Foods 2025 Conference
The 6th International Electronic Conference on Foods
28–30 October 2025 | Online
MDPI

Machine Learning-Based Prediction of Polycyclic Aromatic Hydrocarbon (PAH) Levels in Smoked Fish Using WEKA: Evaluation of Smoking Parameters and Model Performance

Irem Kılınç 1*, Hayal Boyacıoğlu 2, Berna Kılınç 3
1. Katip Çelebi University, Fisheries Faculty, Fish Processing Technology Department, 35610 Çiğli-Izmir, Türkiye
2.Ege University, Science Faculty, Statistics Department, 35100 Bornova-Izmir, Türkiye
3. Ege University, Fisheries Faculty, Fish Processing Technology Department, 35100 Bornova-Izmir, Türkiye

INTRODUCTION & AIM

The most critical food safety concerns is the contamination of smoked fish with polycyclic aromatic hydrocarbons (PAHs), which are toxic, carcinogenic compounds formed during the pyrolysis of organic matter in smoking processes (Nizio et al., 2023).Traditional monitoring methods for PAH concentrations often require extensive laboratory analyses, which are time-consuming, labor-intensive, and costly. Predictive modeling using statistical and machine learning (ML) methods provides a complementary approach, enabling rapid assessment and optimization of processing parameters to minimize PAH levels. Machine learning, in particular, allows modeling of nonlinear relationships and interactions among multiple variables, providing enhanced predictive capability compared to classical regression methods (Zhang et al., 2024). The present study aims to (i) quantify the influence of key smoking parameters on total PAH concentrations using multiple regression analysis, (ii) evaluate various ML algorithms in WEKA for predicting PAH levels, and (iii) compare classification models for fish species based on PAH-related characteristics. This integrated approach is designed to support evidence-based process optimization and risk reduction in smoked fish production.

METHOD

Fish fat content, smoking temperature, and wood type were used as input variables for predictive modeling of total PAH concentrations. Multiple linear regression was performed in SPSS 25.0 (SPSS Inc., Chicago, IL, USA) to assess correlations (Tabachnick, B. G., & Fidell, L. S. (2013)). Machine learning analyses were conducted in WEKA, applying Linear Regression, SMOreg, Multilayer Perceptron, M5P, Random Forest, and IBk for regression tasks, and Logistic Regression, RandomTree, and J48 for classification tasks. Model performance was evaluated using Correlation Coefficient, MAE, RMSE, RAE, and RRSE, with 10-fold cross-validation ensuring validation and robustness. (Aksu, G., & Doğan, N. (2010); Arora, R., & Suman, S. (2012); Hemlata. (2018).

RESULTS & DISCUSSION

☐ Multiple Regression Analysis

Table1. The regression coefficients of the dataset

Predictor	β	B	p-value	Interpretation
Fish fat content	0.444	0.685	0.002	Strong positive effect; each 1% increase in fat → +0.685 units PAH
Smoking temperature	0.384	0.035	0.005	Each 1°C rise in temperature significantly increases PAH formation
Wood type	0.251	2.058	0.036	Wood composition and density moderately affect PAH levels

1) **Fish fat content** emerged as the most influential factor. Higher lipid content promotes PAH absorption during smoking, consistent with previous studies indicating that lipophilic PAHs accumulate preferentially in fatty tissues (Nizio et al., 2023).

2) **Smoking temperature** was also a strong determinant. Elevated temperatures accelerate pyrolysis of wood and fish tissues, generating more PAHs (Savin et al., 2024). These results align with prior findings that controlling temperature is critical for reducing carcinogenic PAHs.

3) **Wood type** had a moderate but statistically significant impact. Different wood species contain varying lignin and cellulose levels, which influence smoke composition and PAH generation (Zhang et al., 2024).

- High-fat fish smoked at elevated temperatures using certain wood types are particularly prone to higher PAH levels, emphasizing the need for integrated control strategies in processing.

☐ Machine Learning Analysis

- **Regression Models**

1) M5P and Random Forest algorithms provided the most accurate predictions, with lower RMSE and MAE values compared to linear models. This indicates that tree-based ensemble methods effectively capture nonlinear interactions between fat content, temperature, and wood type.
2) Multilayer Perceptron performed moderately, suggesting that neural networks can model complex relationships but require larger datasets for optimal accuracy.
3) Linear Regression and SMOreg achieved satisfactory results, supporting the regression analysis in SPSS and confirming the strong linear relationships among predictors.

☐ Classification Models

RandomTree: Overall accuracy 84.3%
J48: Overall accuracy 86.7%
Logistic Regression: Overall accuracy 92.1%

1. Logistic Regression outperformed tree-based methods, indicating strong capability in classifying species with PAH-related characteristics and handling nonlinear relationships effectively.
2. The results suggest that species-specific PAH accumulation patterns are predictable using a combination of chemical composition (fat content) and processing parameters (temperature, wood type).
3. Accurate species classification is critical for traceability, food authenticity, and targeted risk

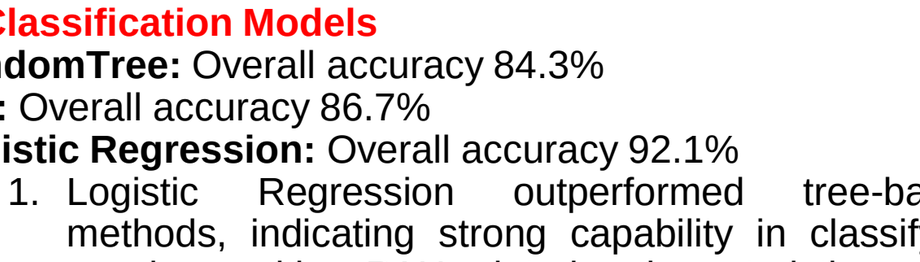

Figure1. The WEKA analysis result of the dataset's

The integration of ML allows both regression and classification analysis, enabling simultaneous estimation of PAH levels and identification of high-risk species or processing conditions.

CONCLUSION

The approach offers a reproducible framework for monitoring and mitigating PAH contamination in smoked fish, which can guide seafood producers in optimizing processing conditions. Importantly, this framework establishes a foundation for future studies, as it can be expanded with larger datasets, additional variables (e.g., wet smoking, dry smoking), and new fish species, thereby contributing to more comprehensive risk assessment models and enhanced food safety strategies.

FUTURE WORK / REFERENCES

This study is an innovative and pioneering study in the process of smoking foods and will provide support for smart technologies in the future.

- Aksu, G., & Doğan, N. (2010). An analysis program used in data mining: WEKA. Journal of Measurement and Evaluation in Education and Psychology, 10(1), 80-95 https://doi.org/10.21031/epod.399832
- Arora, R., & Suman, S. (2012). Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications, 54(13), 21-25. https://doi.org/10.5120/8626-2492
- Hemlata. (2018). COMPREHENSIVE ANALYSIS OF DATA MINING CLASSIFIERS USING WEKA. International Journal of Advanced Research in Computer Science, 9(2), 718723. https://doi.org/10.26483/ijarcs.v9i2.5900
- Nizio, E., Czwartkowski, K., & Niedbała, G. (2023). Impact of Smoking Technology on the Quality of Food Products: Absorption of Polycyclic Aromatic Hydrocarbons (PAHs) by Food Products during Smoking. Sustainability, 15(24), 16890. https://doi.org/10.3390/su152416890
- Savin, R.-L., Ladoși, D., Ladoși, I., Păpuc, T., Becze, A., Cadar, O., Török, I., Simedru, D., Mariș, Ș. C., & Coroian, A. (2024). Influence of Fish Species and Wood Type on Polycyclic Aromatic Hydrocarbons Contamination in Smoked Fish Meat. Foods, 13(12), 1790. https://doi.org/10.3390/foods13121790
- Tabachnick, B. G., & Fidell, L. S. (2013). Using Multivariate Statistics (6th ed.). Boston, MA: Pearson.
- Zhang, X., Wang, X., Wu, F., Liang, W., Wang, S., Liang, J., Zhao, X., & Wu, F. (2024). Machine learning models to predict the bioaccessibility of parent and substituted polycyclic aromatic hydrocarbons (PAHs) in food: Impact on accurate health risk assessment. Journal of Hazardous Materials, 480, 136102. https://doi.org/10.1016/j.jhazmat.2024.136102