DOI: https://doi.org/10.68732/fc791edc1b8b4843

ISBN: 978-9989-2053-7-8

Application of Machine Learning techniques for estimating sinkhole parameters caused by underground mining

Alicja Szmigiel
Strata Mechanics Research Institute Polish Academy of Sciences, Kraków, Poland
alicja.szmigiel@imgpan.pl
0000-0001-6770-0421

Krzysztof Tajduś
Strata Mechanics Research Institute Polish Academy of Sciences, AGH University of Kraków, Kraków, Poland
0000-0003-2014-0900

Janusz Rusek
AGH University of Kraków, Kraków, Poland
0000-0003-0368-2580

Abstract

Sinkholes induced by underground mining activities create a significant risks to infrastructure, the environment, and human safety. Reliable prediction of sinkhole development and its size parameters is therefore crucial for effective risk assessment and land-use planning in mining regions. This study explores the applicability of machine learning techniques for predicting the geometric characteristics of sinkholes formed above abandoned and active mining excavations. A comprehensive database was compiled from several European countries, integrating geological and mining information, including the parameters of underground workings and recorded sinkhole dimensions. Multiple supervised learning algorithms were trained and evaluated to predict sinkhole size parameters. Model performance was assessed using standard statistical metrics to identify the most accurate predictive approach. Feature importance analysis was conducted to determine which parameters most strongly influence sinkhole size predictions. The findings contribute to advancing data-driven methods for hazard assessment in mining areas and provide a foundation for developing early warning and risk-mitigation tools supporting sustainable land management in regions affected by underground mining.

Keywords

Machine learning, Ground deformation, Risk prediction, Post-mining areas, Sinkholes.

1 Introduction

Sinkholes are among the most disruptive geohazards because they can develop suddenly and cause severe damage to buildings, infrastructure, and land use. Although sinkholes are often associated with natural karst terrains, it is well established that underground mining can also generate collapse features, sometimes long after extraction has ended and often in densely urbanized areas (Waltham et al. 2005; Gutiérrez et al. 2014). Case histories from post-mining regions show recurring mechanisms involving underground voids, progressive deterioration of workings, and water-driven internal erosion. Examples from Poland include reactivation of shallow or poorly documented excavations, where rainfall infiltration or leakage can promote suffusion and localized collapse (Konieczny and Słowik 2019; Roman Ścigała et al. 2023). Comparable post-mining instability has also been described in France and Germany, reinforcing that mining-related sinkholes are a long-lasting hazard controlled by excavation geometry, rock-mass properties, and groundwater conditions (Didier 2009; Kretschmann et al. 2017).

Predicting where and how large a sinkhole will form remains difficult because controlling factors are numerous and often poorly constrained. Traditional approaches, empirical relations, analytical formulas, numerical geomechanical modelling, and site investigations, are essential for understanding local mechanisms and designing mitigation, but they are often challenging to apply efficiently over large post-mining areas with heterogeneous conditions and incomplete documentation (Szmigiel and Tajduś 2025). In parallel, machine learning (ML) has increasingly been used in sinkhole studies, mainly for susceptibility mapping and classification. Random Forest, Artificial Neural Networks models have been applied to LiDAR/DEM, hydrogeological variables, and InSAR-derived deformation to identify sinkhole-prone areas and relate sinkhole occurrence to morphometric, lithological, and groundwater-related predictors (Zhu and Pierskalla 2016; Taheri et al. 2019; Bianchini et al. 2022). More recently, gradient-boosting tree ensembles have gained popularity and have begun to appear in sinkhole susceptibility studies (Bilgilioğlu et al. 2025).However, most applications focus on natural karst sinkholes, and comparatively few studies address mine-induced sinkholes or attempt to predict the geometry of individual collapse events.

This study addresses that gap by testing whether supervised ML models can predict two key sinkhole dimensions: depth (Dz) and length/diameter (Lz), from mining and overburden parameters. The analysis uses a database of sinkholes from two well-documented European post-mining regions. We compared a set of regression models representing both ensemble and neural approaches: Random Forest (RF), Artificial Neural Networks (ANN), and two gradient-boosting methods, XGBoost and CatBoost. The objectives are to (i) assess how accurately sinkhole dimensions can be predicted from available mining and geological data, (ii) identify the most influential parameters, and (iii) examine differences in predictability between vertical and lateral sinkhole dimensions. By focusing on mining-induced sinkholes this work contributes to post-mining hazard assessment and supports the development of practical tools for risk evaluation above abandoned underground workings.

2 Database source, decryption and processing

The database used in this study combines sinkhole cases from two well-documented European post-mining regions: the Upper Silesian Coal Basin in southern Poland (885 cases) and former coal basins in central France (123 cases). In Poland, the dataset includes cases from areas affected by historical hard-coal mining, particularly around Jan Kanty, Siersza, and Kazimierz-Juliusz mines, where shallow and abandoned workings have contributed to long-term post-mining instability. In France, the cases come from former coal basins such as Messeix and Ahun, where localized collapses are recognized as important post-mining hazards.

2.1 Database structure

The compiled database integrates information on sinkhole location, geometry, mining context, overburden conditions, and surface effects. For each case, available attributes include coordinates, mine identification, sinkhole dimensions, date of occurrence, mining depth and thickness, duration of exploitation, selected geological indicators, and simplified hydrogeological information such as aquifer presence. Because the data were assembled from different sources and time periods, not all variables were consistently available. Therefore, before model development, a feature-selection stage was carried out to retain parameters that were sufficiently complete, informative, and physically relevant to sinkhole formation. The final predictor set focused mainly on mining depth, mining thickness, exploitation duration, breaks in exploitation, overburden lithology, country, and aquifer presence.

2.2 Handling missing values

Missing values were present in several parameters that are essential for this study, especially mining depth, mining thickness, sinkhole depth, and sinkhole length/diameter. A K-nearest-neighbour (KNN) imputation approach was adopted because it allows missing values to be estimated from the most similar available cases. This was particularly important here, as the dataset combines cases from different mining and geological environments. The method was applied in a two-step procedure, tailored to the nature of the missing variables.In the first step, missing mining parameterswere imputed using spatial similarity. This was based on the assumption that sinkholes located close to one another are likely to be related to the same or similar exploitation field. In the second step, missing sinkhole depth and length/diameter were estimated using similarity in geological and mining feature space, including variables such as mining parameters, lithology of the overburden, and aquifer presence. To support model robustness and avoid bias the imputation workflow was accompanied by distribution checks. Specifically, we compared summary statistics (e.g., quartiles and extrema) before and after imputation to confirm that the overall range, including extreme values, was not artificially distorted. In practice, using the imputed dataset also led to more stable and improved predictive performance than analyses restricted to complete cases.

3 Machine Learning training and evaluation

3.1 Machine learning models

Four supervised learning algorithms were tested in this study: Random Forest (RF), Artificial Neural Networks (ANN), and two gradient-boosting tree methods, XGBoost and CatBoost. These models were selected because they represent complementary approaches that are widely used for predictive modelling on engineering datasets. RF is well suited to heterogeneous and nonlinear datasets (Kumar Singh and Chakravarty 2023), ANN can capture more advanced nonlinear patterns (Gharbi et al. 2025; Kim and Lee 2020; Le et al. 2024) and gradient boosting is a strong family of methods for structured/tabular data. XGBoost is a scalable implementation of tree boosting that has demonstrated high performance across many applied prediction tasks (Chen and Guestrin 2016), while CatBoost introduces algorithmic strategies (including ordered boosting) designed to improve robustness and reduce prediction shift (Prokhorenkova et al. 2018).

3.2 Model training and hyperparameter tuning

Input variables were aligned with the selected predictors, and binary features were encoded as 0/1. The dataset was then split into training (70%) and testing (30%) subsets using a fixed random seed, ensuring an independent hold-out test set for final evaluation. A consistent workflow was applied across all models. For the ANN, multiple network configurations were explored (number of hidden layers, neurons per layer, and activation functions), and dropout and early stopping were used to limit overfitting. For the tree-based models (RF and gradient boosting), hyperparameters were optimized using a systematic search combined with K-fold cross-validation, providing a more stable estimate of generalization than a single split and enabling transparent, reproducible model selection. After tuning, the best-performing configuration for each target variable was refit on the full training set and evaluated on the independent test subset.

3.3 Performance metrics and feature importance

Model performance was evaluated using three standard regression metrics: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). MAE and RMSE quantify the prediction error in the same units as the target variable, while R^2 indicates how much of the variability in the observed data is explained by the model.

To better understand which input variables influenced the predictions most strongly, permutation feature importance was applied. In this approach, each predictor is randomly shuffled one at a time, and the corresponding drop in model performance is measured. Larger decreases indicate stronger dependence of the model on that variable. However, the results should be interpreted as relative model dependence rather than direct causal influence.

4 Results

Test-set performance metrics for sinkhole depth (Dz) and length/diameter (Lz) are summarized in Table 1 and Table 2, respectively. Overall, the best performance was achieved by gradient-boosting tree ensembles, with CatBoost providing the strongest results for Dz and XGBoost performing best for Lz, while RF and ANN remained as baselines.

Table 1. Performance metrics for Dz prediction (test set).

	RF	ANN	XGBoost	CatBoost
MAE	1.36	1.48	1.51	1.31
RMSE	2.29	2.35	2.63	1.99
R^2	0.47	0.44	0.29	**0.66**

Table 2. Performance metrics for Lz prediction (test set).

	RF	ANN	XGBoost	CatBoost
MAE	3.31	3.70	3.36	3.15
RMSE	5.00	5.62	4.75	4.71
R^2	0.66	0.57	**0.69**	0.32

For Dz, the strongest test-set performance was achieved by CatBoost, with MAE = 1.31 m, RMSE = 1.99 m, and R^2 = 0.66 (Table 1), indicating substantially improved skill in reproducing depth variability compared with the other models. Random Forest and ANN showed moderate performance for Dz (RF: R^2 = 0.47; ANN: R^2 = 0.44), while XGBoost performed noticeably worse for this target (R^2 = 0.29).

For Lz, the best results were obtained with XGBoost (MAE = 3.36 m, RMSE = 4.75 m, R^2 = 0.69 (Table 2), closely followed by Random Forest (R^2 = 0.66). The ANN achieved useful predictive skill (R^2 = 0.57), whereas CatBoost generalized poorly for Lz (R^2 = 0.32). Given that the boosting models delivered the highest explained variance on the independent test set (CatBoost for Dz and XGBoost for Lz), the detailed result interpretation in the following section focuses on these models.

Figure 1 summarizes the test-set predictions obtained with CatBoost for sinkhole depth (Dz). The model reproduces the dominant population of shallow sinkholes well, with predicted values closely tracking the main band of observations. As expected for a heterogeneous dataset with few very deep events, performance decreases toward the upper values and the largest depths tend to be underestimated.
Figure 2 presents the corresponding test-set predictions obtained with XGBoost for sinkhole length/diameter (Lz).

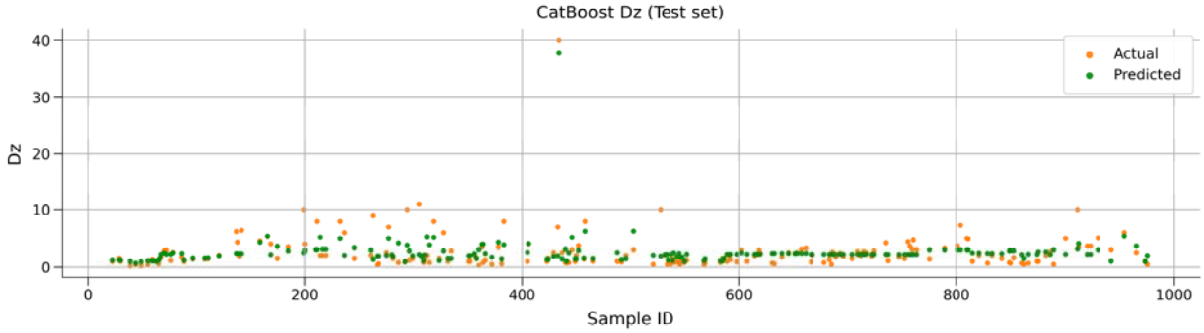


Fig. 1 Observed vs. predicted sinkhole dimensions (Dz) using the CatBoostmodel

The model captures the overall level and variability of the common small-to-moderate sinkholes, while the rare largest events are again smoothed and typically underpredicted. Importantly from a hazard perspective, high-magnitude cases for both Dz and Lz still receive elevated predicted values relative to the population, meaning that even when the exact size is not matched, the model continues to flag potentially large sinkholes as high events.

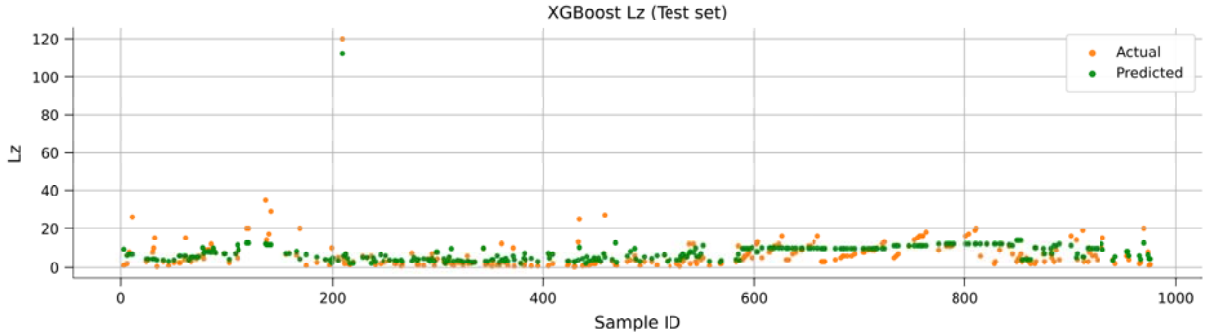


Fig. 2 Observed vs. predicted sinkhole dimensions (Lz) using the XGBoost model.

Figure 3 and 4 summarizes the feature-importance patterns for the best-performing boosting models. CatBoost assigns dominant importance to the mining geometry parameters, with working thickness and excavation depth clearly ranking highest, followed by a smaller contribution from exploitation duration and a modest influence of aquifer presence. The remaining predictors, including lithology indicators and country/break variables, contribute little additional explanatory power. This ranking is consistent with geotechnical expectations that void geometry is a dominant control on collapse propagation and the resulting depth of sinkholesFor Lz, the XGBoost importance distribution shifts toward overburden characteristics. The strongest contributions are associated with lithology indicators, particularly puddingstone, sand, and clay, with additional influence from the break in exploitation indicator and aquifer presence. Mining depth and thickness remain relevant but are less dominant than in the Dz model, suggesting that while mining geometry provides the fundamental precondition for collapse, the lateral extent of sinkhole expression is more sensitive to material properties and site-specific conditions.

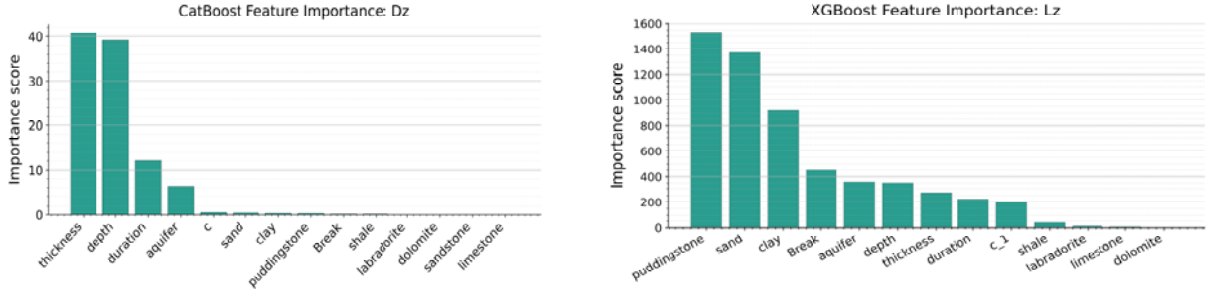


Fig. 3 and 4 Feature importance for the best-performing boosting models: CatBoost (Dz) and XGBoost (Lz).

5 Discussion and conclusions

This study evaluated supervised machine-learning models for predicting two key geometric parameters of mining-induced sinkholes: depth (Dz) and length/diameter (Lz), using mining parameters together with geological and hydrogeological descriptors. The expanded model comparison shows that no single algorithm was optimal for both targets. Instead, performance depended on the predicted parameter: CatBoost provided the best results for Dz (test $R^2 = 0.66$), while XGBoost performed best for Lz (test $R^2 = 0.69$). Random Forest remained a strong and stable baseline (Dz: $R^2 = 0.47$, Lz: $R^2 = 0.66$) and ANN achieved the lowest predictive performance. Overall, these results confirm that modern ensemble methods can recover meaningful nonlinear relationships between mining/overburden conditions and sinkhole geometry, although a non-negligible fraction of variance remains unexplained.The test-set plots for CatBoost (Dz) and XGBoost (Lz) show good agreement for the common small-to-moderate cases, while the rare extremes are typically underestimated. Importantly, the largest events are still assigned noticeably higher predicted values than the background population, so they are flagged as high-magnitude cases even when the exact size is not matched. The feature-importance patterns are physically interpretable and align with established geomechanical understanding, while also highlighting that different geometric dimensions respond to different controls.

Despite encouraging performance, the models should be viewed as first-step predictive tools, suitable for preliminary assessment rather than as decision systems. Current limitations are driven primarily by data constraints: heterogeneous site conditions, incomplete descriptors of structural geology and hydrogeological dynamics, and limited representation of extreme events. In particular, because the database integrates multiple post-mining settings, model performance and feature-importance rankings should be interpreted as dataset-level trends, and transfer to a new mining environment may require

recalibration or the inclusion of additional site-specific descriptors.Future work should prioritize (i) further expansion and standardization of the sinkhole database across additional mining districts, (ii) enriching inputs with higher quality geological and water-related variables, and (iii) incorporating uncertainty quantification (e.g., prediction intervals) to better communicate confidence in predicted sinkhole sizes. Methodologically, hybrid strategies, such as stacked ensembles, target-specific model selection, and multi-task learning, may further improve performance. Overall, this study demonstrates that boosted tree ensembles can provide accurate and interpretable predictions of mining-induced sinkhole geometry, while clearly indicating where improved data coverage and modelling strategies are needed to move toward more reliable predictions.

Acknowledgment

This research was funded by RESEARCH FUND FOR COAL AND STEEL (RFCS), grant number 101157400 and by the POLISH MINISTRY OF EDUCATION AND SCIENCE, grant number 5997/ FBWiS/2024/2 (no W 109/FBWiS/2024) under the project SIRIMA "Sinkhole hazard and risk management in post-mining areas". The authors gratefully acknowledge the funding provided.

References

Bianchini S, Confuorto P, Intrieri E, et al (2022) Machine learning for sinkhole risk mapping in Guidonia-Bagni di Tivoli plain (Rome), Italy. Geocarto Int 37:16687–16715. https://doi.org/10.1080/10106049.2022.2113455

Bilgilioğlu SS, Gezgin C, Iban MC, et al (2025) Explainable Sinkhole Susceptibility Mapping Using Machine-Learning-Based SHAP: Quantifying and Comparing the Effects of Contributing Factors in Konya, Türkiye. Applied Sciences 15:3139. https://doi.org/10.3390/app15063139

Chen T, Guestrin C (2016) XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp 785–794

Didier C (2009) Postmining Management in France: Situation and Perspectives. Risk Analysis 29:1347–1354. https://doi.org/10.1111/j.1539-6924.2009.01258.x

GHARBI A, AYARI M, Touati Y El, et al (2025) Prediction of Mining-Induced Subsidence in Saudi Arabia Phosphate Mines Using ANN Method. International Journal of Advanced Computer Science and Applications 16:. https://doi.org/10.14569/IJACSA.2025.0160852

Kim Y, Lee SS (2020) Application of Artificial Neural Networks in Assessing Mining Subsidence Risk. Applied Sciences 10:1302. https://doi.org/10.3390/app10041302

Konieczny K, Słowik L (2019) An Example of Failure of an Office Building in Upper Silesia. MATEC Web Conf 284:3003. https://doi.org/10.1051/matecconf/201928403003

Kretschmann J, Efremenkov AB, Khoreshok AA (2017) From Mining to Post-Mining: The Sustainable Development Strategy of the German Hard Coal Mining Industry. IOP Conf Ser Earth Environ Sci 50:012024. https://doi.org/10.1088/1755-1315/50/1/012024

Kumar Singh S, Chakravarty D (2023) Efficient and Reliable Prediction of Dump Slope Stability in Mines using Machine Learning: An in-depth Feature Importance Analysis. Archives of Mining Sciences. https://doi.org/10.24425/ams.2023.148157

LE TTT, NGUYEN TG, PHAM C Van, et al (2024) Prediction of Underground Mine's Surface Subsidence using a Recursive Multi-Step Forecasting Model with an Artificial Neural Network. InżynieriaMineralna 2:. https://doi.org/10.29227/IM-2024-01-116

Prokhorenkova L, Gusev G, Vorobev A, et al (2018) CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, pp 6639–6649

Roman Ścigała, Katarzyna Szafulera, Marek Jendryś (2023) Assessment of sinkhole hazard in the post-mining area using the ERT method and numerical modeling. 75 Scientific Journals of the Maritime University of Szczecin 147:. https://doi.org/10.17402/570

Szmigiel A, Apel DB (2022) Predicting the stability of open stopes using Machine Learning. Journal of Sustainable Mining 21:241–248. https://doi.org/10.46873/2300-3960.1369

Szmigiel A, Tajduś K (2025) A Comprehensive Review of Sinkhole Modeling and Prediction Methods. Archives of Mining Sciences 293–312. https://doi.org/10.24425/ams.2025.154664

Taheri K, Shahabi H, Chapi K, et al (2019) Sinkhole susceptibility mapping: A comparison between Bayes-based machine learning algorithms. Land Degrad Dev 30:730–745. https://doi.org/10.1002/ldr.3255

Zhu J, Pierskalla WP (2016) Applying a weighted random forests method to extract karst sinkholes from LiDAR data. J Hydrol (Amst) 533:343–352. https://doi.org/10.1016/j.jhydrol.2015.12.012