

Machine Learning-Based Land Suitability Classification Using NPK and Environmental Factors

Submitted 16 August 2026, Revised 22 September 2026, Accepted 24 September 2026

Fathiah Alatas^{1*}, Sujiyo Miranto², Brian Abdurafi Rambu Basae³, Khansa Tatsbita⁴, Elvan Yuniarti⁵

^{1,3}Department of Physics Education, Faculty of Tarbiya and Teaching Sciences,
UIN Syarif Hidayatullah Jakarta, Tangerang Selatan, Indonesia

²Department of Biology Education, Faculty of Tarbiya and Teaching Sciences,
UIN Syarif Hidayatullah Jakarta, Tangerang Selatan, Indonesia

^{4,5}Department of Physics, Faculty of Science and Technology,
UIN Syarif Hidayatullah Jakarta, Tangerang Selatan, Indonesia
Corresponding Email: *fathiah.alatas@uinjkt.ac.id

Abstract

An accurate estimation of soil fertility is needed for the purposes of nutrient management and the practice of precision agriculture. The objective of this research is to build and test a machine learning model for the classification of soil fertility status using seven parameters which are nitrogen (N), phosphorus (P), potassium (K), pH, EC, soil temperature and soil moisture content. Data were collected via measuring using an Internet of Things (IoT) based NPK Meter system. After the pre-processing of data, 91 observations were collected with 7 predictors, and they were grouped into three classes of soil fertility status, which are infertile, moderately fertile, and fertile. The three machine learning algorithms tested were the Artificial Neural Network (ANN), Gradient Boosting Machine (GBM) and CatBoost. The data was divided into 90% training data and 10% test data, whereas the performance of the model was evaluated using accuracy, precision, recall, F1 score, confusion matrix, and Area Under the ROC Curve (AUC). The test results showed that the ANN achieved an accuracy of 100 per cent, whilst the GBM and CatBoost each achieved 90 per cent using the data split employed. Feature contribution analysis indicates that the NPK parameters, moisture content and electrical conductivity provide predictive information for distinguishing soil fertility classes, although the magnitude of their contribution varies across the different algorithms. These results demonstrate the potential of combining IoT-based soil sensors and machine learning to support the classification of soil conditions.

Keywords: Soil Type Suitability, Artificial Neural Networks (ANN), Gradient Boosting Machine (GBM), CatBoost (Categorical Boosting), NPK.

INTRODUCTION

Global food demand continues to rise as a result of population growth, climate change, and soil degradation, increasing the need for more efficient and sustainable agricultural practices. Soil nutrient availability plays an important role in crop growth, productivity, and fertiliser-use efficiency (Budiman & Suhendi, 2024). Inadequate nutrient management can reduce crop yields, increase production costs, and contribute to environmental problems associated with excessive fertiliser application (Kabato et al., 2025). Therefore, rapid and accurate assessment of soil nutrient status is essential for supporting the implementation of precision agriculture.

The fertility of the soil is one of the most important factors that contribute to agricultural productivity, as nutrient availability and soil physicochemical properties affect plant growth and fertilizer usage effectiveness (Khechba et al., 2021; Sharma & Tomar, 2025; Toselli et al., 2023). The soil evaluation is essential in order to define nutrient management

practices and avoid ineffective fertilizer use (Najdenko et al., 2024; Reza et al., 2025). Parameters like nitrogen (N), phosphorus (P), potassium (K), pH, electric conductivity, temperature and moisture are used for evaluation of soil condition (Dattatreya et al., 2024; Rachakonda & Stasiewicz, 2024). As a result, fast measurements and calculations of those parameters have become an essential part of developing precision agriculture.

Conventional soil nutrient analysis is still dominated by laboratory testing based on physical and chemical parameters, such as nitrogen (N), phosphorus (P) and potassium (K) content, pH, organic carbon content and electrical conductivity. Although these methods produce accurate data, they are costly to carry out, require a relatively long analysis time and demand specialised expertise. Furthermore, due to spatial variations in soil properties, results obtained from specific sampling points often do not accurately reflect the conditions of the entire area. It is these factors that have driven the development of faster and more cost-effective classification methods, which in turn enable real-time decision-making (Shin et al., 2025; Najdenko et al., 2024; Reza et al., 2025).

Machine learning technology has been widely utilised in agricultural data management to accurately analyse non-linear parameters, such as generating accurate soil classification and fertility models, predicting harvests, detecting plant diseases and so on. The use of machine learning is particularly recommended due to its adaptive nature in decision making (Liakos et al., 2018). Some machine learning methods used for agricultural modelling include: Artificial Neural Networks (ANN) (Castillo-Girones et al., 2025), Gradient Boosting Machines (GBM)(Ennaji et al., 2024), Extreme Gradient Boosting (XGBoost) (Anugraha et al., 2026), Light Gradient Boosting Machine (LightGBM), CatBoost (Fadly, 2026), Logistic Regression (LR) (Panigrahi et al., 2022), Naïve Bayes (NB) (Oktapiani & Agustin, 2024), k-Nearest Neighbour (KNN) (Yuliyanto et al., 2025)(Yuliyanto, Y., Decision Tree (DT)(Nila Aulia et al., 2025)., Support Vector Machine (SVM) (Tohari et al., 2023)., Random Forest (RF) (Mahmood et al., 2025) which can be used as classification models.

Several previous works indicate that machine learning algorithms outperform the traditional approach in terms of enhancing the soil fertility classification. Artificial Neural Networks, Gradient Boosting Machines, and CatBoost (Categorical Boosting) have been shown to be efficient on agricultural data (Sujatha & Jaidhar, 2024). However, even those studies demonstrate variation in their results depending on the type of the used dataset, pre-processing techniques and limitations of the model. Moreover, most of the previous work considers comparison of only one or two algorithms and focuses mainly on accuracy; evaluation of other metrics (accuracy, completeness, F1 score, recall, precision, and

interpretability) is rather limited. Thus, this research investigates the area in more detail. On the basis of the above literature review, the research gap found in this work is creation and evaluation of three models. Objectives of this study include determination of the parameters' characteristics and evaluation of the three models, as well as identification of features responsible for classification.

Soil fertility is determined by several criteria, including soil nutrients such as organic carbon content which consists of decomposed plant and animal matter that can influence the physical, chemical and biological properties of the soil. Soil nutrient content includes nitrogen, phosphorus and potassium, as well as other elements required by plants (Dattatreya et al., 2024; Reza et al., 2025). Nitrogen is easily soluble in water and air and is therefore easily lost. A nitrogen (N) content of less than 10 ppm is considered infertile; 10-30 ppm is classified as slightly infertile; and 30-50 ppm as moderately infertile. Phosphorus (P) content is classified as infertile if below 10 ppm, moderately infertile at around 10-20 ppm, and fertile at 20-40 ppm. Meanwhile, potassium content is classified as infertile at less than 150 ppm, moderately infertile at 150-250 ppm, and fertile at over 250 ppm. In addition to nutrients, soil fertility can be assessed by the soil's acidity level (pH) (Neina, 2019); an ideal soil pH is between 6 and 7.5. Soil electrical conductivity (EC) does not directly determine soil fertility but is related to the concentration of elements and ions in the soil (Simanjuntak et al., 2026).

Previous research on the classification of soil conditions using NPK, pH and soil moisture parameters, employing GBM and Light GBM machine learning models, achieved accuracies of 96.14 per cent and 96.82 per cent respectively. The use of ANN machine learning for predicting soil nutrient content has also been investigated, yielding an R^2 value of 0.96–1 for NPK content (Witaningsih, Sulistiyanti S.R., 2025). The CatBoost model has been applied as a crop water monitor to optimise wheat irrigation, yielding excellent performance with an R^2 of 0.992 (Bingyan Dong, Shouchen Ma, 2025)(Bingyan Dong, Shouchen Ma, 2025).

An Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) refers to a computational technique designed to mimic the human biological nervous system. The architecture of an ANN comprises of nodes or neurons organized into three major layers; input layer, hidden layer and the output layer. Weights and biases are the connections between neurons and they act as storage devices in the ANNs. ANN techniques have been researched to detect complex non-linear patterns hence they are widely used in classification and regression techniques. Learning in ANNs occurs through the forward propagation and backpropagation (LeCun et al., 2015). In the forward

propagation phase, the data passes through all the layers to produce an output. In the model, the error is computed using the loss function and it is backpropagated to the previous layer for the adjustment of the weight parameters through the gradient descent algorithm. In this model, the activation functions, such as the non-linear sigmoid or ReLU functions make sure that the model can detect complex patterns. The factors affecting the rate of convergence in this model include learning rate, the number of epochs and batch size. The metrics used to evaluate classification include Accuracy, Precision, Recall and F1-score. In ANN models, a confusion matrix can also be calculated by comparing actual data with predictions (Ramzan et al., 2024).

Gradient Boosting Machine (GBM)

Method for classification of objects by creating a predictive model using decision trees in tandem. But unlike Random Forest that creates decision trees independently, GBM creates them sequentially in such a way that each tree improves the error prediction to get a better result (Friedman, 2001). The trees already created do not change but rather act as an error basis and new trees are formed according to the error of the previous one. This goes on until the condition for termination is fulfilled. GBM works on the principle of gradient descent by using a loss function for optimizing the model direction. The creation of new decision trees depends on the average of previous learning models until the model reaches its optimum level and manages missing values (Qwaide et al., 2026).

CatBoost (Categorical Boosting)

CatBoost is a computational model that applies the gradient boosting algorithm. This model uses ordered boosting and statistical coding methods to avoid overfitting problems (Prokhorenkova et al., 2018). This model is intended to increase prediction accuracy. CatBoost builds a symmetric tree structure with the intention of reducing memory consumption. The steps in the process of building CatBoost are as follows: the first step is the initialization of the model; the second step is determining the loss function to show prediction errors; the third step is calculating the value of the negative gradient of the loss function in order to find the right direction for model decision-making by means of ordered boosting method, learning from the previous model; the fourth step is constructing decision trees symmetrically (Hancock & Khoshgoftaar, 2020).

Model Evaluation

The measures of accuracy (1), precision (2), recall, F-measure (4), confusion matrix (5) and area under the curve (AUC) (6.b) (Sokolova & Lapalme, 2009) are computed in the process of data analysis. The variables considered here are as follows: true positive (TP), false

positive (FP), true negative (TN), false negative (FN), true positive rate (TPR), and false positive rate (FPR).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$Confusion\ Matrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (5)$$

$$TPR = \frac{TP}{TP + FN} \quad , \quad FPR = \frac{FP}{FP + TN} \quad (6.a)$$

$$AUC = \int_0^1 TPR(FPR) d(FPR) \quad (6.b)$$

(Latuconsina & Rahardi, 2026)

METHOD

This research adopted the use of data collected using IoT-based NPK meter. The samples measured. The parameters that were considered to measure soil fertility included nitrogen (N), phosphorus (P), potassium (K), soil temperature, humidity, pH, and soil conductivity. A total of 91 data samples were used in this research. The data adopted had been subjected to data preparation process. The data included three categories of soil fertility: 20 data samples for ‘infertile’ soil, 35 data samples for ‘poorly fertile’ soil, and 36 data samples for ‘fertile’ soil. There was a total of 91 data samples. The classes were categorized based on the standard set by the Ministry of Agriculture of the Republic of Indonesia. The classes included seven parameters. The class data formed the target data (Y). On the other hand, the prediction data involved seven parameters including nitrogen (N) content, phosphorus (P), potassium (K), soil temperature, soil moisture, pH, and soil conductivity.

Data splitting in this study utilised the hold-out method, with a training-to-test data ratio of 90:10, comprising 81 training data points and 10 test data points. The next stage involved data standardisation. The data was standardised using the StandardScaler normalisation process (Pedregosa et al., 2011). Labelling and encoding were then carried out, after which the data was tested using ANN, GBM and CatBoost models. The models were built using the Python programming language. The models were evaluated by making decisions based on patterns in the training data from the seven predictors.

Building the ANN Model

The artificial neural network (ANN) model was implemented using the TensorFlow-Keras library. It consists of a network architecture having one input layer that is made up of 7 neurons, representing the 7 attributes. The model contains 2 hidden layers with ReLU activation functions used to learn the non-linear relationships, apart from one output layer that has 3 neurons. The soft-max activation function is used at the output layer in order to classify the three soil fertility classes. In terms of training, the model uses the Adam optimizer along with categorical cross-entropy loss function for 100 epochs. Building the GBM Model.

Building the GBM Model

This model is generated with the GradientBoostingClassifier algorithm of the Scikit-learn library (Pedregosa et al., 2011; Friedman, 2001). This model generates decision trees in iterations. Every newly generated decision tree improves the prediction mistakes of its predecessor. This model uses `n_estimators` equal to 200 decision trees, with a learning rate of 0.1, `max_depth` equal to 3 and `random_state` is 42. Once the model is generated, the training and testing phases of the model takes place. In this phase, the correlation between the number of estimators and accuracy of the training and test data is measured. The next step for the model is the evaluation phase.

Building the CatBoost Model

The model is developed using the CatBoost Classifier method using the CatBoost library with the help of Ordered Boosting technique (Prokhorenkova et al., 2018). In the construction of decision trees in this model, there are 200 iterations, the learning rate of 0.1, tree depth of 6, MultiClass loss function, and random state equal to 42. To evaluate the correlation between the number of estimators and the accuracy of the training and test data, the process of model training is conducted. The model is trained using the training and test data.

Model Evaluation

The next step after creating the three models was their evaluation on the basis of training and test data. As part of the training procedure, the interaction between the number of epochs and training and validation accuracy was examined. The performance of the model was determined based on such parameters as Accuracy, Precision, Recall, and F1-Score. At the same time, a Classification Report was compiled for each class. Then, the Confusion Matrix was computed in order to evaluate the performance of soil fertility classification. The model's discrimination power was assessed based on the Receiver Operating Characteristic (ROC) Curve with One-vs-Rest (OvR) approach for multi-class classification (Hand & Till,

2001). For each class, the Area Under the Curve (AUC) was also calculated and subsequently averaged on the basis of the Macro Average AUC in order to estimate the capability of the model to differentiate the three soil fertility classes. Moreover, the importance of each variable N, P, K, pH, soil temperature, soil moisture and soil conductivity was evaluated based on Permutation Feature Importance (Fisher et al., 2019).

RESULTS AND DISCUSSION

A description of the dataset can be seen in Figure 1 and Table 1. Figure 1 shows a histogram of seven parameters nitrogen, phosphorus, potassium, pH, conductivity, soil temperature and soil moisture which exhibit varied patterns. The blue curve represents the density (Kernel Density Estimation) and the red dotted line indicates the mean. All parameters exhibit a multimodal distribution pattern. The multimodal distribution pattern indicates the presence of variations in soil characteristics among the observed samples. However, the distribution of individual features alone cannot directly determine soil fertility classes, as classification is performed based on the combination of multiple parameters using machine learning models. Therefore, distribution analysis was conducted as an initial step to understand dataset characteristics and potential patterns among different classes.

Nitrogen content ranges from 0 to 14,500 mg/kg, with an average of 6.383 mg/kg. Phosphorus content ranges from 0 to 38.3 mg/kg, with an average of 19.2 mg/kg. Potassium content ranges from 0 to 37.8 mg/kg, with an average of 18.5 mg/kg. The pH ranged from 3.6 to 9, with an average of 5.6. The measured soil conductivity ranged from 0 to 88.3 $\mu\text{S}/\text{cm}$, with an average of 48.5 $\mu\text{S}/\text{cm}$. Soil temperature ranged from 26.3 to 30.4 $^{\circ}\text{C}$, with an average of 27.75 $^{\circ}\text{C}$. Moisture content ranged from 9–57.4 per cent, with an average of 38.92 per cent. The dataset distribution shows multiple peaks in the N, P, K, pH, and electrical conductivity parameters. This pattern indicates the presence of different soil characteristic groups within the dataset. However, the relationship between individual feature distributions and fertility classes requires further evaluation through classification models. Similar research has also been conducted to measure NPK levels in rice plants, for N (0–5 mg/kg), P (0–3 mg/kg) K (0–4 mg/kg), as well as nitrogen (54 mg/kg), phosphorus (30 mg/kg) and potassium (40 mg/kg), which fall within the optimal range (Sinuraya et al., 2025). The relatively homogeneous temperature distribution indicates limited variability compared with other parameters. Therefore, the contribution of temperature to classification may depend more on its interaction with other features, such as moisture, EC, and nutrient content. The humidity data were predominantly in the average range, indicating that the measurements were taken under good conditions.

Table 1. Description of Soil Parameters

Statistic	Nitrogen (mg/kg)	Phosphorus (mg/kg)	Potassium (mg/kg)	pH	Electrical Conductivity ($\mu\text{S}/\text{cm}$)	Soil Temperature ($^{\circ}\text{C}$)	Soil Moisture (%)
Number of Samples	91	91	91	91	91	91	91
Mean	6.384	19.178	18.530	5.538	48.522	27.748	38.923
Standard Deviation	4.786	11.424	11.481	1.412	24.051	1.097	12.971
Min	0.000	0.000	0.000	3.600	0.000	26.300	9.000
25%	2.700	10.900	10.100	4.200	31.500	26.900	32.700
50%	5.400	17.200	16.500	5.800	44.600	27.400	40.800
75%	10.150	28.050	27.500	6.450	67.150	28.400	50.100
Max	14.500	38.300	37.800	9.000	88.300	30.400	57.400

Based on Table 1, each soil parameter exhibits different levels of variability. Nitrogen, phosphorus, potassium, and electrical conductivity show relatively higher standard deviations compared with soil temperature, indicating variations in soil chemical characteristics among samples. In contrast, soil temperature has the lowest standard deviation (1.097°C), suggesting relatively stable temperature conditions during measurement.

Liakos

Distribution Histograms of Soil Parameters

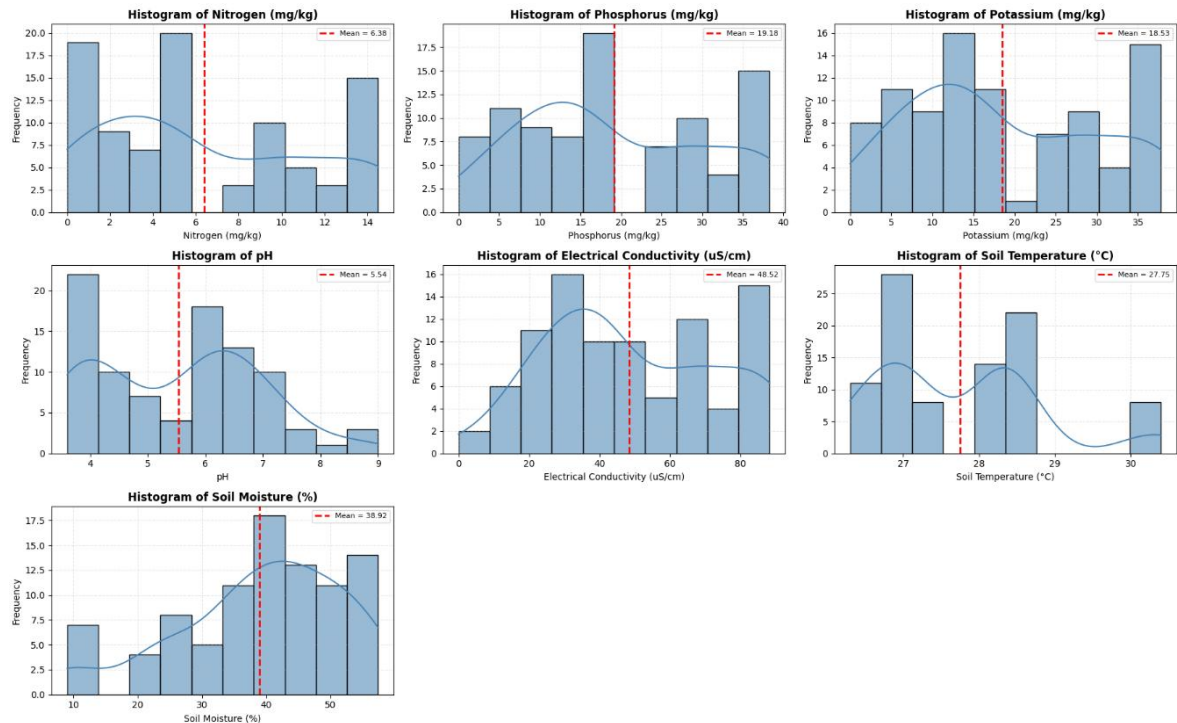


Figure 1. Distribution Histogram of Soil Parameters

The correlation between parameters (features) is shown in Figure 2. Analysis of the linear relationships between features reveals correlation coefficients ranging from -1 to 1. A

positive value close to 1 indicates a strong positive correlation; similarly, a negative value close to -1 indicates a strong inverse correlation, whilst a value of 0 indicates no correlation. The findings indicate that the correlation coefficients for N, P, K and electrical conductivity show a strong positive correlation ranging from 0.99 to 1. This indicates that NPK macronutrient levels show a strong positive association with soil electrical conductivity in the analyzed dataset. This relationship may occur because higher concentrations of dissolved ions in soil can increase its ability to conduct electrical current. This is because electrical conductivity measures the soil's ability to conduct an electric current carried by ions (Sulaiman et al., 2021; Corwin & Lesch, 2005). Meanwhile, pH exhibits a strong negative correlation ranging from -0.65 to -0.68 with nutrients N, P, K and soil conductivity. The negative correlation between pH and N, P, K, and EC indicates that samples with higher nutrient-related values tend to exhibit lower pH values in this dataset. This represents a statistical association among parameters and does not necessarily imply a causal relationship (Kp & Prasanna, 2025). Temperature has a strong negative correlation (-0.73 to -0.75) with nutrient content and a correlation of -0.76 with electrical conductivity. The higher the temperature, the lower the nutrient content and electrical conductivity. The relationship between temperature and pH shows a strong positive correlation of 0.74. Meanwhile, the correlation between moisture and other features has a correlation value close to 0, meaning there is no relationship with other features.

The distribution of parameters according to the 'infertile', 'semi-fertile' and 'fertile' classes is shown in Figure 3. In the figure, the three data groups can be clearly distinguished. In the 'infertile' class, the median NPK values and conductivity were the lowest compared with the 'semi-fertile' and 'fertile' classes. The highest median NPK content and conductivity values were found in the 'fertile' class. The infertile class has a fairly wide range of soil moisture, indicating that soil moisture conditions in infertile soil vary between 9 and 54 per cent at a temperature of around 29.09. In the moderately fertile class, the range of soil moisture is also relatively wide, at around 23.4–57.4 per cent at a temperature of around 27.75°C. Meanwhile, in the fertile class, the moisture content has a relatively narrow range of 32.4–44.8 per cent at a temperature of 27°C. In infertile conditions, soil pH ranges from 6.3 to 9; in less fertile conditions, pH ranges from 4.2 to 8; whilst in fertile conditions, it ranges from 3.6 to 7. This is consistent with previous research indicating the existence of fertility categories (Malo, 2025).

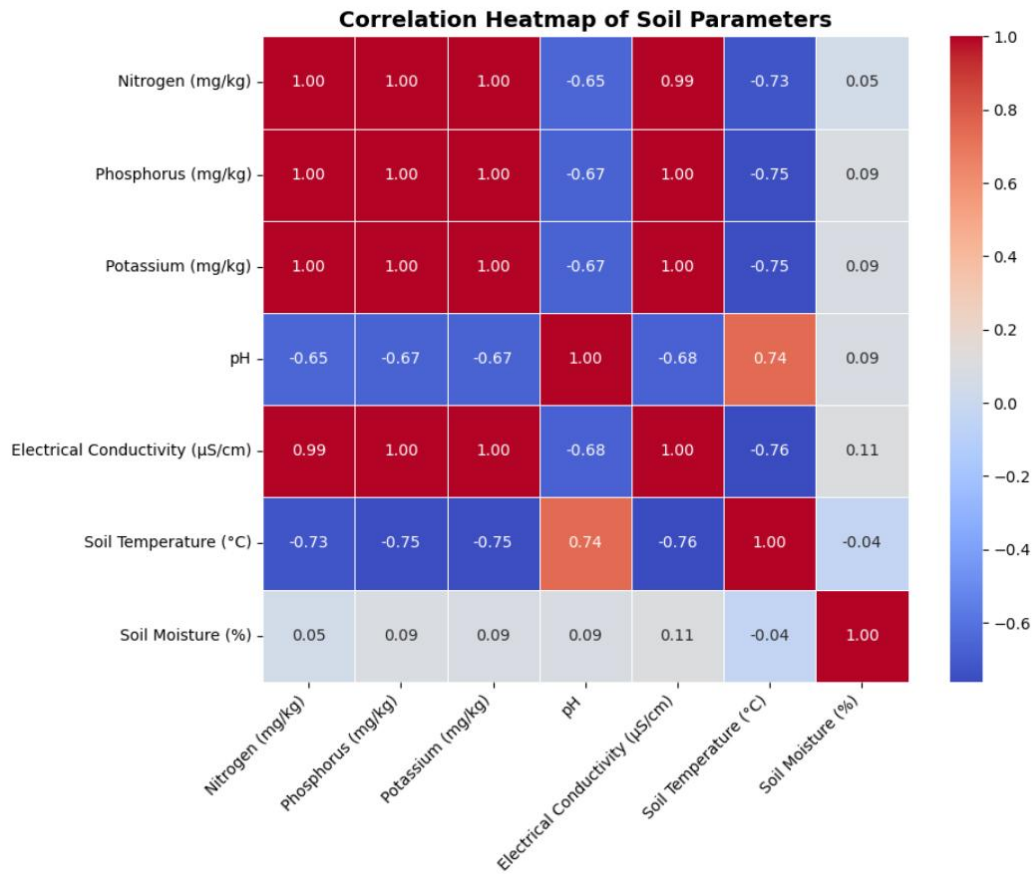


Figure 2. The Linear Relationships of Soil Parameters

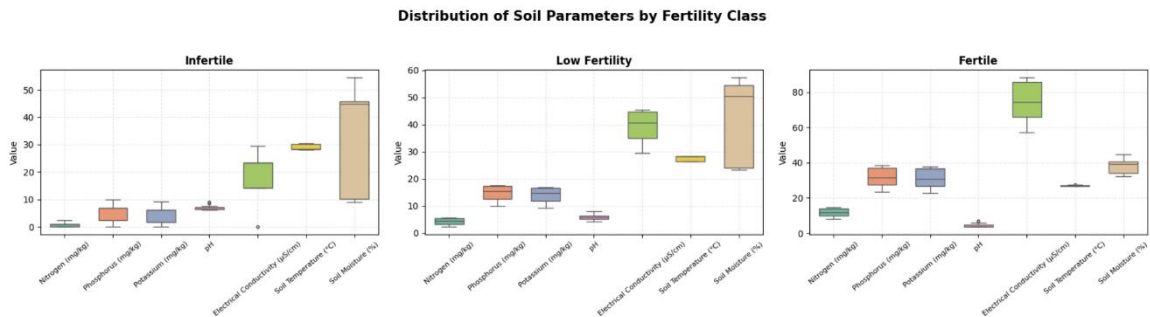


Figure 3. Distribution of Soil Parameters by Fertility Class

Figure 4 depicts an analysis on the relationship between the number of epochs and the estimator concerning accuracy. In Figure 4.a, it can be observed that there is an increase in accuracy for training as well as test data sets in case of ANN. It can be seen that as the number of epochs increases from 0 to 20, there is an increase in accuracy from about 0.35 to 0.99 at 20. As the number of epochs increases for training data, there is stabilization of accuracy at 1, whereas in case of test data, the accuracy tends to stabilize at 0.99 for epochs greater than 20. Similar results were obtained when number of epochs was tested for training as well as validation data (Uribe-tarazona & Ruiz-diaz, 2025). The relationship between the estimator and accuracy for the GBM model (Figure 4b) and for the CatBoost model (4.c) shows that changes in the estimator do not affect accuracy. The GBM model exhibits stable

accuracy values of 1 for the training data and 0.9 for the test data. Similarly, stable accuracy values of 0.9 were obtained for the CatBoost data.

The confusion matrix results for the ANN, GBM and CatBoosting algorithms are shown in Figure 5. A confusion matrix can be used to evaluate a model's performance in classifying plant classes by comparing actual data with predicted data (Sokolova & Lapalme, 2009). The ANN, GBM and CatBoosting models were able to classify the training data accurately. Figures 2a, 2c and 2e show that all training data values on the diagonal of the confusion matrix indicate there are 91 data points for each soil class, and all elements outside the main diagonal have a value of 0 for the GBM and CatBoosting models. This indicates that there were no errors in the training classification process. Consequently, the Accuracy, Precision, Recall and F1-Score values are all 1.00 (100 per cent). In contrast, for the ANN model, there are 90 data points on the diagonal and one outlier, i.e. FN = 1. As for the test data, as shown in Figure 2b for the ANN test data model, there are 10 data points on the main diagonal, indicating that there were no errors in the test classification process. This achieved accuracy, precision, recall and F1-score values of 1.00 (100 per cent) for the test data. In contrast, for the GBM and CatBoosting models, there were still errors in the test data; no false positives (FP) were found in the test data for the GBM and CatBoosting models.

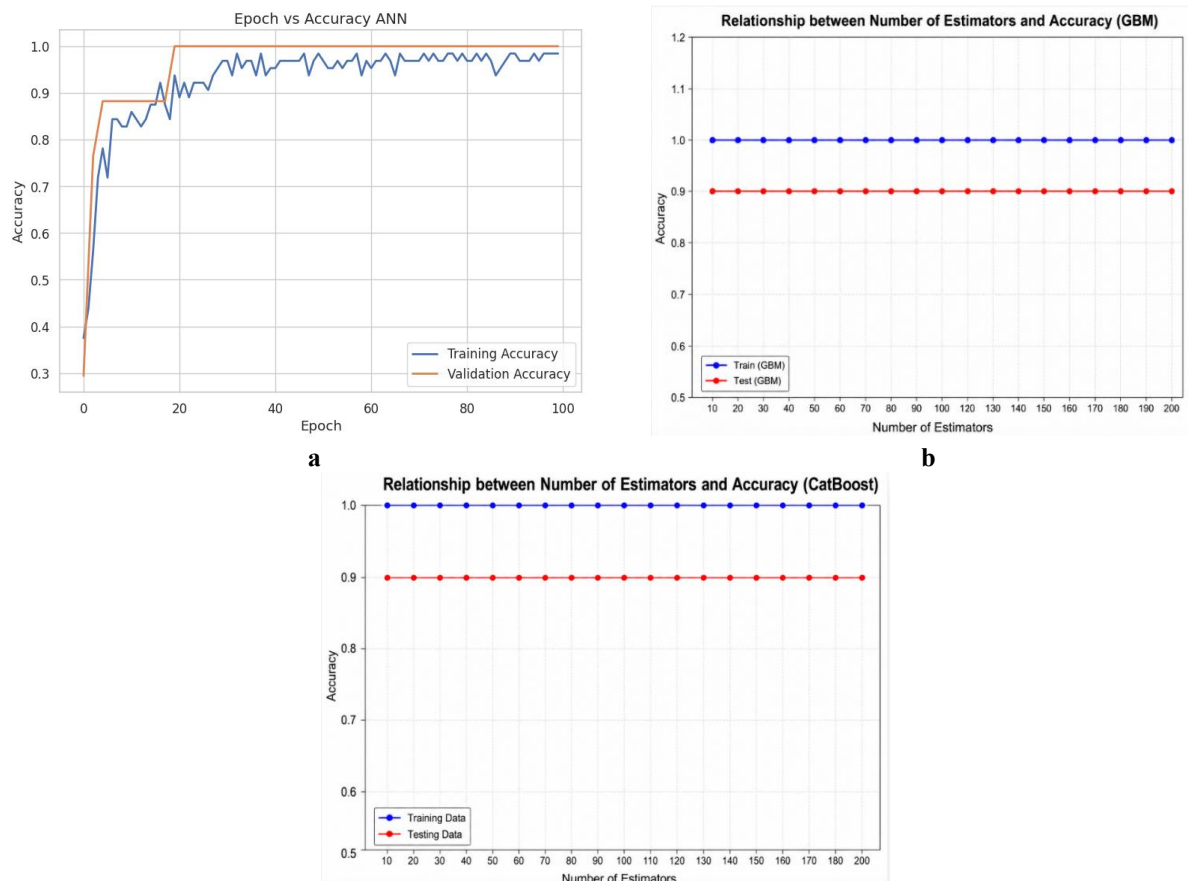


Figure 4: Relationship between Epochs and Estimators and Accuracy

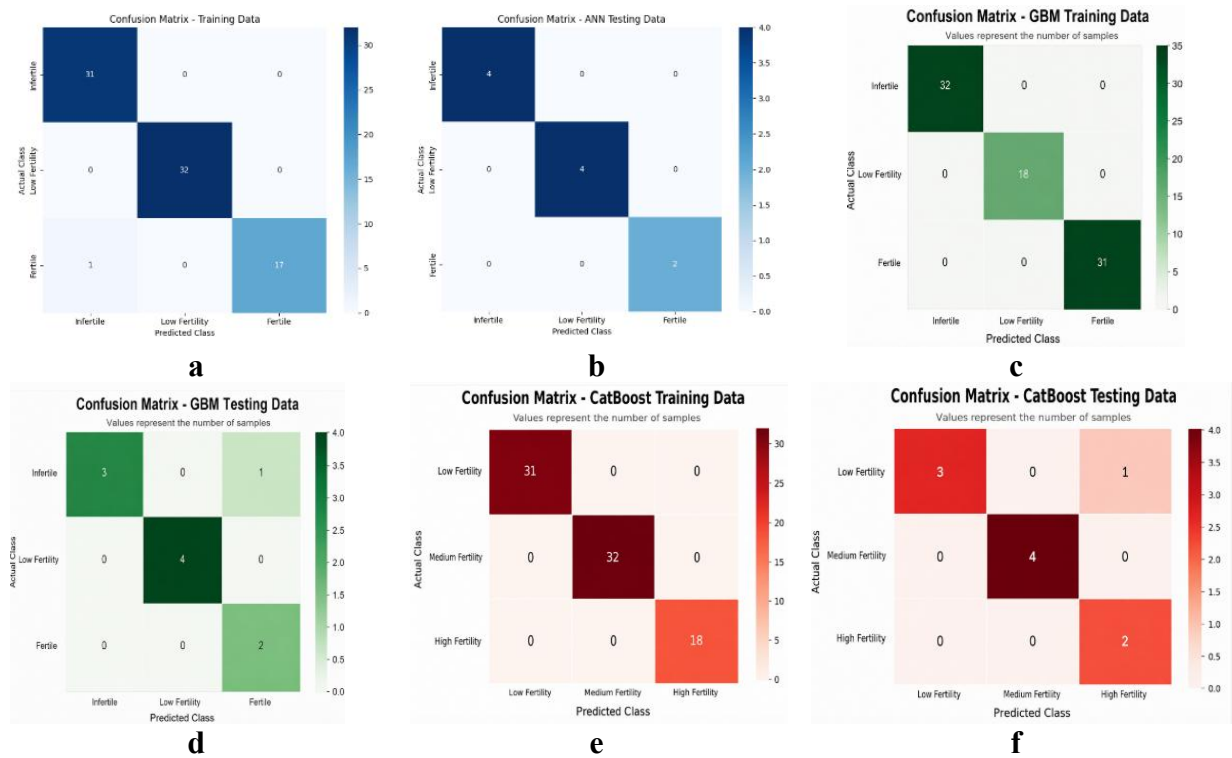
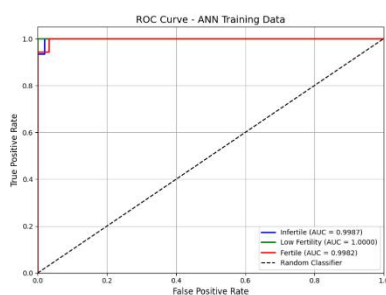


Figure 5. Confusion matrix for a. ANN training data, b. ANN test data, c. GBM training data, d. GBM test data, e. CatBoosting training data and f. CatBoosting test data

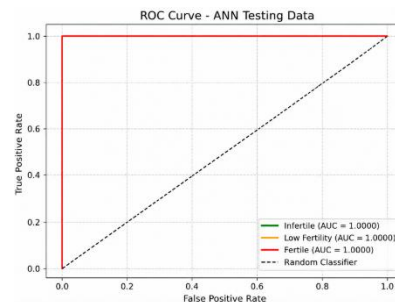
The evaluation results are clearly shown in Table 2. The findings of this study indicate that all training and test data can be classified accurately, with the highest accuracy achieved by the GBM and CatBoosting models, which recorded accuracy, precision, recall and F1-score values of 1.00 (100%). Meanwhile, the ANN model achieved an accuracy of 0.98 (98%). As for the test data, the ANN model achieved an accuracy where the distribution of NPK parameters and environmental factors could clearly distinguish between classifications. The parameters N, P, K, pH, temperature, humidity and rainfall are effective features in classifying soil fertility suitability. An evaluation of the ANN, GBM and CatBoost models shows that they can distinguish between the classifications of each plant; the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) is illustrated by the ROC curves, and the Area Under the Curve (AUC) values can be seen in Figures 3.a, b, c, d, e and f. In these figures, the ROC curves are situated in the top-left corner of the graph for all plant classes; this indicates a very high True Positive Rate, with an AUC value of 1.00 for each class and an average micro-AUC of 1.00. This demonstrates that all models can distinguish between classes perfectly. These AUC values are consistent and align with the results for Accuracy, Precision, Recall and F1-Score, all of which are 1.00 (100 per cent). These three models have also been tested and found to have good accuracy (Ileri, 2025).

Table 2. Model Evaluation

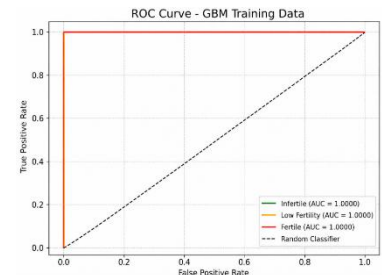
Model	Dataset	Class	Accuracy	Precision	Recall	F1-Score	Support
ANN	Training Data	Infertile		0.97	0.97	0.97	31
ANN	Training Data	Low Fertility		1	1	1	32
ANN	Training Data	Fertile		0.94	0.94	0.94	18
ANN	Training Data	Macro Avg	0.98	0.98	0.97	0.97	81
ANN	Training Data	Weighted Avg	0.98	0.98	0.98	0.98	81
ANN	Testing Data	Infertile		1	1	1	4
ANN	Testing Data	Low Fertility		1	1	1	4
ANN	Testing Data	Fertile		1	1	1	2
ANN	Testing Data	Macro Avg	1	1	1	1	10
ANN	Testing Data	Weighted Avg	1	1	1	1	10
GBM	Training Data	Infertile		1	1	1	31
GBM	Training Data	Low Fertility		1	1	1	32
GBM	Training Data	Fertile		1	1	1	18
GBM	Training Data	Macro Avg	1	1	1	1	81
GBM	Training Data	Weighted Avg	1	1	1	1	81
GBM	Testing Data	Infertile		1	0.75	0.86	4
GBM	Testing Data	Low Fertility		1	1	1	4
GBM	Testing Data	Fertile		0.67	1	0.8	2
GBM	Testing Data	Macro Avg	0.89	0.89	0.92	0.89	10
GBM	Testing Data	Weighted Avg	0.93	0.93	0.9	0.9	10
CatBoost	Training Data	Infertile		1	1	1	31
CatBoost	Training Data	Low Fertility		1	1	1	32
CatBoost	Training Data	Fertile		1	1	1	18
CatBoost	Training Data	Macro Avg	1	1	1	1	81
CatBoost	Training Data	Weighted Avg	1	1	1	1	81
CatBoost	Testing Data	Infertile		1	1	1	4
CatBoost	Testing Data	Low Fertility		1	1	1	4
CatBoost	Testing Data	Fertile		1	1	1	2
CatBoost	Testing Data	Macro Avg	1	1	1	1	10
CatBoost	Testing Data	Weighted Avg	1	1	1	1	10



Macro Average AUC : 0.999



Macro Average AUC: : 1.000



Macro Average AUC : 1.000

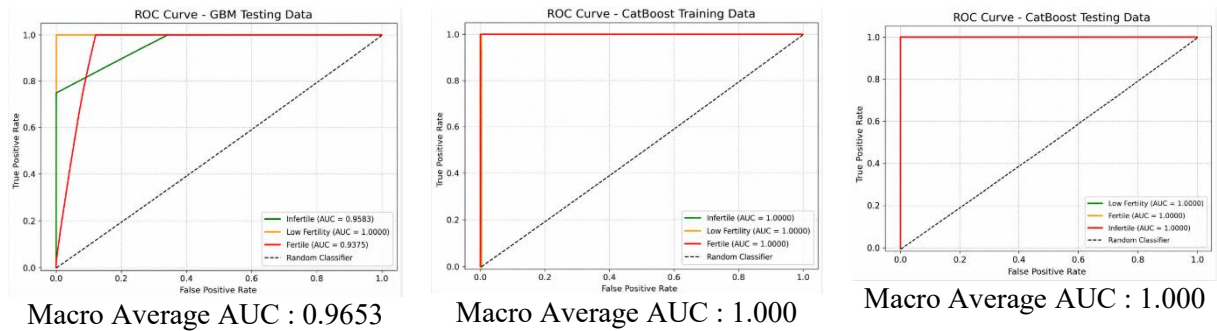


Figure 6. ROC Curves for the Training and Test Data for the ANN, GBM and CatBoost models

Based on the results of the analysis of contributing features from the three models shown Table 3. it was found that soil moisture, nitrogen, phosphorus, potassium, and EC are features that influence soil fertility levels; however, pH is a feature that has no influence, as can be clearly seen in Table 3. Each algorithm exhibits differences in predicting soil classification (Yulianto & Astuti, 2026).

Table 3. Feature Contributions

ANN			GBM		CatBoosting	
Index	Feature	Importance	Feature	Importance	Feature	Importance
0	Soil Temperature	0.035	pH	0.000	Soil Temperature	2.643
1	Nitrogen	0.045	Potassium	0.139	pH	3.273
2	Phosphorus	0.080	Phosphorus	0.135	Soil Moisture	4.923
3	pH	0.100	Soil Temperature	0.165	Temperature	6.927
4	Potassium	0.105	Nitrogen	0.152	Phosphorus	24.368
5	EC	0.115	EC	0.183	Potassium	25.832
6	Soil Moisture	0.185	Soil Moisture	0.185	Nitrogen	29.359

CONCLUSION

Based on the research results, it can be concluded that the ANN, GBM, and CatBoosting models demonstrate good performance in classifying soil fertility. Each model exhibits a high level of accuracy, ranging from 0.9 to 1, with soil moisture and nitrogen identified as the most influential features across the models.

SUGGESTIONS

Future research should utilize a larger and more diverse dataset. It is also necessary to incorporate additional measurement parameters, such as plant species, climate, soil type, and various physical and chemical properties. While the current model performs well, its performance could potentially be enhanced through a hybrid modeling approach.

ACKNOWLEDGEMENT

This research was sponsored by UIN Syarif Hidayatullah Jakarta using BOPTN funds for the 2024 fiscal year. Acknowledgements are extended to the Research Institute of UIN Syarif Hidayatullah Jakarta and the leadership of UIN Syarif Hidayatullah Jakarta for the support provided during the conduct of this research.

REFERENCES

- Anugraha, Y. S., Yenni, H., Agustin, W., & Asnal, H. (2026). Model Prediksi Produktivitas Padi Menggunakan Xgboost Dan Random Forest. *Jurnal Sistem Informasi Dan Informatika (Simika)*, 9(1), 67–78. <https://doi.org/10.47080/simika.v9i1.4169>
- Bingyan Dong , Shouchen Ma, Z. G. and A. Q. (2025). CatBoost Improves Inversion Accuracy of Plant Water Status in Winter Wheat Using Ratio Vegetation Index. *Applied Sciences*, 15, 11363, 1–20. <https://doi.org/https://doi.org/10.3390/app152111363>
- Budiman, L., & Suhendi, D. (2024). Resiliensi Penguatan Ketahanan Pangan Daerah Di Indonesia. *Jurnal Perlindungan Masyarakat: Bestuur Praesidium*, 1(2), 63–71. <https://ejournal.ipdn.ac.id/index.php/jpa/article/view/5113>
- Castillo-Girones, S., Munera, S., Martínez-Sober, M., Blasco, J., Cubero, S., & Gómez-Sanchis, J. (2025). Artificial Neural Networks in Agriculture, the core of artificial intelligence: What, When, and Why. *Computers and Electronics in Agriculture*, 230(October 2024). <https://doi.org/10.1016/j.compag.2025.109938>
- Corwin, D. L., & Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture*, 46(1–3), 11–43. <https://doi.org/10.1016/j.compag.2004.10.005>
- Dattatreya, S., Khan, A. N., Jena, K., & Chatterjee, G. (2024). Conventional to Modern Methods of Soil NPK Sensing: A Review. *IEEE Sensors Journal*, 24(3), 2367–2380. <https://doi.org/10.1109/JSEN.2023.3334243>
- Ennaji, O., Baha, S., Vergutz, L., & El Allali, A. (2024). Gradient boosting for yield prediction of elite maize hybrid ZhengDan 958. *PLoS ONE*, 19(12), 1–16. <https://doi.org/10.1371/journal.pone.0315493>
- Fadly, R. (2026). Perbandingan Algoritma Catboost dan Extra Trees Classifier Untuk Prediksi Tingkat Keberhasilan Panen Padi Berdasarkan Faktor Cuaca dan Kondisi Tanah Rio Fadly. 4(1), 430–444. <https://ejournal.edutechjaya.com/index.php/komprehensif> Perbandingan
- Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81. <https://jmlr.org/papers/v20/18-760.html>
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, 7, 94. <https://doi.org/10.1186/s40537-020-00369-8>

- Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2), 171–186. <https://doi.org/10.1023/A:1010920819831>
- Ileri, K. (2025). Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks. *International Journal of Machine Learning and Cybernetics*, 16(9), 6937–6956. <https://doi.org/10.1007/s13042-025-02654-5>
- Kabato, W., Hailegnaw, N., Mutum, L., & Molnar, Z. (2025). Managing soil health for climate resilience and crop productivity in a changing environment. *Science of the Total Environment*, 1000(May), 180460. <https://doi.org/10.1016/j.scitotenv.2025.180460>
- Kp, Y., & Prasanna, S. M. (2025). *Correlation and regression analysis of key soil fertility parameters in agricultural soils*. 7(6), 26–30.
- Latuconsina, M. S., & Rahardi, M. (2026). Comparison of LightGBM and CatBoost Algorithms for Diabetes Prediction Based on Clinical Data. *Journal of Applied Informatics and Computing*, 10(1), 1058–1065. <https://doi.org/10.30871/jaic.v10i1.12179>
- Khechba, K., Laamrani, A., Dhiba, D., Misbah, K., & Chehbouni, A. (2021). Monitoring and Analyzing Yield Gap in Africa through Soil Attribute Best Management Using Remote Sensing Approaches: A Review. *Remote Sensing*, 13(22), 4602. <https://doi.org/10.3390/rs13224602>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, 18(8), 2674. <https://doi.org/10.3390/s18082674>
- Mahmood, T., Usman, M., & Conrad, C. (2025). Selecting Relevant Features for Random Forest-Based Crop Type Classifications by Spatial Assessments of Backward Feature Reduction. *PFG - Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 93(2), 173–196. <https://doi.org/10.1007/s41064-024-00329-4>
- Malo, S. K. (2025). Assessment of the soil fertility index (SFI) for sustainable nutrient management in rice- based cropping systems of Barind tract of West. *Discov. Soil* 2, 96 (2025). <https://doi.org/10.1007/s44378-025-00130-8>
- Najdenko, E., Lorenz, F., Dittert, K., & Olf, H.-W. (2024). Rapid in-field soil analysis of plant-available nutrients and pH for precision agriculture—a review. *Precision Agriculture*, 25(6), 3189–3218. <https://doi.org/10.1007/s11119-024-10181-6>
- Neina, D. (2019). The Role of Soil pH in Plant Nutrition and Soil Remediation. *Applied and Environmental Soil Science*, 2019, 5794869. <https://doi.org/10.1155/2019/5794869>
- Nila Aulia, Muhammad Rizky, & Chaerul Shaleh. (2025). Memprediksi Jenis Tanaman Berdasarkan Parameter Lingkungan Menggunakan Algoritma Decision Tree dan Random Forest Berbasis Web. *Jurnal Teknik Informatika Dan Teknologi Informasi*, 5(3), 578–596. <https://doi.org/10.55606/jutiti.v5i3.6435>

- Oktapiani, V., & Agustin, Y. H. (2024). Implementation of Naïve Bayes Algorithm to Predict Food Crop Production Results in Garut Regency. *Jurnal Sistem Cerdas*, 7(3), 366–376. <https://doi.org/10.37396/jsc.v7i3.455>
- Panigrahi, B., Kathala, K. C. R., & Sujatha, M. (2022). A Machine Learning-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning with Regression Models. *Procedia Computer Science*, 218(2022), 2684–2693. <https://doi.org/10.1016/j.procs.2023.01.241>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.
- Qwaid, M. A., Sarker, M. T., Shawon, S. M., & Zubair, H. T. (2026). AI-enabled smart farming framework for sustainable date palm cultivation in arid regions using machine learning and IoT integration. *Scientific Reports*, 16(1). <https://doi.org/10.1038/s41598-026-36106-z>
- Rachakonda, L., & Stasiewicz, S. (2024). SanaSolo 2.0: Edge-Based Monitoring and Management of Soil Fertility Using IoT. In *2024 IFIP/IEEE 32nd International Conference on Very Large Scale Integration (VLSI-SoC)*. <https://doi.org/10.1109/VLSI-SoC62099.2024.10767818>
- Ramzan, S., Ali, B., Raza, A., Hussain, I., Fitriyani, N. L., Gu, Y., & Syafrudin, M. (2024). An innovative artificial neural network model for smart crop prediction using sensory network based soil data. *PeerJ Computer Science*, 10, 1–28. <https://doi.org/10.7717/peerj-cs.2478>
- Reza, M. N., Lee, K.-H., Karim, M. R., Haque, M. A., Bicomakuba, E., Dey, P. K., Jang, Y. Y., & Chung, S.-O. (2025). Trends of Soil and Solution Nutrient Sensing for Open Field and Hydroponic Cultivation in Facilitated Smart Agriculture. *Sensors*, 25(2), 453. <https://doi.org/10.3390/s25020453>
- Sharma, A. K., & Tomar, T. S. (2025). Influence of Integrated Nutrient Management on Growth, Yield and Quality of Wheat (*Triticum aestivum* L.) in Uttar Pradesh, India. *Journal of Experimental Agriculture International*, 47(7), 301–312. <https://doi.org/10.9734/jeai/2025/v47i73569>
- Shin, S. K., Lee, S. J., & Park, J. H. (2025). Prediction of Soil Properties Using Vis-NIR Spectroscopy Combined with Machine Learning: A Review. *Sensors*, 25(16). <https://doi.org/10.3390/s25165045>
- Simanjuntak, R. R., Manurung, S., & Djaingsastro, A. J. (2026). Analisis Kandungan N, P, K, dan C-Organik pada Tanah Datar dan Tanah Miring di Kebun Silau Dunia Afdeling III PT Perkebunan Nusantara IV Regional I. *Ranah Research : Journal of Multidisciplinary Research and Development*, 8(3), 1951–1961. <https://doi.org/10.38035/rj.v8i3.2112>

- Sinuraya, J., Weno Syechu, Wiwin Sry Adinda Banjarnahor, & Muhammad Riki Atsauri. (2025). Penerapan Aplikasi Sistem Rekomendasi Pemilihan Varietas Tanaman Berbasis Parameter Kelembaban Tanah, pH, dan Kandungan Nutrisi Pada Gapoktan Sri Karya Menggunakan Algoritma Random Forest. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD)*, 4(6), 1393–1402. <https://doi.org/10.53513/jursi.v4i6.12348>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Sujatha, M., & Jaidhar, C. D. (2024). Machine learning-based approaches to enhance the soil fertility—A review. *Expert Systems with Applications*, 240, 122557. <https://doi.org/10.1016/j.eswa.2023.122557>
- Sulaiman, A. S. S., Ahmad, M. A., Hassim, S. A., & Azman, M. S. A. (2021). Evaluation of fertilizer electrical conductivity (Ec) and temperature distribution via vertical farming system under plant factory. *Basrah Journal of Agricultural Sciences*, 34(special issue 1), 63–72. <https://doi.org/10.37077/25200860.2021.34.SP1.7>
- Tohari, H., Harini, S., Yaqin, M. A., Santoso, I. B., & Crysdian, C. (2023). Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Produktivitas Padi. *Journal of Computer System and Informatics (JoSYC)*, 5(1), 175–183. <https://doi.org/10.47065/josyc.v5i1.4538>
- Toselli, M., Baldi, E., Ferro, F., Rossi, S., & Cillis, D. (2023). Smart Farming Tool for Monitoring Nutrients in Soil and Plants for Precise Fertilization. *Horticulturae*, 9(9), 1011. <https://doi.org/10.3390/horticulturae9091011>
- Uribe-tarazona, D. Y., & Ruiz-diaz, C. M. (2025). Machine learning technique for the identification of two-phase (oil-water) flow patterns through pipelines. (July), 1–18. <https://doi.org/10.3389/fmech.2025.1522120>
- Witaningsih, Sulistiyanti S.R., D. (2025). Artificial Neural Network Backpropagation Method for Predicting Soil Nutrient Content. *Jurnal Teknik Pertanian Lampung*, 14 o 6. <https://jurnal.fp.unila.ac.id/index.php/JTP/article/view/11460/6766>
- Yulianto, P. R., & Astuti, Y. P. (2026). Perbandingan Kinerja Algoritma CatBoost, XGBoost, LightGBM dan Random Forest Dalam Memprediksi Risiko Infeksi Aids Dalam Dataset Kesehatan. *Technology and Science (BITS)*, 7(4), 2383–2393. <https://doi.org/10.47065/bits.v7i4.8975>
- Yuliyanto, Y., Sahibu, S., Imran, T., Arisha, A. O., & Munawirah, M. (2025). Crop Recommendation Based on Soil and Weather Conditions Using the K-Nearest Neighbors Algorithm. *Journal of System and Computer Engineering (JSCE)*, 6(3), 211–219. <https://doi.org/10.61628/jsce.v6i3.1955>