International Journal of Environmental Science and Green Technology
https://journals.ayu.edu.kz

Machine Learning Models in Ecological Data Analysis: A Comparative Review

Anuarbek Amanov[1], Asan Daryn[2]

[1] Faculty of Engineering, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan
[2] Faculty of Engineering, MSc student, Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan

*Corresponding Author e-mail: anuarbek.amanov@ayu.edu.kz

Keywords

Machine learning
Ecological data analysis
Environmental monitoring
Predictive modeling
Data mining

Abstract

This article provides a comparative review of the effectiveness of machine learning models used in environmental data analysis. The aim of the study is to systematize the areas of application, advantages, limitations, and effectiveness of Linear Regression, Random Forest, Support Vector Machine, Neural Networks, and Deep Learning models in working with environmental data. Since environmental data are often multidimensional, nonlinear, spatial, and temporal, traditional statistical methods do not provide sufficient results in all cases. The results of the literature analysis prove that ensemble models such as Random Forest and XGBoost show high accuracy in many environmental forecasting tasks. While Deep Learning models are effective in analyzing complex data such as satellite imagery, biodiversity, animal movements, and time series, they require large data sets and high computational resources. In addition, the explainability of models remains an important issue. Explainable AI methods such as SHAP and LIME allow us to explain the decision-making logic of complex models. Research results show that model selection should consider explainability, data quality, computational efficiency, and ecological relevance in addition to accuracy.

Cite

Anuarbek A., & Asan D. (2026). Machine Learning Models in Ecological Data Analysis: A Comparative Review. International Journal of Environmental Science and Green Technology, 1(1), 1-7. doi: 10.5281/zenodo.20326675

Article Process

Submission Date: 08.01.2026; Revision Date: 17.01.2026; Accepted Date: 22.02.2026; Published Date: 30.03.2026;

INTRODUCTION

In recent years, the volume of data in ecological research has increased dramatically, and their analysis has become more complex. Satellite monitoring, remote sensing, climate monitoring, animal movement recording, acoustic monitoring, and biodiversity data provide a deep understanding of ecological systems. However, such data are often multifactorial, nonlinear, and have spatial-temporal dependencies. Therefore, traditional statistical methods such as Linear Regression cannot provide sufficient flexibility in all cases (Pichler & Hartig, 2022; Tuysuzoglu et al., 2018).

Machine learning models have been widely used in ecology for tasks such as predicting species distribution, assessing habitat suitability, determining forest fire risk, monitoring soil and water quality, and predicting ecosystem productivity (Yudaputra et al., 2019; Huang et al., 2023). In particular, Random Forest, SVM, XGBoost, Neural Networks, and Deep Learning models have shown excellent results in detecting complex ecological patterns (Al-Mukhtar, 2019; Xie et al., 2024).

However, the effectiveness of these models depends on the type, size, quality, and research objective of the data. For example, Random Forest offers high accuracy and relative interpretability, while

Deep Learning models are effective on complex image and spatiotemporal data, but are often described as "black boxes" (Pichler & Hartig, 2022; Southworth et al., 2024). To address this issue, interpretable artificial intelligence methods such as SHAP and LIME are used (Ghafarian et al., 2022; Mammides et al., 2024). The aim of the study is to compare the main machine learning models used in ecological data analysis and determine their accuracy, interpretability, computational efficiency, and usability. Scientific novelty — the article comprehensively compared models used in the analysis of ecological data not only in terms of predictive accuracy, but also in terms of interpretability, stability, adaptation to data types, and practical applicability.

MATERIAL AND METHOD

This study was based on a systematic literature review and comparative analysis. The literature selection included studies on the analysis of environmental data using machine learning published in recent years. The concepts of Linear Regression, Random Forest, Support Vector Machine, Neural Networks, Deep Learning, XGBoost, Explainable AI, SHAP and LIME were used as key words during the search and analysis. The initial search resulted in 122 articles, and 107 articles were identified through additional citation chaining, and 50 highly relevant studies were selected from a total of 229 candidate papers.

The models were evaluated according to five main criteria: predictive accuracy, explainability, computational efficiency, data type sensitivity and stability. These criteria are systematized in Table 1.

Table 1. Criteria for evaluating machine learning models

Criterion	Definition	Importance in Ecological Research
Predictive accuracy	The model's ability to correctly predict a specific ecological phenomenon	Important for tasks such as species distribution, fire risk assessment, and ecosystem productivity prediction
Interpretability	The ability to scientifically explain the model's decision or output	Helps identify the influence of ecological factors
Computational efficiency	The amount of time and computational resources required to train the model	Important when working with large datasets
Adaptability to data types	The model's ability to work with tabular, spatial, temporal, or image-based data	Enables appropriate model selection for different types of ecological data
Robustness	The model's ability to produce reliable results under different conditions	Important when applying the model to a new region or a new time period

RESULTS

The results of the literature analysis revealed that ensemble models, especially Random Forest and XGBoost methods, perform well in ecological prediction tasks. These models handle nonlinear relationships, complex interactions between factors, and multidimensional data structures well (Ghafarian et al., 2022; Koreň et al., 2021; Liu et al., 2021).

Although the Linear Regression model has high interpretability, it cannot fully describe nonlinear patterns in complex ecological systems. The SVM model is effective in working with small and medium-sized data and performs well in remote sensing and classification tasks (Ibrahim et al., 2023; Kupidura et al., 2024). Although Deep Learning models are effective in analyzing complex data such as satellite images, biodiversity indices, and animal movements, they require large data sets and high computing power (Garcia-Quintas et al., 2023; Chang, 2023).

Table 2. *Comparative characteristics of models used in the analysis of ecological data*

Model	Advantage	Limitation	Area of Application
Linear Regression	Simple, clear, and easy to interpret	Weak in detecting nonlinear relationships	Initial analysis, trend assessment
Random Forest	High accuracy, robustness, and ability to show variable importance	Limited ability to provide full causal explanation	Species distribution, ecosystem monitoring, habitat assessment
SVM	Effective with small datasets and performs well in classification tasks	Depends on kernel and parameter selection	Remote sensing, forest type classification, invasive species detection
Neural Networks	Learns complex relationships	Low interpretability and requires many resources	Time series analysis, ecosystem productivity
Deep Learning	Effective for images, satellite imagery, and large datasets	Requires large datasets, GPU resources, and explainable AI methods	Satellite monitoring, biodiversity assessment, animal movement analysis

As shown in Table 2, Random Forest stands out as a model that strikes a balance between accuracy and explainability. While Deep Learning achieves high accuracy, it needs to be complemented with SHAP, LIME, or other Explainable AI methods to be used in environmental decision-making.

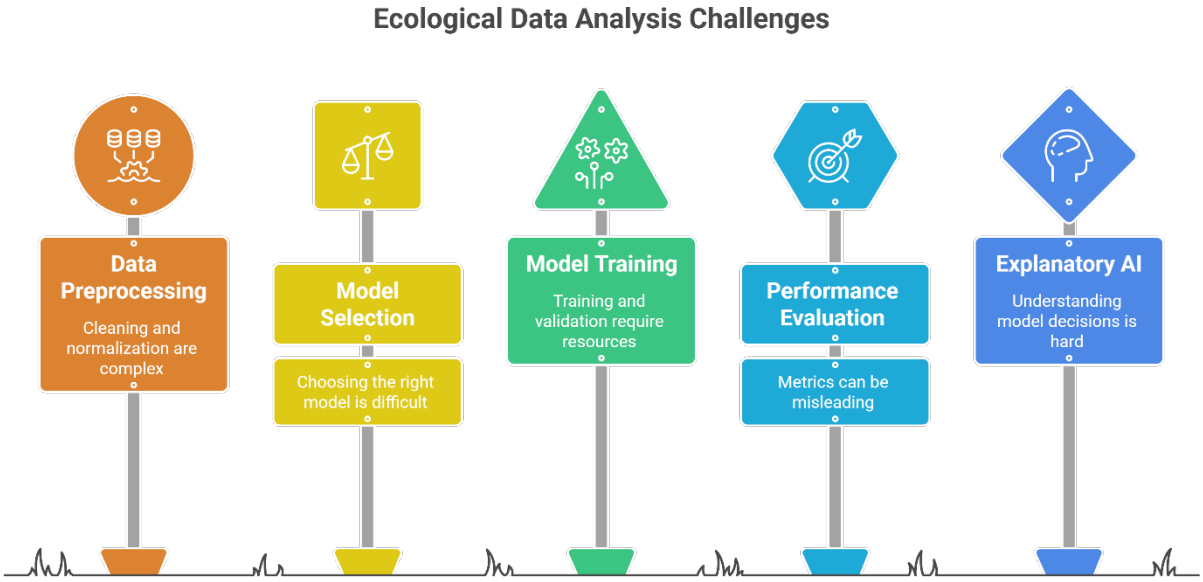

Fig 1. *General scheme of the process of analyzing environmental data using machine learning*

As shown in Figure 1, the analysis of environmental data using machine learning consists of several stages: data collection, preprocessing, model selection, training, evaluation, and interpretation of the results. This scheme allows us to consider the model not only as a technical tool, but also as an analytical system that helps in environmental decision-making.

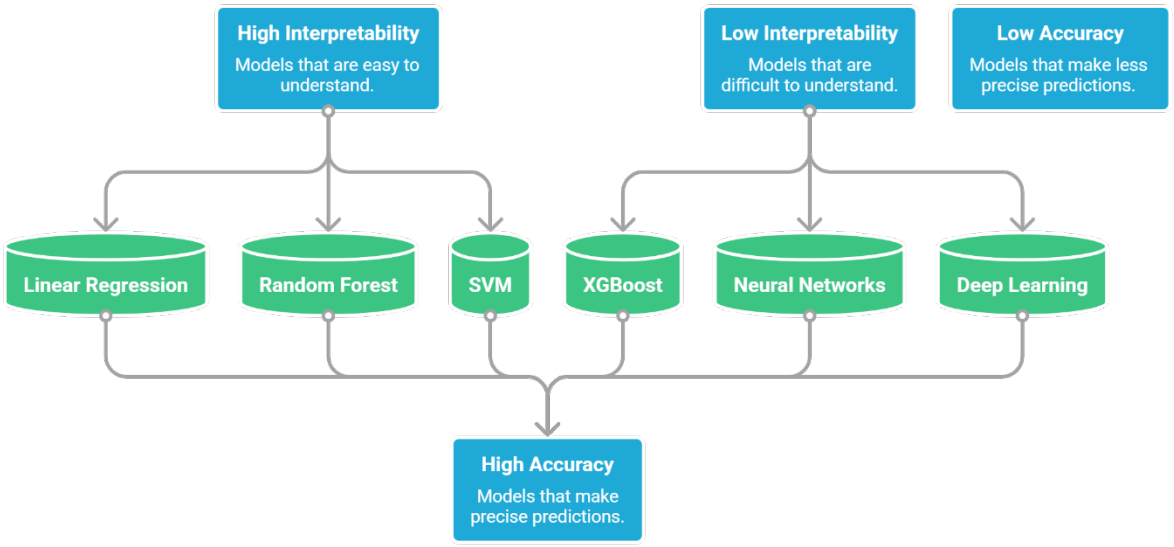

Fig 2. *Model selection logic: the relationship between accuracy and interpretability*

As shown in Figure 2, Linear Regression is strong in terms of explainability, but its accuracy may be limited in complex data. Random Forest and SVM strike a balance between accuracy and explainability. Although Deep Learning and Neural Networks achieve high accuracy, their explainability is low, so additional Explainable AI methods are needed.

DISCUSSION

The results of the study showed that machine learning methods are more flexible than traditional statistical models in analyzing ecological data. Random Forest and XGBoost models provide high accuracy in complex, multifactorial, and nonlinear data (Ghafarian et al., 2022; Liu et al., 2021). These models are effectively used in tasks such as predicting ecosystem productivity, assessing habitat, and determining the distribution of forest pests (Koreň et al., 2021; Huang et al., 2023).

However, in ecology, high model accuracy is not enough. The model output should have ecological meaning and help the researcher to explain natural processes. In this regard, classical statistical models and Linear Regression methods are still important, as they are closer to causal explanations (Pichler & Hartig, 2022; Sun et al., 2024).

The main advantage of Deep Learning models is the ability to detect hidden patterns in complex graphical, spatial and temporal data. However, their "black box" nature makes ecological interpretation difficult (Southworth et al., 2024; Garcia-Quintas et al., 2023). Therefore, the use of methods such as SHAP, LIME and feature importance play an important role in explaining model results (Ghafarian et al., 2022; Hang et al., 2024).

In addition, data quality directly affects model results. If the data are sparse, unbalanced, or spatially biased, sophisticated models may produce high accuracy but unreliable results in novel situations (Bonas et al., 2024; Zbinden et al., 2024). Therefore, cross-validation, external validation, uncertainty assessment, and statistical testing of results are mandatory in ecological machine learning.

CONCLUSION

Machine learning methods have great potential in environmental data analysis. Random Forest and XGBoost models show high accuracy and robustness in many environmental tasks. SVM models are effective for small and medium-sized data, while Deep Learning models are useful for analyzing complex data such as satellite imagery, biodiversity, and time series.

However, it is not correct to rely solely on accuracy when choosing a model. The researcher should consider the type, size, quality of the data, the need for explainability, and computational resources. As shown in Tables 1 and 2, each model has its advantages and limitations. Figure 1 shows the general structure of the environmental machine learning process, while Figure 2 allows us to compare the models in terms of accuracy and explainability.

In future research, it is important to combine environmental machine learning models with Explainable AI, uncertainty quantification, and causal inference methods. This will not only increase the accuracy of environmental predictions, but also increase their scientific validity and reliability in practical decision-making.

REFERENCES

Al-Mukhtar, M. (2019). Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. *Environmental Monitoring and Assessment, 191*, 673. https://doi.org/10.1007/S10661-019-7821-5

Binetti, M. S., Massarelli, C., & Uricchio, V. F. (2024). Machine learning in geosciences: A review of complex environmental monitoring applications. *Machine Learning and Knowledge Extraction*. https://doi.org/10.3390/make6020059

Bonas, M., Datta, A., Wikle, C. K., Boone, E. L., Alamri, F. S., Hari, B. V., Indulekha, K., Simmons, S. J., Jarvis, S., Burr, W. R., Pagendam, D., Chang, W., & Castruccio, S. (2024). Assessing predictability of environmental time series with statistical and machine learning models. *Environmetrics*. https://doi.org/10.1002/env.2864

Chang, J. G. (2023). Biodiversity estimation by environment drivers using machine/deep learning for ecological management. *Ecological Informatics, 78*, 102319. https://doi.org/10.1016/j.ecoinf.2023.102319

Garcia-Quintas, A., Roy, A., Barbraud, C., Demarcq, H., Denis, D., & Bertrand, S. L. (2023). Machine and deep learning approaches to understand and predict habitat suitability for seabird breeding. *Ecology and Evolution, 13*. https://doi.org/10.1002/ece3.10549

Ghafarian, F., Wieland, R., Lüttschwager, D., & Nendel, C. (2022). Application of extreme gradient boosting and Shapley additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. *Environmental Modelling & Software, 156*, 105466. https://doi.org/10.1016/j.envsoft.2022.105466

Hang, H. T., Mallick, J., AlQadhi, S., Bindajam, A. A., & Abdo, H. G. (2024). Exploring forest fire susceptibility and management strategies in western Himalaya: Integrating ensemble machine learning and explainable AI for accurate prediction and comprehensive analysis. *Environmental Technology & Innovation*. https://doi.org/10.1016/j.eti.2024.103655

Huang, C., Chen, B., Sun, C., Wang, Y., Zhang, J., Yang, H., Wu, S., Tu, P., Nguyen, M., Hong, S., & He, C. (2023). Synergistic application of multiple machine learning algorithms and hyperparameter optimization strategies for net ecosystem productivity prediction in Southeast Asia. *Remote Sensing*. https://doi.org/10.3390/rs16010017

Ibrahim, Y., Bagaye, U. Y., & Muhammad, A. I. (2023). Machine learning-based forest type mapping from multi-temporal remote sensing data: Performance and comparative analysis. *Environmental and Climate Remote Sensing*. https://doi.org/10.3390/ecrs2023-15848

Koreň, M., Jakuš, R., Zápotocký, M., Barka, I., Holuša, J., Ďuračiová, R., & Blaženec, M. (2021). Assessment of machine learning algorithms for modeling the spatial distribution of bark beetle infestation. *Forests, 12*(4), 395. https://doi.org/10.3390/F12040395

Kupidura, P., Kępa, A., & Krawczyk, P. (2024). Comparative analysis of the performance of selected machine learning algorithms depending on the size of the training sample. *Reports on Geodesy and Geoinformatics, 118*(1), 53–69. https://doi.org/10.2478/rgg-2024-0015

Liu, J., Zuo, Y., Wang, N., Yuan, F., Zhu, X., Zhang, L., Zhang, J., Sun, Y., Guo, Z., Guo, Y., Song, X., Song, C., & Xu, X. (2021). Comparative analysis of two machine learning algorithms in predicting site-level net ecosystem exchange in major biomes. *Remote Sensing, 13*(12), 2242. https://doi.org/10.3390/RS13122242

Mammides, C., Huang, G., Sree, R. P., Ieronymidou, C., & Papadopoulos, H. (2024). A novel approach for calculating prediction uncertainty when using acoustic indices and machine learning algorithms to monitor animal communities. *Research Square*. https://doi.org/10.21203/rs.3.rs-4494063/v1

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution, 14*(4), 994–1016. https://doi.org/10.1111/2041-210X.1406

Southworth, J., Smith, A. C., Safaei, M., Rahaman, M., Alruzuq, A., Tefera, B. B., Muir, C. S., & Herrero, H. V. (2024). Machine learning versus deep learning in land system science: A decision-making framework for effective land classification. *Frontiers in Remote Sensing*. https://doi.org/10.3389/frsen.2024.1374862

Sun, P., Holden, P. B., & Birks, H. J. B. (2024). Can machine-learning algorithms improve upon classical palaeoenvironmental reconstruction models? *Climate of the Past*. https://doi.org/10.5194/cp-20-2373-2024

Yudaputra, A., Robiansyah, I., & Rinandio, D. S. (2019). The implementation of artificial neural network and random forest in ecological research: Species distribution modelling with presence and absence dataset. *Ecological Modelling*.

Zaka, M. M., & Samat, A. (2024). Advances in remote sensing and machine learning methods for invasive plants study: A comprehensive review. *Remote Sensing, 16*(20), 3781. https://doi.org/10.3390/rs16203781

Zbinden, R., Tiel, N. V., Kellenberger, B., Hughes, L. H., & Tuia, D. (2024). On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning. *Ecological Informatics*, 102623. https://doi.org/10.1016/j.ecoinf.2024.102623