

## Article

# Multi-Output Machine-Learning Prediction of Volatile Organic Compounds (VOCs): Learning from Co-Emitted VOCs

Abdelrahman Eid <sup>1,2,\*</sup> , Shehdeh Jodeh <sup>3,\*</sup> , Ghadir Hanbali <sup>3</sup> , Mohammad Hawawreh <sup>2</sup>, Abdelkhaleq Chakir <sup>4</sup> and Estelle Roth <sup>4</sup> 

<sup>1</sup> Department of Mathematics, An-Najah National University, Nablus P.O. Box 7, Palestine

<sup>2</sup> Data Science Unit, An-Najah National University, Nablus P.O. Box 7, Palestine; m.hawawreh@najah.edu

<sup>3</sup> Department of Chemistry, An-Najah National University, Nablus P.O. Box 7, Palestine; g.hanbali@najah.edu

<sup>4</sup> Groupe de Spectrométrie Moléculaire et Atmosphérique GSMA, UMR CNRS 7331, Université de Reims, Moulin de la Housse B.P. 1039, CEDEX 02, 51687 Reims, France; abdel.chakir@univ-reims.fr (A.C.); estelle.roth@univ-reims.fr (E.R.)

\* Correspondence: abed.eid@najah.edu (A.E.); sjodeh@najah.edu (S.J.); Tel.: +970-598117416 (A.E.); +970-599590498 (S.J.)

## Abstract

Volatile Organic Compounds (VOCs) are important contributors to indoor and occupational air pollution, such as environments involving the extensive use of paints and solvents. The routine measurement of VOCs is often limited by resource constraints, creating a need for indirect estimation techniques. This work presents the need for a predictive framework that offers a practical, interpretable alternative to a full-spectrum chemical analysis and supports early exposure detection in resource-limited settings, contributing to environmental health monitoring and occupational risk assessment. This study explores the capability of machine learning to simultaneously predict the concentrations of five paint-related VOCs using other co-emitted VOCs along with demographic variables. Three models—Multi-Output Gaussian Process Regression (MOGP), CatBoost Multi-Output Regressor, and Multi-Output Neural Networks—were calibrated and each achieved a high predictive performance. Further, a feature importance analysis is conducted and showed that certain VOCs and some demographic variables consistently influenced the predictions across all models, pointing to common exposure determinants for individuals, regardless of their specific exposure setting. Additionally, a subgroup analysis identified the exposure disparities across demographic groups, supporting targeted risk mitigation efforts.

**Keywords:** multi-output regression; CatBoost; neural networks; explainable AI; environmental machine learning; volatile organic compounds (VOCs); indoor air quality; exposome; SDG 3—good health and well-being; SDG 11—sustainable cities and communities



Academic Editor: Peter Brimblecombe

Received: 19 May 2025

Revised: 19 June 2025

Accepted: 25 June 2025

Published: 26 June 2025

**Citation:** Eid, A.; Jodeh, S.; Hanbali, G.; Hawawreh, M.; Chakir, A.; Roth, E. Multi-Output Machine-Learning Prediction of Volatile Organic Compounds (VOCs): Learning from Co-Emitted VOCs. *Environments* **2025**, *12*, 216. <https://doi.org/10.3390/environments12070216>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Volatile Organic Compounds (VOCs) are a diverse group of carbon-based compounds that readily vaporize at room temperature [1]. Commonly emitted from paints, solvents, adhesives, and industrial materials, VOCs are prevalent in both indoor and outdoor environments [2,3]. Exposure to VOCs such as toluene, xylene, and isocyanates has been linked to adverse health outcomes, including respiratory irritation, neurological effects, and carcinogenesis [4,5]. In small-scale occupational settings like carpentry workshops, where ventilation is often inadequate, VOC accumulation poses a heightened risk to the workers and surrounding communities [6,7].

Even in the absence of direct chemical use, indoor environments, especially those with inadequate ventilation, can show elevated levels of VOCs due to their proximity to polluted microenvironments [8]. This problem is especially pronounced in confined spaces like carpentry workshops, where volatile emissions from products such as wood coatings, thinners, and adhesives tend to accumulate and linger, aggravated by a lack of proper ventilation and insufficient environmental monitoring [9].

Monitoring VOC exposure is essential for occupational safety but remains a technical and financial challenge. The standard method, active sampling, relies on mechanical pumps to collect air samples onto sorbent tubes for analysis via gas chromatography–mass spectrometry (GC-MS) [10]. Although highly accurate, this method is labor-intensive, expensive, and impractical for widespread or continuous monitoring.

Alternative approaches have emerged to address these limitations. Passive sampling techniques, which operate without pumps or electricity, allow for time-integrated exposure measurement and are well-suited for field use [10,11]. Biomonitoring using biological matrices such as hair, blood, or urine provides insights into internal and cumulative VOC exposure. However, these methods also present drawbacks, including the variability among individuals, invasive sampling procedures, and complex data interpretation [12].

The research consistently highlights the disparities in VOC exposure across occupational groups. For instance, Romieu et al. reported that service station workers in Mexico City had benzene exposures more than seven times higher than office workers, despite the low benzene concentrations in local fuels [13]. Elevated benzene levels in their blood samples reflected long-term internal accumulation, underscoring how exposure can be influenced by factors beyond direct chemical use, such as traffic-related pollution and ineffective emission controls [14–16].

Recent technological advances have improved the accessibility of exposure monitoring. Low-cost passive samplers and surrogate biological matrices are now helping to assess VOC exposure in resource-limited settings and underserved populations [17–20]. Nevertheless, comprehensive VOC profiling remains constrained by the need for specialized equipment and trained analysts, making it infeasible for routine large-scale assessments [21–23].

Notably, many VOCs originate from shared sources or show correlated behavior in the environment, creating opportunities for surrogate modeling, where harder-to-measure VOCs are estimated using more accessible ones [24]. In this context, machine learning (ML) has emerged as a powerful tool, enabling indirect predictions based on co-occurring compounds. While early applications of ML in environmental science focused on single-output predictions, recent advances in multi-output modeling offer improved accuracy by capturing the relationships between multiple compounds [25].

Among these, Multi-Output Gaussian Process Regression (MOGP) stands out for combining a high predictive performance with interpretability, a crucial feature for informing environmental policy and health risk communication [26]. Studies have demonstrated that ML models like Random Forest, LSSVM, and XGBoost can effectively predict VOC concentrations in indoor settings based on variables such as occupancy, temperature, and humidity [27,28]. Furthermore, ML and AI technologies are increasingly used in broader environmental monitoring applications, including air quality forecasting and pollution hotspot detection [29–32]. Integrating ML and AI into VOC prediction can significantly enhance environmental decision-making. Accurate, data-driven predictions of VOC concentrations enable health agencies and policymakers to assess risks, guide interventions, and develop targeted regulations [33,34]. These tools also optimize the allocation of monitoring resources, especially in economically disadvantaged or under-monitored regions [35,36].

This study proposes a machine-learning-based framework to predict the concentrations of five paint-related VOCs using co-occurring VOCs and demographic information. In

contrast to traditional single-pollutant models, this work applies a multi-output regression approach to capture compound interactions and enable joint predictions.

The objective is to evaluate the predictive performance of different ML models and explore the environmental implications of the most informative predictors. By offering a cost-effective and scalable alternative to conventional methods, this approach supports data-driven environmental health management in resource-constrained settings. In addition, the study interprets the most influential predictors through a feature importance analysis and investigates subgroup-specific exposure patterns by identifying potential high-risk groups. Together, these analyses enhance the practical relevance of the models for environmental health management and targeted risk mitigation.

## 2. Literature Review

Recent developments in machine learning (ML) have introduced new approaches to predicting multiple volatile organic compounds (VOCs) at the same time, offering a broader and more integrated approach to modeling air quality. Unlike traditional models that typically focus on one pollutant at a time, multi-output (multi-target) ML models take advantage of the correlations that often exist between co-emitted VOCs, allowing for better predictive performance and reduced processing complexity.

Masmoudi et al. (2020) proposed a multi-target regression framework that forecasts several air pollutants simultaneously [22]. Their approach used ensembles of regressor chains with random-forest base models and showed that accounting for dependencies among the output variables improves accuracy compared with modeling each pollutant separately [37]. Similarly, Ye et al. (2022) applied random forests, support-vector regression (SVR), and XGBoost to estimate the concentrations of multiple VOCs released in a pharmaceutical production line; they reported  $R^2$  values from 0.40 to 0.93 and highlighted the value of time-lagged variables for prediction [38].

Deep-learning models have also been explored. A convolutional neural network coupled with an optical-absorption sensor predicted benzene, toluene, ethylbenzene, and xylenes (BTEX) from composite spectral signals with  $R^2 > 0.96$  for all targets [39]. Kang et al. (2024) employed a multi-output neural network to predict several performance metrics of a VOC-removal system, demonstrating the ability of NNs to capture complex nonlinear dependencies between inputs and multiple VOC outputs [40].

Beyond sensor-level demonstrations, multi-output ML has also proved valuable in real-world indoor environments. Liu et al. (2023) used random forests, AdaBoost, XGBoost, and least-squares SVM to predict human-generated indoor VOCs (6-MHO and 4-OPA) from occupancy and environmental parameters, achieving mean absolute percentage errors below 5% [28]. Zhang et al. (2021) modeled emissions from coated wood furniture with an artificial neural network approach, obtaining mean prediction errors below 10% and outperforming traditional emission models [41].

Multi-output modeling offers clear advantages: it helps identify relationships among pollutants that share sources or participate in the same chemistry, and it enables the simultaneous prediction of several VOCs—useful for real-time monitoring and early-warning systems in industrial and urban settings. Ye et al. (2022) noted that such models can support more dynamic emission-control strategies in manufacturing [38]. In addition, ensemble methods and Gaussian-process models provide feature-importance measures and uncertainty estimates, adding interpretability crucial for health-risk assessment and policy-making [37,42,43].

Challenges remain. Multi-output models require large, diverse datasets and can be computationally demanding and sensitive to hyper-parameter tuning. Masmoudi et al. (2020) reported that their regressor-chain model needed considerable resources and was sensitive to extreme events unless the data were transformed appropriately [22]. Although interpretability tools such as SHAP values help explain individual predictors, complex models still lack the transparency of simpler regressions. Generalization is another issue: a model trained in one workshop or city may perform poorly in another setting with different sources and materials, unless it is retrained or adapted [28].

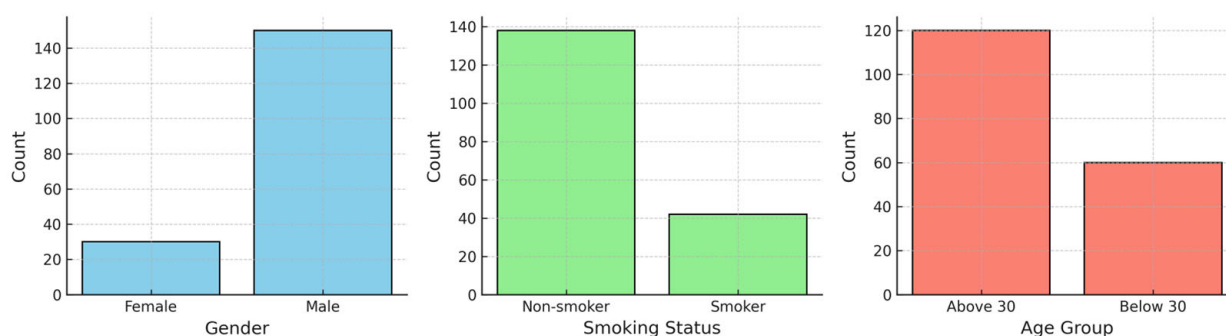
### 3. Materials and Methods

#### 3.1. Dataset

The dataset includes measurements of VOC concentrations obtained from blood samples collected from individuals located in or near a carpentry workshop. This setting is known for elevated levels of indoor pollutants due to the extensive use of paint-based and related materials. The data were originally gathered as part of a previous biomonitoring study [44]. The sampling effort aimed to capture variability in VOC exposure among individuals working in or living close to the workshop. A total of 180 participants were included in the study, consisting of both workshop employees and residents from the surrounding community. For each participant, a single blood sample was collected and analyzed using gas chromatography (GC) equipped with quadrupole mass spectrometry (MS), as outlined in detail in the original study [44].

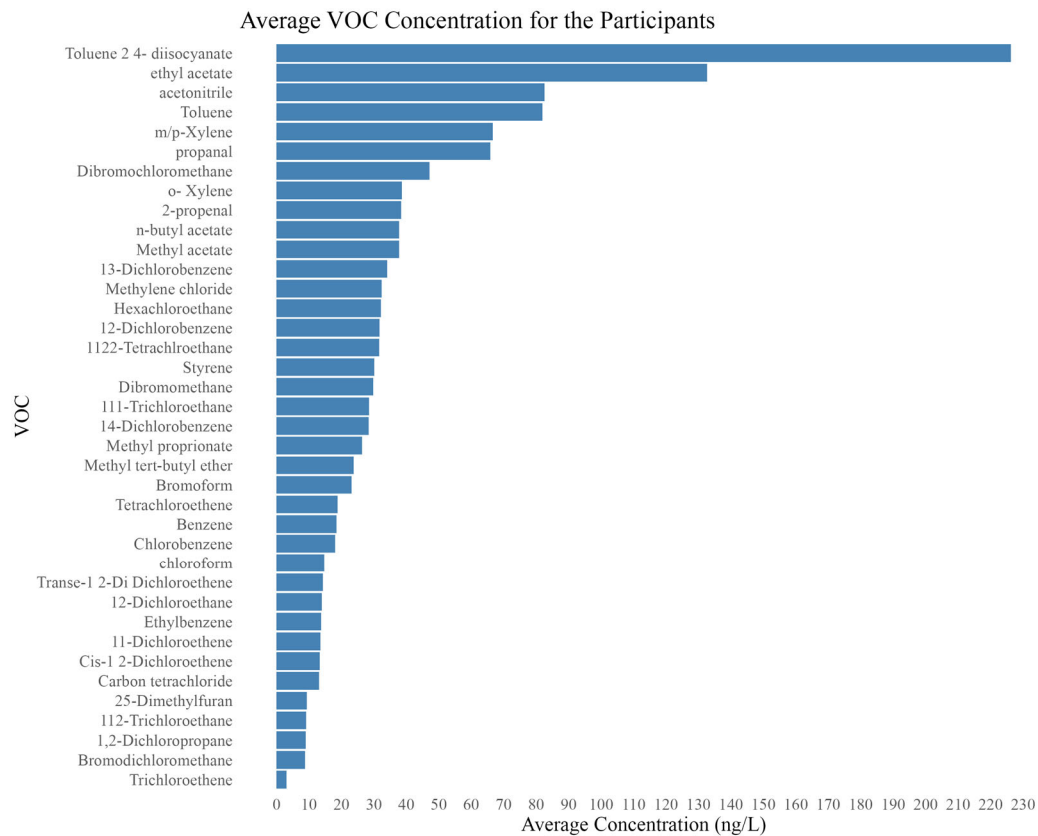
In total, 38 VOCs were identified and quantified. Additionally, demographic information including age, gender, and smoking status was recorded to assess potential influences on VOC levels. Each participant represents one observation in the dataset, resulting in 180 data entries. Each entry includes the concentrations of 38 VOCs and the 3 demographic variables, yielding 41 variables in total. These variables were used as input features or prediction targets, depending on the modeling approach. Figures 1 and 2 provide an overview of the participant characteristics and the general VOC exposure profile in the dataset.

**Participant Count by Gender, Smoking Status, Age**



**Figure 1.** The count of participants according to gender, smoking status, and age group.

The average concentration levels of the 38 quantified VOCs across all 180 participants are presented in Figure 2. This visualization shows the variability in VOC exposure and provides important context for the subsequent modeling and interpretation of multi-VOC predictions.

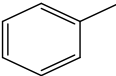
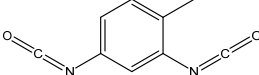
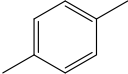
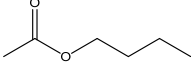



**Figure 2.** The average concentrations of the 38 measured VOCs across the 180 participants.

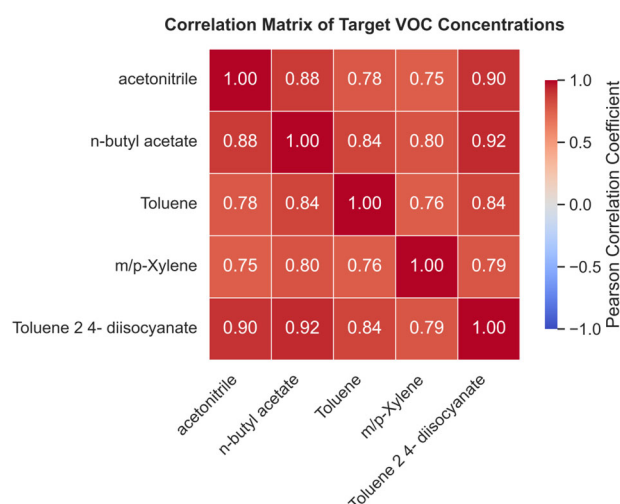
3.2. Target VOCs

The selected target VOCs, presented in Table 1, are significant components of paint-related emissions and were chosen due to their environmental persistence and potential health hazards. Toluene, a common solvent, is commonly linked with risks to the central nervous system with prolonged exposure [45]. Toluene 2,4-diisocyanate, widely utilized in polyurethane-based coatings, is a highly reactive compound known as a potent respiratory sensitizer and a major contributor to occupational asthma [46]. p-Xylene, frequently present in varnishes and fuels, can irritate mucous membranes and neurotoxic effects [47]. n-Butyl Acetate, found in lacquers, may lead to eye, skin, and respiratory irritation upon short-term exposure [48]. Acetonitrile, though less frequently monitored, is relevant due to its use in adhesives and its ability to metabolize into toxic cyanide compounds [49].

**Table 1.** Chemical structures and corresponding CAS numbers of selected paint-related VOCs.

Compound	Structure	CAS Number
Toluene		108-88-3
Toluene 2,4-diisocyanate		584-84-9
p-Xylene		106-42-3
n-Butyl Acetate		123-86-4
Acetonitrile		75-05-8

As seen in Figure 3, a strong positive association among all five target VOCs. Such high inter-correlations confirm that these compounds are typically co-emitted during paint application and curing processes and subsequently co-accumulated in exposed individuals. This covariance justifies the use of multi-output learners that explicitly model cross-target dependence, because information gained from one compound can help refine predictions for the others. Considering these shared exposure signatures is, therefore, expected to improve accuracy, enhance the interpretability of common emission sources, and support more reliable screening of occupational and environmental health risks.



**Figure 3.** The correlation heatmap of the five target paint-related VOCs.

### 3.3. Predictors

The predictors combine the three demographic variables (age, gender, and smoking status) with 33 co-occurring VOCs that typify a carpentry workshop where spray paints, lacquers, cleaning solvents, and adhesives are used daily. These compounds span light nitriles and aldehydes, oxygenated solvents such as ethyl and n-butyl acetate, high-volatility aromatics like benzene, toluene, and the xylene isomers, and numerous chlorinated species that originate from paint thinners and degreasers. Because many of these chemicals evaporate together during sanding, spraying, and drying, their indoor air profiles are strongly correlated, so considering this natural covariance allows the multi-output models to borrow strength across targets and predict the target VOCs, at the same time, more accurately than treating each in isolation. The chemical structures and systematic names, as well as the corresponding CAS numbers of the 33 VOCs, are provided in Appendix A.

### 3.4. ML Models

Selecting suitable machine-learning models is very important for predicting multiple correlated outputs in air quality and environmental datasets. In this research, Multi-Output Gaussian Processes (MOGP), multi-output Neural Networks (NNs), and CatBoost were selected to predict five VOCs simultaneously. MOGP was considered when modeling the correlations between the multiple outputs. Neural networks were employed due to their flexibility in capturing complex nonlinear relationships while considering output correlations through shared hidden layers. CatBoost is preferred over traditional boosting methods such as XGBoost because of its ability to handle the categorical features in our dataset. Further, this method is highly effective for structured datasets such as VOC measurements. Moreover, CatBoost's use of ordered boosting reduces overfitting, which enhances the model stability. Finally, CatBoost was adapted in this work for multi-output prediction by employing a MultiOutputRegressor wrapper, allowing the simultaneous



prediction of the five VOCs. Although CatBoost internally fits one model per output, this approach enables handling all outputs together within a unified training and prediction framework while preserving the model's strong performance on categorical and numerical inputs. All machine learning models were implemented using Python 3.10.6. CatBoost version 1.2.2 and GPyTorch version 1.10 were used for CatBoost and MOGP, respectively.

### 3.4.1. Multi-Output Gaussian Process Regression (MOGP)

Gaussian Process (GP) regression is a non-parametric Bayesian approach for modeling and predicting unknown functions [50]. In the standard setting, GP regression models a single-output variable by assuming that the function values at any set of input points are jointly Gaussian distributed. Given a set of inputs  $X = \{x_i\}_{i=1}^n$  and corresponding scalar outputs  $y = \{y_i\}_{i=1}^n$ , where each  $y_i \in \mathbb{R}$ , a GP defines the following:

$$y(x) \sim GP(m(x), K(x, x'))$$

where  $m(x)$  is the mean function, often assumed to be zero, and  $k(x, x')$  is the covariance function (kernel) that measures the similarity between input points.

However, many real-world problems involve multiple correlated outputs that should be predicted together. Modeling each output separately using independent GPs ignores the relationships between outputs and may lead to suboptimal predictions.

To address this, Multi-Output Gaussian Processes (MOGPs) extend the standard GP framework to jointly model multiple outputs. In the MOGP setting, for vector-valued outputs  $Y = \{y_i\}_{i=1}^n$  where  $y_i \in \mathbb{R}^D$ , the assumption becomes the following:

$$y(x) \sim GP(m(x), K(x, x'))$$

where  $m(x)$  is a vector-valued mean function and  $K(x, x')$  is a matrix-valued covariance function that captures both input similarities and output correlations.

One widely used form for  $K(x, x')$  is the Linear Model of Coregionalization (LMC), where

$$K(x, x') = \sum_{q=1}^Q B_q k_q(x, x')$$

where  $B_q$  is a positive semi-definite matrix modeling the relationships between outputs, and  $k_q(x, x')$  are standard kernels over the inputs [51,52].

MOGP can improve prediction accuracy by modeling multiple outputs jointly. In this work, we apply MOGP to predict correlated VOCs to evaluate its ability to exploit inter-task correlations for better generalization.

### 3.4.2. Neural Network Multi-Output Regression

Artificial Neural Networks (ANNs) are a class of machine-learning models designed to approximate complex nonlinear functions through layers of interconnected neurons. Their expressive power enables them to represent a wide range of functional relationships between inputs and outputs [53].

In the context of multi-output prediction, neural networks can be adapted to simultaneously forecast multiple target variables. This is achieved by designing a shared architecture where the hidden layers learn common patterns across tasks, while each output neuron corresponds to a distinct prediction. Mathematically, the output vector  $y$  can be described as follows:

$$y = f(x; \theta)$$

where  $x$  is the input vector,  $f(\cdot)$  is the neural network function, and  $\theta$  are the learnable model parameters such as weights and biases.

The standard operation of a fully connected feedforward layer is given by the following:

$$h^{(l)} = \sigma(W^{(l)}h^{(l-1)} + b^{(l)})$$

where  $h^{(l)}$  is the activation at layer  $l$ ,  $\sigma$  is a nonlinear activation function, and  $W^{(l)}$  and  $b^{(l)}$  are the weight matrix and bias vector at layer  $l$ .

In the multi-output design, the final layer produces a vector output  $y \in R^D$ , where  $D$  is the number of target variables to predict.

In previous work, the spatiotemporal neural network architecture proposed in [54] demonstrated significant improvements in forecasting multiple air pollutants simultaneously compared to single-output models. Similarly, for this research, we employ a multi-output neural network to predict correlated VOC levels based on input features in order to study the ability of neural networks to capture complex dependencies among multiple outputs and compare it to other ML models.

### 3.4.3. CatBoost Multi-Output Regression

This is a gradient-boosting algorithm developed to provide high performance with minimal need for extensive data preprocessing [55]. Unlike many traditional boosting methods, such as Gradient Boosting Machines [56], CatBoost is specifically designed to handle categorical features natively without requiring manual encoding, making it particularly attractive for structured datasets commonly used in scientific and industrial applications.

The general structure of a boosted model in CatBoost follows an additive formulation:

$$F(x) = \sum_{m=1}^M \gamma_m h_m(x)$$

where  $h_m(x)$  are decision trees and  $\gamma_m$  are their associated weights.

CatBoost's main innovation, ordered boosting, modifies the standard boosting process to prevent overfitting by calculating residuals without introducing target leakage. It incrementally builds models using only information available at earlier stages, leading to better generalization.

In this work, CatBoost is employed to predict multiple correlated VOCs individually as a strong baseline method. Although it does not explicitly model the correlations among outputs like MOGP, CatBoost provides robust and highly accurate single-task predictions. Comparing CatBoost results with MOGP predictions allows for assessing the practical benefits of modeling output correlations in multi-output regression settings.

## 3.5. Assessment of Predictive Performance

The accurate evaluation of the used prediction models in this study is an important step to confirm their reliability in forecasting VOC concentrations. In this study, three standard performance metrics were used to assess model accuracy: the coefficient of determination ( $R^2$ ), the root mean square error (RMSE), and the mean absolute error (MAE).

### 3.5.1. Coefficient of Determination ( $R^2$ )

It measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated using the following formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$



where  $SS_{res}$  is the sum of squares of residuals, and is the total sum of squares.  $R^2$  provides an indication of goodness of fit and is commonly used to evaluate model accuracy in prediction.

### 3.5.2. Root Mean Square Error (RMSE)

It measures the square root of the average squared differences between predicted and actual values. It is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum (y_{pred} - y_{actual})^2}$$

where  $y_{actual}$  is the observed value of VOC from the dataset, and  $y_{pred}$  is the predicted value generated by the model.  $RMSE$  penalizes large errors more heavily and is used to evaluate the model's prediction accuracy, with lower values indicating better performance.

### 3.5.3. Mean Absolute Error (MAE)

It calculates the average magnitude of errors in a set of predictions, without considering their direction. It is calculated as follows:

$$MAE = \frac{1}{n} \sum |y_{pred} - y_{actual}|$$

where  $y_{actual}$  is the observed value of VOC from the dataset, and  $y_{pred}$  is the predicted value generated by the model.  $MAE$  is a straightforward metric to interpret and is used to measure how far predictions are from actual values on average.

## 3.6. Feature Importance

Understanding the importance of each input VOC for prediction is essential for interpreting machine-learning models and understanding the relationships within the data. In this work, feature importance analysis is conducted to identify which VOCs and categorical variables most strongly influenced the prediction of the five target VOCs. Interpreting feature contributions is particularly important in environmental modeling, as it helps to identify potential interactions between pollutants, detect the predictive VOCs, and support the development of more effective monitoring and control strategies.

Different methods are applied in this work to assess feature importance, including native importance, permutation importance, and ARD kernel analysis. For the Multi-Output Gaussian Process (MOGP) model, the Automatic Relevance Determination (ARD) kernel parameters were analyzed, where lower-length-scale values correspond to more influential features [57]. Permutation importance was used for the neural network model by measuring the decrease in model performance when individual features were randomly shuffled, indicating their influence on predictions [58]. Native feature importance was obtained from the CatBoost model using its built-in method, which evaluates the average contribution of each feature to the predictive performance [55].

## 4. Results

### 4.1. Evaluation of the Ability of Machine-Learning Models to Predict Multiple VOCs Simultaneously

The three proposed machine-learning models were employed to predict the concentrations of five target VOCs simultaneously. Each model was trained using 33 VOC features and 3 categorical variables as inputs. For MOGP, a Linear Coregionalization Model (LCM) was combined with ARD-RBF kernels to capture both the correlations between outputs and the feature relevance. Additionally, for CatBoost, the MultiOutputRegressor wrapper from the scikit-learn library [59] was used to adapt CatBoostRegressor for multi-target

prediction, enabling the simultaneous modeling of five VOC outputs. The CatBoost implementation was based on the CatBoost open-source library [55]. Key hyperparameters included 200 iterations, a learning rate of 0.05, a maximum tree depth of 6, and the use of the *RMSE* loss function. While the *RMSE* loss function was specified, CatBoost internally minimizes the Mean Squared Error (MSE) during training and reports the square root of the loss for interpretability. Finally, for the neural network, a fully connected architecture was applied with two hidden layers (64 and 32 units), ReLU activations, Adam optimizer, and early stopping to prevent overfitting.

The models were evaluated based on the five-fold cross-validation for CatBoost and MOGP, and holdout validation for the neural network. The table below presents the model performance using  $R^2$ , *RMSE*, and MAE assessment metrics.

Table 2 presents a detailed evaluation using the model performance metrics ( $R^2$ , *RMSE*, and MAE) for each individual target VOC, along with the average values across the five VOCs. The predictive performance of the three models presented in Table 2 indicates that all models achieved high  $R^2$  values, which reflects a strong predictive capability. CatBoost achieved the highest  $R^2$ , and also achieved the lowest *RMSE* (3.4788), as well as the lowest MAE (2.4726), which demonstrates a superior predictive accuracy and a minimal average error.

**Table 2.** Model evaluation results for multi-output VOC prediction using ML.

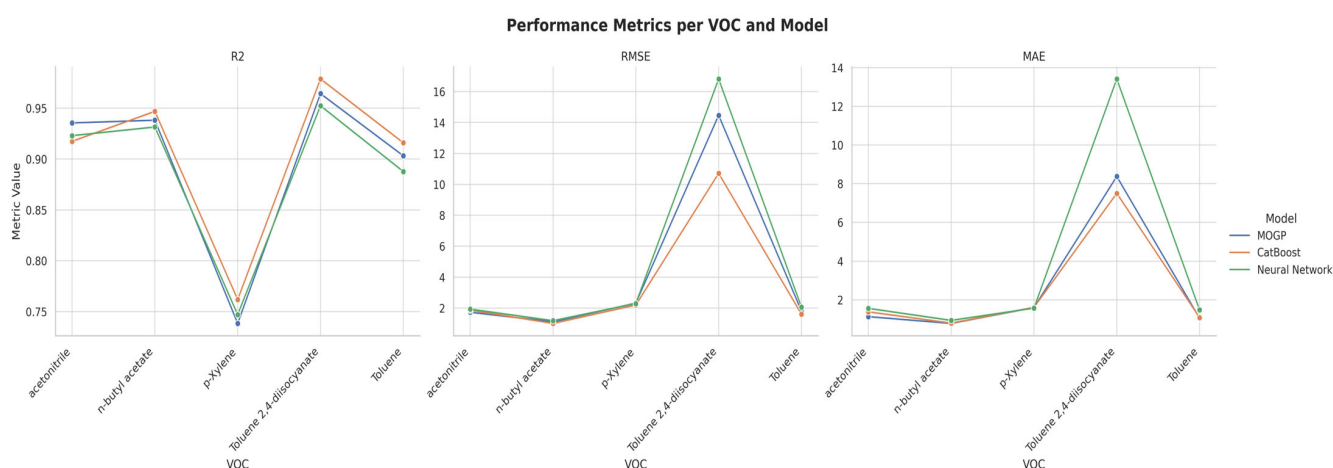
Model	VOC	$R^2$	<i>RMSE</i>	MAE
MOGP	acetonitrile	0.9355	1.7313	1.1284
	n-butyl acetate	0.9382	1.1103	0.7778
	p-Xylene	0.7383	2.3193	1.5984
	Toluene 2,4-diisocyanate	0.9643	14.4662	8.3748
	Toluene	0.9032	1.805	1.0428
	Average	0.8959	4.2864	2.5844
CatBoost	acetonitrile	0.9174	1.87	1.3776
	n-butyl acetate	0.9469	1.0066	0.7956
	p-Xylene	0.7618	2.1983	1.6079
	Toluene 2,4-diisocyanate	0.9788	10.7132	7.5046
	Toluene	0.916	1.606	1.0772
	Average	0.9042	3.4788	2.4726
Neural Network	acetonitrile	0.923	1.9324	1.5573
	n-butyl acetate	0.9315	1.1845	0.9353
	p-Xylene	0.747	2.297	1.5694
	Toluene 2,4-diisocyanate	0.9524	16.8162	13.4165
	Toluene	0.8877	2.0589	1.4734
	Average	0.8883	4.8578	3.7904

The result of Multi-Output Gaussian Process (MOGP) supports its ability to model multiple outputs simultaneously with good generalization. The Multi-Output Neural Network achieved a very close but weaker result. This could be justified by the model's sensitivity to data size and structure, as deep-learning models typically require larger datasets for optimal performance.

Across the five target VOCs, the models generally achieved a strong predictive accuracy for compounds such as Toluene 2,4-diisocyanate and n-butyl acetate, which consistently yielded a high  $R^2$  and low error metrics. In contrast, p-Xylene showed a lower  $R^2$  and slightly higher prediction errors across all models, indicating greater variability or more complex relationships with the input features. These differences highlight that some VOCs are more predictable based on the available co-occurring variables, while others may require additional contextual or environmental data to improve the prediction performance. These patterns may also reflect the characteristics of the study environment, where compounds such as Toluene 2,4-diisocyanate and n-butyl acetate are directly linked

to paint-related activities in the carpentry workshop, resulting in more consistent exposure profiles, while compounds like p-Xylene may be influenced by a broader range of indoor and outdoor sources affecting both workers and nearby residents.

Figure 4 illustrates how the three ML models performed across the five target VOCs. As observed, Toluene 2,4-diisocyanate and n-butyl acetate were consistently well-predicted by all models. Conversely, p-Xylene exhibited a lower  $R^2$  and higher error values across all models, confirming its relatively more challenging predictability. The plot also confirms that, while the models performed well overall, the prediction accuracy varied by VOC, underscoring the value of multi-output modeling approaches that can account for such variability. This visualization provides a clear, VOC-specific view of the models' predictive ability.



**Figure 4.** Performance of ML models on each of the five target VOCs.

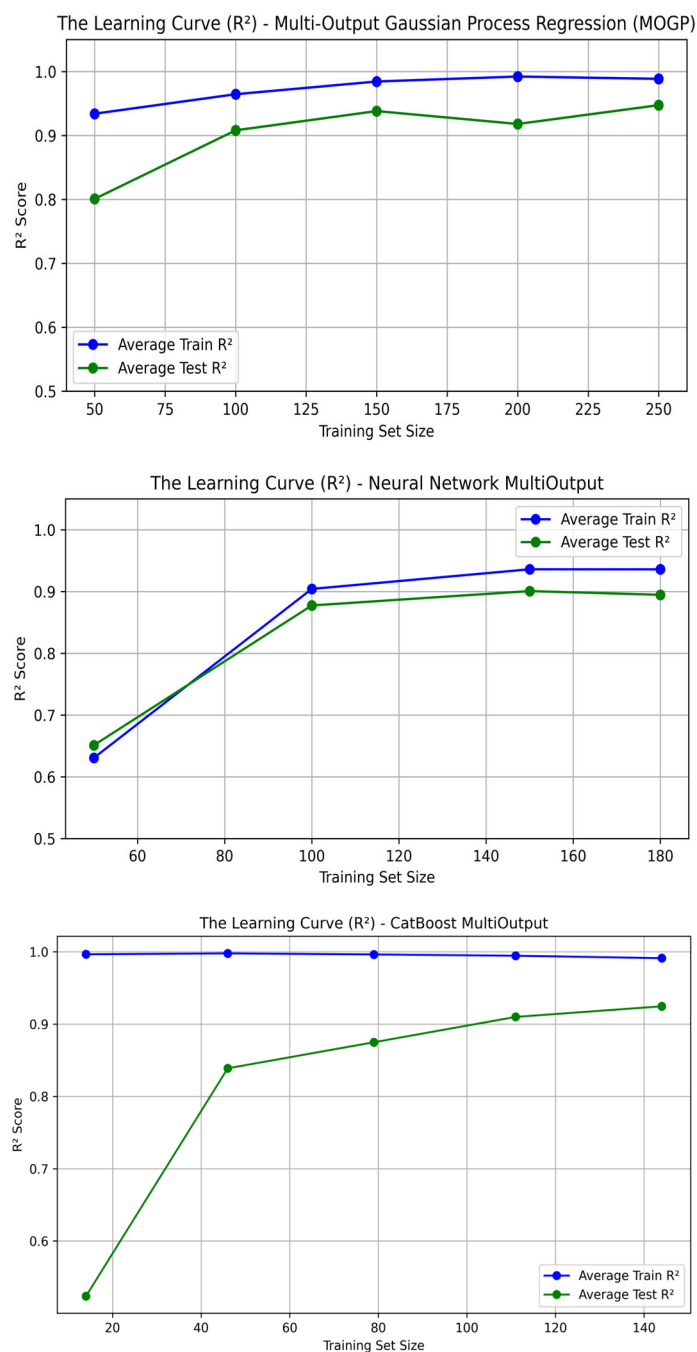
### Overfitting Evaluation

In this step, we verify that our machine-learning models do not suffer from overfitting to ensure their ability to generalize well to unseen VOC data. It is known that overfitting occurs when a model learns the training VOC data too closely, leading to poor performance on new samples. To assess possible overfitting, learning curves were built and analyzed for each model by plotting the training and test  $R^2$  scores as a function of the training set size. A consistent and stable convergence of training and test curves, in general, indicates a good generalization of the model, while large gaps or instability between the two curves may indicate overfitting or underfitting. Figure 1 presents the learning curves used to evaluate the models in order to study the generalization behavior of them.

Based on Figure 5, as the training set size increases, the test  $R^2$  scores improve and stabilize, closely approaching the training  $R^2$  scores. This indicates that all models achieved good generalization without significant overfitting.

### 4.2. Results of Feature Importance Analysis

For each of the machine-learning models, the feature importance values were averaged across the five VOC outputs to obtain a global ranking of the predictors. This allows us to determine the most influential VOCs and categorical variables driving the multi-output prediction, which identifies the pollutant relationships and provides recommendations for future monitoring and control strategies.



**Figure 5.** Learning curves for the MOGP, Neural Network, and CatBoost models, respectively from left to right.

Notably, while several co-occurring VOCs, such as Trichloroethene, o-Xylene, 1,2-Dichlorobenzene, and 2,5-Dimethylfuran, were consistently identified as important predictors across all models, Figure 6 shows that age emerged as the most influential feature overall. This observation is particularly relevant given that the VOC concentrations in this study were measured from blood samples, representing internal exposure. Unlike environmental air monitoring, blood-based biomonitoring reflects both external exposure and individual-level physiological and behavioral factors that influence the absorption, distribution, metabolism, and excretion (ADME) of VOCs. Age is known to impact these processes through changes in metabolic enzyme activity, body fat composition, and organ function, potentially resulting in a slower VOC clearance or greater accumulation in older individuals. Moreover, smoking status also played a significant role for the CatBoost

predictive model, likely due to the presence of numerous VOCs in tobacco smoke, which can elevate the baseline internal VOC levels. These findings suggest that demographic characteristics, particularly, age, may have a stronger influence on blood VOC concentrations than environmental co-exposures alone. This highlights the importance of considering individual biological and lifestyle factors when interpreting VOC exposure models based on internal biomarkers.

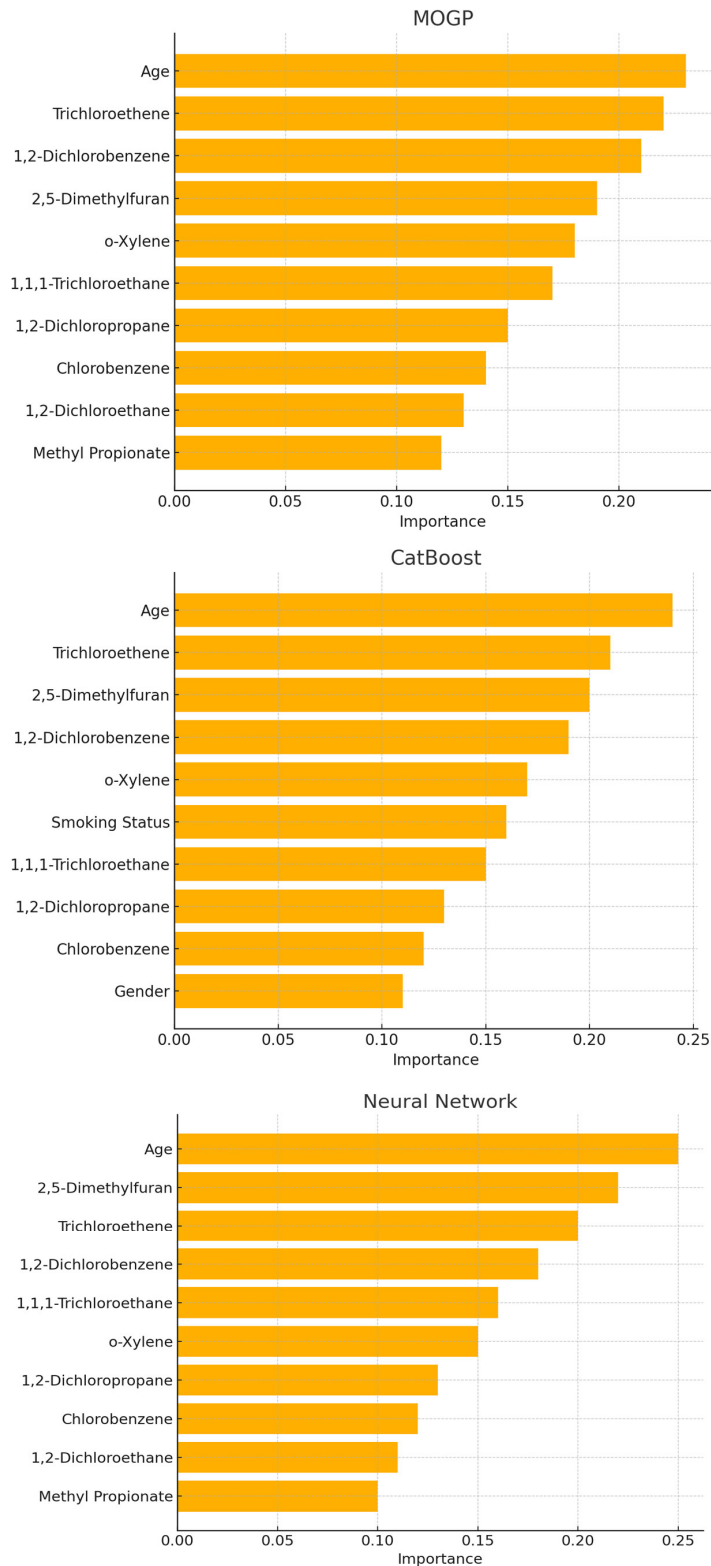


Figure 6. Top 10 influential features identified by the three machine-learning models.

Moreover, the consistency in identifying certain VOCs as influential predictors across all three machine-learning models can be attributed to the characteristics of the collected data and the study environment. Although not all of the measured 38 VOCs originated directly from paint emissions, the environmental setting, the carpentry workshop, where paint-related activities dominate, created a strong exposure signature. VOCs such as Trichloroethene, o-Xylene, 1,2-Dichlorobenzene, and 2,5-Dimethylfuran, which are commonly associated with paint fumes and industrial solvents, appeared consistently among the top influential features because they represent the core components of the air quality profile in and around the carpentry workshop.

Furthermore, the fact that the data were collected from two distinct but environmentally linked groups, workers inside the carpentry workshop and nearby residents, reinforces this pattern. The workers, who were all adult males above 15 years old, experienced continuous and direct exposure to paint-related VOCs. Meanwhile, residents included males and females, smokers and non-smokers, and individuals both younger and older than 15 years. Despite this demographic variability, the proximity to the carpentry workshop ensured that paint-related VOCs remained dominant in shaping the VOC exposure profiles of both groups. As a result, even features not exclusively emitted from paints still co-varied with paint-related compounds due to the shared environmental conditions.

This environmental and demographic structure explains why the same VOCs were consistently important for predicting the five target VOCs. It also shows the real-world complexity captured by the models. Among all predictors, age consistently emerged as the most influential factor across the three ML models. Additionally, smoking status and gender were also ranked among the 10 important predictors, particularly in the CatBoost model, which achieved the best overall performance. These findings highlight that demographic characteristics, especially age, played a major role in shaping blood VOC concentrations, likely reflecting the differences in occupational exposure between groups.

#### *4.3. Identification of High-Risk Groups*

Understanding how VOC exposure varies across population subgroups is important for informing targeted risk mitigation strategies and occupational health interventions. While Section 4.2 determines the most influential VOCs and categorical variables driving the multi-output prediction overall, this section directly addresses an important gap. It provides clear descriptive insights into which participant groups may experience higher predicted VOC exposure levels. This is essential in order to explore the social usefulness of the models by identifying potential high-risk groups. The objective of this section is to stratify and compare the predicted concentrations of the five target VOCs across the participant characteristics: gender, smoking status, and age group ( $\leq 15$  years vs.  $>15$  years). This subgroup analysis complements the model interpretability results and supports the identification of groups that may benefit from targeted exposure reduction measures.

To achieve this, we used the predicted VOC concentrations obtained from the CatBoost model, which demonstrated the best overall predictive performance in this study based on five-fold cross-validation, as the five-fold cross-validation ensures that each of the 180 participants receives a predicted value from the fold where they were included in the validation set. Accordingly, this analysis uses the complete set of 180 predicted VOC values for each participant. Following, for each subgroup (male vs. female, smoker vs. non-smoker, and age  $\leq 15$  vs.  $>15$ ), we calculated and compared the mean and standard deviation of the predicted concentrations for each of the five target VOCs. The results of this stratified analysis using the CatBoost model are presented in Table 3 below. The results obtained using the other two ML models in this study are presented in Appendix A.



**Table 3.** Comparison of mean and standard deviation of predicted concentrations (ng/L) for the five target VOCs stratified by age group, gender, and smoking status based in CatBoost model.

VOC	Age				Gender				Smoking Status			
	$\leq 15$		$>15$		Male		Female		Non-Smoker		Smoker	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
acetonitrile	48.74	1.24	61.18	3.36	56.74	7.07	58.51	1.82	55.14	6.19	63.24	2.49
n-butyl acetate	32.37	0.67	40.46	2.42	37.65	4.71	38.33	0.92	36.37	3.91	42.35	1.49
Toluene	76.27	0.972	84.86	5.25	82.45	6.39	79.73	0.80	79.65	4.10	89.71	4.23
p-Xylene	62.16	1.81	68.76	2.74	66.60	4.29	66.34	1.69	65.19	3.35	71.07	2.07
Toluene 2												
4-diisocyanate	127.14	7.53	274.06	37.36	222.34	82.66	238.86	16.61	199.56	67.75	308.97	21.22

According to Table 3, the results of the predicted VOC concentrations show how paint-related VOC exposures vary across population subgroups within the study population. Importantly, since all five target VOCs originate predominantly from paint emissions used in the carpentry workshop, the differences observed across age, gender, and smoking status reflect the combined effects of occupational exposure, environmental proximity to the source, and individual characteristics.

The most prominent finding is the clear differentiation by age group. For all five VOCs, participants older than 15 years showed higher predicted concentrations compared to those aged 15 or younger. This pattern is fully consistent with the context of the study: all workshop employees are male adults over 15 years old who experience regular and direct exposure to paint-related VOCs through their work activities inside the carpentry workshop. The elevated concentrations among this group strongly reflect their occupational role as the primary factor driving exposure. Smoking status also showed a clear and consistent pattern that aligns well with the study environment. Across multiple VOCs, smokers exhibited higher predicted concentrations than non-smokers. This outcome reflects the population structure of the study: the majority of smokers were male workers employed inside the carpentry workshop, where they experienced regular and direct exposure to paint-related emissions. In contrast, most non-smokers were residents living near the workshop, including women and younger individuals, whose exposure was largely indirect and lower in magnitude. These results highlight that occupational factors, rather than smoking behavior per se, were the primary driver of elevated VOC exposures in this group. This reinforces the broader observation that individuals working directly with paint-based materials in such small-scale environments represent a particularly high-risk subgroup requiring targeted occupational health interventions.

When considering the standard deviation (SD) values across subgroups, an important additional insight emerges for age groups. Across all five target VOCs, participants older than 15 years not only exhibited higher mean concentrations but also higher SD values compared to those aged 15 or younger. This suggests a greater variability of exposure within the  $>15$  group, likely reflecting the differences in individual work tasks, duration of exposure, and varying intensity of occupational activities inside the carpentry workshop. In contrast, for gender and smoking status, the SD patterns are more VOC-specific and less consistent across all compounds. This indicates that, while gender and smoking status contribute to exposure differences, the degree of variability within these subgroups depends on the particular VOC, possibly reflecting individual behavioral patterns, room occupancy, and personal proximity to emission sources.

#### 4.4. Study Limitations

While this study demonstrates the feasibility and scientific value of applying multi-output machine-learning models to VOC exposure prediction, certain limitations must

be acknowledged. First, the dataset was collected from a single occupational setting, a carpentry workshop and its surrounding area, which may limit the generalizability of the models to other environments with different VOC profiles or pollution sources. Second, the relatively small sample size, 180 participants, may constrain the ability of machine-learning and deep-learning models to fully capture the complex nonlinear relationships. Future studies with larger and more diverse datasets could further enhance model robustness. Third, although the multi-output approach effectively leverages the co-occurrence patterns of VOCs, causality between the predictor and target compounds cannot be established from this modeling framework. Finally, the demographic variables included in this study (age, gender, and smoking status) offer only a limited representation of personal factors that may influence VOC exposure. Incorporating additional behavioral covariates (such as time spent indoors versus outdoors, use of personal protective equipment, and ventilation practices), occupational covariates (such as job role, duration and frequency of exposure to paints and solvents, and proximity to emission sources), and environmental covariates (such as room ventilation rate, temperature, humidity, and background outdoor pollution levels) could provide a more comprehensive understanding of the factors driving individual VOC exposure. Including such variables in future studies may improve the model accuracy and help better identify high-risk groups. Addressing these limitations in future work will further improve the applicability and impact of machine-learning models for VOC exposure assessment.

## 5. Conclusions

This study demonstrates that using multi-output regression ML models provides a practical approach for predicting the concentrations of several VOCs based on co-occurring compounds and basic participant characteristics. Through the application of MOGP, Cat-Boost, and Neural Network models, the prediction accuracy was found to be consistently high with all models. A learning curve analysis further confirmed that the models generalized well to unseen data indicating the robustness of the developed predictive frameworks.

A feature importance analysis shows that several VOCs consistently played a dominant role in predicting the target VOCs across all models, which confirms the stability of the identified predictors and reinforces the potential of leveraging a reduced set of environmental and demographic variables for accurate exposure assessment.

The results of this work support the feasibility and scientific value of predicting difficult-to-monitor paint-related VOCs using readily available environmental and participant data. Although the predictor VOCs may originate from different sources, their real-world co-occurrence patterns were effectively captured and utilized, enabling the development of models that can serve as proxies for direct measurements. This offers advantages for the early detection of hazardous exposures and for informing risk management practices, especially in small-scale occupational environments where full chemical monitoring may not be practical.

Moreover, this work addresses a gap in the existing VOC modeling literature by employing a multi-output prediction strategy rather than focusing on single-pollutant models. Further, the incorporation of model interpretability through a feature importance analysis strengthens the applicability of the findings for real-world decision-making and policy development. This research provides a useful solution to the challenge of VOC exposure assessment. It reduces the reliance on expensive and labor-intensive chemical monitoring, offers early warning capabilities, and empowers health and environmental safety officers to make informed decisions. Indeed, this work is particularly helpful for developing regions, where resources for comprehensive air quality monitoring are often limited, and it contributes meaningfully to both the environmental sciences field and the

advancement of applied machine learning for air quality management. Finally, the observed influence of participant characteristics on VOC exposure, as demonstrated by both the feature importance and subgroup analyses, supports the social usefulness of the proposed models for identifying higher-risk groups and guiding occupational health intervention.

**Author Contributions:** A.E. and S.J. contributed to the conceptualization; A.E. contributed to the methodology; A.C. and E.R. contributed to the validation; A.E. and M.H. contributed to the formal analysis and writing; and S.J. and G.H. contributed to the writing, review, and editing. All authors have read and agreed to the published version of the manuscript.

**Funding:** The Al-Maqdisi project (Chargée de mission universitaire et scientifique Consulat Général de France à Jérusalem-SCAC) and the Palestinian Ministry of Higher Education (Grant No. 01/2021) provided financial support for this work.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Acknowledgments:** The authors gratefully acknowledge the financial support from the PHC Al-Maqdisi program and the Palestinian Ministry of Higher Education. We also thank An-Najah National University (Nablus, Palestine) and the University of Reims Champagne-Ardenne (Reims, France) for their valuable logistical and technical assistance.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A

**Table A1.** Analytical parameters for the determination of selected VOCS [44].

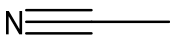
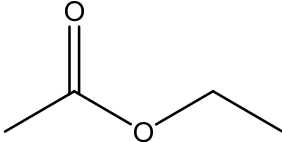
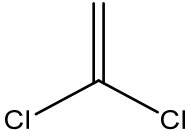
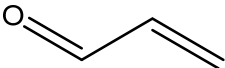
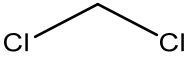
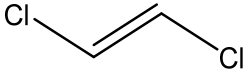

Analyte	Structure	CAS Number
acetonitrile		75-05-8
ethyl acetate		141-78-6
1,1-Dichloroethene		75-35-4
2-propenal		107-02-8
Methylene chloride		75-09-2
Transe-1,2-Di Dichloroethene		156-60-5
propanal		123-38-6

Table A1. Cont.

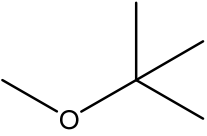
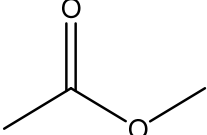
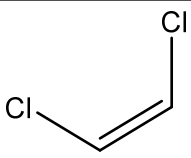
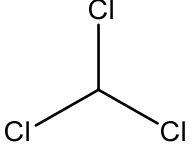

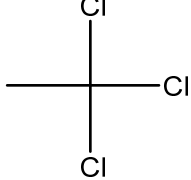
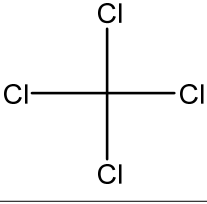
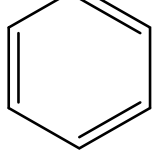
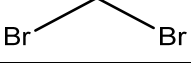
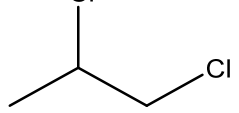
Analyte	Structure	CAS Number
Methyl tert-butyl ether		1634-04-4
Methyl acetate		79-20-9
cis-1,2-Dichloroethene		156-59-2
chloroform		67-66-3
1,2-Dichloroethane		107-06-2
1,1,1-Trichloroethane		71-55-6
Carbon tetrachloride		56-23-5
Benzene		71-43-2
Dibromomethane		74-95-3
1,2-Dichloropropane		78-87-5

Table A1. Cont.

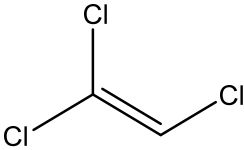
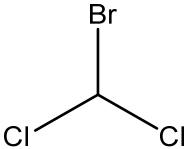
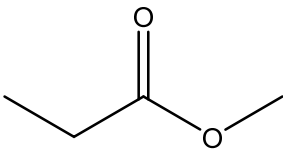
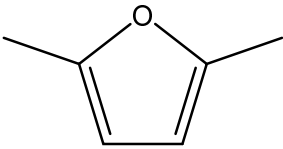
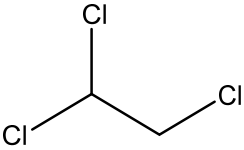
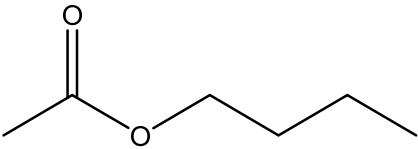
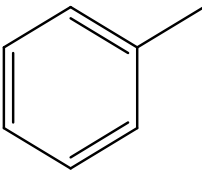
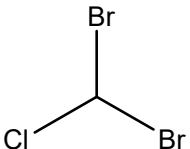
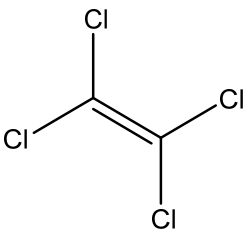
Analyte	Structure	CAS Number
Trichloroethene		79-01-6
Bromodichloromethane		75-27-4
Methyl propionate		554-12-1
2,5-Dimethylfuran		625-86-5
1,1,2-Trichloroethane		79-00-5
n-butyl acetate		123-86-4
Toluene		108-88-3
Dibromochloromethane		124-48-1
Tetrachloroethene		127-18-4

Table A1. Cont.

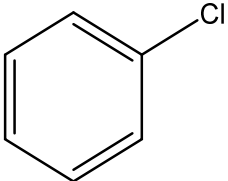
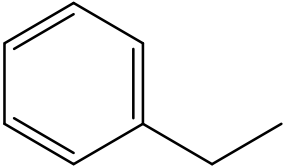
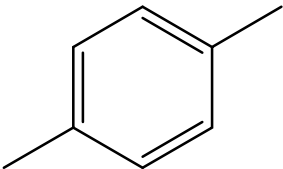
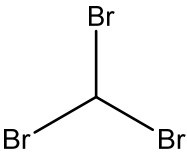
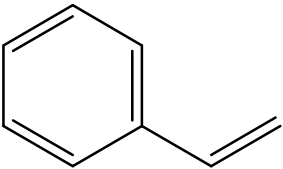
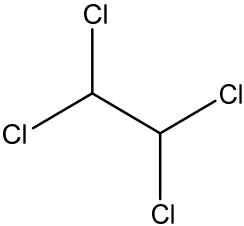
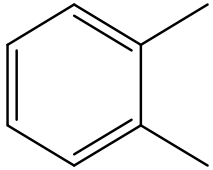
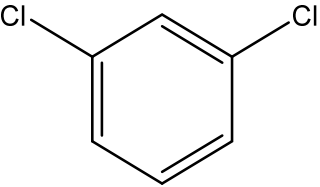
Analyte	Structure	CAS Number
Chlorobenzene		108-90-7
Ethylbenzene		100-41-4
p-Xylene		179601-23-1
Bromoform		75-25-2
Styrene		100-42-5
1,1,2,2-Tetrachloroethane		79-34-5
o- Xylene		95-47-6
1,3-Dichlorobenzene		541-73-1



Table A1. Cont.

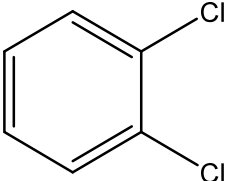
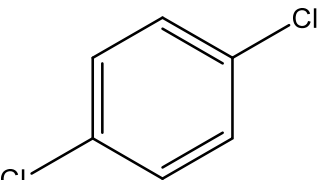
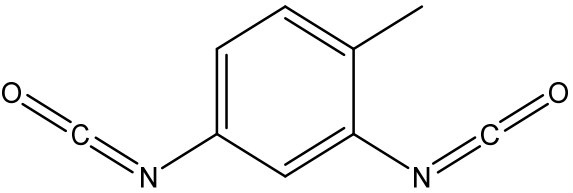
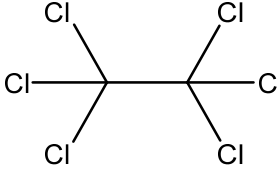
Analyte	Structure	CAS Number
1,2-Dichlorobenzene		95-50-1
1,4-Dichlorobenzene		106-46-7
toluene 2,4- diisocyanate		584-84-9
Hexachloroethane		67-72-1

Table A2. Comparison of mean and standard deviation of predicted concentrations (ng/L) for the five target VOCs stratified by age group, gender, and smoking status based in MOGP model.

VOC	Age				Gender				Smoking Status			
	≤15		>15		Male		Female		Non-Smoker		Smoker	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
acetonitrile	48.64	1.76	61.09	3.78	56.61	7.25	58.56	2.24	55.06	6.30	63.10	3.46
n-butyl acetate	32.50	1.00	40.44	2.64	37.69	4.75	38.31	1.06	36.45	4.03	42.23	1.67
Toluene	76.34	1.02	84.87	5.44	82.57	6.44	79.30	1.31	79.71	4.25	89.63	4.55
p-Xylene	62.44	2.42	68.81	2.93	66.73	4.33	66.46	2.57	65.42	3.65	70.84	2.27
Toluene 2 4-diisocyanate	123.93	11.39	275.35	35.93	222.02	84.46	239.15	16.63	199.36	70.29	308.73	19.89

Table A3. Comparison of mean and standard deviation of predicted concentrations (ng/L) for the five target VOCs stratified by age group, gender, and smoking status based in Neural Network model.

VOC	Age				Gender				Smoking Status			
	≤15		>15		Male		Female		Non-Smoker		Smoker	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
acetonitrile	49.46	1.30	60.71	3.13	56.67	6.434	58.42	1.760	55.28	5.65	62.50	2.55
n-butyl acetate	33.08	0.84	40.27	2.304	37.74	4.26	38.51	0.820	36.61	3.52	42.01	1.52
Toluene	77.10	0.783	84.66	4.79	82.58	5.70	79.95	1.166	80.08	3.68	88.92	4.07
p-Xylene	63.11	1.54	68.88	2.44	66.97	3.74	66.91	1.75	65.78	2.97	70.83	1.93
Toluene 2 4-diisocyanate	135.69	9.93	270.21	33.36	222.88	75.58	237.83	13.48	202.13	62.08	301.73	18.86

## References

- Nawaz, F.; Ali, M.; Ahmad, S.; Yong, Y.; Rahman, S.; Naseem, M.; Hussain, S.; Razzaq, A.; Khan, A.; Ali, F.; et al. Carbon based nanocomposites, surface functionalization as a promising material for VOCs (volatile organic compounds) treatment. *Chemosphere* **2024**, *364*, 143014. [[CrossRef](#)] [[PubMed](#)]
- Gao, M.; Liu, W.; Wang, H.; Shao, X.; Shi, A.; An, X.; Li, G.; Nie, L. Emission factors and characteristics of volatile organic compounds (VOCs) from adhesive application in indoor decoration in China. *Sci. Total Environ.* **2021**, *779*, 145169. [[CrossRef](#)] [[PubMed](#)]
- Halios, C.H.; Landeg-Cox, C.; Lowther, S.D.; Middleton, A.; Marczylo, T.; Dimitroulopoulou, S. Chemicals in European residences—Part I: A review of emissions, concentrations and health effects of volatile organic compounds (VOCs). *Sci. Total Environ.* **2022**, *839*, 156201. [[CrossRef](#)] [[PubMed](#)]
- Khan, A.; Kanwal, H.; Bibi, S.; Mushtaq, S.; Khan, A.; Khan, Y.H.; Mallhi, T.H. Volatile organic compounds and neurological disorders: From exposure to preventive interventions. In *Environmental Contaminants and Neurological Disorders*; Springer: Cham, Switzerland, 2021; pp. 201–230.
- Tsai, W.-T. An overview of health hazards of volatile organic compounds regulated as indoor air pollutants. *Rev. Environ. Health* **2019**, *34*, 81–89. [[CrossRef](#)]
- Kanwal, S.; Abas, N. VOC exposure in Pakistani furniture-painting micro-workshops: Role of inadequate ventilation. *Indoor Air* **2024**, *34*, e13234. [[CrossRef](#)]
- Kim, Y.; Park, J. Measurements of solvent-based coating VOCs in small wood-craft shops. *J. Occup. Health* **2023**, *65*, e12301. [[CrossRef](#)]
- Ranjan, S.; Chaitali, R.; Sinha, S.K. Gas chromatography-mass–mass spectrometry (GC-MS): A comprehensive review of synergistic combinations and their applications in the past two decades. *J. Anal. Sci. Appl. Biotechnol.* **2023**, *5*, 72–85.
- Zhang, J.D.; Le, M.N.; Hill, K.J.; Cooper, A.A.; Stuetz, R.M.; Donald, W.A. Identifying robust and reliable volatile organic compounds in human sebum for biomarker discovery. *Anal. Chim. Acta* **2022**, *1233*, 340506. [[CrossRef](#)]
- Baum, J.L.R. Design of Non-Invasive Systems for Detection of Exogenous and Endogenous Volatile Compounds for Applications in Environmental Exposure and Health Diagnostics. Doctoral Dissertation, University of Miami, Coral Gables, FL, USA, 2020.
- Fanti, G.; Borghi, F.; Spinazzè, A.; Rovelli, S.; Campagnolo, D.; Keller, M.; Cattaneo, A.; Cauda, E.; Cavallo, D.M. Features and practicability of the next-generation sensors and monitors for exposure assessment to airborne pollutants: A systematic review. *Sensors* **2021**, *21*, 4513. [[CrossRef](#)]
- Alyami, A.R. Assessment of occupational exposure to gasoline vapour at petrol stations. Doctoral Dissertation, Cranfield University, Bedford, UK, 2016.
- Romieu, I.; Ramirez, M.; Meneses, F.; Ashley, D.; Lemire, S.; Colome, S.; Fung, K.; Hernandez-Avila, M. Environmental exposure to volatile organic compounds among workers in Mexico City as assessed by personal monitors and blood concentrations. *Environ. Health Perspect.* **1999**, *107*, 511–515. [[CrossRef](#)]
- Vardoulakis, S. Human exposure: Indoor and outdoor. *Issues Environ. Sci. Technol.* **2009**, *28*, 85.
- Kuykendall, J.R.; Shaw, S.L.; Paustenbach, D.; Fehling, K.; Kacew, S.; Kabay, V. Chemicals present in automobile traffic tunnels and the possible community health hazards: A review of the literature. *Inhal. Toxicol.* **2009**, *21*, 747–792. [[CrossRef](#)] [[PubMed](#)]
- Klepeis, N.E. Modeling human exposure to air pollution. *Hum. Expo. Anal.* **2006**, *445*–470.
- Salo, L. The Effects of Coatings on the Indoor air Emissions of Wood Board. Master’s Thesis, Insinöörityöiden korkeakoulu, Espoo, Finland, 2017.
- Christopher, A. Identification and Assessment of Indoor Air Quality Exposures at a Manufacturing Facility. Master’s Thesis, College of Charleston, Charleston, SC, USA, 2024.
- Woolley, T. *Building Materials, Health and Indoor Air Quality*; Routledge: Abingdon, UK, 2024; Volume 2.
- Kelleher, S.; Quinn, C.; Miller-Lionberg, D.; Volckens, J. A low-cost particulate matter (PM 2.5) monitor for wildland fire smoke. *Atmos. Meas. Tech.* **2018**, *11*, 1087–1097. [[CrossRef](#)]
- Epping, R.; Koch, M. On-site detection of volatile organic compounds (VOCs). *Molecules* **2023**, *28*, 1598. [[CrossRef](#)]
- Masmoudi, S.; Elghazel, H.; Taieb, D.; Yazar, O.; Kallel, A. A machine-learning framework for predicting multiple air pollutants’ concentrations via multi-target regression and feature selection. *Sci. Total Environ.* **2020**, *715*, 136991. [[CrossRef](#)]
- Liu, S.; Li, R.; Wild, R.J.; Warneke, C.; De Gouw, J.A.; Brown, S.S.; Miller, M.; Luongo, J.C.; Jimene, J.L.; Ziemann, P.J. Contribution of human-related sources to indoor volatile organic compounds in a university classroom. *Indoor Air* **2016**, *26*, 925–938. [[CrossRef](#)]
- Sheu, R.; Lin, M. Multi-surrogate regression to predict benzene toluene from co-emitted VOCs + meteorology. *Atmosphere* **2023**, *14*, 1145. [[CrossRef](#)]
- Manh, L.H. Machine Learning for Ultraviolet Spectral Prediction. Doctoral Dissertation, The University of Texas at Arlington, Arlington, TX, USA, 2023.
- Huynh, N.; Ludkovski, M. Multi-output Gaussian processes for multi-population longevity modelling. *Ann. Actuar. Sci.* **2021**, *15*, 318–345. [[CrossRef](#)]

27. Eid, A.; Jodeh, S.; Chakir, A.; Hanbali, G.; Roth, E. Machine learning-based analysis of workers' exposure and detection to volatile organic compounds (VOC). *Int. J. Environ. Sci. Technol.* **2025**, *22*, 1–17. [\[CrossRef\]](#)
28. Liu, J.; Zhang, R.; Xiong, J. Machine learning approach for estimating the human-related VOC emissions in a university classroom. In *Building Simulation*; Tsinghua University Press: Beijing, China, 2023; Volume 16, pp. 915–925.
29. Juarez, E.K.; Petersen, M.R. A comparison of machine learning methods to forecast tropospheric ozone levels in Delhi. *Atmosphere* **2021**, *13*, 46. [\[CrossRef\]](#)
30. Bainomugisha, E.; Warigo, P.A.; Daka, F.B.; Nshimye, A.; Birungi, M.; Okure, D. AI-driven environmental sensor networks and digital platforms for urban air pollution monitoring and modelling. *Soc. Impacts* **2024**, *3*, 100044. [\[CrossRef\]](#)
31. Sané, F.C.E.; Mbaye, M.; Gueye, B. Edge-AI for Monitoring Air Pollution from Urban Waste Incineration: A Survey. In *IoT Edge Intelligence*; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 335–363.
32. Yang, J.; Hu, X.; Feng, L.; Liu, Z.; Murtazt, A.; Qin, W.; Zhou, M.; Liu, J.; Bi, Y.; Qian, J.; et al. AI-Enabled Portable E-Nose Regression Predicting Harmful Molecules in a Gas Mixture. *ACS Sens.* **2024**, *9*, 2925–2934. [\[CrossRef\]](#)
33. Koçak, E. Comprehensive evaluation of machine learning models for real-world air quality prediction and health risk assessment by AirQ+. *Earth Sci. Inform.* **2025**, *18*, 1–17. [\[CrossRef\]](#)
34. Shi, G.; Du, H.; Du, L.; Ni, X.; Hu, Y.; Pang, D.; Yao, L. Distribution characteristics of volatile organic compounds and its multidimensional impact on ozone formation in arid regions based on machine learning algorithms. *Environ. Pollut.* **2025**, *373*, 126159. [\[CrossRef\]](#)
35. Popescu, S.M.; Mansoor, S.; Wani, O.A.; Kumar, S.S.; Sharma, V.; Sharma, A.; Arya, V.M.; Kirkham, M.B.; Hou, D.; Bolan, N.; et al. Artificial intelligence and IoT driven technologies for environmental pollution monitoring and management. *Front. Environ. Sci.* **2024**, *12*, 1336088. [\[CrossRef\]](#)
36. Rane, N.; Choudhary, S.; Rane, J. Leading-Edge Artificial Intelligence (AI), Machine Learning (ML), Blockchain, and Internet of Things (IoT) Technologies for Enhanced Wastewater Treatment Systems. 2023. Available online: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4641557](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4641557) (accessed on 25 May 2025).
37. Akinosho, T.D.; Bilal, M.; Hayes, E.T.; Ajayi, A.; Ahmed, A.; Khan, Z. Deep learning-based multi-target regression for traffic-related air pollution forecasting. *Mach. Learn. Appl.* **2023**, *12*, 100474. [\[CrossRef\]](#)
38. Ye, H.; Du, Z.; Lu, H.; Tian, J.; Chen, L.; Lin, W. Using machine learning methods to predict VOC emissions in chemical production with hourly process parameters. *J. Clean. Prod.* **2022**, *369*, 133406. [\[CrossRef\]](#)
39. Hashemitaheri, M.; Ebrahimi, E.; de Silva, G.; Attariani, H. Optical sensor for BTEX detection: Integrating machine learning for enhanced sensing. *Adv. Sens. Energy Mater.* **2024**, *3*, 100114. [\[CrossRef\]](#)
40. Kang, M.; Han, J.; Kim, Y.; Kim, S.; Kang, S. Data-driven autonomous operation of VOCs removal system. *Sci. Rep.* **2024**, *14*, 5953. [\[CrossRef\]](#)
41. Zhang, R.; Wang, H.; Tan, Y.; Zhang, M.; Zhang, X.; Wang, K.; Ji, W.; Sun, L.; Yu, X.; Zhao, J.; et al. Using a machine learning approach to predict the emission characteristics of VOCs from furniture. *Build. Environ.* **2021**, *196*, 107786. [\[CrossRef\]](#)
42. Rasmussen, C.E.; Williams, C. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, MA, USA, 2006; Volume 32, p. 68.
43. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
44. Jodeh, S.; Chakir, A.; Hanbali, G.; Roth, E.; Eid, A. Method Development for Detecting Low Level Volatile Organic Compounds (VOCs) among Workers and Residents from a Carpentry Work Shop in a Palestinian Village. *Int. J. Environ. Res. Public Health* **2023**, *20*, 5613. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Win-Shwe, T.-T.; Fujimaki, H. Neurotoxicity of toluene. *Toxicol. Lett.* **2010**, *198*, 93–99. [\[CrossRef\]](#)
46. Frick-Engfeldt, M.; Zimerson, E.; Karlsson, D.; Marand, A.; Skarping, G.; Isaksson, M.; Bruze, M. Chemical analysis of 2, 4-toluene diisocyanate, 1, 6-hexamethylene diisocyanate, and isophorone diisocyanate in petrolatum patch-test preparations. *DERM* **2005**, *16*, 130–135.
47. Tomas, R.A.; Bordado, J.o.C.; Gomes, J.F. p-Xylene oxidation to terephthalic acid: A literature review oriented toward process optimization and development. *Chem. Rev.* **2013**, *113*, 7421–7469. [\[CrossRef\]](#)
48. Wang, Y.; Chen, Z.; Haefner, M.; Guo, S.; DiReda, N.; Ma, Y.; Avedisian, C.T. Combustion of n-butyl acetate synthesized by a new and sustainable biological process and comparisons with an ultrapure commercial n-butyl acetate produced by conventional Fischer esterification. *Fuel* **2021**, *304*, 121324. [\[CrossRef\]](#)
49. McConvey, I.F.; Woods, D.; Lewis, M.; Gan, Q.; Nancarrow, P. The importance of acetonitrile in the pharmaceutical industry and opportunities for its recovery from waste. *Org. Process Res. Dev.* **2012**, *16*, 612–624. [\[CrossRef\]](#)
50. Liu, M.; Chowdhary, G.; Da Silva, B.C.; Liu, S.Y.; How, J.P. Gaussian processes for learning and control: A tutorial with examples. *IEEE Control. Syst. Mag.* **2018**, *38*, 53–86. [\[CrossRef\]](#)
51. Álvarez, M.A.; Rosasco, L.; Lawrence, N.D. Kernels for vector-valued functions: A review. *Found. Trends® Mach. Learn.* **2012**, *4*, 195–266. [\[CrossRef\]](#)
52. Bonilla, E.V.; Chai, K.M.A.; Williams, C.K.I. Multi-task Gaussian Process Prediction. *Adv. Neural Inf. Process. Syst.* **2007**, *20*, 153–160.

53. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learnin*; MIT Press: Cambridge, MA, USA, 2016; Volume 1.
54. Samal, K.; Babu, K.S.; Das, S. Spatio-temporal prediction of air quality using distance based interpolation and deep learning techniques. *EAI Endorsed Trans. Smart Cities* **2021**, *5*, e4. [[CrossRef](#)]
55. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6639–6649.
56. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
57. Neal, R.M. *Bayesian Learning for Neural Networks*; Springer Science & Business Media: New York, NY, USA, 2012; Volume 118.
58. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
59. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.