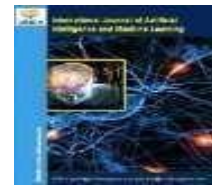




SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

A Comprehensive Review of Modified Recursive Feature Elimination and Related Feature Selection Methods in Precision Agriculture and Soil Health Assessment

Himat Singh¹, Dr. Harsh Sadawarti²

¹Research Scholar, Computer Science & Application, Desh Bhagat University, Mandi Gobindgarh.

²Professor, Department of Computer Science & Application, Desh Bhagat University, Mandi Gobindgarh.

E-mail: 2298145.Researchscholar@deshbhagatuniversity.in

Abstract

High-dimensional agricultural datasets integrate soil measurements, nutrient profiles, climatic records and remote-sensing observations. Selecting informative variables from these datasets is important for developing parsimonious predictive models and supporting interpretation. Recursive Feature Elimination (RFE) is a wrapper-based feature-selection procedure that repeatedly fits a learner, ranks predictors and removes lower-ranked variables. Modified Recursive Feature Elimination (MRFE) and related approaches adapt aspects of this procedure, although MRFE does not denote a single universally standardized algorithm. This review examines applications of direct MRFE, conventional RFE and related feature-selection methods in land-suitability classification, soil-property mapping, soil-fertility modelling, crop-yield prediction and nutrient recommendation. The reviewed studies include an MRFE-bagging application reporting approximately 95% classification accuracy, an SVM-RFE study that reduced 30 environmental variables to five for soil-pH mapping, and a modified greedy feature-selection study that retained nine predictors from an initial set of 392 covariates. Other applications include a hybrid correlation-based filter and RF-RFE approach, cross-validated RFE for maize nutrient recommendation, and comparative evaluations of feature-selection methods for soil-fertility modelling. Reported results vary across prediction targets, datasets, algorithms and validation designs and should not be interpreted as a pooled estimate or direct ranking of methods. Nitrogen, phosphorus, temperature-related variables, topographic indices, soil moisture and vegetation indices recur among the predictors discussed in the reviewed examples. The review also identifies methodological considerations concerning feature-selection leakage, subset-size determination, correlated predictors, spatial and temporal validation, interpretability and computational cost. Future research should emphasize transparent algorithm definitions, leakage-safe validation, independent multi-season testing and reproducible benchmarks.

Keywords: Recursive Feature Elimination(RFE), Modified Recursive Feature Elimination(MRFE), precision agriculture, soil nutrient ranking, crop-yield prediction, explainable machine learning, digital soil mapping

1. Introduction

Modern agriculture faces the combined challenges of increasing production, adapting to climatic variability and limiting the environmental effects of intensive input use. Precision agriculture addresses spatial and temporal variation by tailoring agronomic actions to local conditions [1,2]. Identifying relevant soil nutrients, physicochemical properties and climatic factors can support predictive modelling and the interpretation of agricultural systems. Sensor networks, multispectral imagery and multi-site experiments can produce datasets containing many candidate predictors [3,4].

High dimensionality can increase computational demands and complicate model interpretation. Correlated or noisy variables may also contribute to unstable estimates and overfitting, particularly when the number of observations is limited [5,6]. Feature selection is therefore commonly used to reduce the predictor set and focus modelling on variables that are informative for a specified task [7,8].

Feature-selection methods are often grouped into filter, wrapper and embedded approaches. Filter methods assess variables using criteria independent of a downstream learner; embedded methods perform selection as part of model fitting; and wrapper methods evaluate subsets using a chosen learning algorithm [5,9]. RFE is a wrapper procedure: a model is fitted, features are ranked using a model-dependent importance measure, and the least important feature or features are removed iteratively until a stopping condition is met [10,12].

Modified RFE approaches seek to address limitations of a basic recursive procedure. Modifications may involve the elimination schedule, importance measure, base learner or stopping criterion. Validation and interpretation procedures, including nested cross-validation, bootstrap stability analysis, SHAP or LIME explanations, and spatial diagnostics, can be combined with RFE, but should not automatically be treated as defining components of every

MRFE algorithm [13,14,17]. The precise method must be established from the original study rather than inferred from the label MRFE.

The studies considered in this review do not all implement the same feature-selection algorithm. They include direct applications of Modified Recursive Feature Elimination (MRFE), conventional Recursive Feature Elimination (RFE), cross-validated RFE (RFECV), hybrid filter-wrapper procedures and modified greedy feature selection. These approaches share the broad objective of identifying informative predictor subsets, but differ in their selection mechanisms and algorithmic design. Accordingly, this review distinguishes studies that explicitly report MRFE from those that apply RFE or other related methods. This distinction is important when interpreting reported performance and assessing the applicability of individual approaches to agricultural prediction tasks.

1.1 Scope and Nature of the Review

This study adopts a narrative review approach to examine the applications and methodological developments of Recursive Feature Elimination (RFE) and its modified variants in precision agriculture and soil health assessment. Relevant studies are discussed to identify key methodological approaches, application areas, research trends, and existing knowledge gaps. The review focuses on the role of feature selection in improving model interpretability, reducing dimensionality, and supporting data-driven agricultural decision-making. As the literature search and study selection were not conducted using a predefined, reproducible systematic protocol, this review is not intended to be exhaustive or classified as a systematic review. Instead, it provides a thematic synthesis of selected research contributions and highlights potential directions for future investigation.

2. Artificial Intelligence and Feature Selection in Precision Agriculture

Artificial intelligence and machine learning are used in agricultural applications involving prediction of yield and soil properties, classification of land suitability or management zones, and analysis of influential variables [1]. Random Forest, gradient-boosting methods, support-vector machines and neural-network architectures can be applied to different data types and prediction tasks. Remote-sensing imagery and time-series observations may provide spatial and temporal information, while sensor networks can support repeated monitoring.

Feature selection is one component of a broader modelling workflow. Its role is to identify a subset of predictors for a particular target and learner; it does not by itself establish causal influence or guarantee improved agronomic outcomes. Explainability methods such as SHAP and LIME can help interpret model predictions, but their outputs are not interchangeable with the selection criterion used by RFE [14,17,18]. Similarly, spatial diagnostics can assess residual spatial structure or the validity of spatial predictions, but should be reported as a separate evaluation step.

Potential operational benefits of well-validated agricultural models include more focused data collection and decision support. Their practical performance depends on data quality, calibration, sampling design, temporal coverage and the suitability of validation for the intended deployment setting. Feature selection should therefore be evaluated alongside the complete modelling pipeline rather than as an isolated guarantee of improved performance.

3. Methodological Foundations of RFE and Its Modifications

3.1 Classical RFE and its constraints

Classical RFE begins with a candidate feature set, fits a base learner, ranks predictors using an importance measure, removes the lowest-ranked feature or group of features, and repeats the process. The procedure terminates when a prespecified subset size or another stopping rule is reached [10–12]. The selected subset is conditional on the data, learner, importance measure, elimination step and stopping rule.

Several limitations are relevant to agricultural datasets. The preferred subset size may not be known in advance; rankings can vary with sampling and model choice; correlated predictors may substitute for one another; and repeated model fitting can be computationally expensive. These issues are particularly important when observations are spatially clustered or collected across seasons. A compact subset is not necessarily stable, causal or transferable to a new region.

3.2 Modified and related recursive approaches

Modifications to recursive feature selection may alter the importance measure, the number of predictors eliminated at each iteration, the base learner or the stopping criterion. However, the term Modified Recursive Feature Elimination (MRFE) is not associated with one universally standardized procedure. Its implementation should therefore be described according to the algorithmic details reported in the original study.

The literature considered in this review includes several approaches that are related to, but distinct from, direct MRFE. Conventional RFE iteratively eliminates predictors according to a model-dependent ranking, whereas recursive feature elimination with cross-validation (RFECV) uses cross-validation to help determine the retained feature subset. Hybrid filter-wrapper methods combine an initial filter-based screening stage with a subsequent

wrapper procedure. Modified greedy feature selection (MGFS), in contrast, uses a greedy selection strategy and should not be treated as an MRFE variant solely because it reduces the predictor set.

The choice of base learner is another important design decision. Tree ensembles, boosting methods and support-vector machines can produce different importance estimates and selected subsets. Nested cross-validation can be used to keep feature selection and hyperparameter tuning within the training data, while bootstrap or repeated-sampling procedures can assess selection stability. These validation and evaluation procedures may accompany recursive feature selection, but they are not defining components of MRFE unless explicitly incorporated into the algorithm under study.

3.3 General workflow and validation

A defensible workflow begins by defining the prediction target, sampling unit and intended deployment setting. Preprocessing, feature selection, model fitting and hyperparameter tuning should be conducted within the training data. Where model selection is part of the analysis, nested cross-validation can provide an outer estimate of generalization while inner folds are used for selection and tuning. An independent test set is useful when available, particularly if it reflects the intended spatial or temporal deployment setting.

Depending on their design, modifications to RFE may improve feature-selection efficiency or ranking stability. However, their effects on overfitting and generalization to new agro-climatic conditions depend on the data, learning algorithm and validation strategy and therefore require empirical evaluation.

After selection, researchers should report the initial feature count, retained feature count, elimination rule, importance measure, base learner, stopping criterion and validation design. Performance should be reported using metrics appropriate to the task. Classification accuracy should not be compared directly with regression R^2 , and metrics from different datasets should not be pooled without a justified synthesis method. Where interpretability is a goal, selected predictors and post-hoc explanations should be reported separately.

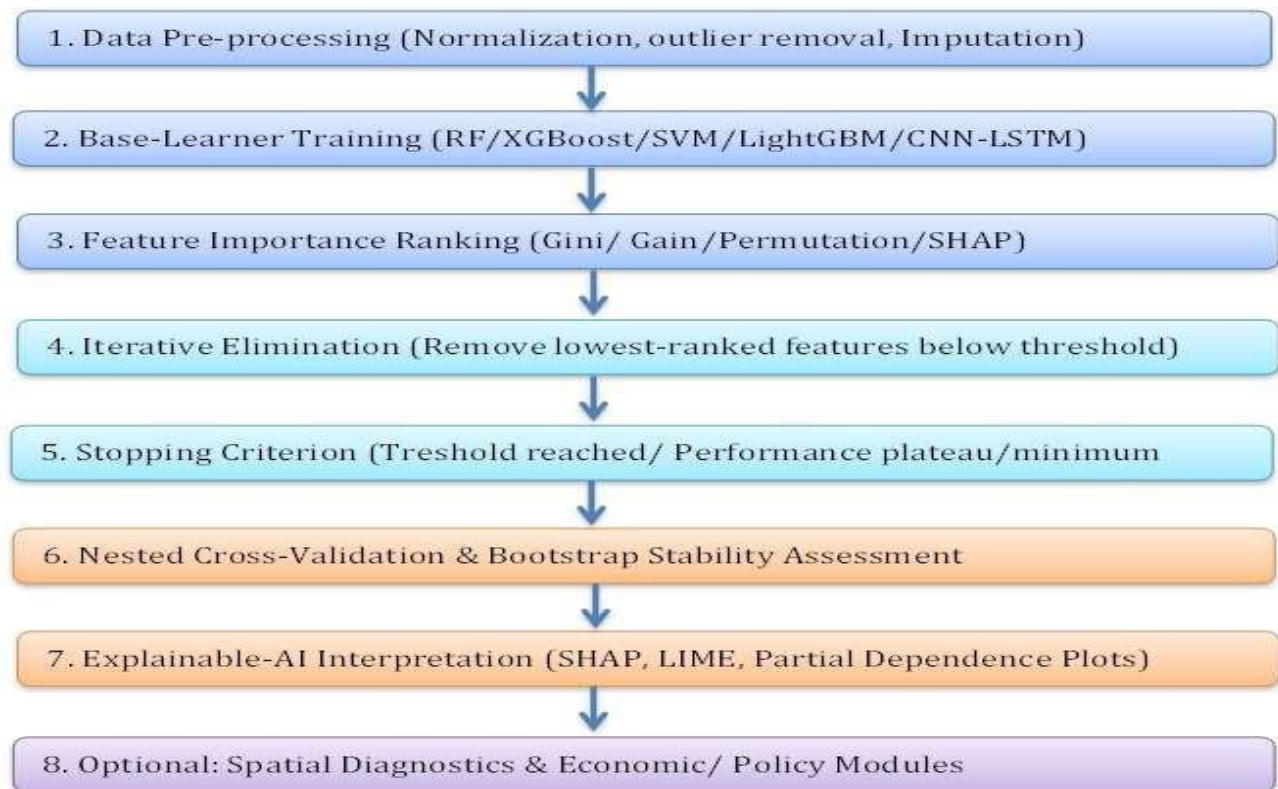


Fig. 1 Schematic outline of the MRFE workflow

4. Applications in Precision Agriculture

The reviewed studies apply direct MRFE, conventional RFE and related feature-selection approaches to several agricultural prediction tasks, including land-suitability classification, soil-property mapping, soil-fertility modelling, crop-yield prediction and nutrient recommendation. Although these methods share the objective of identifying informative predictor subsets, they differ in their selection procedures and modelling frameworks. To avoid conflating distinct algorithms, the applications are discussed according to the method reported in each study.

Direct MRFE applications are distinguished from conventional RFE, RFECV, hybrid filter-wrapper methods and modified greedy feature selection.

4.1 Soil health and nutrient-related modelling

In soil-pH mapping, Guo, Wang and Jin [12] applied SVM-RFE to thirty environmental variables derived from 134 samples in Fuzhou, China. Five variables were retained: mean annual temperature, slope, Topographic Wetness Index, MSAVI and Band 5. Modeling with XGBoost produced an R^2 of 0.68, MAE of 0.16 and RMSE of 0.26, an improvement over the full-variable baseline.

Ferhatoglu and Miller [8] compared six feature-selection methods for five soil-fertility properties across ten fields using 1,049 environmental covariates. The cited summary indicates that wrapper and embedded methods frequently produced optimal models when paired with decision-tree algorithms, and that feature selection improved cross-validation and independent-validation measures. Exact property-wise values should be included only after checking the original article.

Huang et al. [19] investigated the use of recursive feature elimination (RFE) for mapping soil organic carbon (SOC), soil total phosphorus (STP) and soil available phosphorus (SAP) in the Haihun River sub-watershed, China. Using 180 soil samples and 40 environmental covariates, the authors identified property-specific predictor subsets containing 12 variables for SOC, eight for STP and four for SAP. The selected-feature models achieved R^2 values of 0.79 and 0.84 for SOC, 0.86 and 0.89 for STP, and 0.70 and 0.77 for SAP using Random Forest and XGBoost, respectively. These results illustrate the application of RFE to multiple soil-mapping targets, with the selected predictors and model performance varying by target property and learning algorithm.

4.2 Land suitability and crop-yield prediction

Mariammal et al. [13] applied a Modified Recursive Feature Elimination (MRFE) approach combined with bagging to land-suitability prediction. The study reported approximately 95% classification accuracy and identified nitrogen, rainfall and temperature-related variables among the leading predictors. This work represents a direct MRFE application among the studies considered in this review.

Elavarasan et al. [15] developed a hybrid feature-selection approach combining correlation-based feature selection (CFS), a filter method, with random-forest recursive feature elimination (RF-RFE), a wrapper method, for agricultural crop-yield prediction. The study compared predictive models based on different feature sets, including the proposed hybrid subset. This approach is related to RFE but should be classified as a hybrid filter-wrapper method rather than as a direct MRFE implementation.

4.3 Soil organic carbon and related recursive selection

Zhang et al. [11] investigated modified greedy feature selection (MGFS) for digital soil mapping of soil organic carbon density. The method selected nine predictors from an initial set of 392 covariates. The resulting model achieved an R^2 of approximately 0.60. Although MGFS and RFE both aim to identify compact predictor sets, MGFS uses a modified greedy selection strategy and is algorithmically distinct from conventional RFE and direct MRFE. It is therefore included as a related feature-selection approach rather than classified as an MRFE implementation.

4.4 Nutrient recommendation and remote-sensing applications

Liben et al. [20] applied an integrated machine-learning framework combining LightGBM, recursive feature elimination with cross-validation (RFECV), and SHAP for spatially explicit nutrient recommendation in maize production systems in Malawi. The study used 14,625 georeferenced observations collected across multiple fertilizer-trial seasons and reported a test-set coefficient of determination (R^2) of 0.71. Nitrogen, phosphorus, soil moisture and solar radiation were identified among the principal predictors contributing to model outputs. In this framework, RFECV provided a cross-validated feature-selection procedure, LightGBM was used for prediction, and SHAP supported interpretation of model outputs. This study is therefore classified as an RFECV application with post-hoc model interpretation, rather than as a direct MRFE implementation.

5. Comparative Synthesis of Representative Studies

The studies reviewed demonstrate the use of direct MRFE, conventional RFE and related feature-selection approaches across several agricultural prediction tasks (Table 1). The methods differ in their selection mechanisms, base learners and evaluation procedures. Mariammal et al. [13] reported a direct MRFE-bagging application for land-suitability classification, whereas Guo et al. [12], Huang et al. [19] and Biru et al. [22] applied RFE-based procedures to soil-property mapping or crop-related prediction. Liben et al. [20] used RFECV for feature selection in a LightGBM-based nutrient-recommendation framework. Elavarasan et al. [15] combined filter and wrapper procedures, while Zhang et al. [11] evaluated a modified greedy feature-selection strategy. Ferhatoglu and Miller [8] provided a comparative assessment of different feature-selection methods for soil-fertility modelling.

The number of selected predictors and the reported model performance varied among the studies. For example, SVM-RFE reduced the predictor set from 30 variables to five in a soil-pH mapping study, while MGFS selected nine

predictors from 392 covariates in digital soil mapping. These results illustrate differences in selection outcomes but do not establish the superiority of one method across applications. Reported performance metrics also depend on the prediction target, dataset and validation design, limiting direct comparison across studies.

Nitrogen, phosphorus, temperature-related variables, topographic indices, soil moisture and vegetation indices were among the predictors identified in the reviewed applications. Their selection reflects the characteristics and objectives of individual studies and should not be interpreted as evidence of universal importance across agricultural environments. Similarly, the use of explainability methods alongside feature selection can provide additional insight into model predictions, but does not independently establish causal relationships or transferability. Overall, the reviewed evidence illustrates the diversity of recursive and related feature-selection approaches rather than demonstrating a universally superior method.

Table 1: Summary of representative RFE and related feature-selection applications

Study	Application and study region	Initial and Selected Features	Feature-selection Method	Reported Model Performance	Selected Predictors or Key Findings
Mariamammal et al. [13] (2021)	Land-suitability classification, India	Not reported	Direct MRFE with bagging	Classification accuracy: approximately 95%	Nitrogen, rainfall and temperature-related variables
Guo et al. [12] (2022)	Soil-pH mapping, China	30 → 5	SVM-RFE; XGBoost prediction	$R^2 = 0.68$; MAE = 0.16; RMSE = 0.26	Mean annual temperature, slope, Topographic Wetness Index (TWI), MSAVI and Band 5
Ferhatoglu and Miller [8] (2022)	Soil-fertility modelling, Iowa, USA	1,049 environmental covariates; retained feature counts not reported	Comparative evaluation of six feature-selection methods	Comparative cross-validation and independent-validation results reported	Wrapper and embedded methods frequently performed well when paired with decision-tree models
Elavarasan et al. [15] (2020)	Agricultural crop-yield prediction	Not reported	Hybrid CFS filter and RFE wrapper	Comparative evaluation of alternative feature sets	Climate, soil and groundwater variables; hybrid filter-wrapper approach
Zhang et al. [11] (2023)	Soil organic-carbon density mapping, China	392 → 9 (MGFS); RFE selected 22, VIF selected 30 and Boruta selected 76	Modified Greedy Feature Selection (MGFS)	$R^2 = 0.60$; RMSE = 13.80 t ha ⁻¹ ; LCCC = 0.74	Demonstrated a compact predictor subset using a method related to, but distinct from, conventional RFE
Huang et al. [19] (2024)	Multi-property soil mapping, China	40 covariates; selected 12	RFE with Random Forest and XGBoost	Final R^2 values of 0.79–0.89	Soil organic carbon, total phosphorus and available phosphorus
Liben et al. [20] (2026)	Spatially explicit maize nutrient recommendation, Malawi	Multi-source predictors; selected subset size not reported	RFECV with LightGBM and SHAP-based interpretation	Test $R^2 = 0.71$	Nitrogen, phosphorus, soil moisture and solar radiation
Biru et al. [22] (2025)	Summer wheat leaf area index (LAI) prediction, Ethiopia	29–49 retained predictors; initial feature count not reported	RFE with ensemble modelling	Validation $R^2 \approx 0.88$	Remote-sensing and other environmental predictors

6. Limitations and Research Prospects

The reviewed examples raise several methodological issues. First, selected subsets can depend on the base learner, importance measure, elimination step and stopping rule. Second, correlated predictors may lead to different but similarly predictive subsets, so a single ranking should not automatically be interpreted as a unique ordering of agronomic importance. Third, spatially clustered observations can make random cross-validation optimistic when the intended use is prediction in new locations. Temporal and multi-year validations are also important where sensor relationships and crop responses change across seasons.

Feature selection must be performed within training folds to avoid leakage into validation data. Nested validation can reduce bias from tuning and selection, although it increases computational cost. Reports should state whether preprocessing, feature selection and hyperparameter optimization were repeated independently within each fold. Stability analysis across resamples or sites can complement predictive metrics, but its results should be quantified rather than asserted.

Further research could evaluate adaptive selection under streaming observations, lightweight implementations for resource-constrained devices, and process-informed constraints that may improve plausibility beyond the observed data range. Multi-region and multi-season benchmarks, open data and reproducible pipelines would make comparisons more transparent. Potential links to nutrient-use efficiency, carbon accounting or subsidy mechanisms should be investigated empirically rather than assumed to follow directly from feature-selection performance. These priorities require collaboration among soil scientists, agronomists, machine-learning researchers and extension specialists.

7. Conclusion

Direct MRFE, conventional RFE and related feature-selection methods have been applied to land-suitability classification, soil-property mapping, soil-fertility modelling, crop-yield prediction and nutrient recommendation. The reviewed studies demonstrate that these approaches can produce compact predictor sets and, in some study-specific settings, achieve useful predictive performance. Examples include an MRFE-bagging application reporting approximately 95% classification accuracy, an SVM-RFE study reducing 30 environmental variables to five for soil-pH mapping, and a modified greedy feature-selection study selecting nine predictors from 392 covariates. Other studies applied hybrid filter-wrapper selection or RFECV in agricultural prediction workflows.

These results should not be interpreted as evidence of a universal performance gain for MRFE or RFE. The methods differ in their algorithmic design, base learners, selection criteria and validation procedures. In particular, modified greedy feature selection, hybrid filter-wrapper methods and RFECV should be identified separately from direct MRFE unless the original study explicitly defines them otherwise. Reliable assessment requires leakage-safe feature selection, appropriate spatial and temporal validation, transparent reporting of selection rules and attention to ranking stability. SHAP and LIME can support model interpretation, but they do not replace validation or establish causal effects. Overall, the reviewed studies support the use of recursive and related feature-selection procedures as components of agricultural machine-learning workflows, while broader claims about transferability and operational impact require further independent evaluation.

8. References

- [1] Gebbers, R. and Adamchuk, V.I. (2010). Precision agriculture and food security. *Science*, 327(5967), 828–831.
- [2] Kamilaris, A. and Prenafeta-Boldú, F.X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90.
- [3] Liakos, K.G., Busato, P., Moshou, D., Pearson, S. and Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 2674.
- [4] Chlingaryan, A., Sukkarieh, S. and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151, 61–69.
- [5] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
- [6] Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28.
- [7] Li, Z., Zhou, X., Cheng, Q., Zhai, W., Mao, B., Li, Y. and Chen, Z. (2023). An integrated feature selection approach to high water stress yield prediction. *Frontiers in Plant Science*, 14, 1289692.
- [8] Ferhatoglu, C. and Miller, B.A. (2022). Choosing feature selection methods for spatial modeling of soil fertility properties at the field scale. *Agronomy*, 12(8), 1786.
- [9] Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York.

- [10] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, 389–422.
- [11] Zhang, X., Chen, S., Xue, J., Wang, N., Xiao, Y. et al. (2023). Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping. *Geoderma*, 432, 116383.
- [12] Guo, J., Wang, K. and Jin, S. (2022). Mapping of soil pH based on SVM-RFE feature selection algorithm. *Agronomy*, 12(11), 2742.
- [13] Mariammal, G., Suruliandi, A., Raja, S.P. and Poongothai, E. (2021). Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modified recursive feature elimination technique with various classifiers. *IEEE Transactions on Computational Social Systems*, 8(5), 1132–1142.
- [14] Lundberg, S.M. and Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- [15] Elavarasan, D. and Vincent, P.M.D.R. (2020). A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modelling. *Agriculture*, 10(9), 400.
- [16] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.
- [17] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD*, pp. 1135–1144.
- [18] Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Munich: Christoph Molnar. ISBN 9798411463330.
- [19] Huang, J., Liu, J., Ye, Y., Jiang, Y., Lai, Y., Qin, X., Zhang, L. and Jiang, Y. (2024). Mapping soil properties in the Haihun River sub-watershed, Yangtze River Basin, China, by integrating machine learning and variable selection. *Sensors*, 24(12), 3784.
- [20] Liben, F., Kihara, J., Abera, W., AbebeMesfin, T., Homann-kee Tui, S. and Munthali, M. (2026). From prediction to prescription: Explainable machine learning-driven spatially explicit nutrient recommendations for maize systems in Malawi. *Field Crops Research*, 347, 110658.
- [21] Sapkal, K. and Kadam, A.B. (2025). Optimal Feature Selection for Soil Data Analysis with Enhanced Crop Recommendation using Machine Learning. *2025 International Conference on Sustainable Technologies for Humanity and Smart World (HSWTech)*, IEEE.
- [22] Biru, D., Gessesse, B. and Abebe, G. (2025). Recursive feature elimination for summer wheat leaf area index using ensemble algorithm-based modeling: The case of central Highland of Ethiopia. *Environmental Challenges*, 19, 101113.