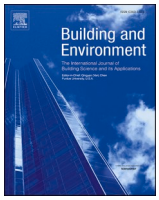
Building and Environment 245 (2023) 110879

Contents lists available at ScienceDirect

Building and Environment

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



Indoor radon interval prediction in the Swedish building stock using machine learning

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

[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:
Indoor radon
Predictive modeling
XGBoost
Deep learning
Radon exposure estimation
Regional building stock

ABSTRACT

Indoor radon represents a health hazard for occupants. However, the indoor radon measurement rate is low in Sweden because of no mandatory requirements. Measuring indoor radon on an urban scale is complicated, machine learning exploiting existing data for pattern identification provides a cost-efficient approach to estimate indoor radon exposure in the building stock. Extreme gradient boosting (XGBoost) models and deep neural network (DNN) models were developed based on indoor radon measurement records, property registers, and geogenic information. The XGBoost models showed promising results in predicting indoor radon intervals for different types of buildings with macro-F1 between 0.93 and 0.96, whereas the DNN models attained macro-F1 between 0.64 and 0.74. After that, the XGBoost models trained on the national indoor radon dataset were transferred to fit building registers in metropolitan regions to estimate the indoor radon intervals in non-measured and measured buildings by regions and building classes. By comparing the prediction results and the statistical summary of indoor radon intervals in measured buildings, the model uncertainty and validity were determined. The study ascertains the prediction performance of machine learning models in classifying indoor radon intervals and discusses the benefits and limitations of the data-driven approach. The research outcomes can assist preliminary large-scale indoor radon distribution estimation for relevant authorities and guide onsite measurements for prioritized building stock prone to indoor radon exposure.

1. Introduction

Indoor radon is a universal health hazard and the second leading cause of lung cancer worldwide. Approximately 15% of lung cancers in Sweden are induced by indoor radon in dwelling buildings, corresponding to 500 lung cancer cases every year [1,2]. The exposure to residential radon is particularly severe in cold climates, given the long time spent indoors in buildings with insufficient ventilation. To address the health risk of indoor radon and monitor its exposure in the indoor environment, most European countries adopt three indoor radon reference thresholds: (i) 200 Bq/m^3 for residential and public buildings and as the highest acceptable level for new buildings, (ii) 400 Bq/m^3 for existing buildings, (iii) above 1,000 Bq/m^3 for immediate decontamination [3]. From available measurement records, it is estimated that around 16% of single-family houses and 19% of workplaces exceed the indoor radon reference level in the Swedish building stock [2]. No requirements on the measured frequency have been put in place nowadays; measurements are, however, recommended every ten years or after an extensive renovation that may affect the indoor radon concentration. For buildings whose indoor radon concentrations exceed the reference limit of 200 Bq/m^3, their indoor radon sources must be identified before decontamination. The indoor radon measurements and remediation are the responsibility of property owners and are supervised by the county's and municipality's environmental and health protection committees.

In light of the new Swedish National Action Plan [4], the nationwide average of indoor radon levels and the extent of the buildings in Swedish dwellings and workplaces should be determined. It is estimated that around 400,000 dwellings exceed the reference limit nowadays [5]. A recent report by the Swedish Radiation Safety Authority (SSM) [2] analyzed indoor radon measurement records from the past two decades and compared the results with the previous surveys: the ELIB in 1991/1992 [6], the Radon Survey in 2000 [6], and the BETSI study in 2007/2009 [7]. The concluding outcomes from the latest report and

* Corresponding author. Sven Hultins Plats 5, 412 58, Gothenburg, Sweden.
E-mail addresses: pei-yu.wu@ri.se (P.-Y. Wu), tim.johansson@ri.se (T. Johansson), claes.sandels@ri.se (C. Sandels), mikael.mangold@ri.se (M. Mangold), kristina.mjornell@ri.se (K. Mjörnell).

https://doi.org/10.1016/j.buildenv.2023.110879
Received 5 June 2023; Received in revised form 21 September 2023; Accepted 24 September 2023
Available online 25 September 2023

surveys were that indoor radon concentration above 200 Bq/m^3 occurs more frequently in single-family houses (ELIB: 16–18%; Radon Survey: 35%; SSM: 19%) than in multifamily houses (ELIB: 5–8%; Radon Survey: 28%). The average indoor radon in single-family houses (ELIB: 141 Bq/m^3; BETSI: 124 Bq/m^3; SSM: 128–136 Bq/m^3) is almost twice as high as the concentration in multifamily houses (ELIB: 75 Bq/m^3). The varying values reported by former studies reflect the complexity of indoor radon estimation in heterogeneous building stocks.

Although a descriptive overview of the indoor radon situation in the building stock can be obtained from these cross-sectional investigations, the uncertainty of the statistical inference for specific building typologies and regions is still high due to the low share of buildings in which indoor radon measurements have been conducted, which represents around 16% single-family houses and 17% apartments in multifamily houses in 2020 [2]. The low indoor radon measurement rates could possibly be attributed to a lack of health risk awareness, ignorance of conducting measurements, affordability to decontamination costs, house ownership, and concern for future property selling [8]. Nevertheless, the limited amounts of empirical indoor radon measurements hinder the accurate evaluation of indoor radon levels in the existing buildings. The current knowledge of indoor radon statistics remains on the national and individual building scales for regulatory and decontamination purposes. However, the information on the detailed distribution of the indoor radon level in non-residential regional buildings is lacking [2]. To encourage effective indoor radon monitoring and mitigation interventions by local municipalities, predictive analysis at a lower aggregation level for unmeasured buildings is needed. The prediction outcomes could also advise property owners of radon risk-prone buildings to conduct thorough and consecutive indoor radon measurements.

2. Literature review

Indoor radon prediction powered by statistics and machine learning has shown promising results and gained a growing interest in recent years. Machine learning, derived from statistical modeling, was explored in indoor air quality assessment and indoor radon is one of the most studied substances [9–12]. The primary focus of indoor radon prediction in previous studies can be classified into short-term active monitoring using time series models and long-term concentration estimation based on regression models. Studies that represented the former research purpose are, for example, indoor radon concentration development forecasts [13] and real-time monitoring along a short timeframe [11]. By utilizing recurrent neural networks, more specifically, long short-term memory networks, indoor radon level evolvement over time could be projected [11,13]. The latter objective was exemplified by employing artificial neural networks for determining the influence of environmental variables on indoor radon concentrations [14], mapping indoor radon-prone areas using Bayesian spatial quantile regression [15] and extreme learning machine [10], kernel regressions [16] and ensemble regression trees [17]. Besides the use of deep learning and machine learning models, statistical methods were also explored to obtain spatial inference of indoor radon levels. For instance, interpolation techniques were investigated to predict the mean indoor radon concentration of spatial grids. Among different interpolation techniques, regression kriging showed the best performance and was used to develop a European indoor radon map [18]. A summary of the literature on indoor radon prediction is presented in respective model categories and algorithm types in Table 1.

To date, some progress has been achieved in predicting indoor radon proxy on an urban scale; however, more refinements should be made to improve the models' generalization. Previous studies tended to use aggregated data, such as postcode areas or DeSO (demographic statistical areas)/ RegSO (regional statistical areas) areas but lacked individual property or building information due to privacy regulations [20], which led to a limited opportunity for data coupling between indoor

Table 1
Summary of state-of-the-art studies on indoor radon prediction.

Reference	Purpose	Data size	Model	Performance
Statistical models				
1. Spatial inference				
[18]	Predict indoor radon concentration at the ground-floor level of buildings	1.2 million indoor radon records in Europe	Interpolation techniques, i.e., inverse distance weighting, ordinary kriging, collocated cokriging, regression kriging	R^2_{IDW} = 0.1001 R^2_{OK} = 0.3457 R^2_{CCK} = 0.3512 R^2_{RK} = 0.3687
Machine learning models				
1. Regressions				
[16]	Predict and map national indoor radon concentrations	238,769 indoor radon records from 148,458 houses in Switzerland	Kernel regression, probability estimation	$R^2 = 0.28$
[15]	Delineate spatial clusters of radon-prone areas	2,382 indoor radon records from the Abruzzo, Italy	Bayesian spatial quantile regression, stepwise analysis	N/A
[19]	Estimate the indoor radon concentrations	123,000 indoor radon records from Sweden	Multivariate adaptive regression splines	$R^2_{All} = 0.13$ $R^2_{Singlefamily} = 0.14$ $R^2_{Multifamily} = 0.13$ $R^2_{School} = 0.08$ $R^2_{Others} = 0.03$
2. Decision trees				
[17]	Classify lithological units automatically and improve radon prediction	238,769 indoor radon records from 148,458 houses in Switzerland	Random forest, Bayesian additive regression trees, k-medoid clustering	$R^2_{RF} = 0.33$ R^2_{BART} = 0.29
[20]	Estimate the indoor radon concentrations	123,000 indoor radon records from Sweden	Random forest	$R^2_{All} = 0.24$ $R^2_{Singlefamily} = 0.21$ $R^2_{Multifamily} = 0.28$ $R^2_{School} = 0.06$ $R^2_{Others} = 0.02$
3. Artificial neural networks				
[14]	Predict and benchmark indoor radon	192 indoor radon records from tertiary institutions in Nigeria	Feed-forward backpropagation neural network	AVE = 0.05 MAE = 0.02 RMSE = 0.04 MAPE = 3.64% G = 83.71%
4. Recurrent neural networks				
[13]	Forecast indoor radon in Canadian and Swedish dwellings by 2050	Indoor radon records from 25,489 Canadian and 38,596 Swedish properties	Long short-term memory	N/A
[11]	Predict indoor radon based on the current Rn	12,000 indoor radon records from a building in Spain	Long short-term memory	RMSE = 28 Bq/m^3

(*continued on next page*)

Table 1 (*continued*)

Reference	Purpose	Data size	Model	Performance
[10]	Map geogenic radon potential	1,452 indoor radon records in Danyang-Gun, South Korea	Long short-term memory, Extreme learning machine, Random factor function link	AUC = 0.824 RMSE = 0.209 StD = 0.207

radon measurements, geogenic factors, and dwelling characteristics [10]. Models trained with indoor radon measurements sampled from a substantial variety of buildings may not be accurate and interpretable; on the other hand, models built on input data from single buildings are barely transferable. Hence, careful data stratification should be considered in developing large-scale predictive models for various building types.

Indoor radon prediction is traditionally formulated as regression problems; however, the prediction performance of models is not satisfactory. Kropat et al. [16] reported $R^2 = 0.28$ using kernel regression and probability estimation methods and Wu et al. [19] reported $R^2 = 0.14$ using multivariate adaptive regression splines. Subsequently, their attempts with indoor radon concentration prediction using random forest models attained $R^2 = 0.28$ [19] and $R^2 = 0.33$ [17]. To enhance the model's performance while elevating the model's granularity to the property level, more advanced machine learning algorithms were required to untangle the complexity [18]. Taking the existing regulative indoor radon thresholds into account, i.e., 200 Bq/m^3 and 400 Bq/m^3, the problem formulation of the study investigated multi-class classification for indoor radon interval prediction.

The Extreme Gradient Boosting (XGBoost) and Deep Neural Network (DNN) algorithms were considered given their capability for efficiently handling large amounts of high dimensional, non-linear, imbalanced datasets with mixed data types [21,22]. Therefore, they were regarded as promising to model large and intricate national indoor radon measurements and have the potential to overcome the limited performance of simplified supervised learning models used in former studies for long-term indoor radon prediction. The XGBoost algorithm exploits the ensemble boosting method of multiple decision trees and adjusts parameters iteratively [23]. Thus, it is less likely to be overfitting than the ensemble bagging method owing to the sequential training procedure, and also more regularized than the gradient boosting approach. On the other hand, deep learning features deep neural networks with multiple hidden processing layers for learning highly abstracted data representations and relationships [24]. It has a self-learning capability of automatic feature generation and selection. By utilizing hidden layers and supplying parameterized weights in DNN in model optimization, the inputs (X) can be mapped out to the outputs (Y = f (X)) with assigned functions automated in a one-directional data flow. Although the ways of fitting data are different, tuning tree booster parameters in XGBoost models and hyperparameters in the DNN model have some similarities and both can be done through grid search. Besides, instead of predicting indoor radon concentrations in regression, classifying indoor radon intervals can be a research opportunity given the complicated synergies between climate, geogenic, and anthropic factors. Exploiting prediction approaches for the multi-class classification problem have not yet been investigated in the context of indoor radon prediction. The study aims to fill the gap by probing the application of advanced machine learning models trained and validated on a comprehensive and high-granular indoor radon dataset in Sweden.

3. Scope of the paper

The research aims to screen existing building stock with onsite indoor radon measurement priority by predicting indoor radon intervals (level indicators) as the proxy for long-term exposure estimation. Considering this research objective, the prediction granularity of indoor radon intervals based on the current legislative requirement was considered sufficient. The dependent variable of "indoor radon intervals" was clustered from the long-term estimated annual average indoor radon concentrations, which are aggregated values from individual measurements in the same dwellings according to the method description of indoor radon measurements in buildings [25,26]. Thus, multi-class classification was more suitable than regression in the context of long-term indoor radon exposure assessment given the inherited data uncertainty. The machine learning and deep learning multi-class classification models were trained on the national indoor radon measurement records, property registers, and geogenic information to estimate the indoor radon intervals for buildings without indoor radon measurements. The research outcomes contribute to an improved understanding of variable dependency on indoor radon in various building types and an overview of the present indoor radon situation in the building stocks for relevant authorities. To realize the overarching research goals, research questions are formulated as follows:

RQ 1: How accurately can extreme gradient boosting and deep neural network models predict the indoor radon intervals on the property scale?
RQ 2: What are the estimated shares of buildings prone to high indoor radon intervals in the Swedish metropolitan building stocks?

4. Materials

The dataset comprised multiple data sources, including indoor radon measurement records, property registers, and geogenic and geographical information. These data were extracted and linked at the property level to obtain predictive variables for modeling.

4.1. Indoor radon measurement records

The indoor radon measurement records were retrieved from the Swedish Energy Performance Certificates (EPCs) and municipalities' open databases and APIs. The latest EPCs of the 2022 version contain up-to-date information on property usage, building features, energy consumption and sources, ventilation types, and indoor radon records, where 167,468 random indoor radon measurements from different building classes across Sweden were also included. Afterward, an additional 23,084 indoor radon measurements from municipalities' databases were appended to the indoor radon subset and ensured that each observation represented an individual building. These indoor radon records provide comprehensive information on the measured dates, periods, locations, methods, and annual average indoor radon concentrations to guide the retrieval of valid observations. Most measurements were conducted after 1999 during the heating season using passive alpha-track detectors [3].

4.2. Property registers and geogenic information

The property register database from the Swedish Cadastral and Land Registration Authority contains the municipal cadastral register, of which unique building IDs and parameters are documented, described in Appendix A. Then the spatial joins between SWEREF 99 TM (Swedish geographical reference system) from property registers, a standard plan coordination system used among many public authorities, and the Geological Survey of Sweden (SGU) databases, i.e., geophysical aerial measurements and soil types. The radiometric grids of K-40 (potassium), U-238 (uranium), and Th-232 (thorium) concentrations were requested from the geophysical aerial measurement database. The concentrations were calculated from the measurements made by aircraft at low altitudes, where the emitted gamma radiation is made with 200 m line spacing with denser and sparser measurements. The interpolation had been made into a grid with squares of 200 m times 200 m with a

blanking distance of 1000 m to avoid gaps in the image where the line distance is 800 m. The measured values in a measuring point represented the weighted average of three mean values over a larger area on the ground using the inverse distance weighting interpolation [27]. Similarly, the soil types were acquired from the SGU soil type database in the spatial measurement formats scaling between 1:25 k-1:100 k. These polygons were described by soil type code, specification, and collection methods for upper and foundation layers [28]. The matching uncertainty was minimized by evaluating the number of measurements when combining the geological data and the property registers.

5. Methods

The acquired data were processed according to a three-fold procedure presented in Fig. 1.

5.1. Data assembling and preprocessing

Assembling and preprocessing the raw data were performed with the FME (Feature Manipulation Engine) and Python's libraries Numpy and Pandas, which is a data integration platform enabling spatial data transformation [29]. Using the real estate index and the address in the indoor radon records, the property registers and the geogenic information of the measured buildings could be extracted and matched. After that, based on the building use type code from the municipal cadastral register, the data were grouped into four building classes – single-family houses, multifamily houses, school buildings, and other buildings (i.e., commercial and office buildings) – to ensure relatively homogeneous building characteristics in each subgroup. When evaluating the validity of the indoor radon measurements, the method descriptions of indoor radon measurement for residential buildings, workplaces, and public premises [25,26,30] were referred to. The indoor radon measurements that did not conform with the guideline, such as those measuring less than two months or during the non-heating seasons and those conducted before 2000, were removed. Buildings built before 1930 and after 2020 were also eliminated to decrease the uncertainty of the measurement data.

Moreover, Swedish building stock built between 1930 and 1980 containing possibly radioactive concrete, which releases 20–25 times more radon gas than ordinary concrete, is a limitation of this study due to insufficient records in property registers [3]. The 270 properties known to be built with radioactive concrete based on the inspection records in municipality indoor radon datasets were removed, but the situation of radioactive concrete in most of the buildings in the compiled indoor radon dataset remained unknown. Afterward, the interquartile range (IQR) method was carried out on the radioactive substances concentrations to detect the outliers in the data analysis and visualization. This results in 114,857 observations, and a subset of 34,983 indoor radon measurements after 2015 with higher certainty in registers matching were retrieved for modeling. The rest of the data were categorized into three intervals: low (0–200 Bq/m^3), medium (200–400 Bq/m^3), and high (above 400 Bq/m^3), corresponding to the existing regulatory indoor radon thresholds [3].

5.2. Machine learning models development

Machine learning models for multi-class classification tailored for different building classes were built to predict indoor radon intervals using Python skikit-learn, imbalanced learn, and H2O AutoML libraries. In both XGBoost and DNN models, softmax was used as the output activation function, where logits were turned into probabilities summing to 1 and the class with the highest probabilities became the prediction results. The objective of the training was to minimize the categorical cross-entropy loss function calculated based on the softmax outputs. Fig. 2 presents the general architecture of the machine learning model, including data subgroups, input variables, hidden layers or bootstrap samples, and prediction outputs. The predictors for the indoor radon levels involved geological attributes, i.e., radioactive substances concentrations and lithological units, and anthropic parameters, i.e., building information. The motivations for such feature selection were based on the literature [14–17] and the results from a previous study concerning multivariate adaptive regression splines and random forest regression [19]. The strong positive correlation between gamma radiation from uranium and indoor radon concentration was confirmed [20], yet the impacts of potassium and thorium required further investigation. Soil types also appeared to have a subtle influence on indoor radon; hence, the 13 most common soil types were included [2]. Among building parameters, exhaust and balanced ventilation regulated indoor radon concentration downwards substantially, while natural ventilation had the opposite effect [31–33]. The second influencing factor was the foundation [16,17,34]. However, this information was unavailable in the registers and could only be approximately inferred from the number of basements (or the number of floors below ground) and the ground types where the buildings were situated. Since the annual average indoor radon concentration was calculated from indoor radon measurements in the ground floor, basement, and selective upper floors in

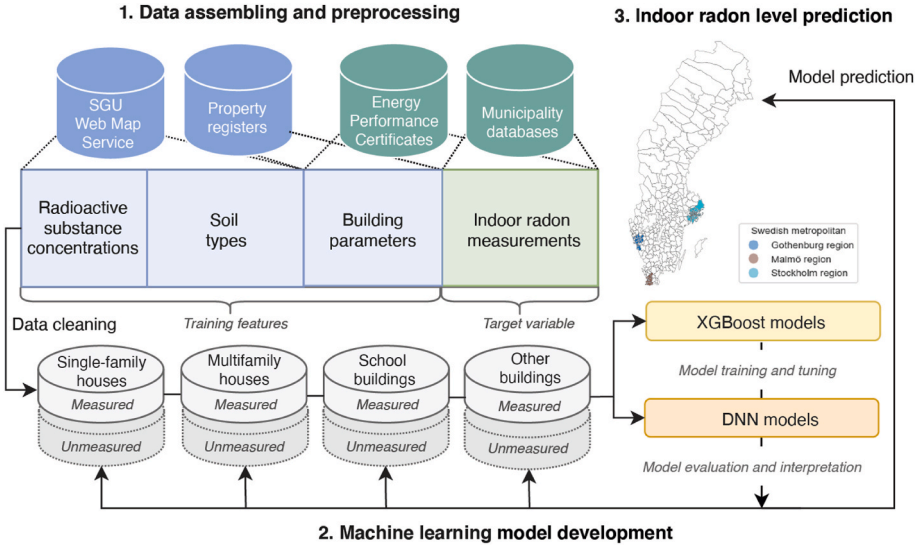


Fig. 1. Study outline.

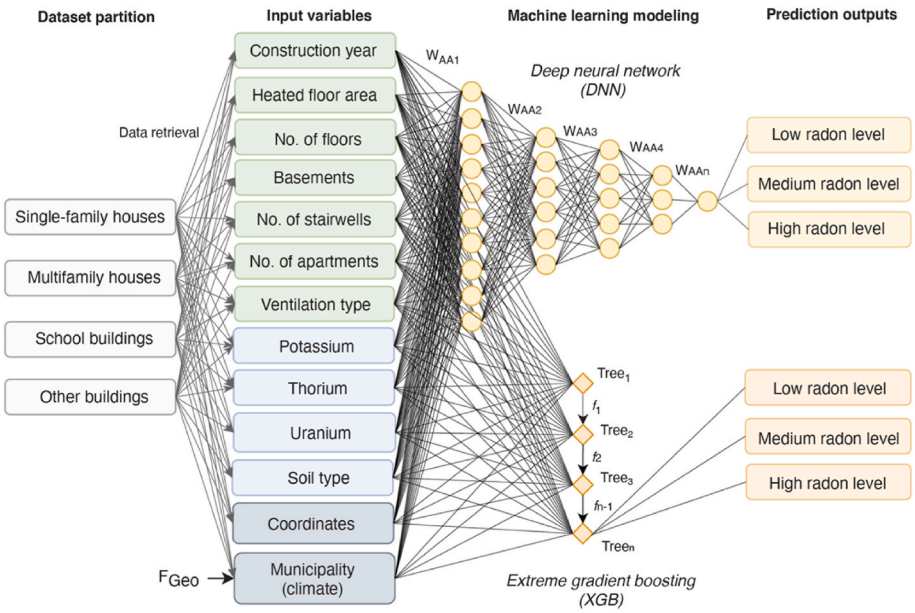


Fig. 2. The general architecture of the machine learning models with input data from building parameters (green), geogenic attributes (blue), and geographic factors (gray). The dataset was partitioned according to the building class and a DNN and a XGB model were created for each data subgroup. *Indoor radon level indicators = f (construction year, floor area, basements, number of floors, stairwells, and apartments, ventilation type, potassium, thorium, and uranium concentrations, soil type, coordinates, geographical adjustment factors)*. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

multifamily houses and all measurements in single-family houses [26], the number of floors should be considered. Construction year and floor area indicated the use of building materials and construction practices, but also outdoor air ventilation that affected the indoor environment [13].

In view of the broad sampling of indoor radon measurements across the geographical stretch, time horizon, and the unavailability of indoor environment climate data of the observed buildings, the impact of meteorological attributes, such as temperature and humidity, was surrogated through coordinates and geographical adjustment factors (FGeo) [35]. The foundation of the geographical adjustment factors was based on Sweden's municipal division and the four existing climate zones to harmonize climate conditions in various locations of the country with 12 scales ranging from 0.8 to 1.9. By applying the geographical adjustment factors to the models, the spatial dependence across data could be controlled, and the regional climatic variance could be minimized on the national scale.

Each of the building class subgroups was first split into 80% training and 20% validation subsets by stratifying similar label proportions of the dependent variables. To extract key features for indoor radon interval prediction, raw and derived variables were employed in feature selection algorithms based on the F-statistics, which measured the ratio of paired variances and correlation between labels and features [24]. The number of apartments and stairwells were not used as features in modeling school and other building subgroups due to large numbers of missing values. To address uneven class distribution, sample weights adjusted inversely proportional to class frequency in the input data were attached to XGBoost and DNN algorithms, and the imbalanced labels were resampled with Synthetic Minority Oversampling TEchnique (SMOTE) technique to oversample the minority classes for cost-sensitive learning. Meanwhile, missing values of the features were imputed by the five nearest neighboring values, then trained with 5-fold cross-validation with random grid search for optimal hyperparameter configurations. The training stopped when the log loss (cross-entropy loss, a measure to quantify the difference between predicted probabilities and actual values) started to increase again in the tuning process for model fit evaluation between training, cross-validation, and validation.

Then lead XGBoost and DNN models with the highest macro-F1 score (the unweighted arithmetic mean of the F1 scores calculated per class in imbalanced datasets for an objective model evaluation [36]) and the lowest mean per class errors from the confusion matrix were employed in the validation subset to estimate their prediction performance for the unseen data. Then the ROC AUC scores (area under the receiver operation curve plotting true positive against false positive rates) were computed using the one-vs-rest classification method, a heuristic technique splitting a multi-class dataset into multiple sets of binary problems, to determine the degree of separability of the labels by all possible thresholds for each indoor radon interval. Lastly, the feature importance based on prediction outputs was plotted for each building class to improve model interpretation.

5.3. Indoor radon interval prediction

The lead prediction models were applied to the rest of the building registers to estimate indoor radon intervals for each building class in buildings not yet measured. Performing such an inference was possible because the predictive models were trained with indoor radon measurements from large quantities of dwellings over decades; thus, the measured buildings were regarded as representative of the regional building stock. The metropolitan regions of Stockholm, Gothenburg, and Malmo regions, where most indoor radon measurements were collected, were used as a case study to demonstrate indoor radon interval prediction. The 257,781 property registers of the 48 municipalities were retrieved from the latest EPCs, of which duplicated and invalid entries were removed and supplemented with the geogenic information. The features of the buildings that had not yet conducted indoor radon measurements were supplied to the developed models for prediction. Further on, the models were also applied to the buildings with indoor radon measurements to ascertain the model uncertainty by comparing the prediction outcomes with the actual statistics. Compiling the statistical and estimated shares, the overall indoor radon interval distribution was summarized by building classes for regional building stocks.

6. Results

The results are presented in three consecutive parts: (1) data analysis

and visualization, (2) predictive model evaluation, and (3) indoor radon interval estimation.

6.1. Data analysis and visualization

The spatial and statistical distribution of the indoor radon measurement records were investigated to determine the representativeness of the training dataset in relation to the Swedish building stock. Fig. 3 below illustrates the spatial characteristics of the indoor radon dataset, including the distribution of measurements against geographical adjustment factors, the mean annual average indoor radon concentrations, and the ground uranium concentration across municipalities. The results showed that the indoor radon measurements in the study were distributed approximately according to the density of the built-up areas. The majority of the samples were derived from the Stockholm region, followed by southwest coast regions and some along the east coastline to the north. Further comparing the nationwide mean indoor radon concentrations and the average uranium concentrations, their spatial association was confirmed in the locations of corresponding color patches. Municipalities with a higher ground uranium concentration could possibly imply higher indoor radon levels, which could be seen in some municipalities located in the middle south part of Sweden with larger sample sizes. Nevertheless, some exceptions existed and thus uranium concentration could not be regarded as a single indicator for indoor radon inference.

The subsequent multivariate analysis explored variable interaction and underlying patterns in the dataset. Table 2 presents a statistical description of the 114,857 indoor radon measurements based on their building classes. The confidence interval of the annual average indoor radon concentration lay at 110 ± 1 Bq/m^3 and around 12% of observations exceeded the reference limit of 200 Bq/m^3. Among building classes, single-family houses were measured at a higher mean indoor radon concentration of 118 Bq/m^3 and the share above limit was also the largest, nearly 14%. The largest standard deviations of the mean indoor radon concentration were found in other buildings. In general, the statistics of indoor radon measurements in multifamily houses aligned with those of the total observations. The mean indoor radon concentration in school buildings tended to be the lowest, which is probably related to constant indoor radon monitoring under the Act of Environmental Goals [7].

Fig. 4 below illustrates the average indoor radon concentration for all buildings and respective building classes in relation to the construction year. The distribution of the sample in gray staples conformed to the historical timeline of building production in Sweden with the construction peak between 1960 and 1970 and the 1990s, indicating the samples were representative of the national building stock. Each bin implied a year and the sample size could be used to evaluate the certainty of the calculated mean. The line chart showed a mild downward trend of the mean indoor radon concentration for all buildings since 1960. Among building classes, post-war dwellings built between 1950 and 1980 tended to have higher indoor radon concentrations. The indoor radon concentrations in single-family houses and multifamily houses were nearly parallel to the baseline. For schools and other buildings, the variation of indoor radon concentration was higher, in particular, in hose buildings built before 1945 and after 1995, which entailed a higher measurement uncertainty of these subgroups in the dataset.

The impact of building parameters on the indoor radon concentration for each building class is illustrated in Fig. 5. The sample randomness was examined by plotting the kernel density, quantile summary, and data distribution. The findings showed long-tailed value distributions irrespective of building class. From the distribution of the median values in the boxplots, it was observed that buildings with basements are generally prone to have higher indoor radon concentration than those without, especially in single-family houses and school buildings. Compared to natural ventilation, balanced and exhaust ventilation could slightly mitigate indoor radon concentration. The effect of ventilation on regulating indoor radon concentration was more evident in buildings without the basement, and the findings were consistent across building classes. Lastly, the scale of the heated area did not seem to correlate with indoor radon concentrations given random distribution of the values in the strip plots.

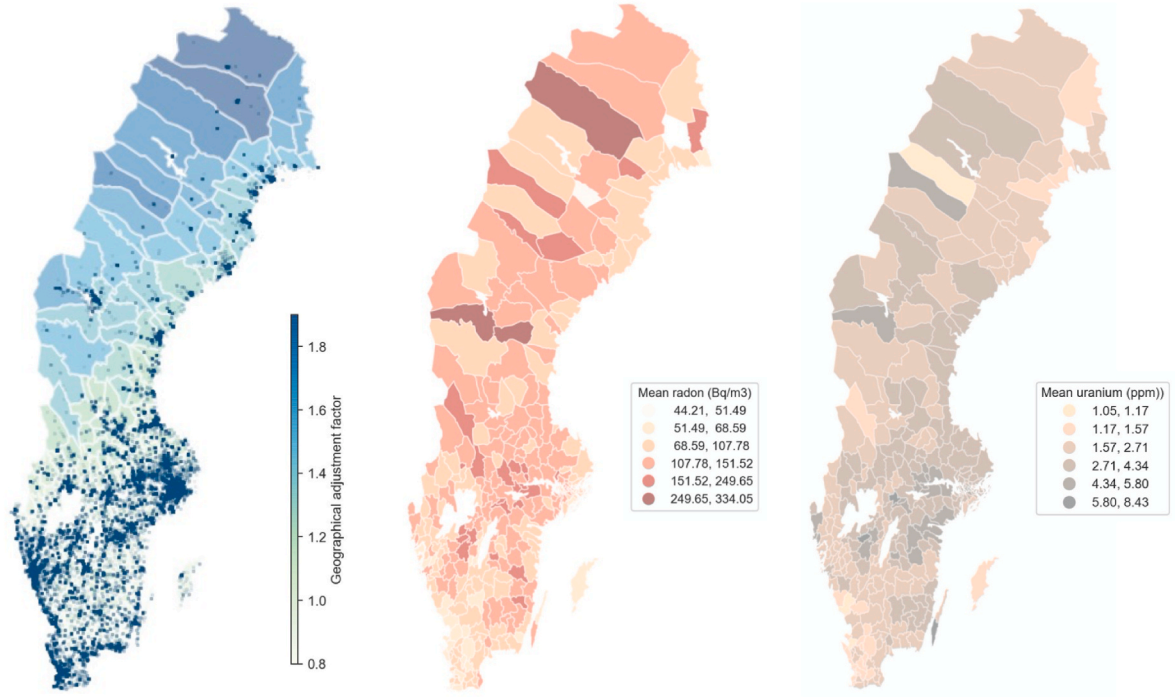


Fig. 3. Spatial characteristics of the available indoor radon measurements: (i) distribution of the indoor radon measurements across geographical adjustment factors; (ii) mean annual average indoor radon concentration in decile intervals; (iii) mean uranium concentration in decile intervals.

Table 2
Statistics of the indoor radon measurements at the property level by building classes.

Building class	Single-family house	Multifamily house	School building	Other building	Total
Count (%)	53,533 (47%)	49,139 (43%)	5,660 (5%)	6,525 (5%)	114,857 (100%)
Radon range (Bq/m^3)	[0–26,025]	[0–21,750]	[1–2610]	[1–65,424]	[0–65,424]
Avg. radon (Bq/m^3)	118 ± 2	105 ± 1	98 ± 3	105 ± 20	110 ± 1
Above limit	13.9%	11.7%	9.4%	8.9%	12.4%
Low level	86.1%	88.3%	90.6%	91.1%	87.5%
Medium level	11.1%	9.3%	7.0%	6.2%	9.9%
High level	2.8%	2.4%	2.4%	2.7%	2.6%

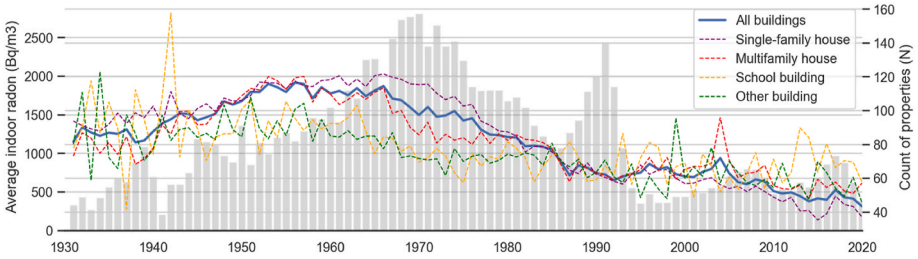


Fig. 4. The development of the mean annual average indoor radon concentration in Sweden by building classes with data count at the property level.

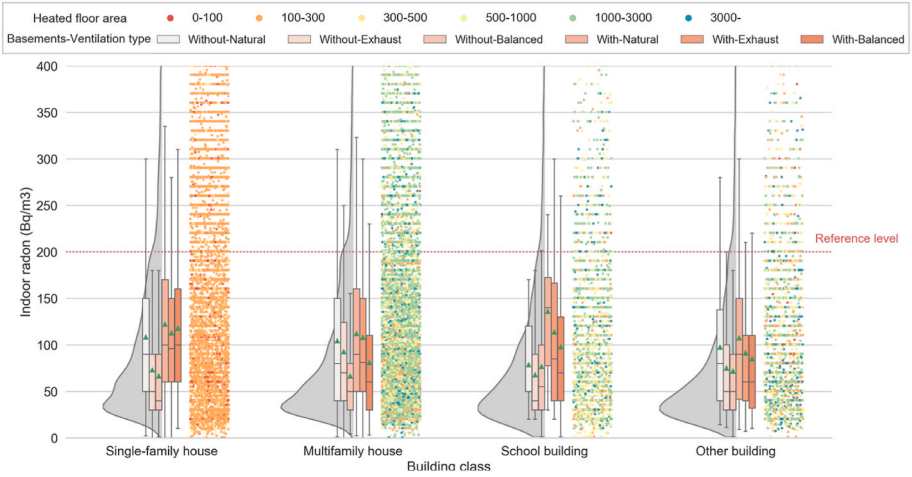


Fig. 5. The impact of building parameters on the indoor radon concentrations by building class: (i) density distribution of indoor radon concentration; (ii) the quantile value distribution of indoor radon concentration clustered by basements and ventilation types; (iii) the distribution of the heated floor area intervals.

6.2. *Predictive models evaluation*

Fig. 6 summarizes the performance evaluation of the predictive models from cross-validation and validation. Overall, the XGBoost models had significantly better performance than the DNN models with an average of 0.93 macro-F1 for single-family houses, 0.95 for multi-family houses, 0.94 for school buildings, and 0.96 for other buildings. The corresponding log loss and mean-per-class errors lay between 0.15-0.20 and 0.05–0.07. In comparison, DNN models only attained 0.64–0.74 macro-F1, and their log loss and mean per class error were

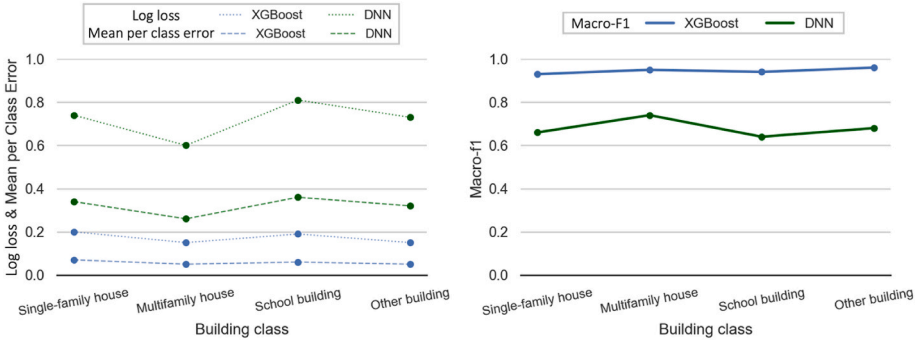


Fig. 6. Average performance of the prediction models from cross-validation and validation subsets adjusted with sample weights and resampling.

substantially higher. Both algorithms performed coherently better in predicting indoor radon intervals in multifamily houses and other buildings than in single-family houses and school buildings. Higher mean-per-class error rates were observed for medium labels than the average error rates in the confusion matrix. Considering the model robustness and performance, the subsequent upscaling prediction was conducted with the lead XGBoost models with the hyperparameter settings in Appendix B.

To determine the effect of resampling on the model's efficiency for label distinguishment, the probability distribution and the AUC were computed for the original and the oversampled datasets in Fig. 7. The AUC improved considerably after applying resampling that implied effective adjustment of the class imbalance, i.e., single-family houses (AUC = 0.77, resampled AUC = 0.98), multifamily houses (AUC = 0.84, resampled AUC = 0.99), school buildings (AUC = 0.70, resampled AUC = 0.97), other buildings (AUC = 0.72, resampled AUC = 0.98). The highest AUC in the original dataset was found in multifamily houses (≈ 0.85), followed by single-family houses (≈ 0.77), other buildings (≈ 0.72), and school buildings (≈ 0.70), of which more low and high labels are predicted correctly than the medium label.

The key features for indoor radon interval classification were ranked according to the aggregated feature importance of XGBoost models in Fig. 8. Building physical footprint (area per floor), floor area, and construction year were common critical features for all building classes. Other features such as latitude, exhaust ventilation, longitude, uranium concentration, basements, geographical adjustment factors, and natural ventilation contributed less but still played important roles. Natural ventilation was by far the most crucial variable for single-family houses, while construction year and building physical footprint were essential for multifamily houses. On the contrary, the key features associated with school buildings were basements, area, and building physical footprint. The features related to other buildings were less pronounced, i.e., area, geographical adjustment factor, building physical footprint, and exhaust ventilation.

6.3. Indoor radon interval estimation in metropolitan buildings

The models trained on the national radon measurements were regarded as the base models for evaluating model generability when transferring to the regional scale. Due to the non-existence of certain soil types in the metropolitan regions, the fine-tuned XGBoost models had to be re-adapted to fit the metropolitan dataset containing 14,419 observations before predicting the unmeasured buildings in the same municipalities. To ascertain model sensitivity, the prediction performance was provided: single-family houses (macro-F1: 0.93), multifamily houses (macro-F1: 0.95), school buildings (macro-F1: 0.95), and other buildings (macro-F1: 0.96). These model performances were nearly equal to the previous results in Fig. 6 and thus the models were adopted for indoor radon interval estimation in metropolitan building stocks.

To verify the prediction outcomes, a statistical summary of the historical indoor radon measurement records in the Swedish metropolitan building stocks was compiled. The retrieved properties from the Stockholm, Gothenburg, and Malmo metropolitan regions accounted for circa 32% of the declared building stock in terms of the number of properties. The average indoor radon measurement rate was around 23%, where the Stockholm region building stock had the highest indoor radon measurement rate (31%) while the Malmo region building stock had the lowest and the findings were consistent in all building classes. Around half of the multifamily houses (51%) and school buildings (49%) were measured, yet only 15% of single-family houses and 18% of other buildings had indoor radon records. The low measurement rate in single-family houses was in good agreement with the literature [4], confirming the representativeness and validity of the prediction dataset to the entire building stock.

Table 3 below shows the indoor radon interval distribution by building classes and metropolitan regions: the statistical shares from historical measurements, the predicted shares for measured buildings, and the predicted share for non-measured buildings. The findings showed that the distribution of predicted shares for measured buildings was close to the ground truth of statistical shares. However, the models tended to misclassify 2–3% medium labels to low labels. Possible reasons for misclassification could be the low number of high indoor radon labels in the training data, outliers owing to measurement errors, and the lack of radioactive concrete records in the feature sets, thus causing prediction errors for certain classes. This uncertainty should be considered when evaluating the estimated label distribution in the non-measured buildings, whose indoor radon situations were predicted less serious than those measured ones. The results were regarded as reasonable since measurements could be more likely conducted in indoor radon-prone buildings.

7. Discussion

The first section discusses the result implications regarding model training and prediction, then continues with data limitations and methodological challenges. Subsequently, the results of the study were compared with the previous study to discuss the model's performance and generability. The last part of the section highlights the benefits, limitations, and contribution of the research outcomes to indoor radon interval estimation.

Based on the nationwide indoor radon measurement records and property registers, critical variables to the indoor radon concentration and their distribution in buildings were delineated. In the past decades, indoor radon measurements were conducted more frequently in residential and public buildings, representing circa 95% of observations in the training dataset. The multivariate analysis provided an in-depth statistical overview of the measured properties by building classes that complement the current indoor radon knowledge in dwelling buildings and contributed to a new understanding of indoor radon status in non-residential buildings. The measured indoor radon ranges, the average indoor radon concentrations, and the share above 200 Bq/m^3 limit of the single-family houses and multifamily houses groups in Fig. 1 were in good agreement with the literature, i.e., the ELIB survey [6], the BETSI survey [7], and the SSM report [2]. Further computing indoor radon interval distributions between the training, the validation, and the metropolitan subset showed similar patterns, indicating that the training data, despite the concern of potential sampling bias, were still representative of the Swedish building stock.

In this multi-class classification task, both machine learning models were able to predict the indoor radon interval to various extents depending on building classes. The cross-validation and validation loss decreased approximately in parallel to each other, which entailed the models were neither overfitting nor underfitting. The XGBoost attained macro-F1 of 0.93–0.96, while the DNN models only reached macro-F1 of 0.64–0.74. The highest F1 and AUC were found in multifamily houses and the patterns were consistent across models. This may be explained by the local loss minimization in small datasets that were not deficient for neural network training. Several previous studies about the algorithm comparison between XGBoost and DNN also reached a similar conclusion in regression and image classification tasks [37]. Compared to the XGBoost models, the DNN models required more training iterations to reach comparable loss values [22].

To the authors' knowledge, the paper is the first study approaching indoor radon prediction with multi-class classification. Former studies focused on indoor radon regression and could unfortunately not be used as a baseline for comparison. Due to this reason, two types of machine learning models were developed to benchmark their performance, and the trained models were transferred to predict indoor radon interval distribution for measured and non-measured buildings in comparison with the actual distributions. In model evaluation, the confusion matrix showed around 8–11% medium-level misclassification, whereas the corresponding error rates for low-level and high-level are 2–8% and

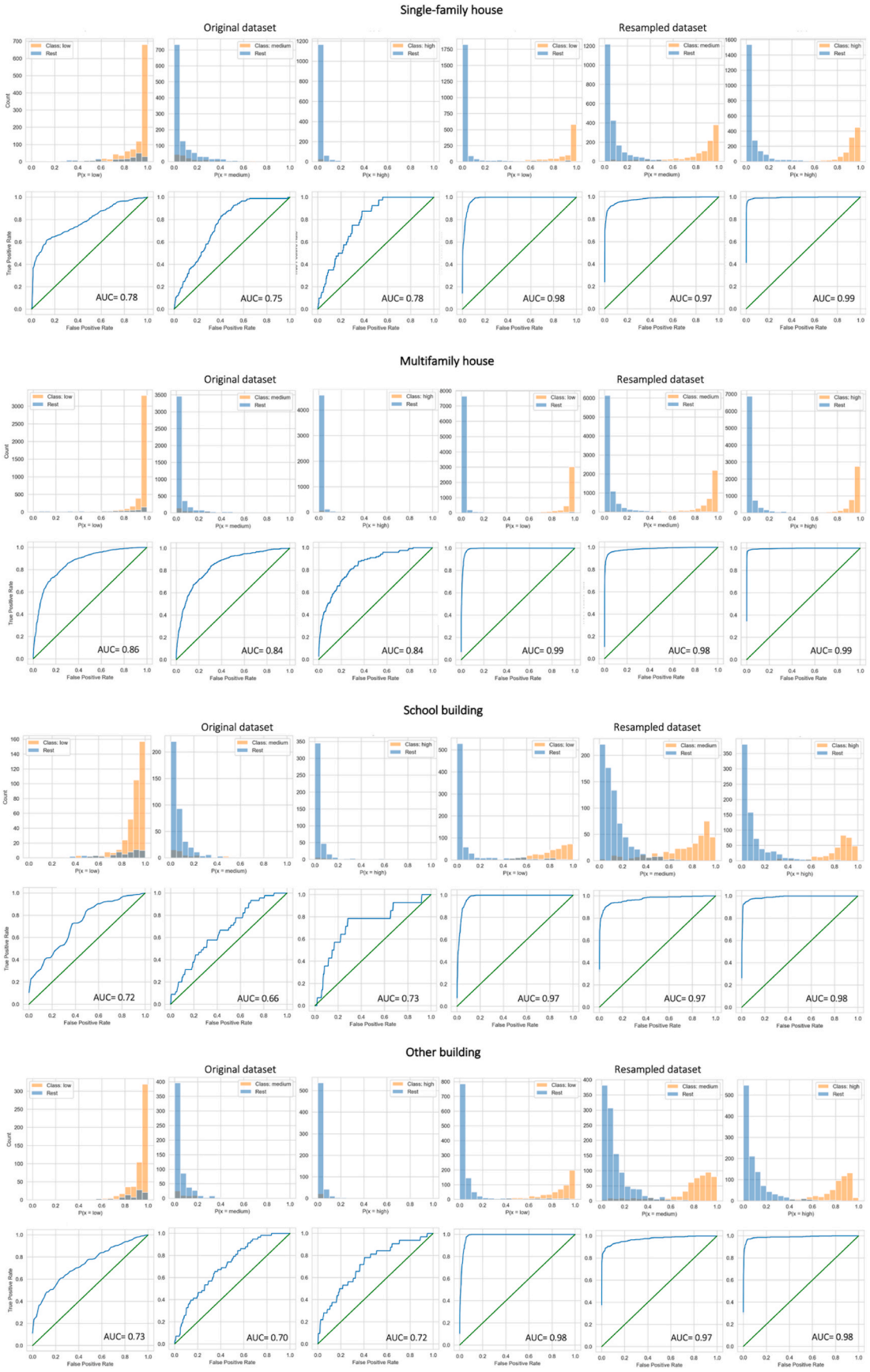


Fig. 7. Probability distribution and AUC scores for one-vs-rest classification by building classes, where the three plots on the left were computed with the original imbalanced dataset and the three plots on the right with the resampled dataset.

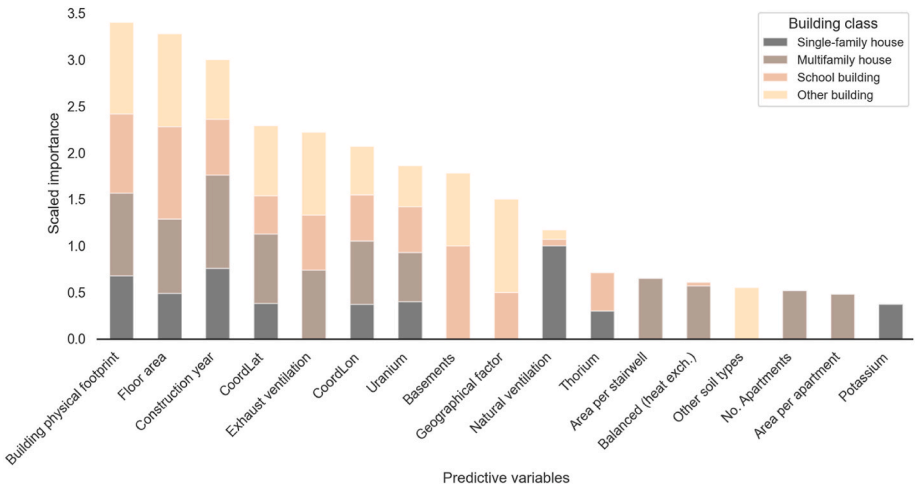


Fig. 8. Variable importance of indoor radon interval classification by building classes.

Table 3
Estimation of the indoor radon interval distribution in the Swedish metropolitan regions by building classes in three parts-the statistical shares from historical measurements, the predicted shares for measured buildings, and the predicted share for non-measured buildings.

Region	Stockholm	Gothenburg	Malmo	Stockholm	Gothenburg	Malmo	Stockholm	Gothenburg	Malmo
Class	Single-family house								
Radon	Actual measured			Predicted measured			Predicted non-measured		
%count	0.21	0.08	0.06	0.21	0.08	0.06	0.79	0.92	0.94
N	20,769	4,014	2,580	20,769	4,014	2,580	77,000	44,457	38,868
%low	87	94	94	91	97	95	94	99	97
%medium	11	5	5	7	2	4	4	1	2
%high	2	1	1	2	1	1	1	0	1
	Multifamily house								
	Actual measured			Predicted measured			Predicted non-measured		
%count	0.66	0.38	0.22	0.66	0.38	0.22	0.34	0.62	0.78
N	18,819	4,685	2,076	18,819	4,685	2,076	9,634	7,639	7372
%low	91	88	97	94	85	99	97	94	100
%medium	8	10	2	5	13	1	2	6	0
%high	1	2	1	1	2	0	1	0	0
	School building								
	Actual measured			Predicted measured			Predicted non-measured		
%count	0.63	0.52	0.30	0.63	0.52	0.30	0.37	0.48	0.70
N	1,820	695	324	1,820	695	324	1,072	634	739
%low	88	95	97	90	97	100	95	99	100
%medium	8	4	2	7	2	0	3	1	0
%high	4	1	1	3	1	0	2	0	0
	Other building								
	Actual measured			Predicted measured			Predicted non-measured		
%count	0.25	0.18	0.10	0.25	0.18	0.10	0.75	0.82	0.90
N	1,977	670	310	1,977	670	310	5,860	2,988	2815
%low	89	95	98	93	96	100	96	99	100
%medium	8	4	2	5	3	0	2	1	0
%high	3	1	0	2	1	0	2	0	0

2–3%. These results may be attributed to the fact that models were trained on the training set assuming no presence of radioactive concrete; however, radioactive concrete buildings existed in the metropolitan prediction dataset and might explain the lower shares of medium levels in prediction. A previous study found that the average indoor radon was 63% higher in radioactive concrete-containing buildings than those built without depending on the amount of radium content and the extent of radioactive concrete in buildings [13]. Other errors could be due to a lack of crucial features, such as foundation types and the presence of radioactive concrete [3], or inaccuracy of the response variable, such as measurement errors in the data source, and unavailable information about measurement places. To address the former issue, more extensive building parameters need to be collected in the registers; while for the latter issue, multiple measurement records and details should be made available for further data quality check. A possible solution to improve the model fit while reducing the misclassification is adding regional factors in data partition and model training, i.e., regional boundary, and more measurement records of medium and high indoor radon in model training. Plotting the indoor radon intervals in the metropolitan regions later verified the assumption as the patterns varied significantly between regions than building classes, which may be attributed to regional governance supervising and providing subsidies for indoor radon monitoring and remediation. This in turn resulted in varied indoor radon measurement frequency and data availability causing inevitable

selection bias in the training dataset.

The proposed approach for estimating the indoor radon intervals on the urban scale from the register records has its limitations and benefits that are mostly related to data uncertainty. As the indoor radon measurements were reported along with the EPC, the records were aggregated, and the annual average indoor radon measurement may include several buildings in which energy performance was calculated together. Such data quality problem was more likely to occur in complex building, for instance, multifamily houses and school buildings, but less affecting single-family houses. To avoid modeling duplicated measurement records, the prediction was determined at the property level to keep unique measurements by address and real estate index. On the other hand, the benefit of including indoor radon measurements over the years for all building types was maintaining a comprehensive and representative training dataset that reflected the diversity of indoor radon measurements in reality. A large amount of training data also prevented the model from overfitting and retaining sufficient samples for the minority class. The cost-sensitive algorithms adjusted with sample weights and trained on the resampled dataset were found effective in addressing the class imbalanced problems [21,22] [37], that were reflected by high macro-F1 and AUC and validated through comparable distributions of indoor radon intervals between estimated and statistical shares.

The indoor radon problem is contextually diverse. Researchers of various domains have tried to model the geogenic and indoor radon development for risk screening, monitoring, and remediation. Various data-driven approaches were explored for specific purposes with required prediction performance to satisfy their application scope. Screening the indoor radon levels was usually performed on the national or regional scale, while active indoor radon monitoring and remediation mandates immediate action was executed within a particular building. Former research explored indoor radon concentration prediction using kernel regression [16] and random forest [17] and explained 28% and 33% variations based on more than 238,000 observations. With more than 1.2 million data [18], obtained a 37% coefficient of determination with regression kriging interpolation. In the study, around 34,983 latest indoor radon measurements were split into four building class subsets to train XGBooost and DNN models and obtained macro-F1 of 0.93–0.96 and AUC of 0.97–0.99 in classifying indoor radon intervals. Although these evaluation metrics for classification could not be directly compared with R^2, it was still undeniable that the models attained exceptionally high performance. Since indoor radon prediction was never positioned as a classification problem, the model performance could not be benchmarked with those regression models in literature but rather provided a reference of the classification models for future studies. The limitations of existing regression methods were low coefficients of determination and limited data granularity on geographical regions, whereas the proposed classification methods did not predict the exact values but rather a range for building properties. Whether to approach indoor radon prediction from classification or regression should be determined by prediction purposes and the use of prediction outcomes. With much fewer data amounts than the previous studies and less extensive feature sets, the results were fairly satisfactory for estimating indoor radon exposure assessment. To further improve the model performance for long-term prediction, models could be revised for specific municipality or district applications using input data from a shorter measurement timeframe. This was expected to reduce measurement or evaluation errors related to heterogeneous regional and building attributes from the aggregated measurements in registers.

Another workaround is to combine human experts and soft probability outputs from machine learning models in indoor radon level assessment [38]. More specifically, the prediction outcomes returned prediction probabilities of the response variables over the indoor radon intervals and allowed human experts to evaluate the observations with even probability distribution close to class thresholds. The drawback of the hybrid solution is that it demands extra resources to control 2–3% uncertain samples, however, the classification results will be more reliable with double evaluation. This study contributed to the initial assessment of the current indoor radon situation and enhanced understanding of the indoor radon-related attributes in each building class. Given any individual building registers from the indoor radon unmeasured buildings, the developed models can generate a suitable level indicator, and even the probability distribution across three level indicators with subtle alteration, to guide prioritizing onsite measurements for the indoor radon-prone buildings for national and local authorities. The proposed approach can be replicated in countries where indoor radon measurements, property registers, and geogenic information are available.

8. Conclusions

Low awareness of indoor radon concentrations simply because of low shares of Swedish building stock with indoor radon measurement records and no mandatory requirements on indoor radon measurements, but also the complexity and cost of technical measures to reduce radon concentrations, are two major reasons for comprehensive indoor radon remediation. This paper proposed and investigated the applicability of machine learning multi-class classification models adjusted with minority class resampling for the identification of indoor radon patterns in the Swedish building stock based on historical measurement records. The novelty of the paper lies in demonstrating the machine learning modeling workflow and enabling the estimation of the shares of building stock potentially prone to indoor radon exposure for prioritizing onsite measurements. By enhancing the prediction granularity to the property scale and refined machine learning models on the basis of building classes, the study contributed to a diagnostic overview of the status quo of indoor radon conditions, and also the predictive analysis for properties that did not have indoor measurements. According to the tree booster models, the building physical footprint, heated floor area, and construction year were key factors for indoor radon interval prediction. Other factors such as latitude, exhaust ventilation, longitude, uranium concentration, basements, geographical adjustment factors, and natural ventilation were also contributing factors. The XGBoost outperformed DNN with high macro-F1 and AUC in all building classes based on the results from the cross-validation set and validation set; thus, they were used for upscaling prediction in the regional building stocks. To verify the models' reliability, the prediction outcomes of the measured building registers were benchmarked with their actual label distributions in statistical summary, and the results were approximately consistent with minor variance. Furthermore, employing the lead XGBoost models to unmeasured buildings in the Stockholm, Gothenburg, and Malmo regions, the estimation of indoor radon intervals was generated and compared with the rest of the historical measurements. Future studies are suggested to develop machine learning models tailored for smaller geographical areas, i.e., municipality or county, to reduce input data uncertainty of indoor radon measurements, as well as apply models in case studies for real-world validation.

Funding

This work has received funding from the Swedish Foundation for Strategic Research (SSF) [FID18-0021] and the Maj and Hilding Brosenius Research Foundation.

Institutional review board statement

Not applicable.

Informed consent statement

Not applicable.

CRediT authorship contribution statement

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

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.

Data availability

The authors do not have permission to share data.

Acknowledgments

The research work is part of the PhD project "Prediction of Hazardous Materials in Buildings using Machine Learning", supported by RISE Research Institutes of Sweden.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.buildenv.2023.110879.

Appendix A

Table A1
Overview of the indoor radon database.

Category	Data specification	Measurement type
1. Matching keys and sorting	Real estate index	String + Nominal
	EPC index	Nominal [7-digit]
	Coordinates	Nominal [Long, Lat]
	Address	String
2. Cadestral informaiton	County code	Nominal []
	County name	String
	Municipality code	Nominal [1–93]
	Municipality name	String
3. Building characteristics	Building age	Scale variable [Year]
4. Building usage	Building usage category code	Nominal [1–7]
	Building usage type code	Nominal [1–99]
	Building category	Nominal [Single- or two-family house, Multifamily house, Other building]
	Detailed usage of the building	Share of the building used for the 12 most common types
5. Building area	Building size (living space)	Scale [m^2]
	Heated floor area (Atemp)	Scale [m^2]
	Number of floors	Ordinal
	Number of stairwells	Ordinal
	Number of apartments	Ordinal
	Number of floors below ground	Ordinal
6. Ventilation	Ventilation type	Nominal [Exhaust, balanced, balanced with heat exchanger, Exhaust with heat pump, natural ventilation]
7. Indoor adon measurement	Indoor radon measurement date	Timestamp [Year-month-day]
	Indoor radon measurement type	Scale [Long time measurement, other methods]
	Indoor radon concentration	Scale [Bq/m^3]

Appendix B

Table B1
Configuration of the hyperparameters of the lead XGBoost classifiers for indoor radon interval prediction in the metropolitan building stocks.

Hyperparameters	Single-family house	Multifamily house	School building	Other building
Learning rate	0.3	0.3	0.3	0.3
Gamma	0	0	0	0
Max depth	10	15	15	9
Subsample	0.6	0.6	0.8	0.6
No. trees	124	159	109	128

References

[1] Swedish National Board of Housing Building and Planning, Radon in the indoor environment (Radon i inomhusmiljö). https://www.folkhalsomyndigheten.se/livsvillkor-levnadsvanor/miljohalsa-och-halsoskydd/inomhusmiljo-allmanna-lokaler-och-platser/radon/, 2010.
[2] T. Rönnqvist, Analysis of Radon Levels in Swedish Dwellings and Workplaces, 2021.
[3] B. Clavensjö, G. Åkerblom, Radonboken. Befintliga Byggnader, Fjärde Utg, Svensk byggtjänst, Stockholm, Sweden, 2020.
[4] Swedish Radiation Safety Authority, Nationell Handlingsplan För Radon, 2018.
[5] Swedish Radiation Safety Authority, Att Mäta Radon, 2022. https://www.stralsakerhetsmyndigheten.se/omraden/radon/att-mata-radon/. (Accessed 26 January 2023).
[6] D. Sedin, I. Hjelte, The Radon Situation in Sweden, 2004, pp. 3–5.
[7] Swedish National Board of Housing Building and Planning, Technical Status in Swedish Buildings - Results from the BETSI Project (Teknisk Status I Den Svenska

Bebyggelsen - Resultat Från Projektet BETSI), 2010. http://www.boverket.se/globalassets/publikationer/dokument/2011/betst-teknisk-status.pdf.
[8] S.M. Khan, S. Chreim, Residents' perceptions of radon health risks: a qualitative study, BMC Publ. Health 19 (2019) 1–11, https://doi.org/10.1186/s12889-019-7449-y.
[9] W. Wei, O. Ramalho, L. Malingre, S. Sivanantham, J.C. Little, C. Mandin, Machine learning and statistical models for predicting indoor air quality, Indoor Air 29 (2019) 704–726, https://doi.org/10.1111/ina.12580.
[10] F. Rezaie, S.W. Kim, M. Alizadeh, M. Panahi, H. Kim, S. Kim, J. Lee, J. Lee, J. Yoo, S. Lee, Application of machine learning algorithms for geogenic radon potential mapping in Danyang-Gun, South Korea, Front. Environ. Sci. 9 (2021) 1–17, https://doi.org/10.3389/fenvs.2021.753028.
[11] D. Valcarce, A. Alvarellos, J.R. Rabuñal, J. Dorado, M. Gestal, Machine learning-based radon monitoring system, Chemosensors 10 (2022), https://doi.org/10.3390/chemosensors10070239.
[12] S. Khan, J. Taron, A. Goodarzi, Machine learning as a next-generation tool for indoor air radon exposure prediction, machine learning as a next-generation tool for indoor air radon exposure prediction. https://doi.org/10.4135/9781529743708, 2020.
[13] S.M. Khan, D.D. Pearson, T. Rönnqvist, M.E. Nielsen, J.M. Taron, A.A. Goodarzi, Rising Canadian and falling Swedish radon gas exposure as a consequence of 20th to 21st century residential build practices, Sci. Rep. 11 (2021) 1–15, https://doi.org/10.1038/s41598-021-96928-x.
[14] O.M. Oni, A.A. Aremu, O.O. Oladapo, B.A. Agboluaje, J.A. Fajemiroye, Artificial neural network modeling of meteorological and geological influences on indoor radon concentration in selected tertiary institutions in Southwestern Nigeria, J. Environ. Radioact. 251–252 (2022), 106933, https://doi.org/10.1016/j.jenvrad.2022.106933.
[15] A. Sarra, L. Fontanella, P. Valentini, S. Palermi, Quantile regression and Bayesian cluster detection to identify radon prone areas, J. Environ. Radioact. 164 (2016) 354–364, https://doi.org/10.1016/j.jenvrad.2016.06.014.
[16] G. Kropat, F. Bochud, M. Jaboyedoff, J.P. Laedermann, C. Murith, M. Palacios Gruson, S. Baechler, Predictive analysis and mapping of indoor radon concentrations in a complex environment using kernel estimation: an application to Switzerland, Sci. Total Environ. 505 (2015) 137–148, https://doi.org/10.1016/j.scitotenv.2014.09.064.
[17] G. Kropat, F. Bochud, M. Jaboyedoff, J.P. Laedermann, C. Murith, M. Palacios, S. Baechler, Improved predictive mapping of indoor radon concentrations using ensemble regression trees based on automatic clustering of geological units, J. Environ. Radioact. 147 (2015) 51–62, https://doi.org/10.1016/j.jenvrad.2015.05.006.
[18] J. Elío, G. Cinelli, P. Bossew, J.L. Gutiérrez-Villanueva, T. Tollefsen, M. De Cort, A. Nogarotto, R. Braga, The first version of the pan-European indoor radon map, Nat. Hazards Earth Syst. Sci. 19 (2019) 2451–2464, https://doi.org/10.5194/nhess-19-2451-2019.
[19] P.-Y. Wu, T. Johansson, M. Mangold, C. Sandels, K. Mjörnell, Evaluating the indoor radon concentrations in the Swedish building stock using statistical and machine learning, in: 13th Nordic Symposium on Building Physics, IOP Journal of Physics: Conference Series, Aalborg, 2023.
[20] B. Olsthoorn, T. Rönnqvist, C. Lau, S. Rajasekaran, T. Persson, M. Månsson, A. V. Balatsky, Indoor radon exposure and its correlation with the radiometric map of uranium in Sweden, Sci. Total Environ. 811 (2022), https://doi.org/10.1016/j.scitotenv.2021.151406.
[21] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, M. Asadpour, Boosting methods for multi-class imbalanced data classification: an experimental review, J Big Data 7 (2020), https://doi.org/10.1186/s40537-020-00349-y.
[22] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J Big Data 6 (2019), https://doi.org/10.1186/s40537-019-0192-5.
[23] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, https://doi.org/10.1145/2939672.2939785.
[24] S. Raschka, V. Mirjalili, Python Machine Learning : Machine Learning and Deep Learning with python, Scikit-Learn, and Tensorflow 2, third ed., Packt Publishing Ltd., 2019.
[25] Swedish Radiation Safety Authority, Measurement of Radon in Workplace - Method Description (Mätning Av Radon På Arbetsplatser - Metodbeskrivning), 2021.
[26] Swedish Radiation Safety Authority, Measurement of Radon in Residential Buildings - Method Description (Mätning Av Radon I Bostäder - Metodbeskrivning), 2013.
[27] GISGeography, Inverse Distance Weighting Interpolation, 2022. https://gisgeography.com/inverse-distance-weighting-idw-interpolation/. (Accessed 31 August 2022).
[28] Geological Survey of Sweden, Soil Types (Jordarter) 1:25 000-1:100 000, vol. 1, 2018, pp. 1–13.
[29] 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.
[30] Swedish Radiation Safety Authority, Radon – Residences and Premises to Which the Public Has Access, Radon – Bostäder och lokaler dit allmänheten har tillträde), 2020.
[31] K. Akbari, R. Oman, Impacts of heat recovery ventilators on energy savings and indoor radon in a Swedish detached house, WSEAS Trans. Environ. Dev. 9 (2013) 24–34.
[32] K. Akbari, J. Mahmoudi, M. Ghanbari, Influence of indoor air conditions on radon concentration in a detached house, J. Environ. Radioact. 116 (2013) 166–173, https://doi.org/10.1016/j.jenvrad.2012.08.013.
[33] G. Axelsson, E.M. Andersson, L. Barregard, Lung cancer risk from radon exposure in dwellings in Sweden: how many cases can be prevented if radon levels are lowered? CCC (Cancer Causes Control) 26 (2015) 541–547, https://doi.org/10.1007/s10552-015-0531-6.
[34] G. Kropat, F. Bochud, M. Jaboyedoff, J.P. Laedermann, C. Murith, M. Palacios, S. Baechler, Major influencing factors of indoor radon concentrations in Switzerland, J. Environ. Radioact. 129 (2014) 7–22, https://doi.org/10.1016/J.JENVRAD.2013.11.010.
[35] Swedish National Board of Housing Building and Planning, Geografiska justeringsfaktorer, 2017. https://www.rockwool.com/se/downloads-tools/bbb-boverkets-byggregler/geografiska-justeringsfaktorer/. (Accessed 20 January 2023).
[36] D. Harbecke, Y. Chen, L. Hennig, C. Alt, Why only micro-F1? Class weighting of measures for relation classification, NLP-power 2022 - 1st workshop on efficient benchmarking in NLP, Proceedings of the Workshop (2022) 32–41, https://doi.org/10.18653/v1/2022.nlppower-1.4.
[37] F. Giannakas, C. Troussas, A. Krouska, C. Sgouropoulou, I. Voyiatzis, XGBoost and Deep Neural Network Comparison: the Case of Teams' Performance, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12677 LNCS, 2021, pp. 343–349, https://doi.org/10.1007/978-3-030-80421-3_37/COVER.
[38] M. Bukowski, J. Kurek, I. Antoniuk, A. Jegorowa, Decision confidence assessment in multi-class classification, Sensors 21 (2021) 1–15, https://doi.org/10.3390/s21113834.