. Gunning, "XAI—Explainable artificial intelligence," *Sci. Robot.*, p. 7120, 2019.

[5] X. Huang, "Assessing chemical exposure risk in breastfeeding infants: An explainable machine learning model for human milk transfer prediction," *Ecotoxicol. Environ. Saf.*, p. 117707, 2025.

[6] G. Jovanović, ""Introducing of modeling techniques in the research of POPs in breast milk – A pilot study," *Ecotoxicol. Environ. Saf*, pp. 341-347, 2019.

[7] A. Stojić, "Shapley additive explanations of indicator pcb-138 distribution in breast milk.," in *international scientific conference on Information technology and data related research - Sinteza, Singidunum University*, Belgrade, Serbia, Belgrade, 2020.

[8] G. Jovanović, "Patterns of PCB-138 Occurrence in the Breast Milk of Primiparae and Multiparae Using SHapley Additive exPlanations Analysis," in *Artificial Intelligence: Theory and Applications*, Springer, Cham, 2021, pp. 191-206.

[9] "Project crAIRsis – Characterizing Crises-Caused Air Pollution Alterations Using an Artificial Intelligence-Based Framework," cience Fund of the Republic of Serbia (Grant No. 7373), PRISMA program, 2024–2027.

[10] G. Jovanović, "The PM 2.5-Bound Polycyclic Aromatic Hydrocarbon Behavior in Indoor and Outdoor Environments, Part III: Role of Environmental Settings in Elevating Indoor Concentrations of Benzo (a) pyrene.," *Atmosphere*, 2024.

[11] A. Stojić, "Artificial Intelligence-Based Framework for Analyzing Crises-Caused Air Pollution.," in *Sinteza 2024-International Scientific Conference on Information Technology, Computer Science, and Data Science. Singidunum University*,, Belgrade, 2024.

[12] G. Mendaš, "Presence of polycyclic aromatic hydrocarbons and persistent organochlorine pollutants in human Milk: Evaluating their levels, association with Total antioxidant capacity, and risk assessment.," *Sci. Total Environ.*, p. 172911, 2024.

[13] D. Klinčić, "Polychlorinated biphenyls and organochlorine pesticides in human milk samples from two regions in Croatia..," *Environ. Toxicol. Pharmacol*, pp. 543-552, 2014.

[14] D. Klinčić, "Organochlorine pesticides and PCBs (including dl-PCBs) in human milk samples collected from multiparae from Croatia and comparison with primiparae.," *Environ. Toxicol. Pharmacol,* pp. 74-79, 2016.

[15] I. Šimić, "Optimization of gas chromatography-electron ionization-tandem mass spectrometry for determining toxic non-ortho polychlorinated biphenyls in breast milk.," *Biomed. Environ. Sci.*, pp. 58-61, 2020.

[16] Y. Freund, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, p. 119–139, 1997.

[17] L. Prokhorenkova, "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[18] G. Ke, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[19] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, p. 1189–1232, 2001.

[20] X.-S. Yang, "Firefly Algorithms for Multimodal Optimization," in *Stochastic Algorithms: Foundations and Applications, O. Watanabe and T. Zeugmann, Eds.*, , Springer Berlin Heidelberg, 2009, pp. 169-178.

[21] A. A. Heidari, "Harris hawks optimization: Algorithm and applications," Future Gener. Comput. Syst., pp. 849-872, 2019.

[22] S. M. Lundberg, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[23] I. Covert, "Understanding global feature contributions with additive importance measures," *Adv. Neural Inf. Process. Syst.*, pp. 17212-17223, 2020.

[24] Y. Wang, "Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization," *Journal of Machine Learning Research,* pp. 1-73, 2021.

[25] L. McInnes, "Umap: Uniform manifold approximation and projection for dimension reduction," *ArXiv Prepr.* , p. 180203426, 2018.

[26] L. McInnes, "HDBSCAN: Hierarchical density based clustering," *J Open Source Softw*, p. 205, 2017.

[27] R. Aerts, "Determinants of persistent organic pollutant (POP) concentrations in human breast milk of a cross-sectional sample of primiparous mothers in Belgium.," *Environ. Int.*, p. 104979, 2019.

[28] J. Komprda, "Dynamics of PCB exposure in the past 50 years and recent high concentrations in human breast milk: analysis of influencing factors using a physiologically based pharmacokinetic model.," *Sci. Total Environ,* pp. 388-399, 2019.

[29] A. Polder, "Levels of chlorinated pesticides and polychlorinated biphenyls in Norwegian breast milk (2002–2006), and factors that may predict the level of contamination.," *Sci. Total Environ.*, pp. 4584-4590, 2009.