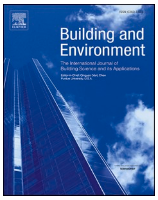
Building and Environment 213 (2022) 108894

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv



Predicting the presence of hazardous materials in buildings using machine learning

Pei-Yu Wu [a,b,*], Claes Sandels [a], Kristina Mjörnell [a,b], Mikael Mangold [a], Tim Johansson [a]

[a] *RISE Research Institutes of Sweden, 412 58, Gothenburg, Sweden*

[b] *Department of Building and Environmental Technology, Faculty of Engineering, Lund University, 221 00, Lund, Sweden*

ARTICLE INFO

Keywords:
Machine learning
Hazardous materials
Asbestos
PCB
Pre-demolition audit
Circular economy

ABSTRACT

Identifying in situ hazardous materials can improve demolition waste recyclability and reduce project uncertainties concerning cost overrun and delay. With the attempt to characterize their detection patterns in buildings, the study investigates the prediction potential of machine learning techniques with hazardous waste inventories and building registers as input data. By matching, validating, and assuring the quality of empirical data, a hazardous material dataset for training, testing, and validation was created. The objectives of the explorative study are to highlight the challenges in machine learning pipeline development and verify two prediction hypotheses. Our findings show an average of 74% and 83% accuracy rates in predicting asbestos pipe insulation in multifamily houses and PCB joints or sealants in school buildings in two major Swedish cities Gothenburg and Stockholm. Similarly, 78% and 83% of recall rates were obtained for imbalanced classification. By correlating the training sample size and cross-validation accuracy, the bias and variance issues were assessed in learning curves. In general, the models perform well on the limited dataset, yet collecting more training data can improve the model's generalizability to other building stocks, meanwhile decreasing the chance of overfitting. Furthermore, the average impact on the model output magnitude of each feature was illustrated. The proposed applied machine learning approach is promising for in situ hazardous material management and could support decision-making regarding risk evaluation in selective demolition work.

1. Introduction

Hazardous materials hamper material recyclability and value recovery from end-of-lifecycle buildings [1]. The contaminants entering the waste stream and re-contaminating other building components during reconstruction, renovation, or demolition challenge the newly formed circular economy chain in the building sector [2,3]. As such, appropriate risk management means that assessing the extent of contamination from the existing building components is necessary to facilitate circular economy-inspired actions [3]. Currently, inventory of hazardous waste is performed during the pre-demolition audits for individual buildings to guide the waste management process [4]. To ensure safe and sustainable management of construction and demolition waste, the EU Construction and Demolition Waste Management Protocol (EC, 2016) and Guideline were introduced to enhance the confidence in the waste management process and the trust in the quality of recycled materials [4,5]. Through early detection, source separation, and onsite collection, selective demolition work that processes the material fractions for high-quality recovery can be planned accordingly [1,2].

Nowadays, the attitude toward in situ hazardous material management has changed from passive monitoring to predictive maintenance by taking precautionary actions into account [6]. By processing the operation data, the gaps in maintenance practice can be addressed to support the facility management avoiding unfavored consequences without costly onsite inspections [7]. Conventionally, the risk of asbestos and polychlorinated biphenyls (PCBs) exposure was usually controlled through regular sampling [8,9]. With respect to the substantial decontamination and high disposal cost, short-term mitigation measures such as transport pathway blocking, concentration dilution, and source removal are usually regarded as viable alternatives [9]. At the crossroad of transition towards the circular built environment, along with the need to implement predictive maintenance, developing new approaches to identify hazardous materials extensively has become a focus in interdisciplinary research fields [1,2,10].

Identifying in situ hazardous materials in the existing building stock can, on the one hand, help property owners manage health exposure risk and possible project disturbance; on the other hand, contribute to a closed material loop for construction and demolition waste as a whole

[*] Corresponding author. RISE Research Institutes of Sweden, Sven Hultins Plats 5, 412 58, Gothenburg, Sweden.

E-mail addresses: pei-yu.wu@ri.se (P.-Y. Wu), claes.sandels@ri.se (C. Sandels), kristina.mjornell@ri.se (K. Mjörnell), mikael.mangold@ri.se (M. Mangold), tim.johansson@ri.se (T. Johansson).

https://doi.org/10.1016/j.buildenv.2022.108894
Received 23 December 2021; Received in revised form 24 January 2022; Accepted 10 February 2022
Available online 16 February 2022

[2]. However, the information concerning the presence of hazardous materials in buildings is usually inadequate, leading to unforeseen events during reconstruction or renovation [10]. Around 20% of additional costs for acute decontamination have been reported in demolishing residential buildings [10]. Therefore, the detection records on hazardous waste inventories from the past demolished and renovated buildings have become a valuable source for studying the potential detection patterns of hazardous materials [2]. The emergent artificial intelligence and digitalization in data capture and visualization can improve information availability and assist decision-making in building material assessments [1,11].

Among all substances in the EU Waste Framework Directives (WFD 2008/98/EC, amended 2018/815), asbestos and PCB-containing materials are rigorously regulated owing to their critical properties and regulations [12]. Hence, quantitative studies on inferring their likely presence in the building stock have been conducted in several countries [13]. For instance, an ontology-based method for estimating the probability of asbestos-containing materials was proposed by Mecharnia et al. [14]. By employing building registers and asbestos detection records in a rule-based logic, the use of asbestos-containing materials at the product, location, structure, and building levels were evaluated. However, lacking complete product descriptions limited the application of inductive logic programming and rule discovery approaches. Besides, attempts have been made to probe other possibilities to retrieve data. Using a self-assessment mobile application, the type of in situ asbestos-containing materials and their condition could be investigated and documented [15,16]. Nevertheless, a small sample size and a self-selected sample population might cause selection bias and make it hard to generalize the results to the national scale [17]. To overcome the sampling bias and characterize the occurrence pattern of asbestos-containing materials, the potential of using a public reconstruction and demolition database containing inspection records of hazardous waste, as input data were explored. Applying data abstractor, visualized plotting, and statistical correlation study, asbestos-containing materials' location, type, amount, and abatement costs were obtained [10]. The abovementioned examples and their promising results devote pioneering efforts in exploring data-driven building material management.

In light of the difficulty to access building-specific environmental data [12] and low adoption of information technologies in construction and demolition waste management [18], machine learning shows a substantial potential for pattern identification in records. By utilizing data labels in the training dataset, predicting the unknown examples through recognizing critical features of contaminants is possible [13]. Due to their promising prediction performance and high model generalization, machine learning classifiers have been widely deployed in image recognition for construction materials, hyperspectral images, and aerial photographs [13]. Several supervised learning classifiers were proved to detect asbestos-containing materials effectively, among all, the convolutional neural networks used for asbestos cement roofing identification [19] as well as the random forest algorithm applied to predict the spatial distribution of asbestos cement roofing [20]. Other algorithms, such as the statistical classifier Naïve Bayes, the distance-based classifier k-nearest neighbor (k-NN), the tree-ensembled classifier random forest (RF), and the support vector machines (SVM), can also achieve higher prediction accuracy than traditional rule-based, object-based image analysis in optical recognition [21]. However, several limitations of applying machine learning in construction and demolition management were also highlighted by previous research, including complex interdependencies between feature representation, type of classifiers as well as hyperparameter tunning and regularization [22], as well as a lack of labeled data to train and validate the models [23].

Despite considerable progress in pioneering studies, adopting the developed models in practice remains slow. This is generally due to low model transferability and generalization from specific buildings to other buildings [22]. The heterogeneous structure and content of pre-demolition audit documents make the detection results comparison challenging [24]. The extensive use of hazardous materials in past construction projects and their variety lead to extreme difficulty characterizing the nature of exposure. Moreover, some materials are not even visually recognizable and require expertise in sampling and analysis in the lab. Therefore, the possibility of using hazardous waste inventories as input data to assess the risk of hazardous materials in the regional building stock was studied by Wu et al. [12]. The machine learning preprocessing work regarding data validation and representativeness control performed in previous study laid a solid foundation for algorithm development. As the pre-demolition audit process has been enforced in several European countries [25], the data-driven method is replicable to ascertain the positive detection rates and delineate building types with quality data.

2. Scope of the paper

Although various predictions have been made for building materials recognition using image data, utilizing records from inventories of hazardous waste for comprehensive hazardous material prediction at the building level remains rather unexplored. The paper aims to explore the prediction potential for the presence of hazardous materials in specific buildings classes with a proposed machine learning pipeline. By verifying expert assumptions in the pre-demolition audit practice, the understanding of the occurrence patterns for residual contaminants can be further enhanced. Two common hazardous materials and their presence in two types of buildings based on previous literature – asbestos-containing pipe insulation in multifamily houses and PCB-containing joints or sealants in school buildings – will be investigated as case studies. Through assembling hazardous waste inventories data from the Swedish cities of Gothenburg and Stockholm as well as their building registers to a hazardous material dataset, the possibility of succeeding in pattern identification with supervised learning algorithms increases. Since the performance of machine learning classifiers is susceptible to the variation of sample size and the inclusion of validated observations, the effect of an increase of training data helps examine the optimal data size for prediction. By tuning and regularizing the potential features, the most appropriate classifiers for the task can be determined. The prediction results from the explanatory machine models can form the basis of contaminant risk management at the building stock level. Thus, the paper contributes to predictive maintenance of the existing buildings in terms of safe managing of demolition waste. To facilitate the objective of the study, three interconnected research questions were formulated as follows:

RQ1: Which machine learning classifier provides the best results for the task?

RQ2: How many training examples are needed to obtain sufficient prediction results?

RQ3: What are the influential factors for predicting the target hazardous materials in the specific building classes?

3. Materials and methods

The section describes the research process and the main tasks performed in the study, illustrated in Fig. 1. First, an overview of the data source regarding data assembling and matching was provided, followed by data validation and processing to delineate the underlying structure. Potential key variables identified in explorative data analysis and feature engineering were subsequently employed in supervised models' development. Tuning the selected features and regularizing various machine learning classifiers can optimize the prediction accuracy for model evaluation. Finally, influential features associated with the prediction outcomes were outlined in result interpretation.

3.1. Data source

A hazardous material dataset consisting of 906 detection records and

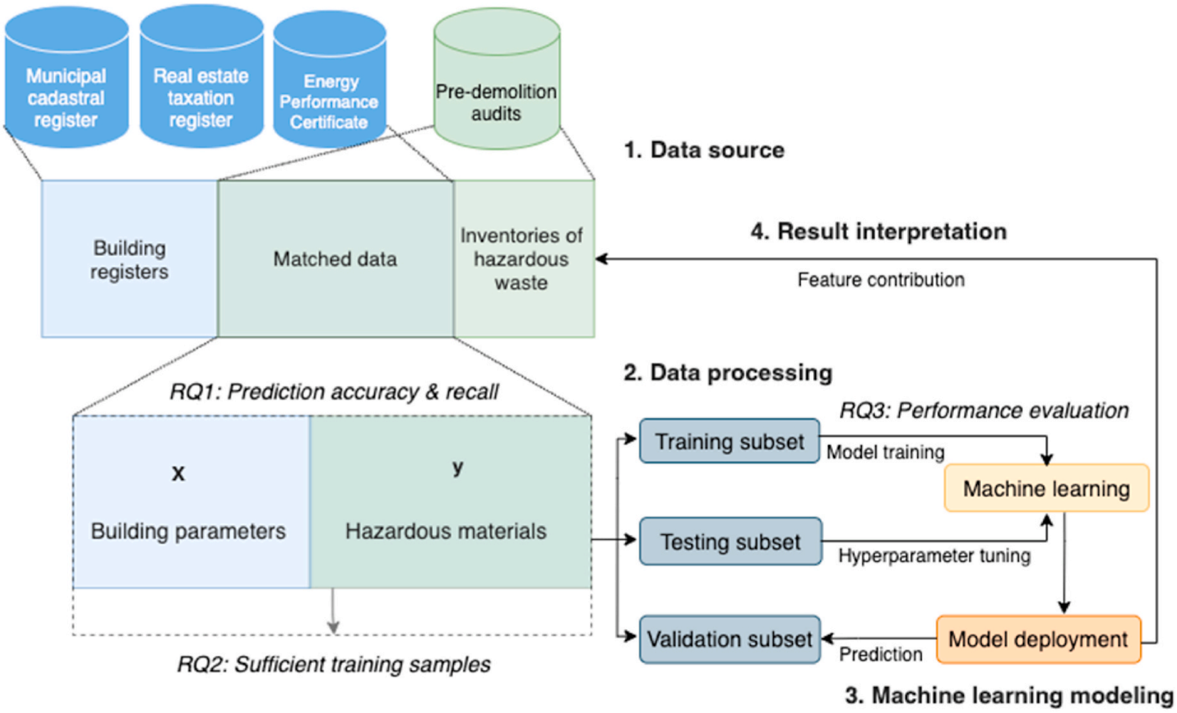


Fig. 1. The study design consists of four parts: (1) data sourcing for acquiring building registers and hazardous waste inventories from authorities, (2) data processing for variables transformation and feature selection, (3) machine learning model development and evaluation using 10-fold nested cross-validation, (4) result interpretation for generating explainable machine learning model and evaluating feature importance.

national building registers were compiled to study the presence of hazardous materials in the Swedish building stock. The detection records encompass hazardous waste inventories of demolished and renovated buildings between 2010 and 2020 retrieved from the Gothenburg and the Stockholm City Archives. Buildings built before 1990 are especially of interest as asbestos and PCB-containing materials were used extensively in construction. Detailed inventories, such as reports and protocols, accounted for most of the acquired documents with specification of the detected hazardous materials and investigated buildings.

Merging the municipality cadastral register, the real estate taxation register, and the Energy Performance Certificates (EPC) data form a basis for a comprehensive national building registers database that contains building parameters for individual buildings. The national building register database was created with GIS Feature Manipulation Engine from the Safe Software, referring to the work by Johansson et al. [26]. As the aggregation level for hazardous waste inventories is at the building level, pairing the registered data at multiple levels may cause uncertainty in data merging. For instance, the Swedish real estate taxation register adopts value units that are in most cases at property levels, while the EPC register is structured according to the EPC index that attaches to one or more properties with one or more buildings. Therefore, extra attributes were introduced to reduce the risk of inaccurate data merging and retrieval, i.e., the total number of possible matching relationships and the potential duplicates. According to national building registers, around 73.3% of buildings in Gothenburg and 84.6% of buildings in Stockholm are constructed before 1990 and thus more likely to contain contaminants.

Further on, the national real estate indexes of the investigated buildings were used as the common key to find the match between the building-specific information and the building registers. The matching between registers and inventory data was performed manually by examining the address, construction year, and floor area in the one-to-many relationship to ensure correct data coupling. The results of each observation were labeled with different matching codes according to information conformity. In case of lacking information on building class or construction year, the observations were dropped out to reduce irrelevant data noise. Considering parameters of construction year, renovation year, and floor area exist in several registers, revised variables harmonizing the inventory data and the building registers were created for subsequent analysis with Python statistical visualization libraries Matplotlib and Seaborn. By the end of the data cleaning, 848 eligible observations remained. An overview of the variable characteristics in the hazardous material dataset is displayed in Table 1.

To clarify which hypothesis was viable for attempted prediction, a cross-validation matrix was created using Python scientific computation package Numpy to evaluate data quality and quantity. The information of the acquired inventories was processed in formula (1) to calculate the assessment scores, which became the basis for prioritizing quality data subgroups for modeling. The scale of y is between 0 and 100 with percentage as the unit. The building classes with comprehensive detection records and large sample sizes were considered. They are for example school buildings, multifamily houses, commercial buildings, offices, and industrial or production buildings [12]. Stratifying data subsets with similar building parameters can prevent the risk of false inference.

$$y = \frac{(Ir \times nr + Ip \times np + Ic \times nc + Id \times nd)}{n} * K \tag{1}$$

y = Assessment score.
I = Inventory type for weighting individual observations. $I = 1$ if is the report (r), $I = 0.75$ if is the protocol (p), $I = 0.5$ if is the control plan (c), and $I = 0.25$ if is the demolition plan (d).
n = The number of observations in the studied subgroup $[0 < n]$.
N = The number of observations in the entire dataset.
K = The number of observations enough for statistical operation.
$K = 1$ if $n \geq (0.05*\mathrm{N})$, $K = 0.5$ if $(0.025*\mathrm{N}) =< n < (0.05*\mathrm{N})$, $K = 0$ if $n < (0.025*\mathrm{N})$.

Table 1
An overview of the variable characteristics in the hazardous material dataset. The modeling part was split into 60% training, 20% testing, and 20% validation subsets.

Data	Value category	Data specification	Measurement types
Model part	Geographics	City	Nominal [Gothenburg, Stockholm]
	Building usage	EPC building category	Nominal [premise building, multifamily house, premise and special building, single and two-dwelling house]
		EPC building type	Nominal [detached, gable, intermediate]
	Building parameter	Construction year	Scale variable [year]
		Renovation year	Scale variable [year]
		Floor areas	Scales [m^2]
		Number of floors	Ordinal [N]
		Number of apartments	Scales [N]
		Number of stairwells	Scales [N]
		Number of basements	Scales [N]
		Ventilation type	Nominal [exhaust, balanced, balanced with heat exchanger, exhaust with heat pump, natural ventilation]
	Hazardous substance	Asbestos	Nominal [positive, negative, NA]
		PCB	Nominal [positive, negative, NA]
	Hazardous material	Building component [a]	Nominal [positive, negative, NA]
Complementary part	Matching keys	National real estate index	String & numeric [index]
		Address	String
		Matching code	Nominal [1–8]
	Permit application	Project description	String
	Building usage	Building class[b]	Nominal [single-family house, multifamily house, temporary dwelling, school, office, commercial building, production building, industrial building, warehouse, and other/infrastructure]
		Municipality category code	Nominal [1–7]
		Municipality type code	Nominal [1-99]
	Inventories of hazardous waste	Scope	Nominal [entire, part of the building]
		Investigation year	Scale variable [year]
		Investigator	String
		Decontamination	Nominal [asbestos, PCB, NA]

[a] The building components imply the building materials that contain human or environmentally hazardous substances. The asbestos-containing materials include pipe insulation, valves, door/windows insulation, cement panel, tile/clinker, carpet glue, floor mat, ventilation channel, switchboard, joint, and other asbestos products. The PCB containing materials encompass joint/sealant, sealed double glazing windows, capacitor, acrylic flooring, door closer, cable with PCB-oil, and other PCB products.

[b] The building class refers to the primary use of the building on the permit application and the building type in building registers and EPC. Ten building classes were created to synthesize different category systems and details between the data sources – single-family houses, multifamily houses, temporary dwellings, schools, offices, commercial buildings, production buildings, industrial buildings, warehouses, others/infrastructure.

3.2. Data processing

The data processing work, such as dataset partitioning, variable transformation, and feature selection, were executed with Pandas and scikit-learn, the data analysis and machine learning libraries for Python. Dataset partitioning was performed first to avoid data leakage, then approximately 20% of the data points were held out for model validation. Then the rest of the observations were randomly separated into 75% of training data and 25% of testing data for model training. This resulted in splitting the dataset to 60% of training, 20% of testing, and 20% of validation subsets. A balance between the size of the training, validation, and testing subsets was determined based on the amount of the training data to prevent generalization errors [27].

Afterward, variable transformation was performed for algorithm optimization, especially for scale-invariant classifiers like the decision tree and the random forest [27]. The labels of categorical variables were encoded and continuous variables were standardized to a comparable scale. The non-null values were quantified for the entire dataset to detect the amounts of missing data, which is essential for determining the useable variables and the data filling strategy. For categorical variables, missing values cannot be computed, but the missing numerical variables were imputed and replaced with mean values. Further filtering the buildings with extreme parameters, such as construction year earlier than 1900 or later than 1990, floor area over 25 000 m^2, and the number of floors above 15 stories, to have a more homogeneous and representative dataset for the Swedish building stock. With these criteria, the outliers, including eight multifamily houses and 43 school buildings, were detected, and removed from the modeling dataset.

The last step before modeling, feature selection, was implemented to improve the accuracy of the models, reduce training time and data dimensionality for better generalization performance [27,28]. Meanwhile, model complexity was regulated, and the risk of overfitting was reduced for unregularized classifiers. The Recursive Feature Elimination (RFE) method was employed to remove variables recursively by evaluating the prediction accuracy and returning the most contributing features based on the optimized model [29]. The optimal number of features was determined through plotting cross-validated accuracy scores and the number of selected features. Next, feature importance was evaluated by tree-based estimators to validate the feature selection results. The Extremely Randomized Trees Classifier (Extra Trees) computes the Impurity-based feature importance based on the averaged impurity decrease from all decision trees in the forest without making assumptions about the data linearity [27]. Combining the results from the Extra Trees classifier and the RFE, the identified key features were then used as predictive variables for model development.

3.3. Machine learning modeling

To address the classification problem of heterogeneous data in a small tabular dataset, various supervised algorithms were tested during model development: logistic regression, kernel support vector machines (SVM), k-nearest neighbors (k-NN), random forest, extreme gradient boosting (XGBoost), and CatBoost. Incorporating classifiers with different strengths and weaknesses helped evaluate the bias and variance trade-off for searching the optimal models. The bias error results from erroneous assumptions in the algorithm and causes model underfitting; whereas the variance error is derived from sensitivity to small fluctuations in the training dataset and leads to model overfitting. An ideal model with a good generalization capability for the unknown data should have a balanced bias and variance trade-off. The chosen classifiers are non-parametric models that can learn from growing parameters in the training data, summarized in Table 2.

Logistic regression is a basic linear classifier that represents the class distribution probability. Computing without assumptions and suitability for the low dimensional dataset is thus used as a base model to compare prediction performance with other classifiers. SVM features margin maximization using the closest data points to maximize the distance between the separating hyperplane [27]. The kernel SVM with radial basis function projects the linearly inseparable training data to higher-dimensional feature space and train a linear SVM to classify the data. SVM was reported powerful in handling high dimensional data, yet it has low bias and high variance, thus requiring careful feature standardization and parameter tuning [30]. In comparison, k-NN is an instance-based algorithm that quickly adapts to new training data by classifying the k samples based on the distance metric and the majority vote. Varied k and Euclidean distance measures were tested to prevent dimensionality-based overfitting as regularization was not applicable for the algorithm. However, a common shortcoming of these algorithms in the study is that they require numerical inputs.

On the other hand, tree ensemble algorithms such as random forest, XGBoost, and CatBoost benefit the information gained in the iterative decision-making process and can vastly improve model interpretability. The random forest algorithm exploits the performance of average multiple trees to counteract the high variance of a single tree without specifying hyperparameter values. XGBoost contains the advantages of the random forest classifier and further evolved with boosting and gradient boosting, where new models to predict the residuals before making a final prediction were created [31]. Featuring a scalable tree boosting system, XGBoost outshines other tree-based algorithms regarding parallelized tree building and pruning for system optimization, as well as regularization, sparsity awareness, and built-in cross-validation in algorithmic enhancements [32]. Further on, CatBoost was adopted to process categorical features and prevent target leakage by implementing ordered boosting [33]. Although the classifier has a wide application in interdisciplinary fields, it is sensitive to hyperparameter change and time-consuming for hyperparameter tuning [34].

Next, the class imbalance between positive (represented as 1) and negative (represented as 0) detections was handled by oversampling the minority groups. Having both classes equally presented in the dataset is essential for dealing with classification problems as most of the machine learning algorithms assume an unbiased dataset. In this way, poor predictive performance for the minority class may be due to the skew dataset, where algorithms are sensitive to classification errors for the minority class than the majority class, can be prevented [35]. These data resampling techniques can also address cost-sensitive learning when false-negative errors are valued differently from the false-positive errors [36]. By changing the composition of the class distribution, the cost proportionate weighting can fulfill the expectation of minimizing the misclassification errors.

Evaluating the algorithms' performance and tuning their hyperparameters enable us to find the optimal model. The training models were assessed in terms of generalization with nested cross-validation, illustrated in Appendix A. Nested cross-validation is a wrapper of an outer loop with numerous training and test folds to train optimal parameters, and the inner loops at the training folds to tune parameters and select the model using k-fold cross-validation [27]. Then the test fold in the outer loop is applied to assess the model performance. Nested cross-validation was preferable due to the limited amount of data. Meanwhile, hyperparameter tunning was performed through the grid search method. Modifying the architecture of the models implies adjusting regularization constant, kernel type, and constants in SVM or k values in k-NN to tackle generalization errors of learning algorithms. The grid search method of testing a wide range of hyperparameter combinations was employed for the purpose. Next, additional data were fed into the models to evaluate the accuracy change. The generated learning curves can measure the bias-variance trade-off and control the risk of overfitting.

Finally, the classifiers' performance was evaluated with the confusion matrix on the accuracy, precision, recall, and the F1-score, illustrated in (2)–(5) in Appendix B Fig. B1. Accuracy measures the number of correct predictions over the number of total predictions, whereas recall estimates the fraction of retrieved relevant instances. The assessment criteria for high accuracy and recall rates were chosen given the defined prediction goal for extracting hazardous material detection. Besides, a binary classifier evaluation system was schemed to assess and tune the ensemble classifier. The receiver operating characteristic (ROC) curve was plotted with various discrimination thresholds to diagnose the trade-off between sensitivity (True positive rate) and specificity (False

Table 2
Characteristics of the selected algorithms used in the study.

Algorithm	Description	Regularization	Strengths	Limitations
Logistic regression	Linear classifiers for the discrete variable as the output is transformed to log-odds	No	• Probability estimation • No assumptions about class distributions in feature space • Coefficient size and direction of the association	• Require no multicollinearity between independent variables • Overfitted in high dimensional datasets
Kernel SVM	A distance-based classifier that maximizes the gap between the projected data	Yes	• Handle high dimensional data • Perform well on a wide range of datasets	• Scale variant and need parameter standardization • Require parameter tuning • Kernel needs to be specified
k-NN	An instance-based classifier that assorts the data based on the distance metric and the majority vote	No	• Incremental learning • Simple implementation	• Scale variant and need parameter standardization • Require parameter tuning
Random forest	A tree-ensembled classifier that ensembles the predictions of multiple trees	No	• Scale-invariant • Reduced overfitting • Parallel processing • No parameter specification required	• Hard to interpret the result • Require parameter tuning
XGBoost	A tree-ensembled classifier that features a scalable gradient boosting system	No	• No regularization required • Handle missing data • Parallel processing • Built-in cross-validation • High accuracy and robust	• No support for categorical feature transformation • Computation intensive and long training time
CatBoost	A tree-ensembled classifier that exploits order boosting with categorical features	No	• Reduced overfitting • Transform categorical feature • Parallel processing • Wide application • High accuracy and robust	• Sensitive to hyperparameters change • Require parameter tuning

positive rate). The goal was to achieve high area values under the roc curve (ROC AUC), where 1 implies high separability and low separability in the case below 0,5. The refined ensembled models were verified on the testing subset for unbiased performance evaluation before model deployment.

3.4. Results interpretation

The prediction results from machine learning models are sometimes hard to understand and explain by human experts. However, developing transparent models interpretable by domain knowledge helps control scientific consistency [37]. Several methods have been proposed to ease the tension between accuracy and interpretability. For instance, SHapley Additive exPlanations (SHAP) was created as a unified framework to explain output from machine learning models. Each feature is assigned a SHAP value to visualize its contributions to the model output [38]. Accordingly, a two-step approach was implemented to facilitate results interpretation: (1) applying the SHAP explainer with different models and inserting the generated SHAP values into summary plots. From here, the information regarding the prediction score, base value, feature magnitude, and direction, as well as classification results, can be obtained; (2) involving domain experts to assess the coherency between the feature importance identified by the SHAP framework and the scientific assumptions from field practice. Adopting the hybrid approach can comprehend the inherent prediction mechanism of machine learning models and evaluate hazardous material risk in specific building classes based on influential features.

4. Results

The results section presents the statistical analyses of the hazardous material dataset, then evaluates prediction results between the models. After that, the impact of the data size on model performance was described, and the influential features were highlighted.

4.1. Presentation of the data

The detailed investigation documented in reports (68.4%) and protocols (16.4%) accounted for the major sources of observations in the hazardous material dataset (N = 848), indicating high data reliability considering the experience of investigators and sampling method. Around 65.6% of environmental investigations concerned the entire buildings, and 55.8% of the observations had undergone renovation. However, most of the decontamination history is unknown. The available data shows that approximately 7.1% and 4.4% of the buildings underwent PCB or asbestos decontamination. Furthermore, the number of missing values and class distribution was plotted to control the data completeness and skewness. It was found that 93% of observations contain information on asbestos and asbestos-containing materials, while the counterpart of information on PCB is 73%. The stacked histograms in Fig. 2 show the detection frequency of asbestos and PCB across varied building types. Surprisingly, 74% and 45% of the buildings are exposed to asbestos and PCB, respectively. Within these contaminated building stock, multifamily houses, schools, and commercial buildings represent significant proportions of positive asbestos or/and

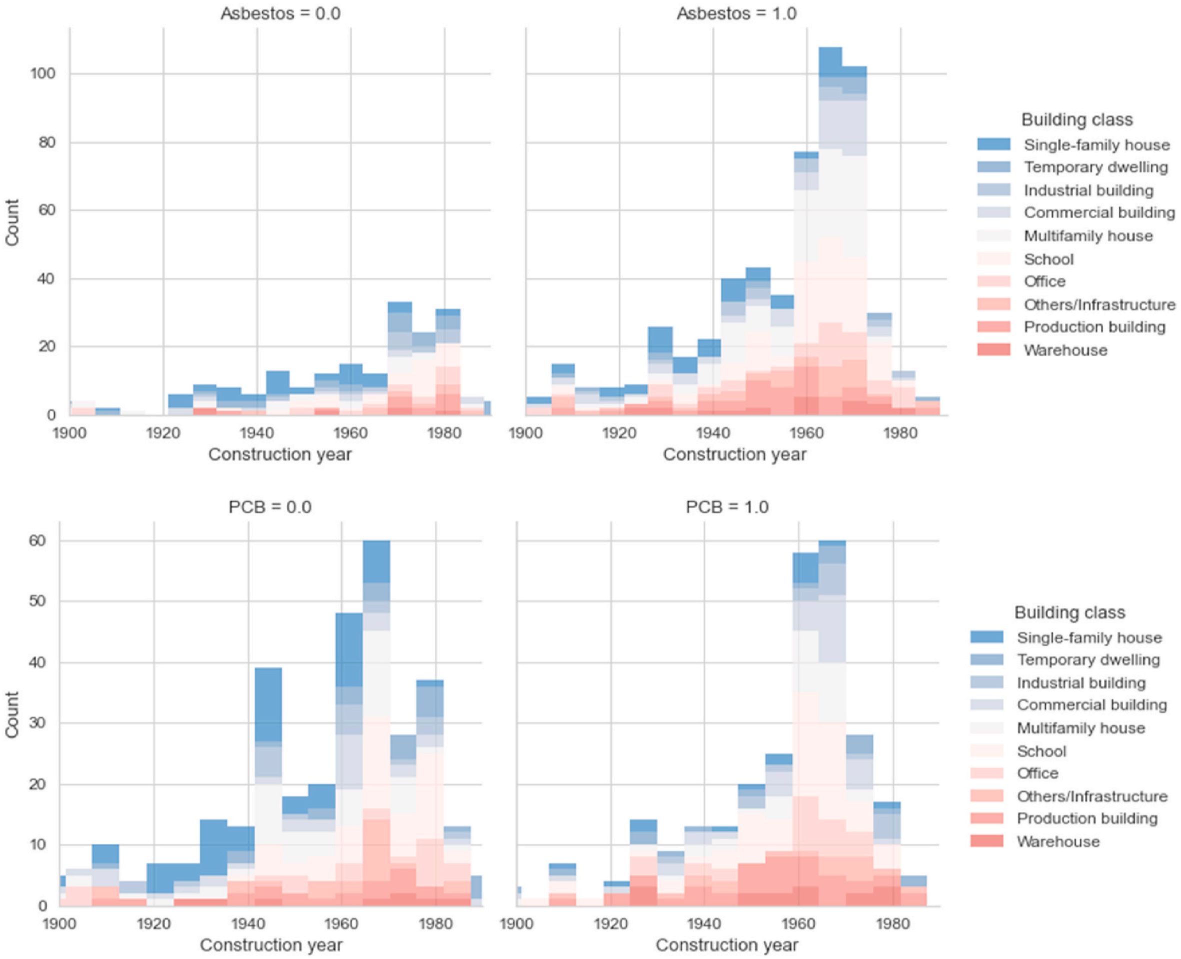


Fig. 2. The stacked histograms display the detection records of asbestos (upper) and PCB (bottom) across building classes in the Gothenburg and the Stockholm building stocks built between 1900 and 1990 (N = 813). Positive detections are denoted as 1, vice versa.

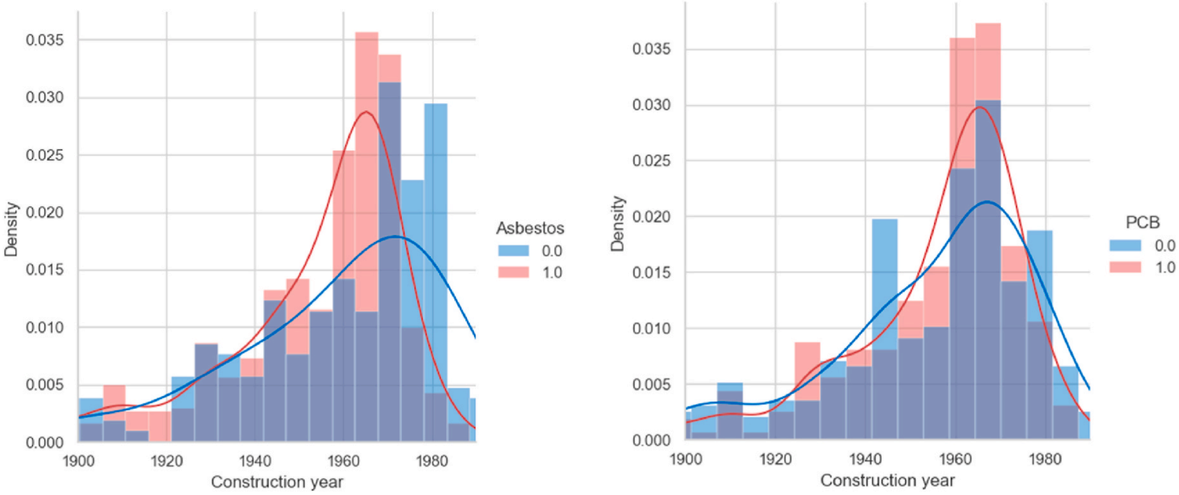


Fig. 3. The aggregated normalized density distribution for asbestos (left) and PCB (right) detection.

PCB detection.

Afterward, the numbers of asbestos and PCB detection were normalized to visualize the aggregated density distribution to assess the contamination likelihood. The distribution tendency in Fig. 3 indicated that buildings constructed between 1955 and 1975 are more likely to contain hazardous materials. This period corresponds to large housing programs – the post-war era (1945–1960) and the Million Homes Program (1965–1974), and most of the Swedish building stock nowadays is inherited from that time. The peaks of the asbestos and PCB risk were observed around 1965 and dramatically decreased after 1975 when the likelihood of negative detections exceeded the positive detections after the use bans. Besides, compared to PCB, the shift of distribution between the positive and negative detection of asbestos along the timeline is more evident. This entails that the construction year may be a relevant factor for determining the presence of asbestos.

Hazardous materials common in specific building classes and thus suitable for machine learning modeling were determined. Table 3 below describes the score ranking of the results from the cross-validation matrix, along with the statistical features of each data subgroup concerning the positive detection rates, the total and available data amounts, and the numbers of missing values. High assessment scores were obtained for asbestos joints, floor mats, ventilation channels, tiles or clinker, carpet

Table 3
The top score ranking of the data subgroups for each hazardous material in particular building classes using the cross-validation matrix. The subgroups with cross-validation scores over 90 were listed.

Rank	Class	Substance	Building part	Score	Rate	Total	NA(%)	N
1	Multifamily House	Asbestos	Joint	99	0.67	153	63	57
2	School	Asbestos	Floor mat	98	0.35	154	45	86
	School	Asbestos	Ventilation channel	98	0.40	154	60	62
3	Multifamily House	Asbestos	Ventilation channel	96	0.47	153	48	79
	Multifamily House	Asbestos	Others	96	0.65	153	65	54
	School	Asbestos	Tile/clinker	96	0.25	154	28	112
	School	Asbestos	Carpet glue	96	0.24	154	37	97
	Commercial building	Asbestos	Door/windows insulation	96	0.65	85	44	48
	Commercial building	Asbestos	Tile/clinker	96	0.28	85	33	57
	Commercial building	Asbestos	Floor mat	96	0.48	85	41	50
4	Multifamily House	Asbestos	Floor mat	95	0.57	153	48	79
	School	Asbestos	Pipe insulation	95	0.46	154	37	98
	School	Asbestos	Door/windows insulation	95	0.37	154	46	83
	School	Asbestos	Joint	95	0.37	154	68	49
	Commercial building	Asbestos	Pipe insulation	95	0.76	85	31	59
	Commercial building	Asbestos	Ventilation channel	95	0.34	85	48	44
	Commercial building	PCB	Sealed double glazing windows	95	0.18	85	48	44
	Production Building	Asbestos	Pipe insulation	95	0.76	75	32	51
5	Multifamily House	Asbestos	Door/windows insulation	94	0.81	153	37	96
	Multifamily House	Asbestos	Tile/clinker	94	0.50	153	29	109
	School	Asbestos	Cement panel board	94	0.41	154	55	70
	School	PCB	Sealed double glazing windows	94	0.10	154	47	82
	School	PCB	Capacitors	94	0.47	154	50	78
	School	PCB	Acrylic flooring	94	0.04	154	52	75
	Commercial building	PCB	Joint/sealant	94	0.27	85	39	52
6	Multifamily House	Asbestos	Cement panel board	93	0.73	153	64	55
	Multifamily House	Asbestos	Carpet glue	93	0.53	153	50	77
	Commercial building	Asbestos	Carpet glue	93	0.47	85	40	51
7	Multifamily House	Asbestos	Pipe insulation	92	0.82	153	18	126
	School	Asbestos	Valves	92	0.09	154	72	43
	School	PCB	Joint/sealant	92	0.18	154	40	93
8	Multifamily House	PCB	Joint/sealant	91	0.28	153	58	65
	Multifamily House	Asbestos	Sealed double glazing windows	91	0.17	153	66	52

*Positive detection rate = Number of Positives/(Total number of observations—Number of NA).

Table 4
The overview of the potential variables for predicting asbestos pipe insulation in multifamily houses and PCB joints or sealants in school buildings.

Potential features	Unit	Feature representation	Average	Feature representation	Average
		Asbestos pipe insulation		PCB joints or sealants	
City	–	(Gothenburg, Stockholm)	–	(Gothenburg, Stockholm)	–
EPC building category	–	(Premise, multifamily building)	–	(Premise, multifamily building)	–
EPC building type	–	(Detached, gable, intermediate)	–	(Detached, gable, intermediate)	–
Construction year	Year	(1903–1977)	1955	(1906–1983)	1962
Renovated	–	(0,1)	–	(0,1)	–
Renovation year	Year	(1959–2018)	1999	(1940–2016)	1993
Floor area	m^2	(174–23297)	3677	(135–17164)	2643
Numbers of floors	N	(1-14)	5	(1–7)	2
Number of basements	N	(0,1,2,3+)	1	(0,1)	1
Number of stairwells	N	(0-22)	3	(0–5)	0
Number of apartments	N	(0–376)	42	–	–
Ventilation type	–	(Exhaust, balanced, balanced with heat exchanger, exhaust with heat pump, natural ventilation)	–	(Exhaust, balanced, balanced with heat exchanger, natural ventilation)	–

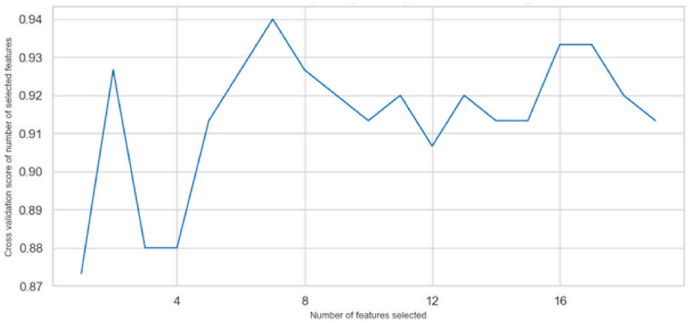
glues, door or windows insulation, and pipe insulation. On the contrary, PCB-containing materials have slightly lower scores than asbestos-containing materials, of which sealed double glazing windows, capacitors, acrylic flooring, and joints or sealants are promising for analysis. Their presence in schools, multifamily houses, and commercial buildings shows potential for modeling with relatively high data amount and quality. Synthesizing the data evaluation results, asbestos pipe insulation in multifamily houses and PCB joints or sealants in school buildings were chosen for further machine learning prediction.

The data characteristics of potential variables for predicting asbestos pipe insulation in multifamily houses and PCB joints or sealants in school buildings were illustrated in Table 4. The asbestos pipe insulation subset contains 139 multifamily houses with an average construction year of 1955 and the renovation year of 1997. On average, the multifamily buildings are 3677 m^2 with five stories, one basements, three stairwells, and 42 apartments. On the other hand, the PCB joints or sealants subset comprises 103 school buildings. The average construction year is later (1962), and the floor area is smaller (2643 m^2) compared to multifamily houses. The height of the school building is generally lower (1 story), and mostly with one basement but without stairwells. By featuring the building parameters of the studied building classes help determine the application boundary and representativeness of the prediction outcomes.

4.2. Performance evaluation between classification models

Feature selection, involving the recursive feature selection (RFE) and the Extra Tree classifier, was performed prior to applying the input data to the machine learning pipeline. Determining the number of features and feature sets can impact predictive models and thus are crucial hyperparameters to be configurated. The automatic tuning of the feature number using RFE was presented in Fig. 4, and the relevant features were ranked according to their relative importance with Extra Trees classifier in Fig. 5. The findings show that the optimal number of features for asbestos pipe insulation based on the cross-validation accuracy is seven and influential features are: floor area, construction year, the number of apartments, the number of stairwells, renovation year, and the number of floors. On the other hand, the favorable number of features for PCB joints or sealants are four, namely, floor area, construction year, balanced ventilation systems, and renovation year as critical features for prediction.

Data resampling was performed to facilitate cost-sensitive learning for imbalanced classes. This entails that the minority class is required to be oversampled to have an equal number in both classes. In this case, the number of negative detections of asbestos pipe insulation was oversampled to 94 and returned 188 observations for predictions. A similar skewed issue in the PCB joints subset was also addressed with the same approach. The fraction of negative detections is four times higher than the positive detections. Hence, the number of positive detections was oversampled to 54 and resulted in 106 observations for identical weights for misclassification. Afterward, iterating the prediction pipeline with varied combinations of features into the six supervised classifiers, the optimal models for specific prediction tasks were highlighted. The confusion matrix summarized in Table 5 showed promising prediction results. The comparable accuracy and recall rates within each classifier indicated a balance of correct retrieval of true positives from all observations and relevant observations. With three features, the presence of asbestos pipe insulation can be predicted to 74% and 78% of average accuracy and recall rates, as well as 87% and 91% of average accuracy and recall rates for tree-ensembled classifiers. Random forest and



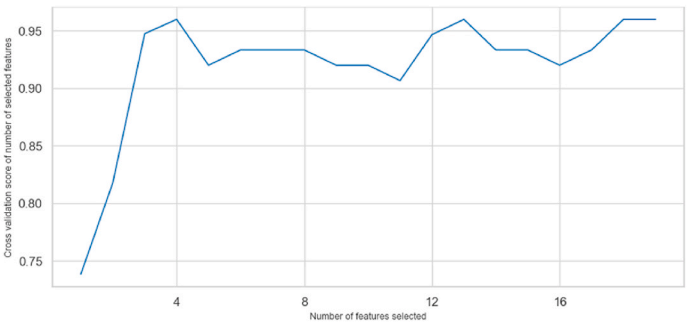


Fig. 4. Cross validation score as function of number of features used for selecting the optimal number of features. The optimal feature number for asbestos pipe insulation prediction (left) was seven and PCB joints or sealants prediction (right) was four.

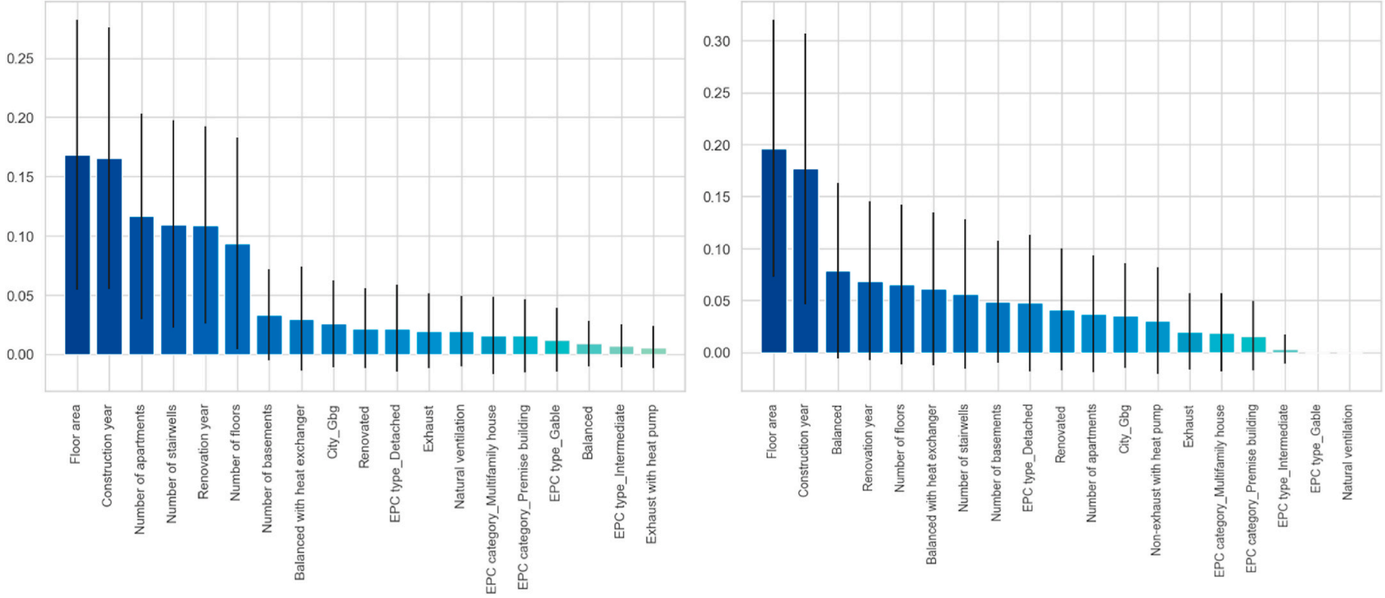


Fig. 5. The relative feature importance ranking for asbestos pipe insulation (left) and PCB joints or sealants (right) are listed in descending orders. The y axis concerns the normalized feature importance values and summed up to 1.0.

Table 5
The performance accuracy and recall rates of the optimal models for asbestos pipe insulation and PCB joints or sealant predictions.

Prediction	Asbestos pipe insulation (N = 188)		PCB joints or sealants (N = 106)	
Selected features	Construction year, Floor area, City		City, EPC category, Exhaust, Balanced, Balanced with heat exchanger, Natural ventilation, Construction year, Number of floors, Floor area, Number of basements, Number of stairwells	
	%accuracy	**%recall**	**%accuracy**	**%recall**
Logistic regression	55	62	67	68
SVM	50	58	67	69
kNN	74	78	81	81
Random forest	89	92	95	94
XGBoost	89	92	90	89
Catboost	84	88	95	94
Average	74	78	83	83
Average (tree-ensembled)	87	91	93	92

XGBoost reached the highest prediction rates. To predict PCB joints or sealants, more features were used and returned 83% and 83% of average accuracy and recall rates, as well as 93% and 92% of average accuracy and recall rates for tree-ensembled classifiers. The random forest and the CatBoost had the optimal prediction performance among the classifiers.

As an alternative performance evaluation matric for classification models, the receiver operating characteristic (ROC) graph plots the true positive rate (TPR) against the false positive rate (FPR) with the shifting decision threshold of the classifier, illustrated in Appendix B Fig. B2. AUC is scale-invariant and classification-threshold-invariant, which implies that it measures how well the predictions are ranked irrespective of what classification threshold is chosen [39]. A diagonal plot of ROC represents random guessing, whereas the preferable classifier will configure top left with a TPR of 1 and an FPR of 0 [23]. The ROC area under the curve (AUC) can thus be used to evaluate a classifier's performance within a range of 0 and 1. Fig. 6 shows the ROC AUC of the selected classifiers for the two prediction hypotheses. In terms of asbestos pipe insulation prediction in multifamily houses, CatBoost (AUC = 0,96) and the random forest classifier (AUC = 0,92) performed similarly well on the validation data subset. These ensemble classifiers, including the random forest, XGBoost, CatBoost classifiers (AUC = 1, 00), also outperformed individual classifiers in predicting PCB joints or sealants in school buildings and achieved a higher average AUC compared to pipe insulation classification. The results from AUC align with the conclusion of the previous confusion matrix that the pattern of

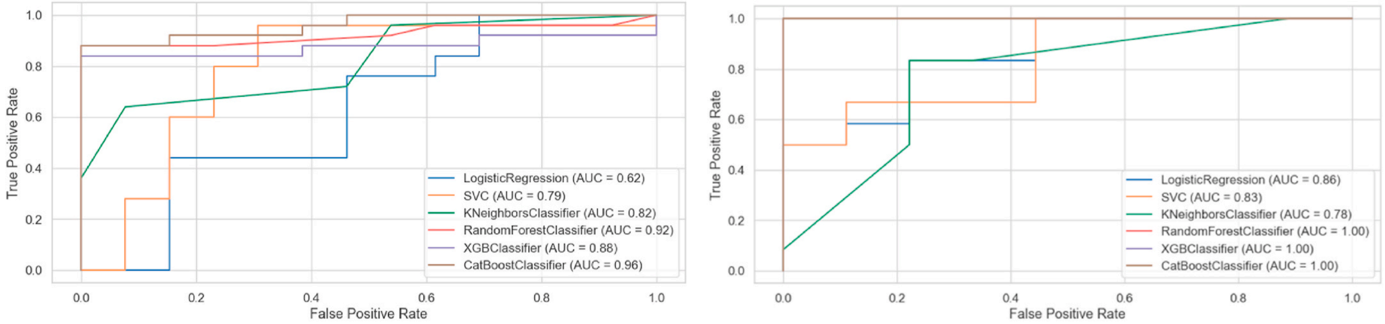


Fig. 6. The receiver operating characteristic curve plots the performance of the selected classifiers across the true positive rate and the false positive rate, measured by the area under the curve (AUC). The ROC curve for asbestos pipe insulation in multifamily houses (left) and PCB joints or sealants in school buildings (right).

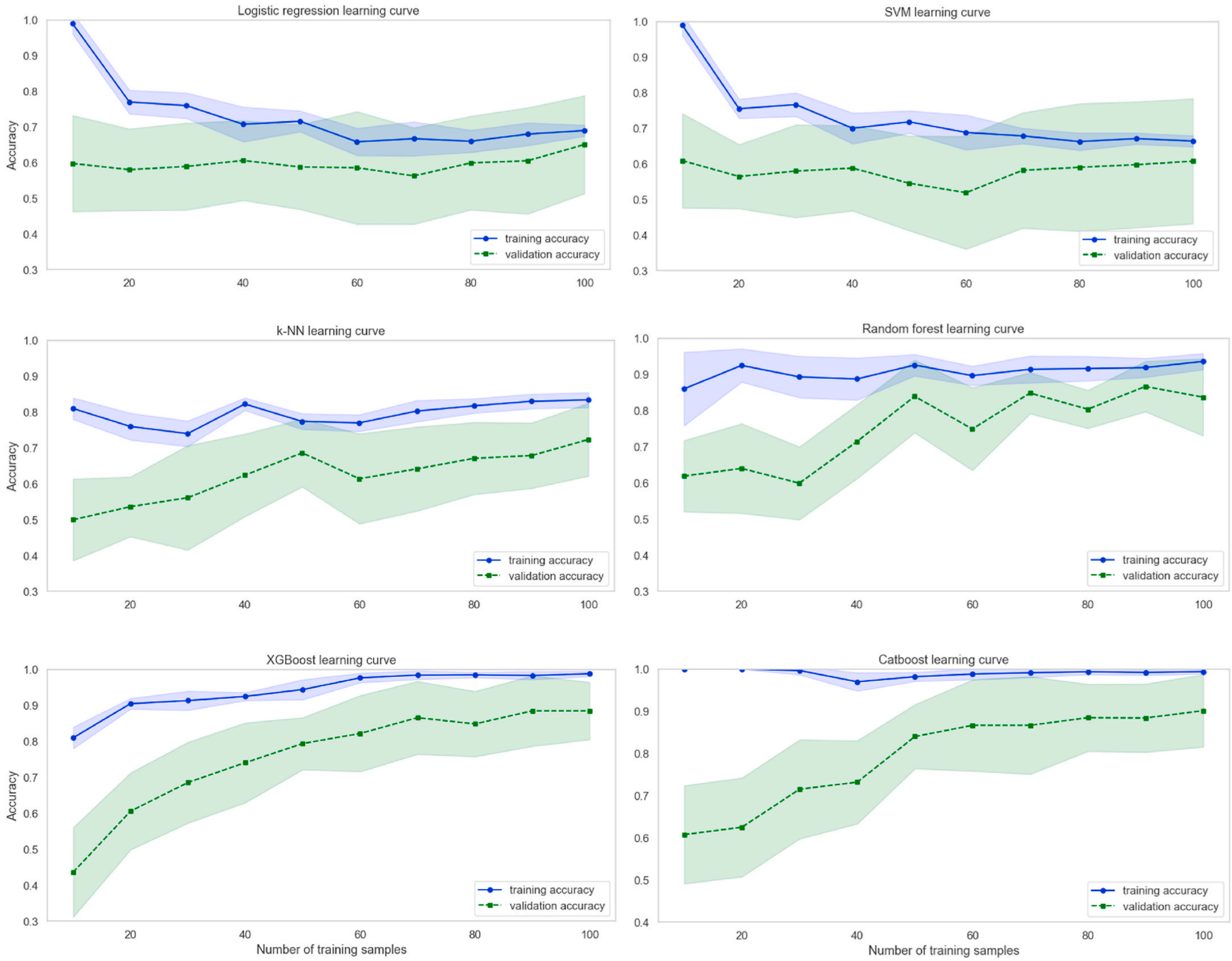


Fig. 7. Learning curves were created for selected classifiers to diagnose the bias and variance problem of the predication models by projecting the average accuracy changes in relation to increasing data the amount for training subset. The shaded areas of the training accuracy (expressed in blue) and validation accuracy (expressed in green) indicate the variance of the estimates.
(7.1) The learning curves of the selected classifiers in predicting asbestos pipe insulation in multifamily houses (the upper six figures).
(7.2) The learning curves of the selected classifiers for predicting PCB joints or sealants in school buildings (the lower six figures). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

PCB joints was easier to be identified, and the tree-ensembled classifiers perform better than other tested classifiers for the predefined prediction scopes.

4.3. *Impact of the data size on model performance*

Learning curves were schemed to diagnose bias and variance trade-off of the developed models and investigate whether more data input is needed to address the issue. In Fig. 7, the training accuracy and the 10-fold cross-validation accuracy with the increasing number of training samples were plotted. In all cases, the standard deviation of the validation scores was larger than the training scores. Besides, the overall validation accuracy rates reached 75–85% for the tree ensembled classifiers with a minimum of 50 data points. The learning curves for asbestos pipe insulation prediction suggest high bias and underfitting in the logistic regression and the SVM models; in contrast, the k-NN, XGBoost, and CatBoost models appear to be high variance with a large gap between the training and cross-validation accuracy. As for PCB sealants or joints prediction in school buildings, the learning curves indicate that the k-NN model is slightly underfitting with low training and cross-validation accuracies, while the tree-ensembled models tend to be overfitting. The logistic regression and the SVM model obtained a balanced bias and variance trade-off. To deal with the underfitting problem, the number of model parameters needs to be increased, or the degree of regularization should be decreased. On the other hand, the overfitting problem can be addressed with more training data, simplified models, increasing the regularization parameters, or reducing the number of features.

4.4. *Influential features for prediction*

The SHAP values were adopted to explain individual predictions and interpreted each feature's impacts. In Fig. 8, the overview of the SHAP values for every sample in the XGBoost and the CatBoost models, along with their feature values, were presented. Although various models may rank features in a different order, the overall influence of the feature

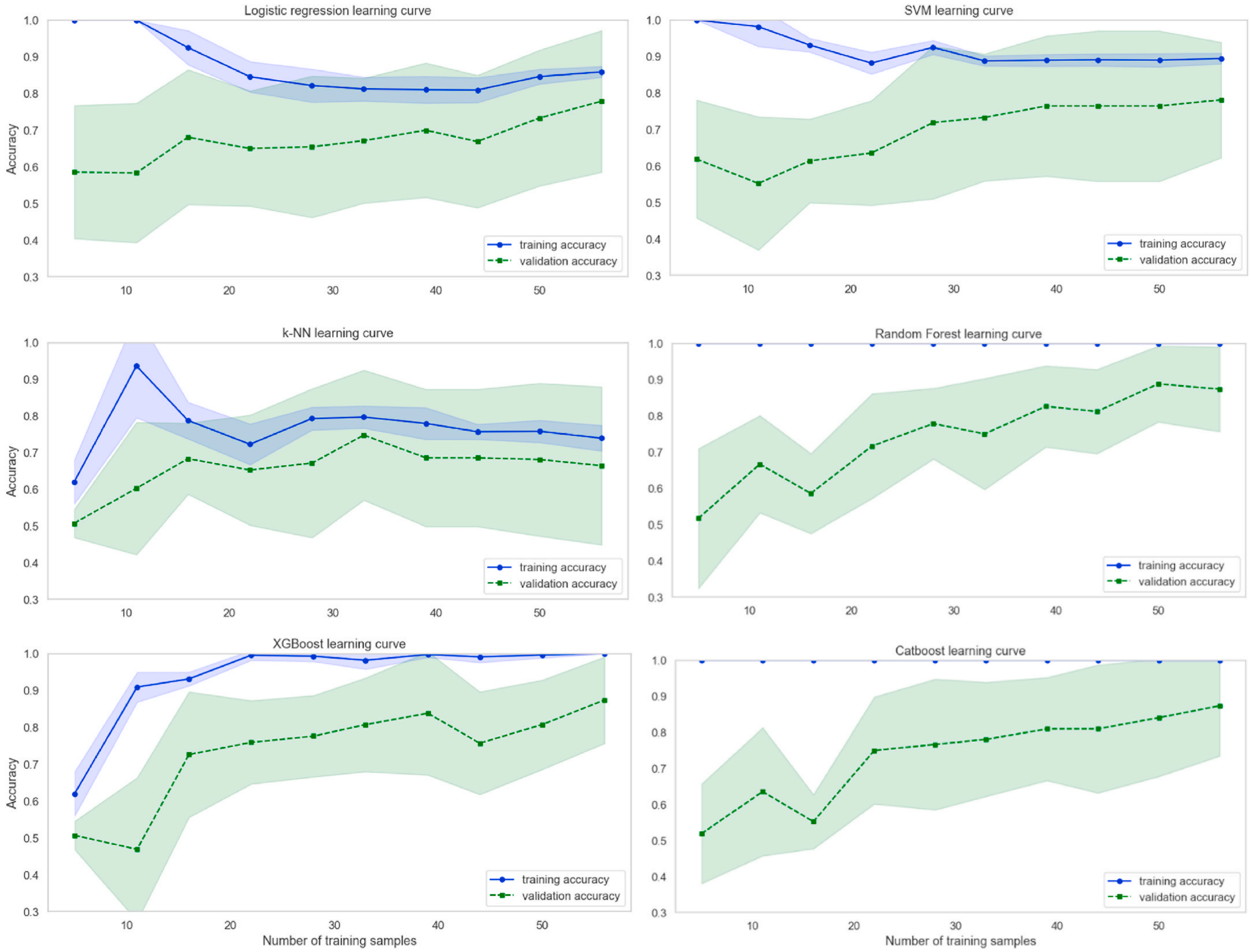


Fig. 7. (*continued*).

values on model output is aligned. If the SHAP values are close to 0, the features have nearly no contributions to the prediction output. For example, the EPC building category and building type had almost no impact on prediction and thus may be ignored by classifiers. The findings also show that the construction year has a significant impact on predicting asbestos pipe insulation in multifamily houses, followed by floor area, the number of stairwells, renovation year, and the number of apartments. The higher values of the construction year, renovation year, the number of basements, the more likely the multifamily houses are predicted with asbestos pipe insulation. In the case of PCB joints or sealants, construction year, balanced ventilation system, and floor area are the most crucial features for both models. If the schools were built in the later decades with balanced ventilation and a larger floor area, they were more probably be associated with the risk of PCB joints or sealants. To sum up the impact magnitudes of each feature, the aggregated sum of the SHAP value over all samples was presented in Appendix C.

5. Discussion

The section discusses the feasibility of using machine learning to predict the presence of hazardous materials in the building stock, specifically, the prediction results of asbestos and PCB-containing materials in two building classes in Gothenburg and Stockholm in Sweden, as well as the uncertainties in data preprocessing.

5.1. *Prediction result interpretation*

Collecting the training data for building stock analysis is effort and resource-demanding [40]. Given the variety of buildings and diverse ways to measure and document information, accessing the eligible and structural building-specific data is somehow challenging [41]. In addition, low data quality and the exclusion of extreme values also lower the available amount of data. Concerning the issue, Althnian et al. [42] explored the impact of dataset size on classification performance. Their results revealed that the overall performance of classifiers depends more on the dataset representation of the original distribution than data size. Also, there is a tendency that particular classifiers perform better than the others for the small tabular datasets with a mix of numerical and categorical data. Irrespective of the different building classes and data sizes in predicting asbestos pipe insulation and PCB joints or sealants, the random forest, XGBoost, and CatBoost classifiers had robust performance across data subsets. Our results agree with the previous study by Cha et al. [43] regarding the development of prediction models based on small datasets. The random forest classifier is proven to predict demolition waste generation with limited input variables effectively; yet, one should be aware of possible dataset overfitting, based on the results of the learning curves, due to the strong adaptability of the tree-ensembled algorithms. Collecting more data or/and intensifying algorithm regularization are desirable approaches to resolve the issue [27].

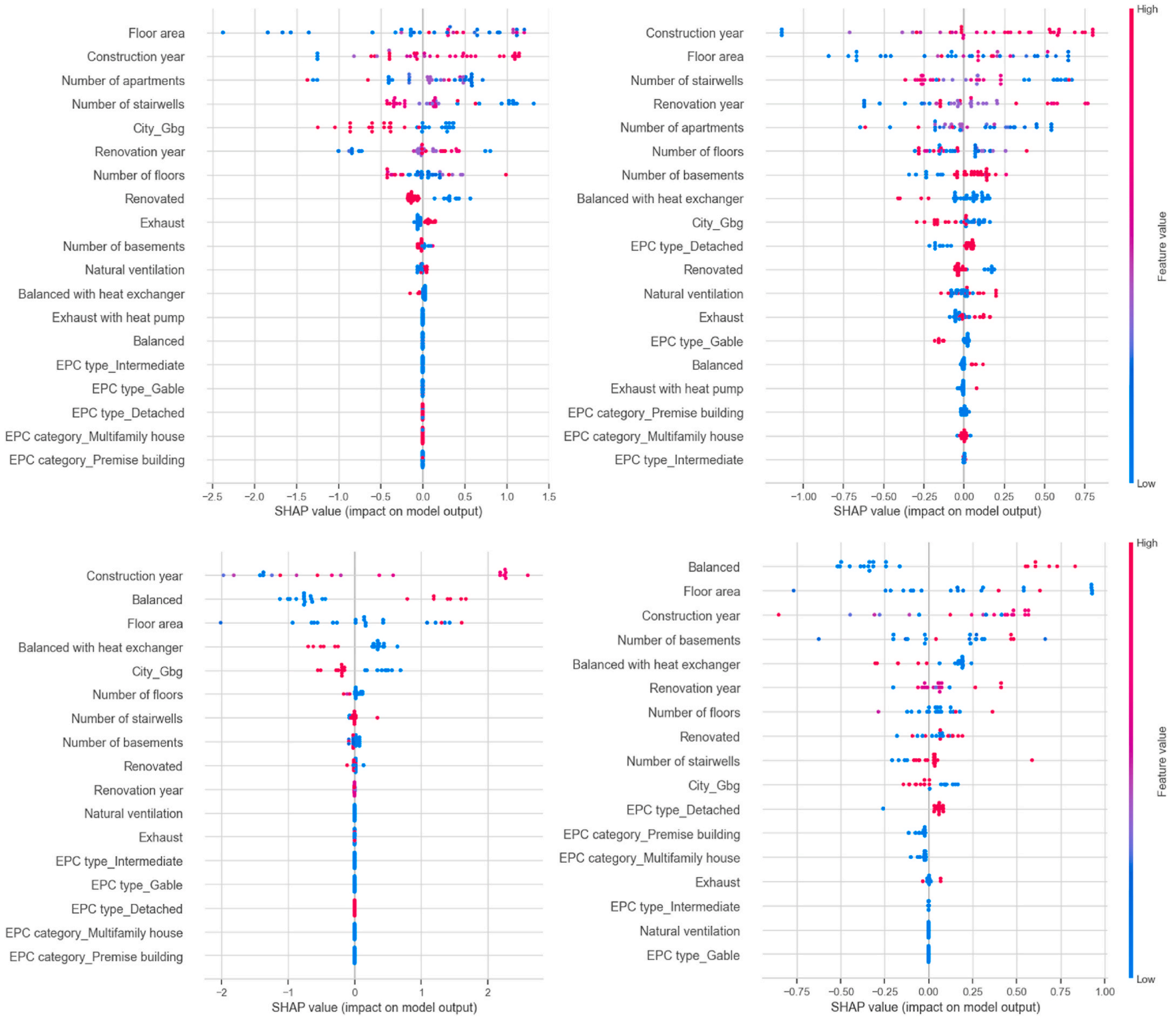


Fig. 8. The summary plots visualize the impact of features on the model output. The observations were represented as a single dot for each feature and densified along the x-axis. The red dots in the feature flow signified high feature values, while the blue dots were low feature values.
(8.1) The feature impacts on asbestos pipe insulation prediction in the XGBoost model (left) and the CatBoost model (right) for multifamily houses (the upper figures).
(8.2) The feature impacts on PCB joints or sealants prediction in the XGBoost model (left) and the CatBoost model (right) for school buildings (the lower figures). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The attempts of estimating the spatial distribution of asbestos cement roofing have been probed by Wilk et al. [20] and Krówczyńska et al. [19]. The potential of using the random forest classifier on aerial images and convolutional neural network on multispectral data to map national asbestos-cement roofing was verified with around 76% and 89% accuracies. In comparison, an average high prediction accuracy of 87% for asbestos pipe insulation in multifamily houses were obtained in the actual study with tree-ensembled classifiers. In fact, other asbestos-containing materials, such as door or windows insulation, cement panel, tile or clinker, carpet glue, floor mat, ventilation channel, and joints, also show high modeling potential for the multifamily houses, schools, and commercial buildings, based on the cross-validation scores. Whether the synergies of detection patterns between these contaminated building components exist requires further investigation.

On the other hand, no previous relevant studies are found regarding the prediction of PCB-containing materials using machine learning techniques; instead, these focus primarily on PCB stock estimation and exposure reduction. The results are not surprising as developing artificial intelligence applications for waste generation estimation, onsite sorting, and collection rather the focus areas in the previous construction and demolition waste management research. For instance, Akanbi et al. [44] and Cha et al. [45] predicted demolition waste generation based on deep learning models and the decision tree method. Bonifazi et al. [46,47] developed asbestos waste recognition tools for onsite sorting using unsupervised learning on hyperspectral images. A combination of image processing and supervised learning classifiers for recycled aggregates identification was explored by Kuritcyn et al. [48] and Anding et al. [22] for waste fraction collection. Nevertheless, studies featuring the in situ hazardous material identification from end-of-life building stock are few and mostly remain on descriptive analysis using statistical approaches. Franzblau et al. [10] and Govorko et al. [15,17] characterized the frequent asbestos-containing materials, asbestos types and conditions, as well as the disturbing likelihood in residential buildings. In light of the broad uptake of pre-demolition audits for enabling circular construction [5], along with its integration with the EU REACH regulation (EC 1907/2006, Registration, Evaluation, Authorization, and Restriction of Chemicals) [3], more data-driven material assessment tools associated with the risk of hazardous materials can be expected to aid selective demolition [11].

When predicting the presence of hazardous materials, the false negative (Type II error) has a more serious consequence than the false positive (Type I error). To address the uneven cost of misclassification, the cost-sensitive learning was explored in this study by resampling the imbalanced class, and the prediction accuracy and recall rates obtained were nearly equally high. There are other methods in theory to achieve the same objective, including cost-sensitive algorithms and cost-sensitive ensembles [36]. The former modifies the class weight, i.e., SVM, decision tree, or appends costs as a penalty for misclassification, i.e., logistic regression. The latter refers to wrapper methods (or meta-learners, ensembles) by relabeling examples in the training dataset to minimize data preprocessing costs [36]. The application of these two techniques is limited to algorithm-specific augmentations and is time-consuming for developing and testing; thus, it is not feasible for the study when multiple classifiers are involved. Empirically, the misclassification costs are determined by domain experts, and the performance varies accordingly. To calculate the optimal cost weights, Lu et al. [49] tested the grid searching and the fitted function strategies. Their finding showed that the fitted functions of the extreme learning machine, a single hidden layer feed-forward neural network learning algorithm, are more effective in achieving a high weighted classification accuracy. This method's adaptability and stability, when used on the hazardous material dataset, remain to be explored.

The current understanding of the factors relating to hazardous material detection is limited. Using product labels to identify the building components' production year or manufacturers is common in practice, especially for PCB-containing sealed double glazing windows and asbestos-containing fire doors [50]. The timeline of the product used in construction is by far the most acknowledgeable feature for pattern identification of hazardous materials. Mecharnia et al. [14] verified the possibility to use this through exploiting temporal data of the use of asbestos products to estimate the presence probability. Wilk et al. [51] also explored the determinants associated with the amount of asbestos-cement roofing and pinpointed the following factors: regional building stock structure, geographical distance to manufacturing plants, construction year, and the economic situation of the municipality. Regardless of distinct building stock compositions in different contexts, our study also highlighted the significant impact of the construction year, renovation year, and the medium-to-low impact of the city the building situated.

The influential features identified by the Extra Tree classifier are quite aligned to the features with high SHAP values in XGBoost and CatBoost models. However, in our comprehensive search for the optimal feature combinations of the highest prediction performance, the best-performed feature set can differ from the suggested number of the features from the REF due to varied computational logic behind algorithms. Through assembling the recognized contributing features from model preprocessing and post-analysis, it is evident that construction year, renovation year, floor area, the number of stairwells and apartments are critical features for asbestos pipe insulation prediction in multifamily houses. Nevertheless, adding these features in complex models led to a drop in overall accuracy. An opposite situation was observed in predicting PCB joints and selants in school buildings. Simple models with few features were advised by REF, yet the highest overall accuracy was attained by modeling on several features. The identified important feature from feature selection is principally in agreement with those with high impact magnitudes. Attaching additional information such as city, EPC building category, and specific building parameters to the PCB joints and selants models can slightly increase the prediction performance.

5.2. Data uncertainty

Building stock data are, by their nature heterogeneous and unstandardized, which poses substantial challenges for data enrichment and analysis. To add building-specific information to the building registers database, many relationships require to be established to couple the empirical and registered data correctly [52]. However, these relationships are sometimes lacking due to poor data quality [53], varied aggregation levels [26], or incomplete information [54]. These issues were also experienced in the data matching between multiple building registers and hazardous waste inventories. The matched observations with one-to-one relationships constitute 69.0% of the dataset.

To deal with the data uncertainty, it is therefore of great importance to stratify eligible data with similar characteristics. Stratifying the observations with the same building class and similar building parameters helps remove outliers prior to modeling. The cross-validation matrix proved to be an effective way to identify feasible hypotheses for testing by considering their available data quality and quantity. Multifamily houses and school buildings present a promising modeling potential, yet some data limitations need to be clarified. For instance, the pre-demolition inventories from multifamily houses were usually created for the specific renovated area; thus, the inventory records may not be fully representable for this building class. In addition, school buildings lack real estate taxation data and are challenging to distinguish from similar buildings on the complex property. Despite these underlying data uncertainties, these two building classes are more homogeneous in terms of building features, the choice of materials, and parallel construction periods [26]. Besides, previous research on characterizing asbestos-containing materials in residential dwellings [10,17], as well as PCB-containing components in residential areas [55] and school buildings [9], are available. This domain knowledge and the expert assumptions from pre-demolition auditors navigate the selection of the potential features and the focus on predicting hazardous materials.

In this study, the inventory data extraction and compiling rely on manual operations as the no query-based, single-format, digital databases for pre-demolition audit documents exist. Also, the registered data were not structured for analysis purposes, making merging and matching difficult. To raise public awareness of asbestos while supporting data collection, Govorko et al. [15,16] developed a mobile phone application to identify asbestos-containing materials in residential dwellings and automatically generate assessment reports. Utilizing the self-assessment yields an advantage for transforming the field data in a machine-friendly manner. Based on the same concept, if the digital protocol for hazardous waste inventories is designed and used to retrieve the building registers data with the respective national real estate index and building ID, a lot of time on the desk study and the risk of potential recording or estimation errors of the building parameters, i.e., construction year, renovation year, and area can be minimized. This bottom-up approach can also contribute to register data synchronization, as well as instant version control.

6. Conclusions

The study demonstrates the possibility of applying machine learning in predicting hazardous materials in particular building classes based on inventories of hazardous waste and building registers. Two prediction hypotheses – asbestos pipe insulation in multifamily houses and PCB joints or selants in school buildings - were tested and evaluated with six supervised classifiers. Using cost-sensitive learning, the tree-ensembled classifiers, i.e., random forest, XGBoost, and CatBoost, performed well in the small, low dimensional datasets. According to the learning curves, high validation accuracies, namely 75–85%, were obtained for the tree-ensembled classifiers after training on a minimum of 50 data points. Construction year, floor area, renovation year, the number of stairwells and apartments were vital features for predicting the presence of asbestos pipe insulation in multifamily houses with the optimal average accuracy and recall rates of 74% and 83%. In comparison, average accuracy and recall rates of 78% and 83% were obtained for predicting PCB joints or sealants in school buildings with construction year, balanced ventilation, floor area, and balanced ventilation with heat exchanger.

Enhancing the quality of the mixed waste is a prerequisite for realizing circular material in construction and relies on accurate hazardous material identification in semi-selective demolition. The study presents the challenges of utilizing the past hazardous waste inventories from pre-demolition audits for data-driven management. These insights from the post-analysis perspective can be valuable for the EU Construction and Demolition Waste Management Protocol concerning data consistency, quality, and completeness. Future research is advised to test the machine learning pipeline on inventory data from building stocks in other municipalities and integrate the learning outputs to improve the model's generalizability on the national and international scale.

Funding

This research was funded by the Swedish Foundation for Strategic Research (SSF), grant number FID18-0021.

Institutional review board statement

Not applicable.

Informed consent statement

Not applicable.

Data availability statement

The environmental inventories used in the study are confidential and regulated by the Gothenburg City Archive and the Stockholm City Archive. The national building registers acquired from different authorities were requested for the specific research purpose. Therefore, they are not available online.

CRediT authorship contribution statement

Pei-Yu Wu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Claes Sandels:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **Kristina Mjörnell:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing. **Mikael Mangold:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Tim Johansson:** Writing – review & editing, Software, Resources, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work is part of the research project Prediction of hazardous materials in buildings using AI and is supported by RISE Research Institutes of Sweden. A special thank is sent to Frida Palstam for the assistance in transforming the raw data into the hazardous material dataset, as well as the Gothenburg City Archive and the Stockholm City Archive for the support of searching hazardous waste inventories in the building permit documents.

Appendix A

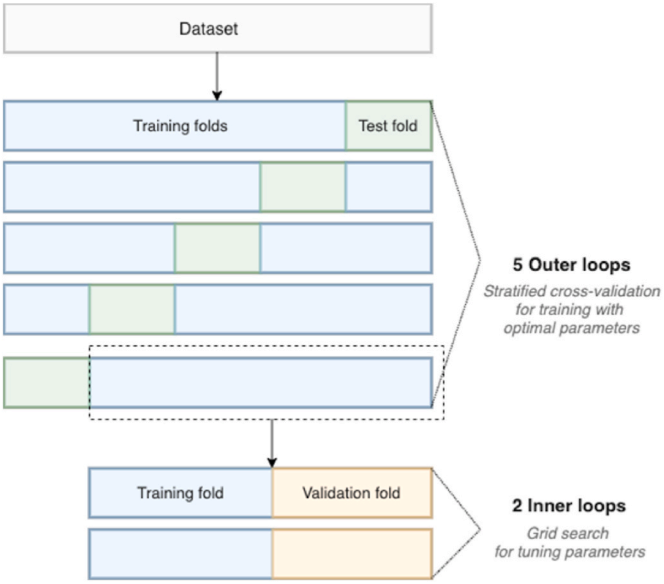

Dataset
Training folds
Test fold
5 Outer loops
Stratified cross-validation for training with optimal parameters
Training fold
Validation fold
2 Inner loops
Grid search for tuning parameters

Fig. A1. A diagram of nested cross-validation for optimal hyperparameter selection, adopted from Raschka and Mirjalili [23]. The outer loops of training folds and test folds were cross-validated with optimal parameters. Then the training folds of the outer loops were further divided into micro training fold and validation fold in the inner loops, where Grid Search was conducted for parameter tuning.

Appendix B

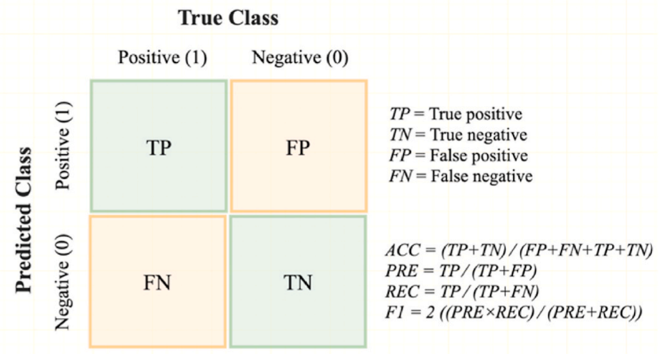


Fig. B1. Confusion matric adopts the performance metrics, such as accuracy (ACC), precision (PRE), recall (REC), and F1-score, to evaluate the model relevance.

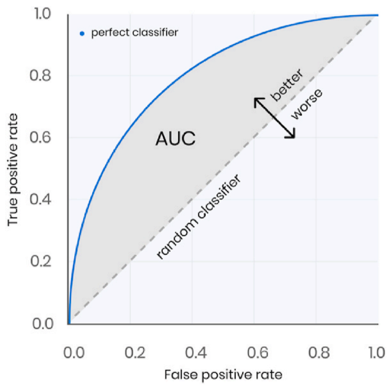


Fig. B2. The AUC is the area under the ROC curve used for measuring the model's classification performance. A higher AUC implies better model performance (Adopted from Evispot [39]).

Appendix C

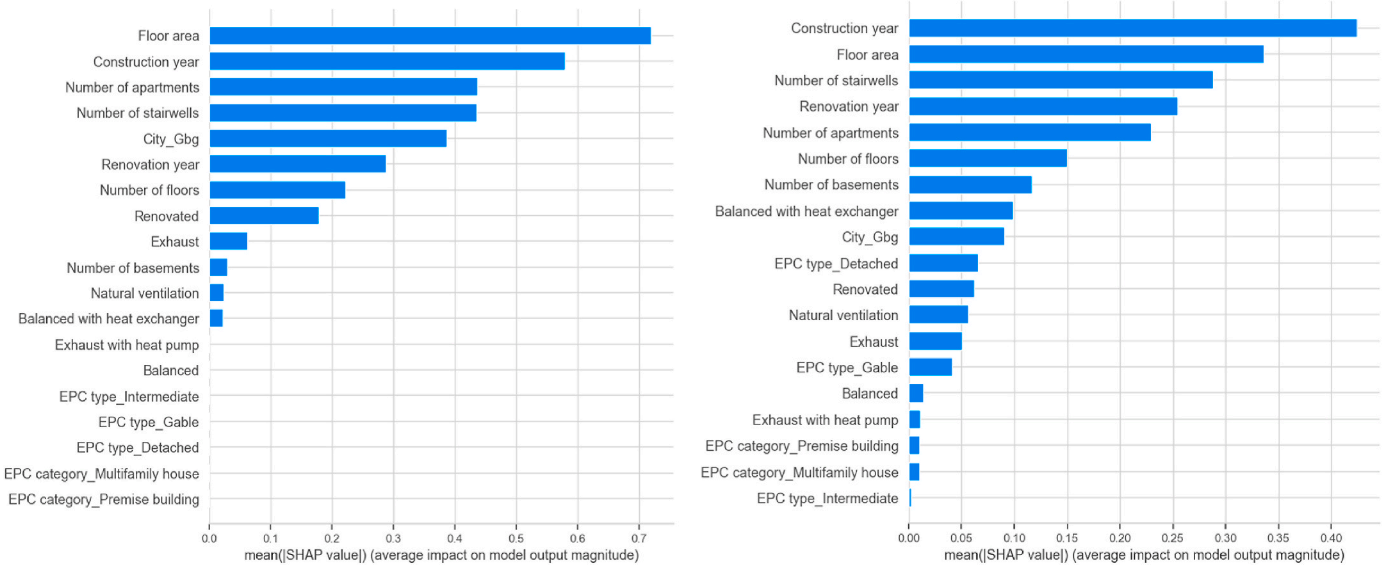


Fig. C1.1. The impact magnitudes of the features on asbestos pipe insulation prediction in the XGBoost model (left) and the CatBoost model (right) for multi-family houses.

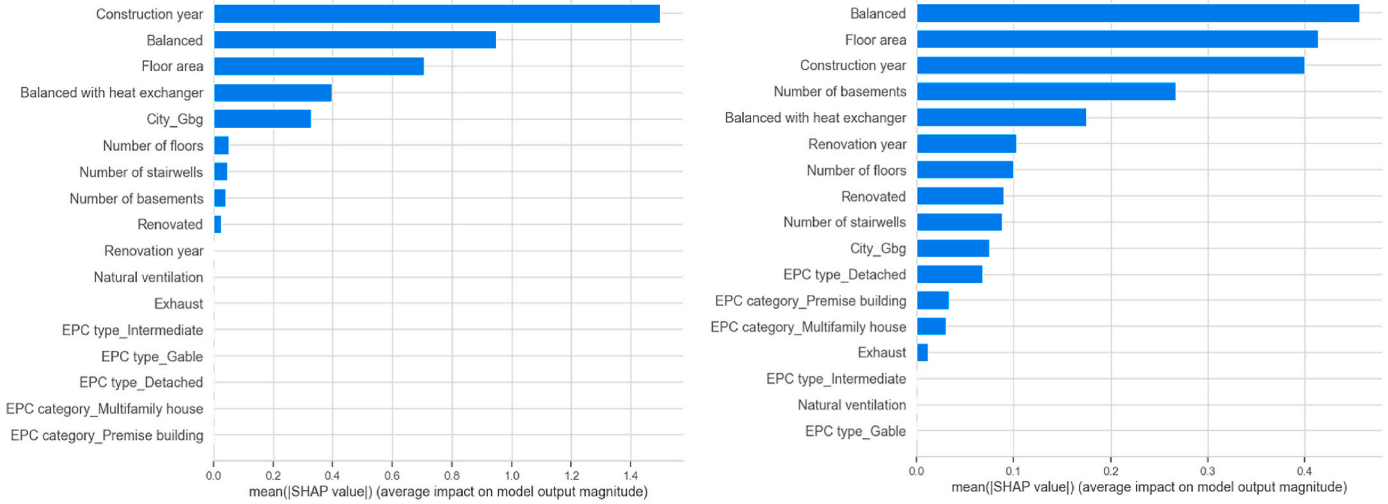


Fig. C1.2. The impact magnitudes of the features on PCB joints or sealants prediction in the XGBoost model (left) and the CatBoost model (right) for school buildings. Fig. C1. The bar plots summarized the average impact of features on the model output magnitudes.

References

[1] M. Wahlström, J. Bergmans, T. Teittinen, J. Bachér, A. Smeets, A. Paduart, Construction and Demolition Waste : Challenges and Opportunities in a Circular Economy, Mol, Belgium, 2020. https://www.eea.europa.eu/publications/construction-and-demolition-waste-challenges/at_download/file.

[2] M. Wahlström, T. Teittinen, T. Kaartinen, van C. Liesbet, Hazardous Substances in Construction Products and Materials: PARADE. Best Practices for Pre-demolition Audits Ensuring High Quality RAw Materials, Esbo, Finland, 2019.

[3] C. Bodar, J. Spijker, J. Lijzen, S. Waaijers-van der Loop, R. Luit, E. Heugens, M. Janssen, P. Wassenaar, T. Traas, Risk management of hazardous substances in a circular economy, J. Environ. Manag. 212 (2018) 108–114, https://doi.org/10.1016/j.jenvman.2018.02.014.

[4] ECORYS, EU Construction & Demolition Waste Management Protocol, 2016. Brussels, Belgium.

[5] European Commission, Guidelines for the waste audits before demolition and renovation works of buildings, UE Construct. Demol. Waste Manag. Ref. Ares (2018), 4724185 - 14/09/2018. (2018) 37.

[6] European Commission, Waste Framework Directive, 2008 n.d.), https://ec.europa.eu/environment/topics/waste-and-recycling/waste-framework-directive_en. (Accessed 5 July 2021).

[7] Y. Bouabdallaoui, Z. Lafhaj, P. Yim, L. Ducoulombier, B. Bennadji, Predictive maintenance in building facilities: a machine learning-based approach, Sensors 21 (2021) 1–15, https://doi.org/10.3390/s21041044.

[8] T.M. Spear, J.F. Hart, T.E. Spear, M.M. Loushin, N.N. Shaw, M.I. Elashhab, The presence of asbestos-contaminated vermiculite attic insulation or other asbestos-containing materials in homes and the potential for living space contamination, J. Environ. Health 75 (2012) 24–29.

[9] K.W. Brown, T. Minegishi, C.C. Cummiskey, M.A. Fragala, R. Hartman, D. L. MacIntosh, PCB remediation in schools: a review, Environ. Sci. Pollut. Res. 23 (2016) 1986–1997, https://doi.org/10.1007/s11356-015-4689-y.

[10] A. Franzblau, A.H. Demond, S.K. Sayler, H. D'Arcy, R.L. Neitzel, Asbestos-containing materials in abandoned residential dwellings in Detroit, Sci. Total Environ. 714 (2020) 136580, https://doi.org/10.1016/j.scitotenv.2020.136580.

[11] M. Rašković, A.M. Ragossnig, K. Kondracki, M. Ragossnig-Angst, Clean construction and demolition waste material cycles through optimised pre-demolition waste audit documentation: a review on building material assessment tools, Waste Manag. Res. 38 (2020) 923–941, https://doi.org/10.1177/0734242X20936763.

[12] P. Wu, K. Mjörnell, M. Mangold, C. Sandels, T. Johansson, A data-driven approach to assess the risk of encountering hazardous materials in the building stock based on environmental inventories, Sustain. Times 13 (2021) 1–26.

[13] P.-Y. Wu, K. Mjörnell, C. Sandels, M. Mangold, Machine learning in hazardous building material management: research status and applications, Recent Prog. Mater. 3 (2021), https://doi.org/10.21926/rpm.2102017, 1–1.

[14] T. Mecharnia, L.C. Khelifa, N. Pernelle, F. Hamdi, An approach toward a prediction of the presence of asbestos in buildings based on incomplete temporal descriptions of marketed products, in: K-CAP 2019 - Proc. 10th Int. Conf. Knowl. Capture, 2019, pp. 239–242, https://doi.org/10.1145/3360901.3364428. Marina del Rey, United States.

[15] M.H. Govorko, L. Fritschi, A. Reid, Accuracy of a mobile app to identify suspect asbestos-containing material in Australian residential settings, J. Occup. Environ. Hyg. 15 (2018) 598–606, https://doi.org/10.1080/15459624.2018.1475743.

[16] M.H. Govorko, L. Fritschi, J. White, A. Reid, Identifying asbestos-containing materials in homes: design and development of the ACM check mobile phone app, JMIR form, Res. 1 (2017) e7, https://doi.org/10.2196/formative.8370.

[17] M. Govorko, L. Fritschi, A. Reid, Using a mobile phone app to identify and assess remaining stocks of in situ asbestos in Australian residential settings, Int. J. Environ. Res. Publ. Health 16 (2019), https://doi.org/10.3390/ijerph16244922.

[18] H. Wu, J. Zuo, G. Zillante, J. Wang, H. Yuan, Status quo and future directions of construction and demolition waste research: a critical review, J. Clean. Prod. 240 (2019) 118163, https://doi.org/10.1016/j.jclepro.2019.118163.

[19] M. Krówczyńska, E. Raczko, N. Staniszewska, E. Wilk, Asbestos-cement roofing identification using remote sensing and convolutional neural networks (CNNs), Rem. Sens. 12 (2020) 1–16, https://doi.org/10.3390/rs12030408.

[20] E. Wilk, M. Krówczyńska, B. Zagajewski, Modelling the spatial distribution of asbestos-cement products in Poland with the use of the random forest algorithm, Sustain. Times 11 (2019), https://doi.org/10.3390/su11164355.

[21] M.B.A. Gibril, H.Z.M. Shafri, A. Hamedianfar, New semi-automated mapping of asbestos cement roofs using rule-based object-based image analysis and Taguchi optimization technique from WorldView-2 images, Int. J. Rem. Sens. 38 (2017) 467–491, https://doi.org/10.1080/01431161.2016.1266109.

[22] K. Anding, E. Linß, H. Träger, M. Rückwardt, A. Göpfert, Optical identification of construction and demolition waste by using image processing and machine learning methods, 14th Jt. Int. IMEKO TC1, TC7, TC13 Symp. Intell. Qual. Meas. - theory, Educ. Train, Held Conj. 56th IWK Ilmenau Univ. Technol. (2011) 126–132, 2011.

[23] T. Hong, Z. Wang, X. Luo, W. Zhang, State-of-the-art on research and applications of machine learning in the building life cycle, Energy Build. 212 (2020) 109831, https://doi.org/10.1016/j.enbuild.2020.109831.

[24] P.-Y. Wu, K. Mjörnell, M. Mangold, C. Sandels, T. Johansson, Tracing hazardous materials in registered records : a case study of demolished and renovated buildings in Gothenburg tracing hazardous materials in registered records : a case study of demolished and renovated buildings in Gothenburg, J. Phys. Conf. Ser. 2069 (2021), https://doi.org/10.1088/1742-6596/2069/1/012234.

[25] Deloitte, Study on Resource Efficient Use of Mixed Wastes, Improving Management of Construction and Demolition Waste - Final Report, Nantes, France, 2017. https://ec.europa.eu/environment/waste/studies/pdf/CDW_Final_Report.pdf. (Accessed 17 January 2021).

[26] T. Johansson, T. Olofsson, M. Mangold, Development of an energy atlas for renovation of the multifamily building stock in Sweden, Appl. Energy 203 (2017) 723–736, https://doi.org/10.1016/j.apenergy.2017.06.027.

[27] S. Raschka, V. Mirjalili, Python Machine Learning - Second Edition: Machine Learning and Deep Learning with Python, Scikit-Learn, and Tensorflow, Packt Publishing Ltd., Birmingham, 2017.

[28] J. Brownlee, Feature selection in Python with scikit learn, Mach. Learn. Mastery. (2020) 1–7. https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/. (Accessed 9 July 2021).

[29] J. Brownlee, Feature selection for machine learning in Python, Mach. Learn. Mastery (2020). https://machinelearningmastery.com/feature-selection-machine-learning-python/. (Accessed 8 July 2021).

[30] J. Von Platten, C. Sandels, K. Jörgensson, V. Karlsson, M. Mangold, K. Mjörnell, Using machine learning to enrich building databases-methods for tailored energy retrofits, Energies 13 (2020), https://doi.org/10.3390/en13102574.

[31] J. Brownlee, A gentle introduction to XGBoost for applied machine learning, Mach. Learn. Mastery. (2021). https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/. (Accessed 13 July 2021).

[32] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system. https://doi.org/10.1145/2939672.2939785 n.d.

[33] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, (n.d.), https://github.com/catboost/catboost. (Accessed 13 July 2021).

[34] J.T. Hancock, T.M. Khoshgoftaar, CatBoost for big data: an interdisciplinary review, J. Big Data. 7 (2020), https://doi.org/10.1186/s40537-020-00369-8.

[35] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Prog. Artif. Intell. 5 (2016) 221–232, https://doi.org/10.1007/S13748-016-0094-0.

[36] J. Brownlee, Cost-sensitive learning for imbalanced classification. https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/, 2020. (Accessed 27 November 2021).

[37] R. Roscher, B. Bohn, M.F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, IEEE Access 8 (2020) 42200–42216, https://doi.org/10.1109/ACCESS.2020.2976199.

[38] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017). https://github.com/slundberg/shap. (Accessed 9 July 2021).

[39] Evispot, Area under the ROC curve (AUC) - Evispot. https://evispot.ai/area-under-the-roc-curve-auc/, 2021. (Accessed 21 December 2021).

[40] G.W. Cha, Y.C. Kim, H.J. Moon, W.H. Hong, New approach for forecasting demolition waste generation using chi-squared automatic interaction detection (CHAID) method, J. Clean. Prod. 168 (2017) 375–385, https://doi.org/10.1016/J.JCLEPRO.2017.09.025.

[41] N. Kohler, P. Steadman, U. Hassler, Building Research & Information Research on the Building Stock and its Applications, 2010, https://doi.org/10.1080/09613210903189384.

[42] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. Bin Dris, N. Alzakari, A. Abou Elwafa, H. Kurdi, Impact of dataset size on classification performance: an empirical evaluation in the medical domain, Appl. Sci. 11 (2021) 1–18, https://doi.org/10.3390/app11020796.

[43] G.W. Cha, H.J. Moon, Y.M. Kim, W.H. Hong, J.H. Hwang, W.J. Park, Y.C. Kim, Development of a prediction model for demolition waste generation using a random forest algorithm based on small datasets, Int. J. Environ. Res. Publ. Health 17 (2020) 1–15, https://doi.org/10.3390/ijerph17196997.

[44] L.A. Akanbi, A.O. Oyedele, L.O. Oyedele, R.O. Salami, Deep learning model for Demolition Waste Prediction in a circular economy, J. Clean. Prod. 274 (2020), https://doi.org/10.1016/j.jclepro.2020.122843.

[45] G.W. Cha, Y.C. Kim, H.J. Moon, W.H. Hong, New approach for forecasting demolition waste generation using chi-squared automatic interaction detection (CHAID) method, J. Clean. Prod. 168 (2017) 375–385, https://doi.org/10.1016/j.jclepro.2017.09.025.

[46] G. Bonifazi, G. Capobianco, S. Serranti, Hyperspectral imaging and hierarchical PLS-DA applied to asbestos recognition in construction and demolition waste, Appl. Sci. 9 (2019) 1–15, https://doi.org/10.3390/app9214587.

[47] G. Bonifazi, G. Capobianco, S. Serranti, Asbestos containing materials detection and classification by the use of hyperspectral imaging, J. Hazard Mater. 344 (2018) 981–993, https://doi.org/10.1016/j.jhazmat.2017.11.056.

[48] P. Kuritcyn, K. Anding, E. Linß, S.M. Latyev, Increasing the safety in recycling of construction and demolition waste by using supervised machine learning, J. Phys. Conf. Ser. 588 (2015), https://doi.org/10.1088/1742-6596/588/1/012035.

[49] H. Lu, Y. Xu, M. Ye, K. Yan, Q. Jin, Z. Gao, Learning Misclassification Costs for Imbalanced Datasets, Application in Gene Expression Data Classification, (n.d.). https://doi.org/10.1007/978-3-319-95930-6_47.

[50] Kretsloppsrådet, Resurs Och Avfallshantering Vid Byggande Och Rivning, 2019.

[51] E. Wilk, M. Krówczyńska, P. Pabjanek, Determinants influencing the amount of asbestos-cement roofing in Poland, Misc. Geogr. 19 (2015) 82–86, https://doi.org/10.1515/mgrsd-2015-0014.

[52] Z. Feng, W. Mayer, M. Stumptner, G. Grossmann, W. Huang, Relationship matching of data sources: a graph-based approach, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 10816 LNCS (2018) 539–553, https://doi.org/10.1007/978-3-319-91563-0_33.

[53] M. Mangold, M. Österbring, H. Wallbaum, Handling data uncertainties when using Swedish energy performance certificate data to describe energy usage in the building stock, Energy Build. 102 (2015) 328–336, https://doi.org/10.1016/j.enbuild.2015.05.045.

[54] Z. Yang, F. Xue, W. Lu, Handling missing data for construction waste management: machine learning based on aggregated waste generation behaviors, Resour. Conserv. Recycl. 175 (2021) 105809, https://doi.org/10.1016/j.resconrec.2021.105809.

[55] M.L. Diamond, L. Melymuk, S.A. Csiszar, M. Robson, Estimation of PCB stocks, emissions, and urban fate: will our policies reduce concentrations and exposure? Environ. Sci. Technol. 44 (2010) 2777–2783, https://doi.org/10.1021/es9012036.