

Article

Evaluation of Residential Indoor Radon Levels in Zagreb Using Machine Learning

Tomislav Bituh ^{1,†} , Marija Jelena Lovrić Štefiček ^{1,*} , Tea Čvorišćec ¹, Branko Petrincec ^{1,2}  and Silvije Davila ¹ 

¹ Institute for Medical Research and Occupational Health, 10000 Zagreb, Croatia; tbituh@imi.hr (T.B.)

² Faculty of Dental Medicine and Health, Josip Juraj Strossmayer University of Osijek, 31000 Osijek, Croatia

* Correspondence: mlovric@imi.hr

† These authors contributed equally to this work.

Abstract

Machine learning (ML) models can complement traditional measurement-based approaches by supporting large-scale screening, spatial analysis, and prioritization of buildings for testing of indoor radon, a leading cause of lung cancer among non-smokers. Originating from uranium decay in soil and rock, radon enters homes via foundation cracks and accumulates indoors, influenced by building characteristics, ventilation, urbanization, and geogenic factors. As part of the Zagreb pilot within the “Evidence Driven Indoor Air Quality Improvement” (EDIAQI) project, this is the first ML application for indoor radon analysis in Croatia. This research evaluates residential indoor radon concentrations in Zagreb using ML applied to a dataset of 80 households. Several linear regression and tree-based ensemble methods were tested. The best-performing model (GBR) achieved an R^2 of 0.99 on the training set and 0.57 on the test set, with an RMSE of 33 Bq/m³ and MAE of 26 Bq/m³. Although predictive performance was moderate and generalization limited, key building characteristics such as construction year, dwelling type, occupancy details, and floor level were identified as relevant variables. The results suggest that machine learning may support radon risk prioritization in urban environments, but cannot replace direct measurements for regulatory purposes.

Keywords: radon; indoor air; machine learning; radioactivity; dwellings

1. Introduction

Machine learning models have emerged as a powerful alternative to traditional mechanistic approaches for the evaluation and prediction of indoor air quality [1,2]. Using these advanced computational algorithms, researchers can analyse complex datasets that include a wide range of environmental, geological, and structural factors influencing hazard concentrations. Radon is a recognized human carcinogen with a long-standing history of regulation in numerous countries. Within the broader framework of indoor air quality (IAQ), it is increasingly assessed in conjunction with other pollutants such as particulate matter (PM_{2.5} and PM₁₀) and carbon dioxide (CO₂), despite its regulatory history being considerably longer [2]. Integrating machine learning approaches can support and optimize measurement campaigns by prioritizing buildings and regions for testing of radon over traditional measurement methods that require extensive fieldwork. Moreover, as these models incorporate larger and more diverse datasets, they hold significant potential to enhance public health initiatives aimed at reducing radon exposure and related risks. Contrary to the conventional U.S. practice, which extensively employs large-scale short-term radon



Academic Editor: Peter Brimblecombe

Received: 27 January 2026

Revised: 2 March 2026

Accepted: 4 March 2026

Published: 6 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

measurements [3], the current study does not advocate for substituting measurement-based radon testing. Rather, machine learning methodologies are explored as decision-support instruments aimed at assisting in risk assessment, mapping, and decision-making.

Radon is a naturally occurring radioactive element originating from the decay chain of uranium and is present in soil and rock formations. Given that a substantial portion of time is spent indoors, exposure to radon through inhalation has emerged as a critical public health concern within indoor environments. Radon gas generated in soil and bedrock migrates towards the surface and enters residential structures through foundation and basement fissures, driven by pressure differences induced by temperature and wind, where it can accumulate to concentrations that pose significant health risks [4–6]. Secondary sources of indoor radon include certain building materials with elevated radionuclide content, groundwater supplied from private wells, natural gas, and outdoor air. Indoor accumulation depends on air exchange [6–9] with usually lower radon outdoor air and is therefore controlled by building air tightness and ventilation rate, leading to systematic vertical gradients in multi-story buildings, where lower floors tend to show higher radon concentrations than upper floors [5,10,11].

Identified by the World Health Organization (WHO) [12] as a major indoor air pollutant and classified as a Group 1 carcinogen by the International Agency for Research on Cancer [13], radon is the leading cause of lung cancer among non-smokers and may also contribute to the incidence of childhood leukemia [14–17]. Gaskin et al., 2018 [18] estimated that global lung cancer mortality attributable to residential radon exposure amounts to 226,000 deaths annually across 66 countries (with sufficient radon data), representing 74% of the global population. Most of these fatalities are caused by the synergistic effects of smoking and radon exposure. Under Directive 2013/59/Euratom, the European Union (EU) established a maximum reference level of 300 Bq/m³ for both residential and indoor workplaces [19].

Recently, Tsapalov et al. [20] presented “The Rational Method”, a protocol that provides a rigorous framework for indoor radon conformity assessment based on measurement uncertainty and decision reliability. Such protocols are essential for regulatory compliance. However, the progress is dependent on the national radiation risk regulators, who need to work towards providing effective services for testing and safeguarding buildings against radon, which would be affordable for the public. Data-driven approaches such as ML can support these measurement frameworks by optimizing survey design and targeting high-risk areas.

Modern machine learning (ML) models for radon prediction typically incorporate different predictor categories to account for the complexity of modelling radon. Building characteristics such as building type (e.g., detached house vs. apartment), construction year, floor level, basement presence, and structural properties, modify radon transmission from soil to indoor air and are repeatedly linked to concentrations in large epidemiological and survey-based studies [5,21–26]. Land use and urban form indicators capture contrasts between rural/metropolitan areas and dense urban zones. Regions with lower imperviousness (often rural or some metropolitan) exhibit higher indoor radon levels due to stronger influence from undisturbed soils [10,27,28]. Alongside building characteristics and urbanization proxies, geogenic factors such as uranium content, soil permeability, geology, and proxies for geogenic radon potential consistently rank as important determinants, reflecting the ground as the primary source of indoor radon [28–32].

Empirical comparisons show that ML approaches generally outperform classical linear regression in radon prediction [17,21]. Decision tree ensembles such as Random Forest (RF) and Extreme Gradient Boosting (XGB) are widely used models because they capture interactions and non-linearities without strong distributional assumptions and handle

mixed-type predictors with relative ease. Studies from several European and Asian settings report that RF and XGB achieve substantially higher validation R^2 values than linear models. For example, a Swiss study [33] observed that RF nearly doubled the explained variance relative to a linear baseline (validation R^2 0.31 vs. 0.15), illustrating the benefit of flexible non-parametric methods for exposure modelling. Random Forest has been used for national-scale radon mapping and household-level concentration estimation, where its ensemble structure helps stabilize predictions across heterogeneous regions and building types [34,35]. In large, data-rich national datasets, XGB-type boosted tree models have also shown promising discriminatory performance for classifying dwellings above regulatory reference levels or within higher radon categories, although such results depend on substantial sample sizes and detailed covariate information [4,34,36].

Even though numerous studies have investigated indoor radon predictions using machine learning models, this study is the first application of such methods for the analysis of indoor radon that has been done within Croatia. As part of the EDIAQI (Evidence Driven Indoor Air Quality Improvement) project, the investigation encompassed radon measurements conducted in residential properties across the city of Zagreb and Zagreb County. The locations of the measurements were selected based on the voluntary participation of individuals interested in indoor air quality through the EDIAQI project [37]. The aims of this study were to measure the activity concentrations of ^{222}Rn within household environments and to use machine learning to explore how various factors, such as floor level, dwelling type, and building age, affect radon levels and consequently model performance. Given the challenges of advanced radon modelling, a household-level dataset with building and occupancy details is needed to develop and compare multiple models. While ML models can explore associations at the household level, predictive performance typically remains insufficient for regulatory identification of individual hazardous buildings. Therefore, the results should be interpreted as probabilistic risk indicators rather than compliance determinations.

The limitations of this study include a short measurement duration (about 90 days), which creates temporal uncertainty in estimating annual average radon concentration [38], and a limited number of dwellings pre-determined by the project pilot study.

2. Materials and Methods

2.1. Sampling and Data Acquisition

Activity concentrations of ^{222}Rn were measured using solid-state nuclear track detectors (SSNTDs), the CR-39 type (Radosys Inc., Budapest, Hungary). The detectors were distributed within the dwellings and were exposed on average for 90 days. They were placed in living rooms or bedrooms on a shelf or in a cupboard at about 1.5–2 m height. After exposure, the detectors were delivered to the laboratory and chemically etched. The etching process is done with 6 M NaOH at 90 °C for 5 h, followed by neutralization (0.1 M HCl) and washing with dH₂O. After drying the detectors, the tracks were counted using an automatic microscope (Radometer, Radosys Inc., Budapest, Hungary). The activity concentration of ^{222}Rn was calculated using Equation (1):

$$\bar{C} = \frac{(n_g - n_b) \cdot F_c}{t \cdot S_{\text{SSNTD}}} \quad (1)$$

where:

- n_g —number of tracks after the exposure
- n_b —number of background tracks
- F_c —calibration factor (in track number per cm²/Bq/h/m²)

- t —exposure time (h)
- S_{SSNTD} —detector surface

The measurement uncertainty of ^{222}Rn activity concentration (C) is calculated according to Equation (2):

$$u(\bar{C}) = \sqrt{\left(n_g + \frac{\bar{n}_b}{N}\right) \cdot \omega^2 + \bar{C}^2 \cdot u_{rel}^2(\omega)} \quad (2)$$

where:

$$u_{rel}^2(\omega) = u_{rel}^2(F_c) + u_{rel}^2(S_{SSNTD}) + u_{rel}^2(t) \quad (3)$$

- $u_{rel}^2(F_c)$ —calibration factor uncertainty (F_c) (5%)
- $u_{rel}^2(S)$ —detector surface uncertainty (S) (5%)
- $u_{rel}^2(t)$ —exposure time uncertainty.

All the results of ^{222}Rn activity concentration are shown with expanded measurement uncertainty with a coverage factor of $k = 2$. Detection limit for 90-day exposure is 5 Bq/m^3 . The method is accredited according to the ISO 11665-4:2021 standard [39], and the laboratory is accredited according to the HRN EN ISO/IEC 17025:2017 standard [40]. As part of the Zagreb pilot within the EDIAQI project, indoor air quality measurements, including indoor radon measurements, in voluntary participation households across Zagreb City and Zagreb County. The detectors were placed in households, considering the uniform coverage of samples in the area. Considering usable samples for detector preservation (lost or damaged detectors) and predictor-variable data availability, 80 households were selected for this study (Figure 1, generated in Jupyter notebook using Urban Atlas and approximate dwelling coordinates) [41].

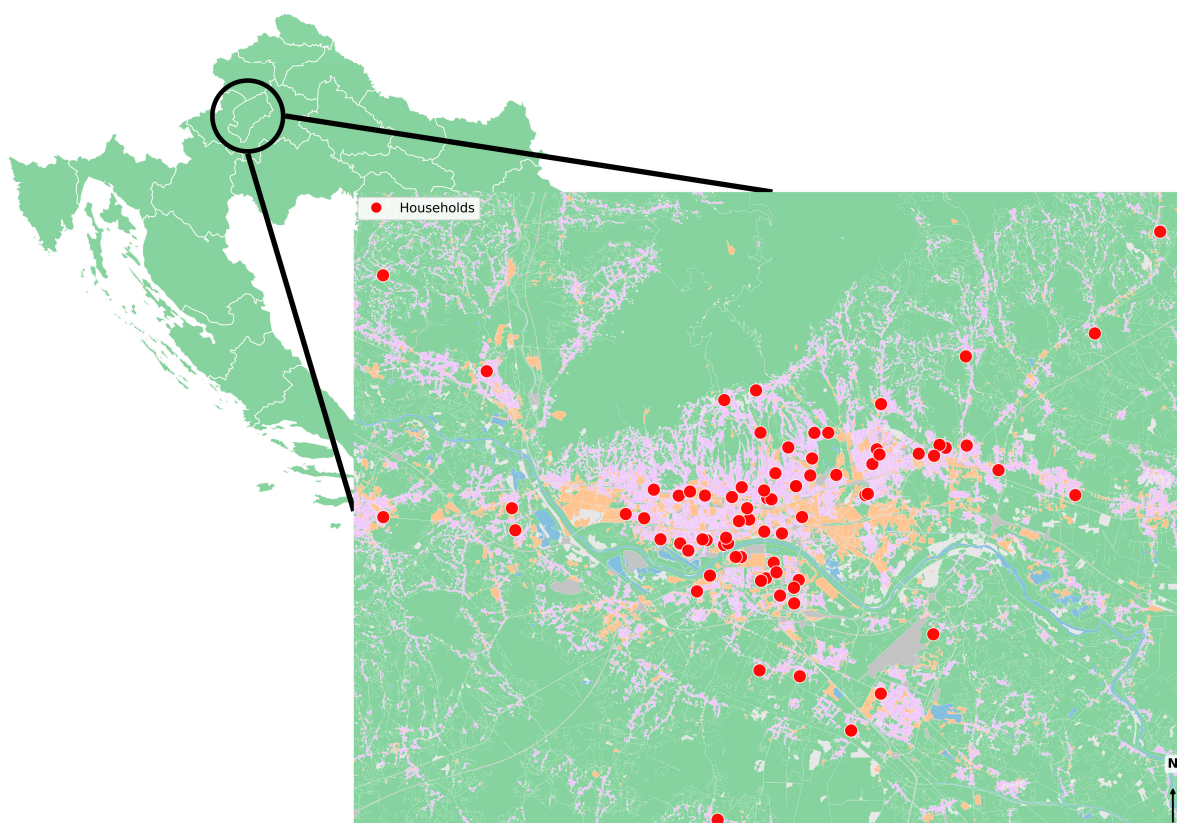


Figure 1. Measurement locations (red dots—sampling sites, urban area—purple; roads—grey; industrial/commercial area—orange; green/nature—green; water—blue).

2.2. Statistical Analysis and Data Pre-Processing

All analyses were performed in Python (v3.11) using the packages pandas [42], NumPy [43], SciPy [44], Statsmodels [45], Matplotlib [46], and Seaborn [47,48], while Scikit-learn [49] was used for modelling. The dataset comprised indoor radon concentrations and contextual predictor variables. Household type was described as a binary `is_apartment` variable categorizing house vs. apartment dwellings, while year represents the construction year of the dwelling. Even though measurements were carried out over about 90 days, there was variability in the duration of measurement, so it was included as a potential exposure proxy. Urban form was characterized using the Copernicus Urban Atlas [41] land cover/land use dataset for the Zagreb Functional Urban Area by spatially joining approximate dwelling coordinates to the corresponding land-use polygons, from which we derived categorical indicators of dominant residential, mixed, and non-residential zones around each dwelling (`ua_roads`, `ua_green_urban_nature`, `ua_urban_highsl`, `ua_industrial_commercial`). These Urban Atlas classes were used solely as contextual, area-level predictors and do not contain information on geogenic radon emission or radon potential. Prior to statistical analysis, the dataset underwent a systematic cleaning procedure to ensure the validity and robustness of the results. Descriptive statistics (mean, standard deviation, quartiles, interquartile range, and percentiles) summarized the central tendency and dispersion of radon activity concentrations. Measured radon values showed pronounced right skew (Shapiro–Wilk $W = 0.811$, $p < 0.001$), indicating non-normality, prompting the use of non-parametric methods. Exploratory nonparametric analyses included Kruskal–Wallis tests for categorical predictors (e.g., household type), Spearman correlations for ordinal or continuous variables (e.g., floor, year), and robust ordinary least squares regression (with heteroscedasticity-robust standard errors). Variables with consistent physical interpretation and stable statistical behavior were retained for multivariable modelling.

Radon concentrations were log-transformed prior to model fitting. This transformation is commonly used in environmental and exposure modelling to address right-skewed distributions and approximate homoscedastic residuals [50]. All numeric predictors were treated as continuous features and scaled within the modelling pipeline, and categorical variables were encoded using one-hot representation so that linear and tree-based models could incorporate group-specific shifts in radon levels. For interpretability of performance, predictions were back-transformed to the original scale, and error metrics were reported on the raw scale.

2.3. Machine Learning

This study investigated two different regression model families, beginning with linear models, namely linear regression, ridge regression with ℓ_2 regularization (optimized via grid search over α), and lasso regression with ℓ_1 regularization (optimized via grid search over α). Linear models provide a transparent baseline and are standard for radon prediction in epidemiological and exposure studies [51]; however, they tend to be more accurate for rooms with low radon concentrations and less so for high concentrations, making them suitable as simple screening tools rather than precise predictors for extreme values. Seasonal measurements have been shown to be representative of annual values in this context, supporting their use in a linear regression framework [52,53]. In addition, Vukotic et al., 2021 [22] implemented a logistic regression approach, selecting five explanatory variables (city, building type, basement, window frame, and construction period) to predict the probability of indoor radon concentrations exceeding 200 Bq/m³ among 734 buildings, a model that could be generalized to other cities.

Alongside linear regression models, tree-based ensemble methods were investigated, including both bagging and boosting approaches. Random Forest Regression (RFR), a bootstrap-aggregated ensemble of decision trees, has the ability to reduce variance in complex datasets by averaging over many decorrelated trees. It is a widely used and well-proven algorithm due to its ease of training and implementation, requiring less pre-processing [35]. In comparison, Gradient Boosting Regressor (GBR) constructs an additive ensemble of shallow decision trees by sequentially fitting each tree to the negative gradients of the squared error loss function, thereby minimizing residuals through step-wise optimization. Similarly, Extreme Gradient Boost Regressor (XGB) [54], a scalable and regularized gradient-boosting implementation, extends this approach with ℓ_1 and ℓ_2 regularization on leaf weights, second-order gradient approximations via Taylor expansion, and tuneable parameters such as maximum tree depth, learning rate (shrinkage), and subsample ratios, making it highly effective for tabular data prediction [16,36,55,56]. Quantile Random Forests (qRF), first introduced by Meinshausen, 2006 [57], give a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables and have been previously used in mapping the condition distribution of outdoor radon [58]. For the purpose of this study, specifically, the median (50th percentile) Quantile Random Forest (qRF-50th) was used.

Model validation is one of the most important parts of building a machine learning algorithm, ensuring robust generalization through sensible data-splitting strategies, with cross-validation (CV) standing as one of the most prevalent methods for model selection and performance estimation [11,59,60]. To mitigate variability from random splits, particularly in small datasets similar to ours, hyperparameter tuning was conducted via repeated 5-fold CV with 5 repetitions (5×5 CV), optimizing negative mean absolute error (neg-MAE) on the log-transformed target variable. Hyperparameter tuning of linear models was conducted with regard to the prevention of memorization and enabling sparsity:

- Ridge {"model__alpha": [0.1, 1, 5, 10, 100]}
- Lasso {"model__alpha": [0.001, 0.01, 0.05, 0.1]}

Hyperparameters for the tree ensembles were tuned using conservative settings. Gradient Boosting used moderate `n_estimators` [150–600], with slow `learning_rates` [0.02–0.08], shallow `max_depths` [2–4], and subsample ratios [0.6–1.0] to stabilize gradients and reduce variance. Random Forest tested `n_estimators` [150–600], constrained depths `model_max_depth` [None, 5, 8, 12, 20], and splits (`min_samples_split` [2–10], `min_samples_leaf` [1–4]), `max_features` ["sqrt", "log2", None], and `max_samples` [0.6–0.9] for bagging stability. XGBoost spanned higher `n_estimators` [300–2000] but with early stopping and ultra-conservative `learning_rates` [0.01–0.05], `max_depth` [2–3], `min_child_weight` [1–8], gamma pruning [0–0.5], row/column subsampling [0.6–0.85], L1/L2 penalties (`reg_alpha` [0–0.01], `reg_lambda` [1–10]), and leaf-constrained growth. Model performance was evaluated on the test set using *RMSE*, *MAE*, and R^2 on the raw target scale, with additional RMSE and MAE reported on the log scale to diagnose overfitting and assess performance in the space where models were trained.

SHAP (SHapley Additive exPlanations) values were applied post-modeling to interpret the machine learning models by quantifying each feature's contribution to individual predictions [4,34,61]. This game-theoretic approach assigns Shapley values representing the average marginal contribution of a feature relative to the model's baseline prediction [62–64]. Global feature importance was estimated by averaging the absolute SHAP values across the dataset. Summary plots visualized feature importance alongside effects, with each point depicting a SHAP value for one instance and feature.

3. Results

Across the 80 dwellings, measured radon ranged from 5.9 to 332.7 Bq/m³ (median $\approx$ 58.9, IQR 39.7–92.1), with a mean of 74.3 Bq/m³ (Table 1). Since the area of interest is not a radon-prone area, this range is expected. The distribution of measured radon is shown in Figure 2.

Table 1. Descriptive statistics for activity concentrations of ²²²Rn (Bq/m³) in dwellings.

Activity Concentrations of ²²² Rn (Bq/m ³)	
count	80
mean	74.29
std	55.44
min	5.9
5%	17.95
25%	39.73
50%	58.91
75%	92.11
95%	167.51
max	332.65

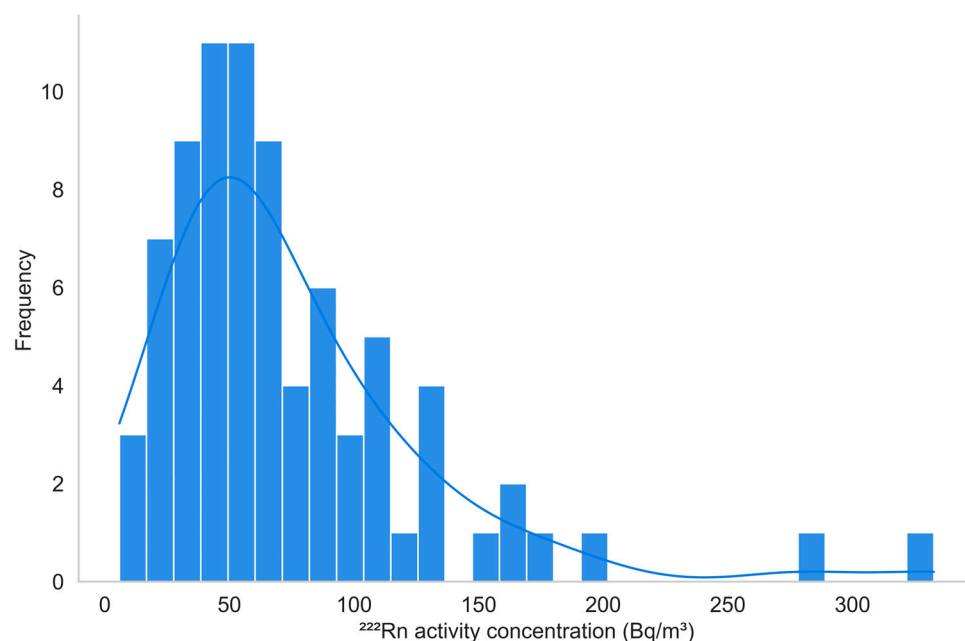


Figure 2. Activity concentrations of ²²²Rn (Bq/m³) in dwellings—frequency distribution (blue line—kernel density estimate (KDE)).

Older buildings tended to have higher radon: median concentrations decreased from 113.6 Bq/m³ in buildings constructed before 1925 to 51.2 Bq/m³ in buildings from 2000 to 2024, with a monotonic downward trend in medians across construction-year categories (Figure 3).

Floor level (Figure 4) showed the expected inverse pattern, with median radon declining from 88.6 Bq/m³ in the basement/ground floor (−1/0) to 42.9 Bq/m³ at $\geq$ 5th floor, although variability was substantial in each group.

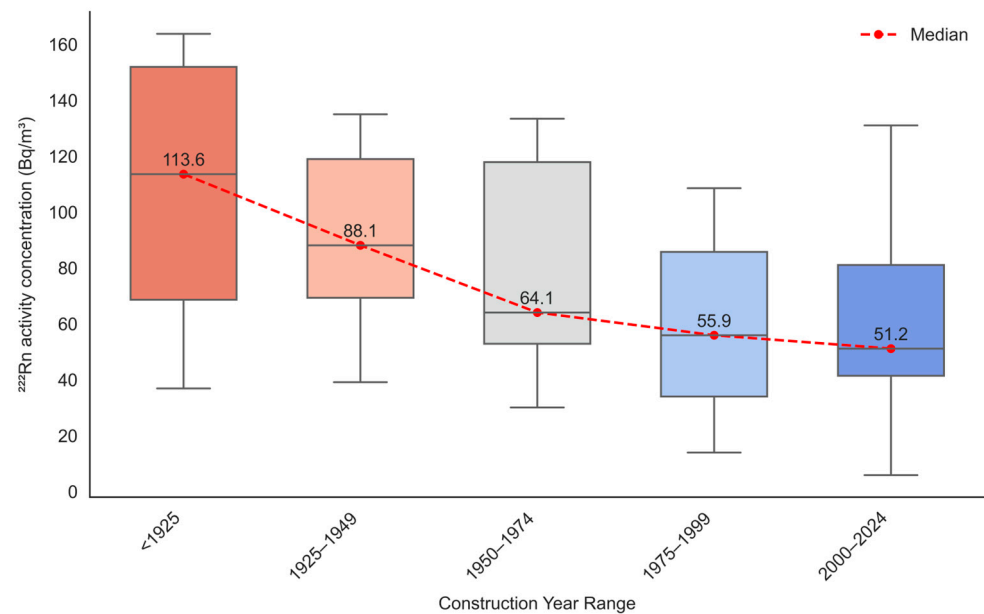


Figure 3. ^{222}Rn activity concentration (Bq/m^3) distributed by building construction year (median, lower quartile, interquartile range, upper quartile, minimum, and maximum values).

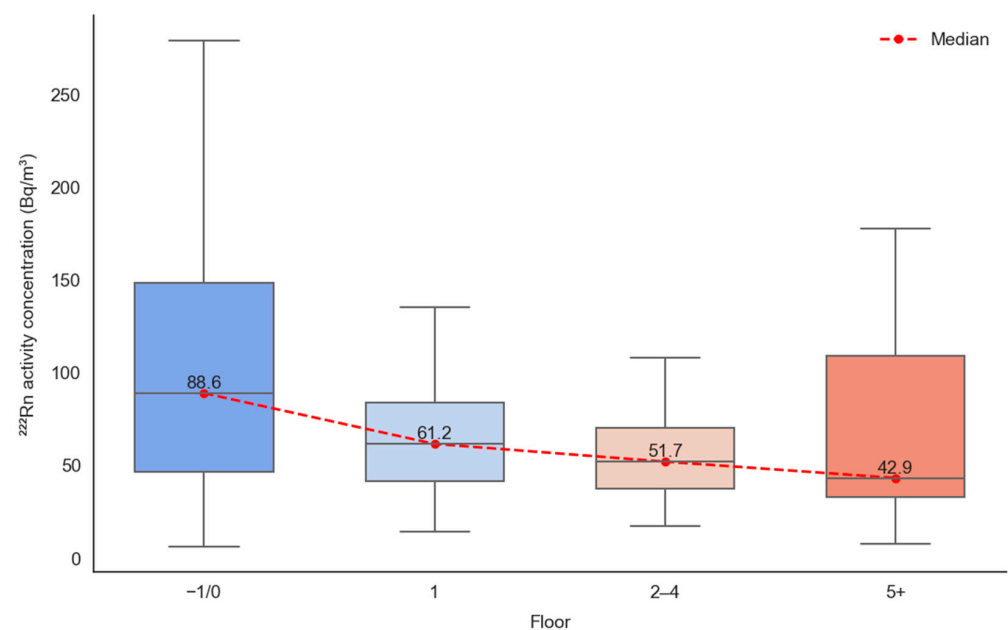


Figure 4. ^{222}Rn activity concentration (Bq/m^3) distributed by floor (median, lower quartile, interquartile range, upper quartile, minimum, and maximum values).

Non-parametric tests revealed that building characteristics dominate household radon variability. Dwellings classified as apartments had significantly different radon distributions compared with non-apartments (Kruskal–Wallis, $p = 0.0049$, $\eta^2 = 0.089$). Floor level showed a modest but significant negative Spearman correlation with radon ($\rho = -0.237$, $p = 0.0341$), while construction year showed a marginal monotonic trend in robust OLS ($R^2 = 0.042$, $p = 0.0545$). No clear associations were observed for duration, window count, or season_heating in these pre-model tests. Urban land use variables were constructed using Copernicus Urban Atlas data [41] by leveraging the longitude and latitude of the chosen dwellings and represented by a compact multi-level variable reflecting the dominant built form. Although geological and soil conditions are key determinants of geogenic radon potential and are therefore conceptually important for radon modelling [65,66], they

were not available with sufficient resolution or completeness at the household level in our dataset, and could not be incorporated as reliable predictors in the final models.

3.1. Hyperparameter Tuning

Hyperparameter optimization for RFR converged on $n_estimators = 150$, $max_depth = 12$, $min_samples_split = 5$, $min_samples_leaf = 2$, $max_samples = 0.9$, and $max_features = None$, representing moderately deep trees regularized by minimum node constraints and sub-sampling, in line with recommended practices for tabular environmental data. In the case of GBR tuning yielded $learning_rate = 0.02$, $max_depth = 4$, $n_estimators = 150$, and $subsample = 1.0$, characterizing a conventional gradient-boosted tree ensemble with shallow trees and a subdued learning rate to mitigate overfitting. XGB with an early stopping consistent with validation error plateaus identified $learning_rate = 0.01$, $max_depth = 3$, $n_estimators = 600$, $subsample = 0.85$, $colsample_bytree = 0.6$, $min_child_weight = 5$, $max_leaves = 8$, $reg_lambda = 10.0$, and $grow_policy = 'depthwise'$ as optimal hyperparameters.

3.2. Overall Predictive Performance

The Gradient Boosting Regressor achieved the best out-of-sample performance, with the lowest test $RMSE = 33.26$ and test $MAE = 26.10$, and the highest test $R^2 = 0.5654$. This indicates that the model explains more than half of the variance in radon levels in the test data and yields the smallest typical prediction error among the evaluated algorithms. GBR used shallow regression trees and a small learning rate, which are known to provide strong performance on structured tabular data with complex but smooth non-linearities. Tree-based models based on bagging and quantile estimation (RFR and qRF) yielded test R^2 values of approximately 0.38, $RMSE = 39$, and MAE values in the range of 31–33, while XGB, as the weakest of the tree-based models, performed with test values of $R^2 = 0.2791$, $RMSE = 42$, and $MAE = 30$. Out of the linear models, linear regression performed slightly better with test values of $R^2 = 0.3615$, $RMSE = 40$, and $MAE = 33$ while Ridge and Lasso failed, suggesting that the relationship between predictor variables and radon in this case study is non-linear and involves interactions that cannot be adequately captured by a purely linear specification. Table 2 summarizes performance on the raw scale for all models.

Table 2. Performance of models on raw target (train and test).

Model *	RMSE Train	MAE Train	R^2 Train	RMSE Test	MAE Test	R^2 Test
GBR	0.5480	0.3825	0.9999	33.2638	26.0993	0.5654
qRF-50th	31.2360	18.5334	0.6909	39.7278	33.1784	0.3801
RFR	37.4508	20.9577	0.5557	39.8401	31.2325	0.3766
LR	50.7884	31.9096	0.1829	40.3175	33.1589	0.3615
XGB	45.5573	26.3556	0.3426	42.8419	30.4625	0.2791
Lasso	54.6211	34.7807	0.0550	43.6915	32.6434	0.2502
Ridge	56.8355	35.9707	−0.0232	48.0360	33.4719	0.0937

* GBR—Gradient Boosting Regressor; qRF-50th—Quantile Random Forests; RFR—Random Forest Regression; LR—Linear Regression; XGB—Extreme Gradient Boosting; Lasso—Lasso regression; Ridge—Ridge regression.

4. Discussion

In contemporary metrology, adherence to national reference levels requires measurement-based conformity assessment incorporating defined uncertainty and decision reliability (e.g., 95%), as outlined in ISO/IEC and JCGM standards [67]. This study does not intend to substitute standardized indoor radon measurement procedures or to establish a conformity assessment protocol. Instead, it investigates the application of ML models as an auxiliary approach for spatial prediction and prioritization.

The limitations of this study include a short measurement duration, about 90 days, which introduces temporal uncertainty when estimating annual average radon concentration [38], and a limited number of dwellings, which were pre-determined by the project pilot study. Seasonal variability can lead to deviations between short-term and true annual means. Additionally, the model does not provide measurement uncertainty in the metrological sense, as well as the fact that individual building levels cannot be determined without direct measurement.

In this study, the GBR model demonstrated superior predictive performance for log-transformed radon concentration activation compared to XGB, qRF-50th, and RF models, particularly in a small-sample regime. The near-perfect training $R^2 \approx 0.9999$ combined with a substantially lower test $R^2 = 0.5654$ indicates that the model fits the training data very closely but still retains reasonable generalization to unseen data. Such overfitting aligns with expectations for tree-based ensembles under constrained sample sizes, where models tightly fit training data but retain reasonable generalization, and is primarily attributed to variations in effective model capacity and the fitting stability. Even though XGB is usually the go-to choice for its flexibility and raw performance in most tasks, GBR pulled ahead here thanks to its simpler, more reliable setup for limited data. The tuned GBR configuration (moderate depth, relatively small learning rate) provided a well-regularized additive model that generalized reliably from limited data, while the XGB pipeline relied on a more complex optimization/hyperparameter space and an early-stopping procedure that further reduced the effective training data via an internal validation split. Considering the number of predictor variables, the GBR's constrained tree growth and smoother boosting trajectory likely reduced variance and overfitting risk, whereas XGB's additional degrees of freedom (e.g., leaf growth policy, regularization, subsampling, column sampling) increased sensitivity to hyperparameter interactions in this low- n setting, resulting in weaker out-of-sample performance. Given the moderate number of predictors and the absence of highly detailed building-physics or geogenic covariates, the remaining error on the held-out test set reflects dataset limitations rather than a failure of the gradient boosting approach itself. More expressive alternatives such as neural networks or kernel Support Vector Machines (SVMs) that show benefits in dense timeseries settings were not explored here because, on small tabular datasets, they generally require substantially larger samples and heavier regularization to avoid overfitting and often do not outperform well-tuned tree ensembles, suggesting that future gains are more likely to come from richer covariate information and larger samples than from more complex model classes [36,68,69].

These findings advance beyond prior benchmarks, including XGB, qRF-50th, and RF, which often relied on larger datasets and reported modest R^2 values. For instance, XGB achieved an R^2 of 0.45 and $RMSE$ of 0.29 on log-transformed radon concentration [4]. Most of the previous research on radon prediction was focused on RF, with results ranging from $R^2 = 0.2$ and $RMSE = 52.76$ for geogenic radon potential prediction [34], and $R^2 = 0.24$ and $RMSE = 110.5$ using qRF for high-resolution indoor radon mapping [5]. For investigating regional residential radon, the performance of RF models ranged from an R^2 of 0.21 to 0.31 [16,21,35], while the best reported RF model for the city level had $R^2 = 0.46$ and $RMSE = 47.8$ [27]. While the RF model performance in our case is lower than the best-reported RF model for city-level radon prediction, the GBR model outperformed that result.

SHAP analysis for the GBR model (Figure 5) indicated that duration, year, and is_apartment were the three most influential predictors of radon according to their mean absolute SHAP values, with importance scores of approximately 0.225, 0.216, and 0.162, respectively. The next tier of contributors comprised household members (≈ 0.055), floor (≈ 0.038), and the urban-attribute indicators ua_roads (≈ 0.019) and ua_green_urban_nature (≈ 0.016),

while `ua_urban_highsl`, `season_heating`, and `ua_industrial_commercial` had smaller but non-zero contributions in the fitted model.

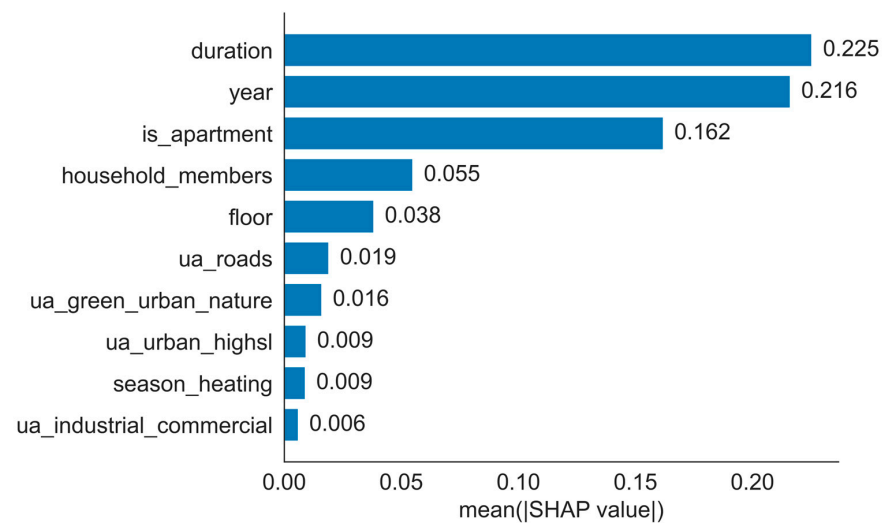


Figure 5. SHAP feature importance for the gradient boosting regressor.

The SHAP summary plot (Figure 6) further showed that longer duration values tended to have positive SHAP values, indicating an associated increase in predicted log radon, whereas shorter durations were associated with negative contributions, consistent with longer measurements capturing higher or more stable radon levels. Higher constriction year values corresponding to the age of the building generally had negative SHAP values, suggesting a temporal decline in modelled radon levels reflecting temporal trends in building practices and advances in construction and energy efficiency of buildings [70]. The floor variable displayed a pattern where lower floors contributed positively and higher floors negatively to the prediction, aligning with physical expectations that radon concentrations decrease with vertical distance from the ground because floor level modulates soil proximity, which is a known geogenic factor [6,22]. Household size is a proxy for occupancy-driven ventilation which affects radon concentrations [71], while urban features may reflect land-use effects on permeability.

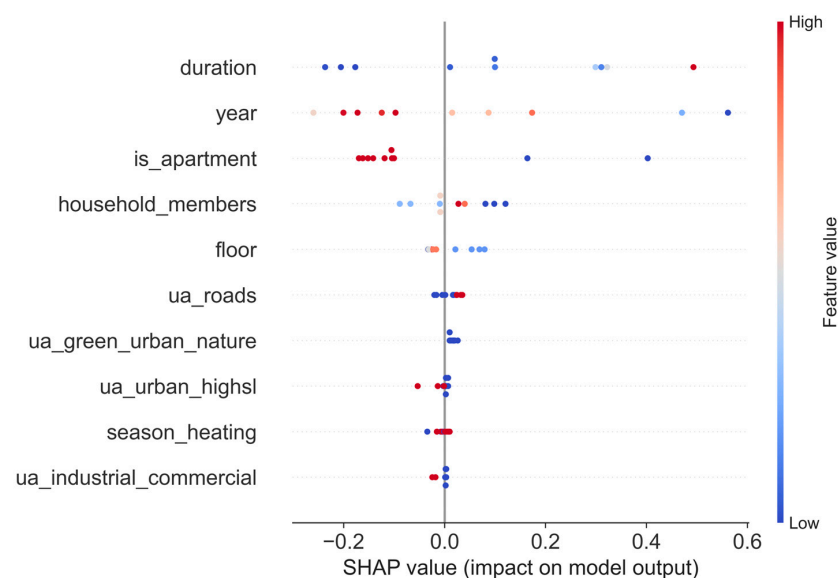


Figure 6. SHAP summary plot for the gradient boosting regressor.

In the SHAP summary plot, all apartment dwellings ($\text{is_apartment} = 1$) have slightly negative SHAP values, while the few non-apartments ($\text{is_apartment} = 0$) have positive SHAP values, indicating that the dwelling type had a systematic influence on model prediction. Looking at the GBR model outcome (Figure 7) based on dwelling type, it is apparent that the predicted Rn values are lower for apartments relative to non-apartment dwellings, which coincides with previous research [16,35]

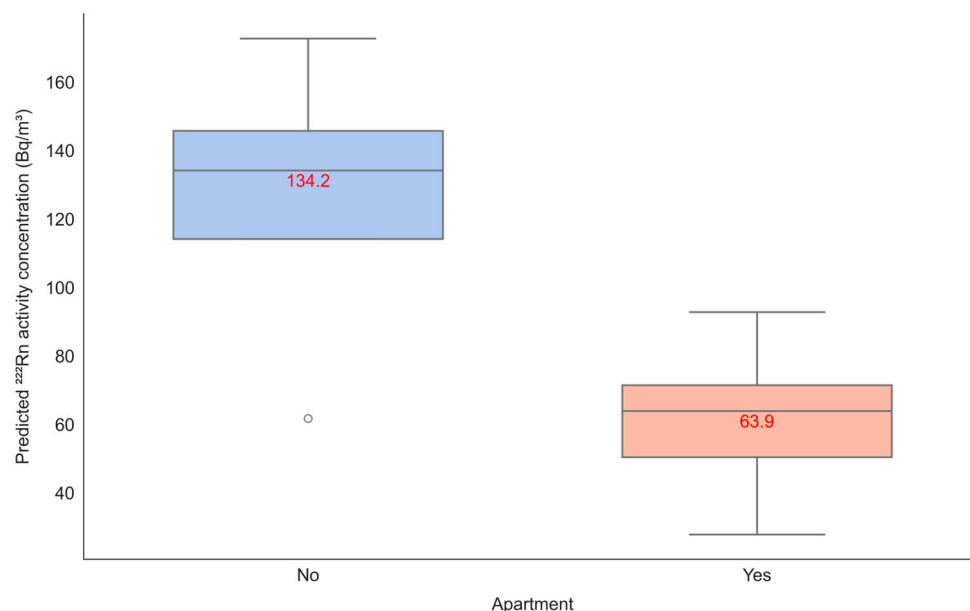


Figure 7. Distribution of predicted Rn by apartment status from the gradient boosting regressor (median, lower quartile, interquartile range, upper quartile, minimum, and maximum values).

5. Conclusions

This study developed and evaluated a supervised regression modelling pipeline for predicting indoor radon concentrations in an urban environment, employing a log-transformed outcome to stabilize variance and mitigate skewness in the highly variable radon data. Among the tested algorithms, Gradient Boosting Regressor (GBR), quantile Random Forest at the 50th percentile (qRF-50th), Random Forest Regressor (RFR), Linear Regression (LR), Extreme Gradient Boosting (XGB), Lasso, and Ridge, GBR demonstrated superior predictive performance. On the test set, GBR achieved an R^2 of 0.5654, a root mean square error (RMSE) of 33.2638, and a mean absolute error (MAE) of 26.0993.

GBR's strong generalization (test R^2 of 0.5654) contrasts with its near-perfect training fit ($R^2 = 0.9999$), suggesting effective handling of overfitting through ensemble techniques, while tree-based competitors like qRF-50th and RFR lagged with test R^2 values around 0.38. Linear models underperformed, likely due to non-linear radon dynamics influenced by spatial, meteorological, and building factors.

This performance exceeds the previously reported R^2 in the literature for household- or spatial-level radon predictions in an urban setting at a zip code level. These results demonstrate that a relatively small set of dwelling characteristics can explain a substantial amount of variability in indoor radon concentrations. At the same time, the remaining unexplained variance and residual error indicate that prediction accuracy is still limited by the absence of detailed predictors related to building physics, ventilation behavior, and high-resolution geogenic radon potential.

The use of 90-day measurements may introduce additional uncertainty in representing annual average concentrations. This affects both measurement-based conformity assessment and model training. Future work should incorporate additional predictors and/or

year-long measurements and extend the investigation to radon-prone areas within Croatia. This study evaluated whether routinely collected household and building characteristics can explain a meaningful proportion of indoor radon variability within this well-defined urban area, rather than produce a fully generalizable spatial prediction model for the wider region. Under these conditions, the model is best interpreted as an interpolator for dwellings that are similar, both in terms of building characteristics and urban context, to those observed in our dataset (i.e., within or immediately adjacent to the Zagreb core)

Comprehensive indoor radon measurements remain sparse in Croatia, lacking nationwide datasets due to high costs, logistical challenges, and low public awareness. This scarcity hinders effective risk mapping, regulatory enforcement, and public health interventions. This study addresses this critical gap by contributing data as well as findings that dwelling characteristics have a substantial contribution to the prediction of indoor radon levels. The ML model developed in this study is not designed for regulatory decision-making at the individual building level. Its predictive uncertainty does not meet the criteria for conformity assessment as specified by metrological standards. Rather, the model is appropriate for identifying radon-prone areas, supporting regional risk mapping, prioritizing buildings for measurement campaigns, and optimizing the allocation of measurement resources.

Author Contributions: Conceptualization, M.J.L.Š. and T.B.; methodology, M.J.L.Š. and T.B.; software, M.J.L.Š.; validation, M.J.L.Š. and T.B.; formal analysis, M.J.L.Š., T.B. and T.Č.; investigation, M.J.L.Š. and T.B.; data curation, M.J.L.Š., T.B. and T.Č.; writing—original draft preparation, M.J.L.Š. and T.B.; writing—review and editing, T.Č., B.P. and S.D.; visualization, M.J.L.Š. and T.B.; supervision, B.P. and S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the research project “Evidence driven indoor air quality improvement” (EDIAQI) funded by the European Union’s Horizon Europe research and innovation program under grant agreement No. 101057497 (EDIAQI). A part of research was performed using the facilities and equipment funded within the European Regional Development Fund project KK.01.1.1.02.0007 “Research and Education Centre of Environmental Health and Radiation Protection—Reconstruction and Expansion of the Institute for Medical Research and Occupational Health” and supported by the Institute for Medical Research and Occupational Health and the European Union—Next Generation EU projects (Program Contract of 8 December 2023, Class: 643-02/23-01/00016, Reg. no. 533-03-23-0006; EBDIZ and EnvironPollutHealth).

Institutional Review Board Statement: This study was approved by the Ethics Committee of the Institute for Medical Research and Occupational Health (Class: 01-18/23-03-2/1, No.: 100-21/23-3) on 6 March 2023.

Informed Consent Statement: Written informed consent was obtained from participants prior to their participating in the study, which was entirely voluntary.

Data Availability Statement: The data presented in this study are available upon reasonable request from the corresponding author because the data are not yet publicly available from the EDIAQI project. This publication has been prepared using European Union’s Copernicus Land Monitoring Service information; <https://doi.org/10.2909/fb4dffa1-6ceb-4cc0-8372-1ed354c285e6> (accessed on 12 January 2026).

Acknowledgments: The authors would like to acknowledge the staff of the Division for Radiation Protection and the Division of Environmental Hygiene of the Institute for Medical Research and Occupational Health who contributed to sampling, sample preparation, and analysis. During the preparation of this manuscript, the authors used Mendeley Desktop, version 1.19.8, for the purposes of reference formatting. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

- Altamirano-Astorga, J.; Gutierrez-Garcia, J.O.; Roman-Rangel, E. Forecasting Indoor Air Quality in Mexico City Using Deep Learning Architectures. *Atmosphere* **2024**, *15*, 1529. [\[CrossRef\]](#)
- Wei, W.; Ramalho, O.; Malingre, L.; Sivanantham, S.; Little, J.C.; Mandin, C. Machine Learning and Statistical Models for Predicting Indoor Air Quality. *Indoor Air* **2019**, *29*, 704–726. [\[CrossRef\]](#) [\[PubMed\]](#)
- US Environmental Protection Agency. *The National Radon Action Plan 2021–2025*; US Environmental Protection Agency: Washington, DC, USA, 2022.
- Gigante, G.; Antignani, S.; Di Carlo, C.; Loret, N.; Bochicchio, F. Decoding Indoor Radon: An Explainable AI Approach to Quantifying Building, Environmental, and Inhabitants' Contributions. *Sci. Total Environ.* **2025**, *998*, 180244. [\[CrossRef\]](#)
- Petermann, E.; Bossew, P.; Kemski, J.; Gruber, V.; Suhr, N.; Hoffmann, B. Development of a High-Resolution Indoor Radon Map Using a New Machine Learning-Based Probabilistic Model and German Radon Survey Data. *Environ. Health Perspect.* **2024**, *132*, 97009. [\[CrossRef\]](#)
- Su, C.; Wang, M.; Yin, Y.; Sun, C.; Zou, Z.; Wang, H.; Dai, Y. Approaches to Estimating Indoor Exposure to Radon—A Systematic Review. *Atmosphere* **2025**, *16*, 286. [\[CrossRef\]](#)
- Andersen, C.E.; Bergsøe, N.C.; Majborn, B.; Ulbak, K. Radon and Natural Ventilation in Newer Danish Single-Family Houses. *Indoor Air* **1997**, *7*, 278–286. [\[CrossRef\]](#)
- Shabaan, D.H.; EL-Araby, E.H.; Yajzey, R.; Azazi, A.; Alzhrani, S. Evaluation of the Radiation Emission of Radon Gas from Various Building Materials. *J. Radiat. Res. Appl. Sci.* **2025**, *18*, 101194. [\[CrossRef\]](#)
- Su, C.; Pan, M.; Zhang, Y.; Kan, H.; Zhao, Z.; Deng, F.; Zhao, B.; Qian, H.; Zeng, X.; Sun, Y.; et al. Indoor Exposure Levels of Radon in Dwellings, Schools, and Offices in China from 2000 to 2020: A Systematic Review. *Indoor Air* **2022**, *32*, e12920. [\[CrossRef\]](#)
- Rezaie, F.; Panahi, M.; Bateni, S.M.; Kim, S.; Lee, J.; Lee, J.; Yoo, J.; Kim, H.; Won Kim, S.; Lee, S. Spatial Modeling of Geogenic Indoor Radon Distribution in Chungcheongnam-Do, South Korea Using Enhanced Machine Learning Algorithms. *Environ. Int.* **2023**, *171*, 107724. [\[CrossRef\]](#)
- Li, L.; Coull, B.A.; Zilli Vieira, C.L.; Koutrakis, P. High-Resolution National Radon Maps Based on Massive Indoor Measurements in the United States. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2408084121. [\[CrossRef\]](#)
- WHO. *WHO Handbook on Indoor Radon: A Public Health Perspective*; WHO: Geneva, Switzerland, 2009; ISBN 9789241547673.
- IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. *Man-Made Mineral Fibres and Radon*; International Agency for Research on Cancer: Lyon, France, 1988; Volume 43.
- Krewski, D.; Lubin, J.H.; Zielinski, J.M.; Alavanja, M.; Catalan, V.S.; Field, R.W.; Klotz, J.B.; Létourneau, E.G.; Lynch, C.F.; Lyon, J.I.; et al. Residential Radon and Risk of Lung Cancer: A Combined Analysis of 7 North American Case-Control Studies. *Epidemiology* **2005**, *16*, 137–145. [\[CrossRef\]](#)
- Darby, S.; Hill, D.; Auvinen, A.; Barros-Dios, J.M.; Baysson, H.; Bochicchio, F.; Deo, H.; Falk, R.; Forastiere, F.; Hakama, M.; et al. Radon in Homes and Risk of Lung Cancer: Collaborative Analysis of Individual Data from 13 European Case-Control Studies. *BMJ* **2005**, *330*, 223. [\[CrossRef\]](#)
- Nikkilä, A.; Arvela, H.; Mehtonen, J.; Raitanen, J.; Heinäniemi, M.; Lohi, O.; Auvinen, A. Predicting Residential Radon Concentrations in Finland: Model Development, Validation, and Application to Childhood Leukemia. *Scand. J. Work Environ. Health* **2020**, *46*, 278–292. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lee, H.; Hanson, H.A.; Logan, J.; Maguire, D.; Kapadia, A.; Dewji, S.; Agasthya, G. Evaluating County-Level Lung Cancer Incidence from Environmental Radiation Exposure, PM2.5, and Other Exposures with Regression and Machine Learning Models. *Environ. Geochem. Health* **2024**, *46*, 82. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gaskin, J.; Coyle, D.; Whyte, J.; Krewski, D. Global Estimate of Lung Cancer Mortality Attributable to Residential Radon. *Environ. Health Perspect.* **2018**, *126*, 057009. [\[CrossRef\]](#) [\[PubMed\]](#)
- European Union. *Council Directive 2013/59/EURATOM Laying down Basic Safety Standards for Protection Against the Dangers Arising from Exposure to Ionizing Radiation*; European Union: Brussels, Belgium, 2014.
- Tsapalov, A.; Kovler, K.; Kiselev, S.; Yarmoshenko, I.; Bobkier, R.; Miklyaev, P. IAEA Safety Guides vs. Actual Challenges for Design and Conduct of Indoor Radon Surveys. *Atmosphere* **2025**, *16*, 253. [\[CrossRef\]](#)
- Vienneau, D.; Boz, S.; Forlin, L.; Flückiger, B.; de Hoogh, K.; Berlin, C.; Bochud, M.; Bulliard, J.-L.; Zwahlen, M.; Rösli, M. Residential Radon—Comparative Analysis of Exposure Models in Switzerland. *Environ. Pollut.* **2021**, *271*, 116356. [\[CrossRef\]](#)
- Vukotic, P.; Antovic, N.; Zekic, R.; Djurovic, A.; Andjelic, T.; Svrkota, N.; Mrdak, R.; Dlabac, A. Influence of Climate, Building and Residential Factors on Radon Levels in Ground-Floor Dwellings in Montenegro. *Nucl. Technol. Radiat. Prot.* **2021**, *36*, 74–84. [\[CrossRef\]](#)

23. Alber, O.; Laubichler, C.; Baumann, S.; Gruber, V.; Kuchling, S.; Schleicher, C. Modeling and Predicting Mean Indoor Radon Concentrations in Austria by Generalized Additive Mixed Models. *Stoch. Environ. Res. Risk Assess.* **2023**, *37*, 3435–3449. [\[CrossRef\]](#)
24. Gruber, V.; Baumann, S.; Wurm, G.; Ringer, W.; Alber, O. The New Austrian Indoor Radon Survey (ÖNRAP 2, 2013–2019): Design, Implementation, Results. *J. Environ. Radioact.* **2021**, *233*, 106618. [\[CrossRef\]](#)
25. Casey, J.A.; Ogburn, E.L.; Rasmussen, S.G.; Irving, J.K.; Pollak, J.; Locke, P.A.; Schwartz, B.S. Predictors of Indoor Radon Concentrations in Pennsylvania, 1989–2013. *Environ. Health Perspect.* **2015**, *123*, 1130–1137. [\[CrossRef\]](#)
26. Widya, L.K.; Rezaie, F.; Lee, J.; Lee, J.; Park, B.R.; Yoo, J.; Lee, W.; Lee, S. AI-Driven Geospatial Analysis of Indoor Radon Levels: A Case Study in Chungcheongbuk-Do, South Korea. *Earth Syst. Environ.* **2025**, *9*, 3615–3633. [\[CrossRef\]](#)
27. Carrion-Matta, A.; Lawrence, J.; Kang, C.-M.; Wolfson, J.M.; Li, L.; Vieira, C.L.Z.; Schwartz, J.; Demokritou, P.; Koutrakis, P. Predictors of Indoor Radon Levels in the Midwest United States. *J. Air Waste Manag. Assoc.* **2021**, *71*, 1515–1528. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Rezaie, F.; Kim, S.W.; Alizadeh, M.; Panahi, M.; Kim, H.; Kim, S.; Lee, J.; Lee, J.; Yoo, J.; Lee, S. Application of Machine Learning Algorithms for Geogenic Radon Potential Mapping in Danyang-Gun, South Korea. *Front. Environ. Sci.* **2021**, *9*, 753028. [\[CrossRef\]](#)
29. Baumann, S.; Petermann, E.; Cinelli, G.; Dehandschutter, B.; Čeliković, I.; Lindner-Leschinski, E.; Bossew, P.; Ciotoli, G.; Gruber, V. Radon Hazard Mapping: Usability of Environmental Predictors Including Atmospheric Radon and Radon Flux and Knowledge Transfer between Regions (Belgium and Germany). *Environ. Earth Sci.* **2025**, *84*, 196. [\[CrossRef\]](#)
30. Park, T.H.; Kang, D.R.; Park, S.H.; Yoon, D.K.; Lee, C.M. Indoor Radon Concentration in Korea Residential Environments. *Environ. Sci. Pollut. Res.* **2018**, *25*, 12678–12685. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Benà, E.; Ciotoli, G.; Petermann, E.; Bossew, P.; Ruggiero, L.; Verdi, L.; Huber, P.; Mori, F.; Mazzoli, C.; Sassi, R. A New Perspective in Radon Risk Assessment: Mapping the Geological Hazard as a First Step to Define the Collective Radon Risk Exposure. *Sci. Total Environ.* **2024**, *912*, 169569. [\[CrossRef\]](#)
32. Kropat, G.; Bochud, F.; Murith, C.; Palacios (Gruson), M.; Baechler, S. Modeling of Geogenic Radon in Switzerland Based on Ordered Logistic Regression. *J. Environ. Radioact.* **2017**, *166*, 376–381. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Martin-Gisbert, L.; Ruano-Ravina, A.; López-Vízcaíno, E.; Barros-Dios, J.; Piñeiro-Lamas, M.; Teijeiro, A.; Casal-Fernández, R.; Kelsey, K.; García, G.; Guerra-Tort, C.; et al. The Galician Radon Map: Determining Indoor Radon Exposure Through Census Tracts. *Indoor Air* **2025**, *2025*, 4176561. [\[CrossRef\]](#)
34. Petermann, E.; Meyer, H.; Nussbaum, M.; Bossew, P. Mapping the Geogenic Radon Potential for Germany by Machine Learning. *Sci. Total Environ.* **2021**, *754*, 142291. [\[CrossRef\]](#)
35. Wu, P.-Y.; Johansson, T.; Mangold, M.; Sandels, C.; Mjörnell, K. Evaluating the Indoor Radon Concentrations in the Swedish Building Stock Using Statistical and Machine Learning. *J. Phys. Conf. Ser.* **2023**, *2654*, 012086. [\[CrossRef\]](#)
36. Wu, P.-Y.; Johansson, T.; Sandels, C.; Mangold, M.; Mjörnell, K. Indoor Radon Interval Prediction in the Swedish Building Stock Using Machine Learning. *Build. Environ.* **2023**, *245*, 110879. [\[CrossRef\]](#)
37. Lovrić, M.; Gajski, G.; Fernández-Agüera, J.; Pöhlker, M.; Gursch, H.; Lovrić, M.; Switters, J.; Borg, A.; Mureddu, F.; Auguštin, D.H.; et al. Evidence Driven Indoor Air Quality Improvement: An Innovative and Interdisciplinary Approach to Improving Indoor Air Quality. *BioFactors* **2025**, *51*, e2126. [\[CrossRef\]](#)
38. Tsapalov, A. Temporal Uncertainty in Rational Method: Reassessment of Data from the Article “Short-Term Temporal Variability of Radon in Finnish Dwellings and the Use of Temporal Correction Factors”. *Open Res. Eur.* **2026**, *5*, 328. [\[CrossRef\]](#)
39. ISO 11665-4:2021; Measurement of Radioactivity in the Environment—Air: Radon-222—Part 4: Integrated Measurement Method for Determining Average Activity Concentration Using Passive Sampling and Delayed Analysis. International Organization for Standardization: Geneva, Switzerland, 2021.
40. EN ISO/IEC 17025:2017; General Requirements for the Competence of Testing and Calibration Laboratories. International Organization for Standardization: Geneva, Switzerland, 2017.
41. European Union’s Copernicus Land Monitoring Service Information Urban Atlas Land Cover/Land Use 2018 (Vector), Europe, 6-Yearly. Available online: <https://sdi.eea.europa.eu/catalogue/copernicus/api/records/fb4dffa1-6ceb-4cc0-8372-1ed354c285e6?language=all> (accessed on 12 January 2026).
42. Pandas Documentation. Pandas 3.0.0 Documentation. Available online: <https://pandas.pydata.org/pandas-docs/stable/index.html> (accessed on 17 February 2026).
43. NumPy Documentation. NumPy v2.4 Manual. Available online: <https://numpy.org/doc/stable/index.html> (accessed on 17 February 2026).
44. SciPy API. SciPy v1.17.0 Manual. Available online: <https://docs.scipy.org/doc/scipy/reference/index.html> (accessed on 17 February 2026).

45. Statsmodels 0.14.6. Available online: <https://www.statsmodels.org/stable/index.html> (accessed on 18 February 2026).
46. Matplotlib Documentation. Matplotlib 3.10.8 Documentation. Available online: <https://matplotlib.org/stable/> (accessed on 17 February 2026).
47. Seaborn: Statistical Data Visualization. Seaborn 0.13.2 Documentation. Available online: <https://seaborn.pydata.org/index.html> (accessed on 17 February 2026).
48. Waskom, M. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [CrossRef]
49. Scikit-Learn. Machine Learning in Python. Scikit-Learn 1.8.0 Documentation. Available online: <https://scikit-learn.org/stable/index.html> (accessed on 17 February 2026).
50. Bossew, P. Radon: Exploring the Log-Normal Mystery. *J. Environ. Radioact.* **2010**, *101*, 826–834. [CrossRef] [PubMed]
51. Hauri, D.D.; Huss, A.; Zimmermann, F.; Kuehni, C.E.; Rössli, M. A Prediction Model for Assessing Residential Radon Concentration in Switzerland. *J. Environ. Radioact.* **2012**, *112*, 83–89. [CrossRef]
52. Stojanovska, Z.; Ivanova, K.; Bossew, P.; Boev, B.; Zunic, Z.; Tsenova, M.; Curguz, Z.; Kolarz, P.; Zdravkovska, M.; Ristova, M. Prediction of Long-Term Indoor Radon Concentration Based on Short-Term Measurements. *Nucl. Technol. Radiat. Prot.* **2017**, *32*, 77–84. [CrossRef]
53. Andersen, C.E.; Raaschou-Nielsen, O.; Andersen, H.P.; Lind, M.; Gravesen, P.; Thomsen, B.L.; Ulbak, K. Prediction of ²²²Rn in Danish Dwellings Using Geology and House Construction Information from Central Databases. *Radiat. Prot. Dosim.* **2007**, *123*, 83–94. [CrossRef] [PubMed]
54. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Jeju, Korea, 13–17 August 2016; pp. 785–794. [CrossRef]
55. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
56. Dicu, T.; Cucos, A.; Botos, M.; Burgehele, B.; Florică, Ș.; Baci, C.; Ștefan, B.; Bălc, R. Exploring Statistical and Machine Learning Techniques to Identify Factors Influencing Indoor Radon Concentration. *Sci. Total Environ.* **2023**, *905*, 167024. [CrossRef]
57. Meinshausen, N. Quantile Regression Forests. *J. Mach. Learn. Res.* **2006**, *7*, 983–999.
58. Petermann, E.; Hoffmann, B. Mapping the Exposure to Outdoor Radon in the German Population. *J. Environ. Radioact.* **2025**, *281*, 107583. [CrossRef]
59. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Int. Jt. Conf. Artif. Intell.* **1995**, *14*, 1137–1145.
60. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2*, 249–262. [CrossRef] [PubMed]
61. Daviran, M.; Masoumi, I.; Ghezelbash, R.; Maggio, S.; De Iaco, S. Machine Learning-Based Mapping of Indoor Radon Potential Using Geogenic Factors. *Environ. Ecol. Stat.* **2025**, *32*, 893–928. [CrossRef]
62. Molnar, C. *Interpretable Machine Learning*, 3rd ed.; Lulu Press: Morrisville, NC, USA, 2025; ISBN 978-3-911578-03-5. Available online: <https://christophm.github.io/interpretable-ml-book> (accessed on 17 February 2026).
63. Welcome to the SHAP Documentation—SHAP Latest Documentation. Available online: <https://shap.readthedocs.io/en/latest/> (accessed on 17 February 2026).
64. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U., Von Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
65. Kropat, G.; Bochud, F.; Jaboyedoff, M.; Laedermann, J.P.; Murith, C.; Palacios, M.; Baechler, S. Major Influencing Factors of Indoor Radon Concentrations in Switzerland. *J. Environ. Radioact.* **2014**, *129*, 7–22. [CrossRef]
66. Kropat, G.; Bochud, F.; Jaboyedoff, M.; Laedermann, J.P.; Murith, C.; Palacios, M.; Baechler, S. Improved Predictive Mapping of Indoor Radon Concentrations Using Ensemble Regression Trees Based on Automatic Clustering of Geological Units. *J. Environ. Radioact.* **2015**, *147*, 51–62. [CrossRef]
67. *JSGM 100*; Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement. International Organization for Standardization: Geneva, Switzerland, 2008; Volume 50, p. 134.
68. Karmoude, M.; Munhungewarwa, B.; Chiraira, I.; McKenzie, R.; Kong, J.; Smith, B.; Ayana, G.; Njara, N.; Mathaha, T.; Kumar, M.; et al. Machine Learning for Air Quality Prediction and Data Analysis: Review on Recent Advancements, Challenges, and Outlooks. *Sci. Total Environ.* **2025**, *1002*, 180593. [CrossRef] [PubMed]
69. Méndez, M.; Merayo, M.G.; Núñez, M. Machine Learning Algorithms to Forecast Air Quality: A Survey. *Artif. Intell. Rev.* **2023**, *56*, 10031–10066. [CrossRef]

70. Hahn, E.J.; Haneberg, W.C.; Stanifer, S.R.; Rademacher, K.; Backus, J.; Rayens, M.K. Geologic, Seasonal, and Atmospheric Predictors of Indoor Home Radon Values. *Environ. Res. Health* **2023**, *1*, 025011. [[CrossRef](#)] [[PubMed](#)]
71. Cucu, M.; Dupleac, D. The Impact of Ventilation Rate on Radon Concentration Inside High-Rise Apartment Buildings. *Radiat. Prot. Dosim.* **2022**, *198*, 290–298. [[CrossRef](#)] [[PubMed](#)]

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