Interpretable machine learning for digital soil pH mapping using an optimized AdaBoost algorithm

Zakiyyan Zain Alkaf[1], A'isya Nur Aulia Yusuf [2], Elsa Sari Hayunah Nurdiniyah[2], Tri Wisudawati[1]
[1]Department of Industrial Engineering and Mechanical Engineering, Faculty of Engineering, Universitas Jenderal Soedirman, Purwokerto, Indonesia
[2]Department of Electrical Engineering, Faculty of Engineering, Universitas Jenderal Soedirman, Purwokerto, Indonesia

Article Info

Article history:

Received Dec 17, 2025
Revised Mar 30, 2026
Accepted May 25, 2026

Keywords:

AdaBoost regressor
Bayesian optimization
Digital soil mapping
Machine learning
Precision agriculture
Shapley Additive exPlanations
Soil pH

ABSTRACT

Soil pH is a fundamental parameter determining nutrient availability, microbial activity, and crop productivity. Unlike previous studies that often prioritize prediction accuracy over explainability, this study proposes an interpretable machine-learning framework integrating hyperparameter-optimized adaptive boosting (AdaBoost) with Shapley Additive exPlanations (SHAP) to unravel the spatial drivers of soil pH. A systematic workflow was implemented to evaluate a diverse set of algorithms, followed by Bayesian optimization to fine-tune the best-performing models. The results demonstrated that the optimized AdaBoost model yielded the largest performance improvement (~7.5%), achieving excellent accuracy on independent test data with a coefficient of determination (R^2) of 0.817 and a mean absolute error (MAE) of 0.293. Furthermore, SHAP analysis identified iron (Fe) and calcium carbonate ($CaCO_3$) as the most influential predictors, revealing that Fe exhibits a strong inverse relationship with pH, while $CaCO_3$ shows a positive association. This framework successfully balances high predictive accuracy with pedological interpretability, offering a robust tool for digital soil mapping and precision agriculture.

Corresponding Author:

Zakiyyan Zain Alkaf
Department of Industrial Engineering and Mechanical Engineering, Faculty of Engineering
Universitas Jenderal Soedirman
Purwokerto, Indonesia
Email: Zakiyyan.alkaf@unsoed.ac.id

1. INTRODUCTION

Soil pH is a fundamental parameter determining soil quality, regulating nutrient bioavailability, microbial activity, and overall crop productivity [1]. To ensure optimal plant health, macronutrients (nitrogen (N), phosphorus (P), and potassium (K)) must remain available while maintaining micronutrients (iron (Fe), manganese (Mn), and zinc (Zn)) at non-toxic levels [2]. Consequently, understanding the spatial distribution of soil pH is critical for environmental monitoring and precision agriculture, enabling site-specific management strategies such as variable-rate liming that minimize environmental risks like groundwater pollution [3]. Accurate digital soil mapping serves as the foundational layer for internet of things (IoT)-based smart farming systems, enabling variable-rate technology (VRT) to adjust inputs precisely based on the spatial distribution of soil properties.

While traditional wet chemistry methods are accurate, they are labor-intensive and costly for large-scale mapping. Alternatively, machine learning (ML) has emerged as a powerful tool for digital soil mapping, with ensemble methods such as random forests (RF), gradient boosting, and support vector regression (SVR) demonstrating the ability to capture complex nonlinear relationships between environmental covariates and soil pH [4], [5] ML optimization has proven effective in various domains, ranging from industrial resource allocation [6] to consumer preference analysis in e-commerce markets [7]. However, existing studies often suffer from two major limitations: the use of default hyperparameters, which leads to suboptimal model performance, and a “black-box” nature that lacks interpretability. Specifically, few studies have proposed optimized ensemble models integrated with interpretability tools (e.g., Shapley Additive exPlanations (SHAP)) for soil pH mapping, creating a gap in understanding the pedological drivers behind prediction [8]-[10].

The primary novelty of this study lies in the synergistic integration of Bayesian-optimized adaptive boosting (AdaBoost) with SHAP-based interpretability. Unlike standard ensemble approaches, this framework not only improves predictive accuracy but also deciphers the ‘black-box’ logic to reveal critical pedological drivers. First, we implement a rigorous preprocessing and feature selection pipeline to ensure high data quality. Second, we systematically evaluate diverse regression algorithms, optimizing the top performers using Bayesian optimization to enhance predictive accuracy. Finally, we employ SHAP to interpret the model outputs. This study holds significant practical relevance for precision agriculture while theoretically bridging the gap between ML optimization and explainable artificial intelligence (XAI) in soil science.

2. METHOD

This study adopts a comprehensive end-to-end ML workflow encompassing data acquisition, preprocessing, feature selection, model benchmarking, hyperparameter tuning via Bayesian optimization, and SHAP-based interpretability analysis to identify key soil pH drivers.

2.1. Dataset description

The analysis utilizes a dataset of 781 soil samples collected from agricultural fields in the Grevena region, Northern Greece, between 2015 and 2019. The samples were taken from a depth of 0–30 cm, covering a wide range of soil textures and chemical properties typical of Mediterranean agricultural zones. The dataset was obtained from the Soil and Water Research Institute (SWRI) and includes key parameters such as pH, organic matter, and micro-nutrients, ensuring sufficient variability for training robust ML models [11]. The predictor set includes electrical conductivity (EC), organic matter (OM), nitrate nitrogen (N-NO_3), P, K, magnesium (Mg), and Fe. Additionally, the dataset incorporates micronutrients including Zn, Mn, copper (Cu), and boron (B), as well as calcium carbonate ($CaCO_3$), which is explicitly included due to its critical role in pH buffering capacity. Soil texture is represented by sand, silt, and clay percentages.

2.2. Research workflow and preprocessing

To ensure reproducibility, the workflow was structured using a Scikit-learn Pipeline (Figure 1). The dataset was partitioned into training (80%) and testing (20%) sets for unbiased evaluation [12]. Missing values were handled using median imputation to accommodate soil heterogeneity, while outliers were capped using the interquartile range (IQR) method based solely on training set statistics to prevent data leakage. Winsorization was subsequently applied at the 1st and 99th percentiles, followed by standard scaling to improve model convergence [13]. All experiments were conducted in Google Colab (Python 3) using Scikit-learn with a fixed random seed of 42 to guarantee reproducibility. Complex non-linear transformations were deliberately avoided, as tree-based ensemble models are invariant to monotonic feature scaling, thereby preserving the physical interpretability of soil measurement units.

2.3. Feature selection

Feature selection was applied to reduce dimensionality and mitigate overfitting [14]. A RandomForestRegressor was trained on preprocessed data, retaining only features with importance scores exceeding the mean threshold [15]. This single-pass approach efficiently captures non-linear feature interactions without the computational overhead of wrapper methods such as recursive feature elimination (RFE), which requires repeated model retraining for each feature subset [16], [17].

2.4. Model development and evaluation

To identify the most effective predictive strategy for soil pH, we benchmarked nine regression algorithms spanning distinct learning paradigms: linear approaches (linear, ridge, and lasso regression), tree-based ensembles (decision tree (DT), RF, gradient boosting, and AdaBoost), and kernel/instance-based methods (SVR and K-nearest neighbors (KNN)). A unified preprocessing pipeline was applied identically across all models. Although tree-based models are invariant to feature scaling, distance-based models such as SVR and KNN require standardized inputs; therefore, all datasets underwent Winsorization and standard scaling prior to training to ensure a consistent evaluation baseline.

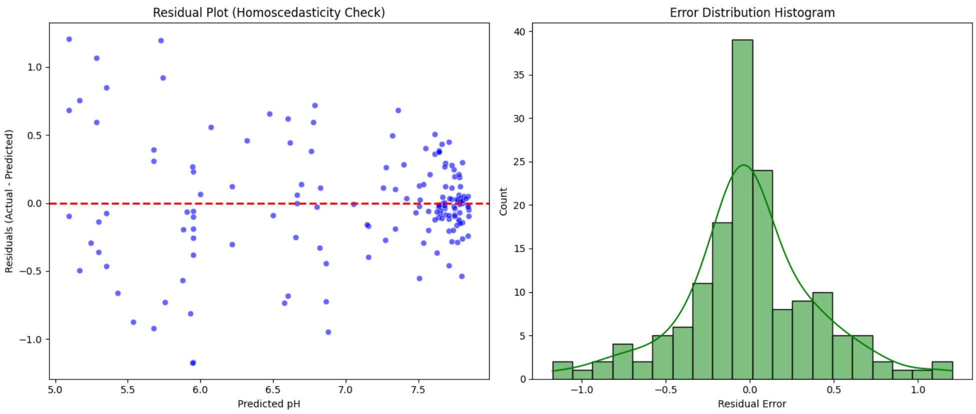

Figure 1. Residual scatter plot and error histogram

Model generalization was assessed using 5-fold cross-validation (CV) on the training set [18]. Performance was quantified using the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE), reported with 95% confidence intervals (CI) [19]–[21]. Hyperparameter tuning was conducted via Bayesian optimization, which efficiently navigates high-dimensional search spaces compared to exhaustive grid search [22], within the ranges defined in Table 1. The optimized model was subsequently interpreted using SHAP, which quantifies each feature's marginal contribution to predictions, providing both global feature rankings and local instance-level explanations grounded in cooperative game theory [23].

Table 1. Result of prediction

Method	Mean R-squared (CV)	MAE (CV)	MSE (CV)
SVR	0.7971	0.275	0.1567
AdaBoost regressor	0.7923	0.29	0.1631
Gradient boosting regressor	0.7898	0.2778	0.164
KNN regressor	0.7709	0.2973	0.1773
RF regressor	0.762	0.2958	0.1859
Linear regression	0.7598	0.304	0.1863
Ridge	0.7598	0.3041	0.1863
DT regressor	0.6387	0.374	0.288
Lasso	-0.0081	0.7181	0.7774

3. RESULTS AND DISCUSSION

A comprehensive evaluation of nine ML algorithms revealed substantial performance variation, providing important insights into the complexity of the relationship between soil properties and pH (Table 1). To ensure statistical robustness, model generalization was assessed using default cross-validation, with performance metrics reported as mean ± 95% CI. The SVR model performed best with $R^2 = 0.7971$, MAE = 0.275, and MSE = 0.1567, indicating superior ability to capture the complexity of the relationship between soil features and pH. The superiority of SVR stems from its ability to map data to a high-dimensional space via kernel functions, enabling the modeling of complex nonlinear relationships without explicitly computing transformations [24]. In the context of soil systems, many pedogenic and biogeochemical processes that influence pH are nonlinear and involve multifaceted interactions such as buffering capacity, which is influenced by interactions between clay minerals, organic matter, and base cations [25]. SVR can capture this complexity more effectively than linear models. The ensemble models AdaBoost regressor ($R^2 = 0.7923$), gradient boosting regressor ($R^2 = 0.7898$), and RF regressor ($R^2 = 0.7620$) also performed very well, consistent with the literature demonstrating the effectiveness of ensemble methods in soil property modeling [26], [27]. Linear models — linear regression, ridge, and lasso showed significantly lower performance ($R^2 \approx 0.76$), with lasso even producing a negative R^2 (-0.0081), indicating it performed worse than predictions based on the target average. This suggests that soil pH prediction cannot be reduced to a simple linear combination of predictor features.

Table 2 presents the comparative performance results of the three best models after hyperparameter optimization. All three models showed significant performance improvements. SVR improved marginally (~0.9%), while AdaBoost achieved the largest gain (~7.5%), improving MSE from 0.1631 to 0.1509. The alignment confirms that the model generalizes well to unseen data and does not exhibit significant overfitting. AdaBoost's performance can be attributed to its sequential learning architecture [28]. Unlike RF, which is robust to default settings, AdaBoost is highly sensitive to the learning rate and estimator count; Bayesian optimization allowed the algorithm to precisely calibrate the step size for correcting predecessor errors, effectively reducing bias without incurring high variance. gradient boosting also showed improvement (~4.7%) but required tuning a more complex hyperparameter space [29]. SVR with an radial basis function (RBF) kernel has a relatively compact hyperparameter space (C, gamma, epsilon) compared to complex ensemble methods, and the default values in scikit-learn may already be close to optimal for this dataset [30].

Table 2. Result of Bayesian optimization

Model	MSE negative (CV)
SVR	-0.1553
AdaBoost regressor	-0.1509
Gradient boosting regressor	-0.1563

Table 3 presents the results of the optimized AdaBoost Regressor on independent test data. The final model achieved an R^2 of 0.8172, indicating that it explained approximately 81.72% of the total variability in soil pH. These results are competitive with recent soil spectroscopy studies [31], which typically report R^2 values in the range of 0.8–0.85, thereby validating the efficacy of our proposed framework. Our model's R^2 of 0.817 notably exceeds typical benchmarks in digital soil pH mapping, which are often in the range of 0.70 in similar pedological studies [32], validating the efficacy of the optimization framework. Although SVR performed well initially, the optimized AdaBoost regressor was selected as the final model because it achieved the lowest (MSE of 0.1509) compared to SVR (0.1553) during the rigorous Bayesian optimization process. From a computational perspective, the optimized AdaBoost model is structurally simpler and more lightweight compared to deep learning (DL) models or heavy ensemble stacks. This low computational complexity makes it highly suitable for deployment on resource-constrained IoT edge devices, enabling real-time on-site soil analysis without relying on cloud connectivity. To further assess reliability, a detailed analysis of the prediction errors reveals that while the model achieves high accuracy across the majority of the pH range, minor deviations are observed at the extreme ends of the distribution. These discrepancies likely stem from local soil heterogeneity or the inherent limitation of regression models in extrapolating beyond the dense regions of the training data. Overall, the metrics exceed the threshold for "excellent prediction" ($R^2 > 0.70$) accepted in digital soil mapping [33].

Table 3. Evaluation of optimized AdaBoost regressor

Metric	Value
R-squared	0.8172
MAE	0.2930
MSE	0.1668

As shown in Figure 1, the residuals are randomly distributed around the horizontal zero line without forming any discernible patterns (e.g., U-shape or funnel shape), which confirms that the model is unbiased across the entire range of soil pH values. Additionally, the error distribution histogram in Figure 1 approximates a Gaussian bell curve, indicating that the prediction errors are consistent and predictable, satisfying the normality assumption required for robust regression analysis.

While SVR demonstrated competitive accuracy, AdaBoost was prioritized due to its inherent compatibility with tree-based feature importance analysis, offering superior transparency over the 'black-box' kernel transformations of SVR. This interpretability is crucial for identifying key soil drivers in precision agriculture. The relative influence of each predictor on soil pH is illustrated in Figure 2 via the SHAP summary plot for the optimized AdaBoost regressor. A standout observation from this visualization is the dominance of Fe as the primary driver of the model's predictions. The distribution of SHAP values for Fe shows a distinctive pattern: high Fe values (red/pink points) are consistently located on the left side (negative SHAP values), indicating that high Fe concentrations are strongly associated with lower (more acidic) predicted pH. Low Fe values (blue points) are located on the right side (positive or near-zero SHAP values), indicating that low Fe is associated with higher (less acidic or neutral) pH. This inverse relationship between Fe and pH is highly consistent with the principles of soil chemistry [34]. $CaCO_3$ emerged as the second most important feature and aligns perfectly with soil chemistry principles. The pattern of SHAP values shows that high $CaCO_3$ values (red/pink points) are predominantly on the right (positive SHAP values), indicating a strong positive association with high pH. Low $CaCO_3$ values (blue points) are on the left side (negative SHAP values), associated with low pH. The positive relationship between $CaCO_3$ and pH is a cornerstone of soil chemistry [32]. Agronomically, these findings suggest that farmers can utilize SHAP-derived insights to create targeted management zones, prioritizing soil testing and liming in areas with high Fe variability and low $CaCO_3$ levels to neutralize acidity and optimize nutrient uptake.

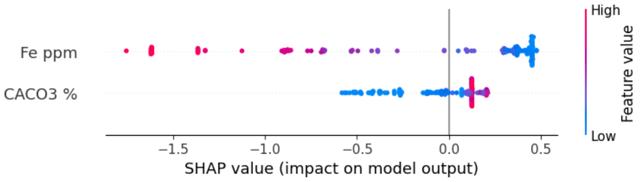

Figure 2. SHAP analysis

4. CONCLUSION

This study demonstrated that nonlinear models consistently outperform linear approaches for digital soil pH mapping. Specifically, the optimized AdaBoost regressor achieved superior predictive accuracy compared to traditional linear baselines, highlighting the complexity of soil-landscape interactions. Methodologically, the integration of Bayesian optimization significantly enhanced model performance, yielding a ~7.5% improvement in MSE for the AdaBoost model. Furthermore, the application of SHAP analysis provided critical transparency, identifying Fe and $CaCO_3$ as the dominant drivers of soil pH spatial variability. The result confirms the framework's ability to produce not only accurate maps but also agronomically meaningful insights. The findings have practical implications for precision agriculture, offering a robust tool for optimizing soil amendment strategies. Current limitations include the reliance on a dataset from a specific geographic region with a moderate sample size and the lack of explicit uncertainty quantification. Future work should focus on integrating satellite-based remote sensing data to improve generalization and deploying the lightweight AdaBoost model into IoT-based edge devices for real-time, in-situ soil monitoring.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the people who contributed their time and effort to help us continue with our research. We also thank the Jenderal Soedirman University for their seamless contribution to our research

FUNDING INFORMATION

This work was supported by the Universitas Jenderal Soedirman BLU Research Grant.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Zakiyyan Zain Alkaf	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
A'isya Nur Aulia Yusuf	✓	✓		✓	✓		✓	✓		✓				
Elsa Sari Hayunah Nurdiniyah	✓			✓	✓	✓		✓	✓					
Tri Wisudawati						✓			✓	✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT
The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY
The data supporting the findings of this study are available in this dataset: https://data.mendeley.com/datasets/r7tjn68rmw/1.

REFERENCES

[1] R. Vizcaíno *et al.*, "Enhanced electrokinetic remediation of polluted soils by anolyte pH conditioning," *Chemosphere*, vol. 199, pp. 477–485, 2018, doi: 10.1016/j.chemosphere.2018.02.038.

[2] H. Singh, N. Halder, B. Singh, J. Singh, S. Sharma, and Y. Shacham-Diamand, "Smart Farming Revolution: Portable and Real-Time Soil Nitrogen and Phosphorus Monitoring for Sustainable Agriculture," *Sensors*, vol. 23, 2023, doi: 10.3390/s23135914.

[3] S. Vogel *et al.*, "Direct prediction of site-specific lime requirement of arable fields using the base neutralizing capacity and a multi-sensor platform for on-the-go soil mapping," *Precision Agric*, vol. 23, pp. 127–149, 2021, doi: 10.1007/s11119-021-09830-x.

[4] J. Schmidinger *et al.*, "LimeSoDa: A Dataset Collection for Benchmarking of Machine Learning Regressors in Digital Soil Mapping," *Geoderma*, vol. 459, p. 117337, Jul. 2025, doi: 10.1016/j.geoderma.2025.117337.

[5] C. V. Patriche, B. Roșca, R. G. Pîrnău, and I. Vasiliniuc, "Spatial modelling of topsoil properties in Romania using geostatistical methods and machine learning," *PLoS One*, vol. 18, no. 8, 2023, doi: 10.1371/journal.pone.0289286.

[6] A. N. A. Yusuf, Z. Z. Alkaf, E. S. H. Nurdiniyah, T. Wisudawati, and M. I. Fawzi, "Classification of Worker Productivity and Resource Allocation Optimization with Machine Learning: Garment Industry," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 5, pp. 2991–3001, 2025, doi: 10.52436/1.jutif.2025.6.5.5263.

[7] Z. Z. Alkaf, M. I. Fawzi, M. W. R. Sastyawan, and R. D. Putera, "Unraveling of men's fragrance preferences on online marketplaces: a machine learning study using DBSCAN clustering and linear regression," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 6, pp. 1913–1920, 2025, doi: 10.52436/1.jutif.2024.5.6.4187.

[8] R. Choudhary, A. Kumar, P. C., M. M. Naik, M. Choudhury, and N. Khan, "Predicting water quality index using stacked ensemble regression and SHAP based explainable artificial intelligence," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-09463-4.

[9] S. Shi, N. Wang, S. Chen, B. Hu, J. Peng, and Z. Shi, "Digital mapping of soil salinity with time-windows features optimization and ensemble learning model," *Ecological Informatics*, vol. 85, 2025, doi: 10.1016/j.ecoinf.2024.102982.

[10] J. Sylvain, F. Anctil, and É. Thiffault, "Using bias correction and ensemble modelling for predictive mapping and related uncertainty: A case study in digital soil mapping," *Geoderma*, vol. 403, 2021, doi: 10.1016/j.geoderma.2021.115153.

[11] P. Tziachris, "Soil data grevena," *Mendeley Data*, vol. 1, 2022.

[12] A. Arabameri, J. P. Tiefenbacher, T. Blaschke, B. Pradhan, and D. T. Bui, "Morphometric Analysis for Soil Erosion Susceptibility Mapping Using Novel GIS-Based Ensemble Model," *Remote Sensing*, vol. 12, p. 874, 2020, doi: 10.3390/rs12050874.

[13] P. Guo, T. Li, H. Gao, X. Chen, Y. Cui, and Y. Huang, "Evaluating Calibration and Spectral Variable Selection Methods for Predicting Three Soil Nutrients Using Vis-NIR Spectroscopy," *Remote Sensing*, vol. 13, p. 4000, 2021, doi: 10.3390/rs13194000.

[14] P. Teo *et al.*, "Determining risk and predictors of head and neck cancer treatment-related lymphedema: A clinicopathologic and dosimetric data mining approach using interpretable machine learning and ensemble feature selection," *Clinical and Translational Radiation Oncology*, vol. 46, 2024, doi: 10.1016/j.ctro.2024.100747.

[15] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma, "A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques," *Sensors*, vol. 23, no. 7, 2023, doi: 10.3390/s23073565.

[16] O. Bulut, B. Tan, E. Mazzullo, and A. Syed, "Benchmarking Variants of Recursive Feature Elimination: Insights from Predictive Tasks in Education and Healthcare," *Information*, vol. 16, no. 6, p. 476, 2025, doi: 10.3390/info16060476.

[17] R. Dunne *et al.*, "Thresholding Gini variable importance with a single-trained random forest: An empirical Bayes approach," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 4354–4360, 2023, doi: 10.1016/j.csbj.2023.08.033.

[18] K. Phinzi, L. Bertalan, G. Chakilu, and S. Szabó, "Improving the spatial prediction of topsoil properties in a typical grazing area using multi-season PlanetScope spectral covariates and data mining techniques," *Earth Science Informatics*, vol. 18, no. 2, 2025, doi: 10.1007/s12145-025-01748-6.

[19] Y. Shen *et al.*, "Big-data and artificial-intelligence-assisted vault prediction and EVO-ICL size selection for myopia correction," *British Journal of Ophthalmology*, vol. 107, no. 2, pp. 201–206, 2021, doi: 10.1136/bjophthalmol-2021-319618.

[20] S. Yang, J. Sohn, S. Lee, J. Lee, and H. C. Kim, "Estimation and Validation of Arterial Blood Pressure Using Photoplethysmogram Morphology Features in Conjunction With Pulse Arrival Time in Large Open Databases," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1018–1030, 2020, doi: 10.1109/jbhi.2020.3009658.

[21] F. Oettl, A. Weinblatt, B. Chalmers, D. Kolin, and G. D. Valle, "Developing and validating machine learning models to predict acetabular cup size in total hip arthroplasty," *Journal of Orthopaedics*, vol. 67, pp. 349–352, 2025, doi: 10.1016/j.jor.2025.07.021.

[22] K. Mae and D. A. Kristiyanti, "Grid search vs Bayesian optimization for intensity scoring classification and channel recommendation prediction," *TELKOMNIKA (Telecommunication Comput. Electron. Control*, vol. 23, no. 4, pp. 1020–1031, 2025, doi: 10.12928/telkomnika.v23i4.26341.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] Y. Han, H. Liang, D. Jiang, Z. Wei, J. You, and G. Zhai, "Electromagnetic Relay Suction Prediction Based on Improved Support Vector Regression," in *2024 IEEE 7th International Electrical and Energy Conference (CIEEC)*, IEEE, May 2024, pp. 1237–1242. doi: 10.1109/CIEEC60922.2024.10583733.

[25] Ł. Uzarowicz *et al.*, "Initial pedogenic processes, mineral and chemical transformations and mobility of trace elements in Technosols on dumps of the former copper mines in Miedziana Góra and Miedzianka, the Świętokrzyskie Mts., south-central Poland," *CATENA*, vol. 245, p. 108293, Oct. 2024, doi: 10.1016/j.catena.2024.108293.

[26] T. A. Assegie, A. O. Salau, and T. U. Badrudeen, "Estimation of concrete compression using regression models," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 5, pp. 2799–2804, Oct. 2022, doi: 10.11591/eei.v11i5.4210.

[27] M. Anjum, K. Khan, W. Ahmad, A. Ahmad, M. N. Amin, and A. Nafees, "Application of Ensemble Machine Learning Methods to Estimate the Compressive Strength of Fiber-Reinforced Nano-Silica Modified Concrete," *Polymers*, vol. 14, no. 18, p. 3906, Sep.

2022, doi: 10.3390/polym14183906.
[28] B. O. Olorunfemi *et al.*, "Efficient diagnosis of diabetes mellitus using an improved ensemble method," *Scientific Reports.*, vol. 15, 2025, doi: 10.1038/s41598-025-87767-1.
[29] M. Fan, K. Xiao, L. Sun, S. Zhang, and Y. Xu, "Automated Hyperparameter Optimization of Gradient Boosting Decision Tree Approach for Gold Mineral Prospectivity Mapping in the Xiong'ershan Area," *Minerals*, vol. 12, no. 12, p. 1621, Dec. 2022, doi: 10.3390/min12121621.
[30] C. Ding, D. Yu, Q. Guo, and S. Zhang, "Parameter optimization of SVM algorithm for predicting physical parameters of SF6–Cu mixture plasma under local thermodynamic equilibrium," *AIP Advances*, vol. 13, no. 8, Aug. 2023, doi: 10.1063/5.0159427.
[31] H.-Y. Huang, Q.-D. Ding, J.-H. Zhang, X. Pan, Y.-H. Zhou, and K.-L. Jia, "Inversion of soil salinity and pH in farmland of the Hetao Plain based on Sentinel-2 and explainable machine learning.," *Ying Yong Sheng Tai Xue Bao*, vol. 36 8, pp. 2407–2419, 2025.
[32] Q. Lu, S. Tian, and L. Wei, "Digital mapping of soil pH and carbonates at the European scale using environmental variables and machine learning," *Science of The Total Environment*, vol. 856, p. 159171, Jan. 2023, doi: 10.1016/j.scitotenv.2022.159171.
[33] A. Lelago and M. Bibiso, "Performance of mid infrared spectroscopy to predict nutrients for agricultural soils in selected areas of Ethiopia," *Heliyon*, vol. 8, no. 3, p. e09050, Mar. 2022, doi: 10.1016/j.heliyon.2022.e09050.
[34] B.-Y. Huang *et al.*, "Machine learning methods to predict cadmium (Cd) concentration in rice grain and support soil management at a regional scale," *Fundamental Research*, vol. 4, pp. 1196–1205, 2023, doi: 10.1016/j.fmre.2023.02.016.

BIOGRAPHIES OF AUTHORS

Zakiyyan Zain Alkaf received his Bachelor's degree in Industrial Engineering from Universitas Islam Indonesia and his Master's degree from Universitas Gadjah Mada. He is currently a lecturer at Universitas Jenderal Soedirman. His research interests include quality management, lean manufacturing, and data analytics, with a specific focus on applying machine learning to industrial problems. He is an AUN/SEED-Net scholarship awardee and was honored with the Astra Satu Indonesia Award for his innovation in IoT-based smart farming solutions. He can be contacted at email: zakiyyan.alkaf@unsoed.ac.id.

A'isya Nur Aulia Yusuf received the Bachelor of Engineering degree in Information Technology from the Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2018, and the Master's degree in electrical engineering from the University of Indonesia, Depok, Indonesia, in 2021. She is a PMDSU scholarship awardee and a Doctoral candidate at the University of Indonesia. She is currently a lecturer in the Department of Electrical Engineering, Universitas Jenderal Soedirman. Her research interest is machine learning, antenna, telecommunication, and IoT. She can be contacted at email: aisya.yusuf@unsoed.ac.id.

Elsa Sari Hayunah Nurdiniyah received her Bachelor's degree in Engineering Physics from Universitas Gadjah Mada, Indonesia, in 2018 and Master Degree in Electrical Engineering from the School of Electrical Engineering and Informatics, Institut Teknologi Bandung, in 2022. She is currently a lecturer in the Department of Electrical Engineering, University Jenderal Soedirman. Her research interests are in the areas of signal processing, biomedical signals, machine learning and data analysis. She can be contacted at email: elsa.nurdiniyah@unsoed.ac.id.

Tri Wisudawati is affiliated with Jenderal Soedirman University, Indonesia. She has published on human system engineering and occupational health and safety. She can be contacted at email: tri.wisudawati@unsoed.ac.id.