ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Machine Learning Model for Predicting the Suitability of Cultivating Alternative Crops in Lower Northern Thailand

Sujitranan Mungklachaiya[1] and Anongporn Salaiwarakul[2]

ABSTRACT

Intensive rice cultivation presents significant environmental and economic challenges. While crop diversification offers potential benefits for agricultural sustainability and financial resilience, farmers face considerable uncertainty when transitioning to alternative crops. This study assessed the prediction efficacy of machine learning (ML) models in identifying suitable crops for cultivation in a specific geographical area considering various factors influencing agricultural viability. Through comprehensive experimentation, a decision tree model, an artificial neural network (ANN), and a Naïve Bayes model were used for predictions and rigorously evaluated for various crops, including rubber, coconut, longan, durian, rambutan, and mangosteen. Various hyperparameter configurations were tested, and multiple evaluation indicators were employed to assess the prediction performance of the models. The results consistently demonstrated the superiority of the decision tree model, which exhibited high accuracy, precision, recall, and F-measure across most crops. Its ability to capture intricate patterns and relationships between crop attributes and suitability levels underscores its value as a decision-support tool in agriculture. While the ANN model performed well for coconut, its effectiveness varied across the other crops, highlighting the need for tailored model selection. This study provides valuable insights into the application of ML in agricultural decision-making processes, suggesting potential avenues for future optimization and enhancement of prediction accuracy.

Article information:

Keywords: Crop Prediction Model, Alternative Crops, Decision Tree, ANN, Naïve Byes, Machine Learning

Article history:
Received: July 25, 2024
Revised: November 21, 2024
Accepted: February 6, 2025
Published: March 15, 2025
(Online)

DOI: 10.37936/ecti-cit.2025192.257686

1. INTRODUCTION

Regional variations in agricultural practices are shaped by local environmental conditions. Farmers select crop varieties based on these conditions, often following conventional cropping patterns and methods. Efficient crop production systems require a holistic approach that accounts for multiple interconnected factors affecting agricultural productivity. Successful agricultural planning depends on a comprehensive understanding of the complex interactions among climatic conditions, soil properties, market dynamics, and biotic influences. The optimal approach to crop cultivation is inherently complex, requiring careful consideration of environmental, economic, and agronomic factors. Numerous variables influence the decision-making process of cultivating specific crops.

In addition to understanding the factors influencing crop selection, farmers must consider their vulnerability to economic volatility caused by fluctuations in prices, demand, and supply [1]. A successful harvest can lead to an oversupply, driving down prices and affecting farmer's decisions. The time interval between planting and harvest known as the lead time, is a critical yet uncontrollable factor.

Intensive rice production in many agricultural regions has resulted in significant challenges, including excessive water consumption, soil degradation, and increased vulnerability to climate change. Additionally, rising production costs and fluctuating market prices pose increasing threats to the economic via-

[1] The author is with the Program in Computer and Information Technology, Loei Rajabaht University, Muang, Loei, 42000, Thailand, Email: sujitranan.mun@lru.ac.th
[2] The author is with the Department of Computer Science and Information Technology, Naresuan University, Muang, Phitsanuloke, 65000, Thailand, Email: anongporns@nu.ac.th
[2] Corresponding author: anongporns@nu.ac.th

bility of rice cultivation. Alternative-crop farming promotes diversification, reducing the risk of market oversupply while enhancing agricultural sustainability and improving producer profitability. However, despite these benefits, many farmers hesitate to transition to alternative crops due to uncertainties regarding feasibility and potential success rates.

Promoting alternative crops and reducing dependency on rice can be facilitated through algorithmic methodologies and computer-based modeling systems, which provide a structured framework for optimizing agricultural-land-use transitions. This systematic approach enables a comprehensive analysis of key factors influencing agricultural production, including soil characteristics, water availability, and climate.

The integration of scientific technologies into crop cultivation is increasingly being explored. Advanced methodologies such as ML and data mining can be utilized to predict and recommend the most suitable crops for specific regions [2, 3]. By leveraging climate data, soil health management techniques, market analysis, and pest control strategies, farmers can optimize the decision-making process and achieve sustainable agricultural outcomes. Additionally, adapting farming operations to local conditions and incorporating traditional knowledge enhances the resilience of farming systems, improving their ability to withstand challenges and generate profits. Cultivating crops in unsuitable conditions often leads to reduced agricultural yields. Prediction models and decision-support technologies can help farmers navigate the complexities of modern agriculture, providing valuable insights to improve agricultural productivity and sustainability.

This article presents a prediction model that integrates multiple factors influencing agricultural production and recommends alternative crop options for a specific geographical region: the Lower Northern Province of Thailand. The primary objective of the research was to identify viable alternatives to rice, which is particularly susceptible to market surplus during periods of abundant harvests. To achieve the objective, alternative crops, including rubber, coconut, longan, durian, rambutan, and mangosteen, were examined. The key contributions of this research are as follows:

- This article introduces an advanced prediction model that integrates multiple critical agricultural parameters, including temperature, rainfall, wind speed, and soil characteristics. Further, limitations of previous research are overcome.
- The study rigorously evaluated optimal hyperparameters for individual and ensemble machine learning (ML) models to enable the accurate forecasting of the suitability levels of alternative-crop cultivation in the study region.

2. RELATED WORKS

The application of ML in agriculture has the potential to revolutionize crop cultivation by providing farmers with precise, data-driven recommendations tailored to their specific conditions. An ML-based recommender system can assist farmers in selecting the most suitable crops by considering various factors influencing agricultural productivity.

Various scientific methodologies have been explored for this purpose. Ontologies serve as structured knowledge bases that provide valuable insights to support cultivation decisions and agricultural recommendations related to crop varieties, growth cycles, and climatic conditions [4], and crop pests [5] that may lead to product losses are included in the information provided by ontologies. By incorporating multiple crop-growth-influencing factors, ontologies facilitate informed crop recommendations for specific regions.

Content-based recommendation systems have also been investigated [6], which use parameters such as soil pH, soil type, and mineral content to suggest suitable crops. While these systems can identify optimal crop-growing regions, they often do not provide recommendations tailored to individual landowners' conditions. To address this issue, improved recommendation systems should offer localized crop advice, enabling farmers to select crops best suited to their land. Additionally, various data mining techniques [7, 8] have been employed to enhance the accuracy and effectiveness of these recommendation systems.

In [9], the application of ML methods, including support vector machines (SVMs), random forest models, Gaussian Naïve Bayes models, and k-nearest neighbors (kNNs), was explored for crop selection and prediction. These algorithms were used to evaluate soil quality, water quality, and agro-climatic variables to optimize crop management.

In the context of expanding precision agriculture, which emphasizes "site-specific" farming, a system reported in [10] employed ML methods, including Naïve Bayes and kNN models, to predict crop yield. These models facilitated the identification of optimal crops based on site-specific factors, enhancing the precision and efficacy of recommendations. The study highlighted the importance of soil properties such as texture, pH, and water retention capacity in crop development. However, a lack of the proper integration of diverse data sources into a cohesive model limited the precision of crop recommendations.

To explore how to utilize ML models in agriculture in greater detail, the study [11] employed diverse datasets sourced from Kaggle, which included critical factors such as soil and climate conditions. This data variety was essential for effectively training ML models, as it captured key influences of the factors on crop growth and yield. The study evaluated several ML algorithms, including a linear regression model, Naïve Bayes model, lasso regression model, support vector regression model, decision tree model, random forest model, k-NN, and gradient boosted regression tree (GBRT) model. However, the performance of these algorithms varied significantly depending on the dataset's specific characteristics, potentially leading to different outcomes when applied to other datasets or agricultural contexts. The bagging technique was employed in the study to enhance prediction accuracy; however, the risk of overfitting remained: the model performed well on training data but poorly on unseen data. Meanwhile, the effectiveness of the modified recursive feature elimination (MRFE) technique and the corresponding ML model was found to heavily depend on the quality and completeness of input data. Inaccurate or incomplete data could lead to poor feature selection and, consequently, unreliable crop predictions.

Previous studies indicated that identifying key soil and environmental factors affecting crop yield could aid in predicting crop productivity. The study [12] introduced a novel MRFE technique to identify key soil and environmental variables involved in predicting crop productivity. The technique aimed to prioritize the most critical dataset elements, improving the precision of predictions. The MRFE technique employed a ranking system to assess the significance of various variables. However, it faced limitations related to regional variability, data integrity, agricultural system complexity, overfitting risk, and heterogeneity in assessed characteristics.

Numerous studies have explored the application of ML methodologies for forecasting crop productivity. The study [13] evaluated various ML algorithms to determine the most effective approach for predicting crop productivity. However, critical challenges remain, including those related to the estimation of the crop yield, impact of soil properties on predictions, and overarching goal of improving farmer profitability. Prior research has demonstrated the effectiveness of ML methods in crop recommendation based on soil characteristics. The study [14] examined ML techniques for crop recommendation, explicitly focusing on soil factors.

The techniques reported in [15] employed a hybrid model integrating decision tree, SVM, and RNN algorithms. Further, a comprehensive analysis of soil variables was incorporated to improve crop-yield projections and provide farmers with valuable guidance. Collectively, the techniques aimed to guide the agricultural decision-making process and increase farmer profitability. However, while the model incorporated several soil characteristics, including nitrogen content, phosphorus content, potassium content, moisture, and precipitation, it did not account for all environmental factors influencing the crop yield. Pests, diseases, and climate change can significantly impact agricultural productivity, and yet, they were not fully integrated into the model's forecasts. Additionally, the hybrid model struggled to generalize its predictions across different geographical regions or crop types. Conditions and farming practices at a location may not apply to other locations, limiting the model's effectiveness in providing universally relevant recommendations.

Furthermore, the study [16] introduced an innovative ML algorithm that simultaneously evaluated meteorological conditions and soil characteristics for optimal crop selection. This approach was distinctive, as previous models often focused on weather or soil data but not both concurrently. The study applied several ML methods, including a k-NN, Naïve Bayes model, random forest model, and long short-term memory (LSTM) RNN. The use of the LSTM RNN for meteorological forecasting was particularly important as it effectively captured temporal relationships in weather patterns, enhancing the prediction accuracy. However, the random forest model used in the second phase required substantial computational resources and processing time due to the construction of multiple trees for output aggregation. While this ensemble approach improved the prediction performance, it reduced interpretability, making it difficult to assess the contribution of individual variables.

The literature highlights significant advancements in applying ML to enhance agricultural operations, particularly in crop selection and production forecasting. Various ML techniques, including decision trees, kNNs, SVMs, random forest models, Naïve Bayes models, GBRTs, and neural networks, have been employed to analyze factors such as soil quality, climatic conditions, and nutrient content. Hybrid models and ensemble methods, such as the MRFE technique and reinforcement learning, have shown promise in enhancing feature selection and prediction accuracy. Additionally, ontology-driven systems and content-centric recommendation frameworks have enhanced the customization of crop suggestions by integrating agricultural knowledge. Despite these advancements, challenges remain, including regional heterogeneity, dataset limitations, and agricultural system complexity, which can lead to overfitting and limited scalability.

This study aimed to bridge this research gap by developing a comprehensive model that integrates soil properties, rainfall, and weather conditions to enhance ML-based predictions of land suitability for cultivation. By addressing the limitations identified in previous studies, the ML-based prediction model shows improved prediction accuracy and reliability. Unlike prior research that primarily classified land as being suitable or unsuitable for specific crops, the proposed methodology introduces a graded assessment of suitability levels. This nuanced approach enables landowners to make more informed decisions, even when land is not optimal but remains viable for agriculture. Moreover, while existing models have demonstrated improved performance, they often lack transparency in their computational processes. This study addresses that gap by ensuring methodological clarity, reproducibility, and adaptability. Additionally, the proposed approach establishes a foundation for future research by prioritizing flexibility in agricultural recommendations, enabling localized, data-driven decision-making processes tailored to the diverse needs of farmers and agricultural stakeholders.

3. RESEARCH FRAMEWORK AND METHODOLOGY

3.1 Research Framework

A comprehensive analysis of physical factors and relevant data from 93 districts in Thailand's Lower Northern region was conducted to predict the most suitable alternative crops for this area. To achieve high accuracy in these predictions, multiple prediction models were evaluated and compared.

Each model underwent a rigorous hyperparameter tuning process for performance optimization. Hyperparameter tuning is a crucial step in ML, where model parameters are adjusted to enhance the forecast accuracy. Through hyperparameter fine-tuning, the models were better aligned with the unique characteristics of the dataset.

The prediction models were then systematically evaluated to identify the most effective one. The evaluation process involved analyzing each model's performance using various criteria, including accuracy, precision, and overall effectiveness in predicting suitable crops. This comparative analysis facilitated the selection of the most robust forecasting model.

The most effective configuration for each model was determined through a systematic assessment of its performance. The selection process involved a detailed analysis of performance indicators to identify the model with the highest prediction accuracy. The most effective model identified from the training phase was then evaluated using the validation dataset.

Through rigorous testing and validation, the study identified the most effective model for predicting suitable crops in the study area. This methodological approach ensured the reliability and effectiveness of the selected model in practical applications, offering valuable insights for agricultural planning and decision-making in Thailand's Lower Northern region.

The research began with data collection from the Provincial Office of Agricultural Extension. A comprehensive analysis of the agricultural dataset was conducted to identify significant trends and insights. The dataset encompassed various variables, including temperature, rainfall, moisture content, wind speed, and soil classification (indicating ten soil types).

A comprehensive preparatory operation was performed to ensure the accuracy and practicality of the data. This process involved data cleansing, handling missing values, and transforming the dataset into a suitable format for analysis. At this stage, key variables influencing agricultural operations were identified and selected.

The study employed various ML models, including a decision tree model, an ANN, and a Naïve Bayes model, to predict the most suitable crop. Each model underwent meticulous hyperparameter fine-tuning and rigorous training to optimize the prediction performance. A thorough comparison of the models was then conducted to evaluate their effectiveness.

The prediction algorithms utilized multiple input variables, including geographical location, soil composition, weather characteristics, and farmer preferences, to recommend the most suitable crop for maximizing agricultural productivity in the given area. This comprehensive approach tailored recommendations to the specific conditions and needs of farmers. The research framework is shown in Fig. 1.

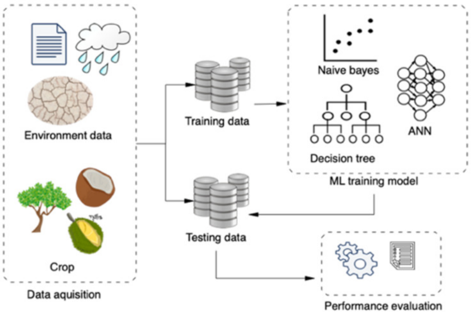


Fig.1: *Alternative-Crop Prediction Framework.*

3.2 Research Methodology

The research employed ML models to generate precise predictions and systematically assessed and analyzed the models to identify the most efficient one. The research process involved data acquisition, data preprocessing, model training, and model evaluation.

Data Acquisition: In this stage, a comprehensive dataset focused exclusively on crop-related information was obtained to ensure accuracy in forecasting and recommendation processes. The dataset encompassed various environmental and soil-related parameters, including precipitation, temperature, moisture content, wind speed, and soil type. These factors significantly influence the growth and productivity of different crop varieties.

The data were obtained from the Agricultural Extension Office of a Thai province to ensure reliability and usefulness. A comprehensive investigation was conducted in the Lower Northern region of Thailand, covering 93 districts. The investigated data sample was meticulously selected to ensure diversity and representativeness across the region. The primary objective of data collection was to capture the wide range of environmental conditions and soil characteristics in various districts. This information was crucial for assessing the viability of cultivating specific crops.

Data Preprocessing: Effective data preprocessing was essential to ensure the accuracy and reliability of predictions. The collected data underwent thorough cleaning and preprocessing to maintain integrity and compatibility, which involved handling missing data, normalizing numerical features, and converting categorical variables into a format appropriate for analysis.

The novelty of our proposed ML approach lies in its comprehensive integration of multiple critical parameters and a unique hyperparameter optimization strategy. Unlike previous studies [8, 13], which primarily focused on the binary classification of land suitability or single-parameter analysis, our study introduces a multilevel classification framework that categorizes cultivation suitability of land into four distinct grades: A, B, C, and D. This granular classification enhances farmers' decision-making capabilities by providing more nuanced recommendations. The model's architecture is specifically designed to process the complex interplay of various agricultural parameters, including soil classification (ten types), meteorological factors (temperature, rainfall, moisture content, and wind speed), and regional characteristics of Thailand's Lower Northern province. Additionally, our approach involves an innovative systematic hyperparameter tuning methodology, wherein each model (decision tree, ANN, and Naïve Bayes model) undergoes rigorous optimization using varying percentage splits of the dataset for training and validation and cross-validation folds.

The hyperparameter optimization methodology employed herein is manual tuning, which offers a systematic and controlled approach well-suited for moderate-scale datasets. This method enhances computational efficiency by avoiding the resource-intensive demands of automated optimization techniques such as grid search or Bayesian optimization. While automated hyperparameter search methods can be effective in specific contexts, they may inadvertently lead to overfitting due to excessive parameter optimization on training data. The manual tuning approach undertaken herein facilitates heuristic-based parameter selection, ensuring model parsimony and improving model generalization. This approach aligns with the principle of optimal model complexity, balancing computational efficiency and prediction performance while mitigating the overfitting risks associated with exhaustive automated searches.

For the decision tree model, we implemented a novel parameter tuning strategy with complexity parameter (C) values of 0.1, 0.25, and 0.5, while the ANN model incorporated a specialized configuration with learning rates ranging from 0.01 to 0.2 and epochs ranging from 50 to 300. This comprehensive

Table 1: *Comparison of Data Quantities Before and After Applying the SMOTE.*

Class / Crop	Number of data (Pre-SMOTE vs. Post-SMOTE)							
	A		B		C		D	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Rubber	31	31	15	30	13	30	19	30
Coconut	8	18	19	19	6	19	15	18
Longan	14	28	28	28	11	27	20	28
Durian	20	20	13	19	14	19	13	19
Rambutan	18	18	9	18	13	17	12	18
Mangosteen	16	20	17	20	21	21	12	21

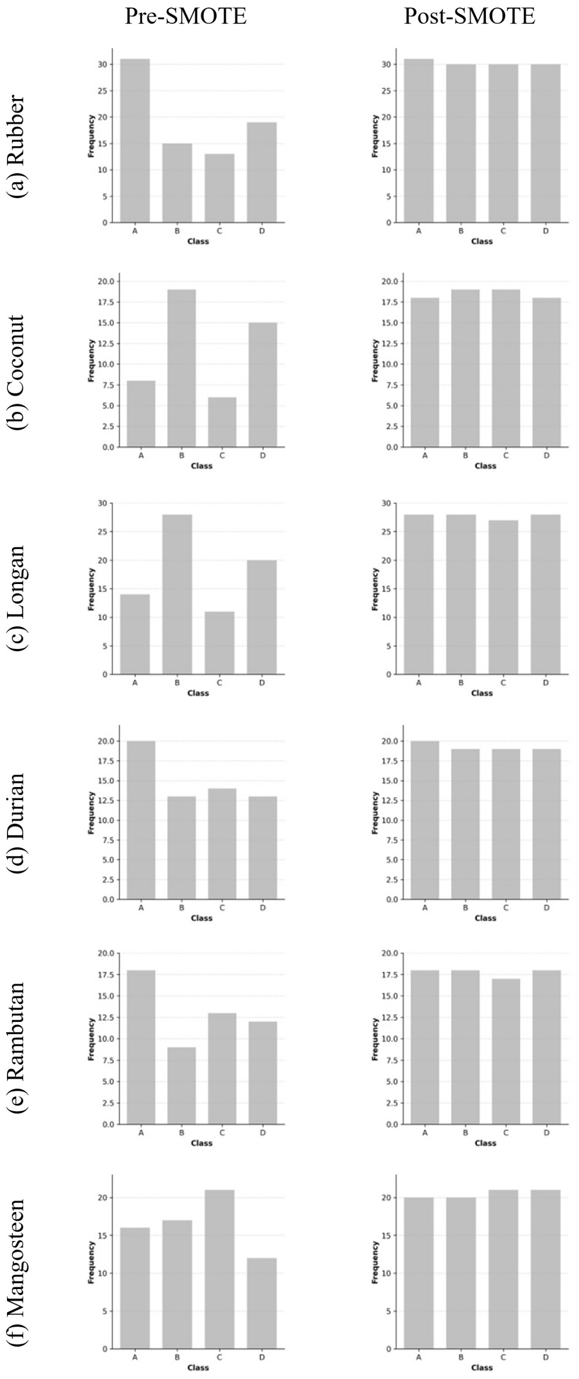
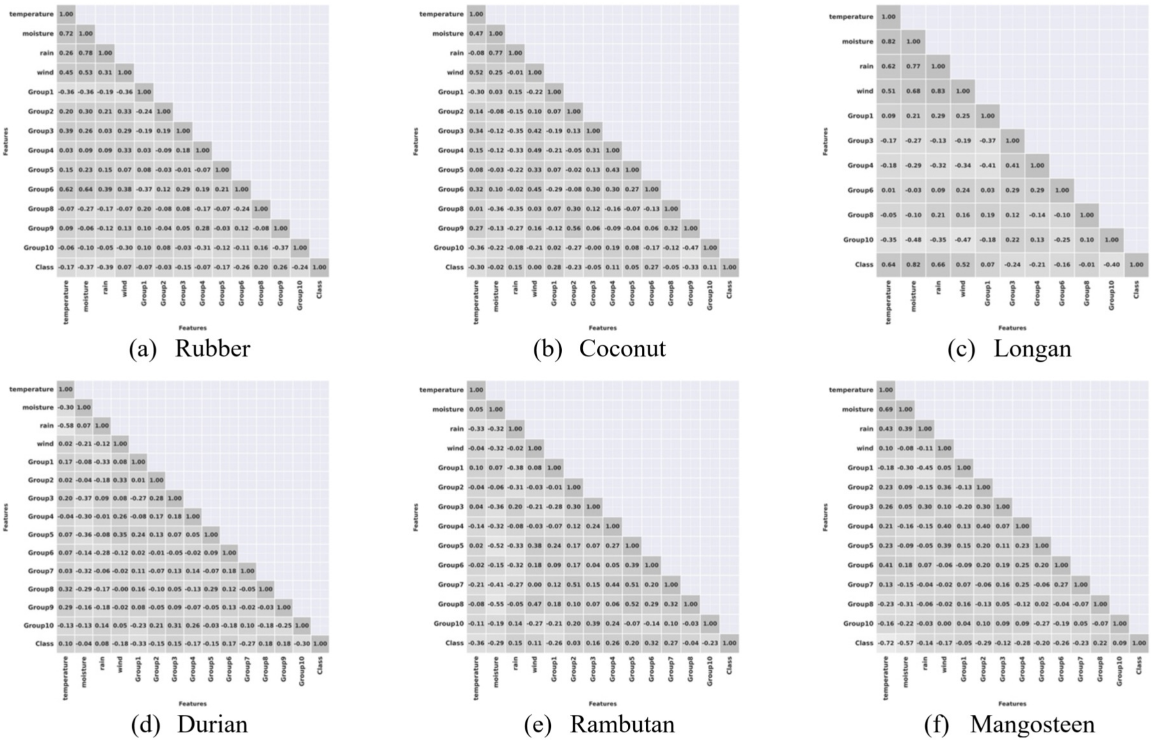


Fig.2: *Pre-SMOTE and Post-SMOTE Frequency Distribution Graph.*



(a) Rubber (b) Coconut (c) Longan

(d) Durian (e) Rambutan (f) Mangosteen

Fig.3: *Variable Correlation Heatmaps.*

optimization approach ensured robust model performance across different crop types and environmental conditions, addressing the limitations of previous models that often used fixed parameter settings.

The dataset was systematically clustered to ensure a standardized distribution for training and validation. A classification system was established to assess the suitability of different regions for specific crops. This system categorized cultivation suitability into four levels: high (A), medium (B), fair (C), and low (D). These classifications were determined based on the region's suitability for the listed crops, considering various environmental and soil conditions.

A significant challenge in this research was addressing dataset imbalance. To resolve this issue, the synthetic minority oversampling technique (SMOTE)—an advanced statistical method—was employed using the Azure ML platform to balance the dataset by generating synthetic instances of minority cases while preserving the original distribution of majority cases.

The instances generated by the SMOTE are not the exact duplicates of existing minority cases. Instead, the method selects samples from the feature space of each target class and its nearest neighbors, then synthesizes new instances by combining features from the target case and its neighbors. This approach enhances the representation of each class's characteristics, improving dataset diversity and resilience.

The SMOTE operates on the entire dataset, focusing on enhancing instances within the minority class. In this study, the number of nearest neighbors was set to one, ensuring a targeted and efficient oversampling strategy. Table 1 presents the dataset before and after applying the SMOTE, demonstrating the impact of this oversampling technique.

This research effectively addressed imbalance dataset by applying the SMOTE in the data preprocessing phase, leading to improved accuracy and reliability of the prediction models. A thorough data preparation process is essential for developing a robust recommendation system that provides accurate and practical insights into agricultural practices.

Fig. 2 presents a frequency distribution graph showing changes in data quantity before and after applying the SMOTE. The graph illustrates alterations in data distribution, providing insights into SMOTE's effectiveness in balancing the dataset.

A comprehensive correlation analysis was conducted to examine relationships between the components for prediction to the prediction model. The Pearson correlation coefficient, calculated using (1), was employed to determine the associations and interdependencies among the variables under investigation. This statistical measure identifies the strength and direction of correlations, providing essential insights into the prediction process of the models.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}. \quad (1)$$

Here, r represents the correlation coefficient between variables x and y. The value of x at the i$^{\text{th}}$

data point is denoted xi, and the mean of x is represented as $\bar{x}$. Similar is the case for y, yi, and $\bar{y}$.

Fig. 3 presents correlation heatmaps illustrating the relationships between the variables in the dataset. Heatmaps are commonly used in statistical analysis to visually assess the strength and direction of correlations. They provided a comprehensive representation of variable relationships, facilitating the identification of variation patterns, variable interdependencies, and potential multicollinearity issues.

4. EXPERIMENTAL RESULTS AND EVALUATION

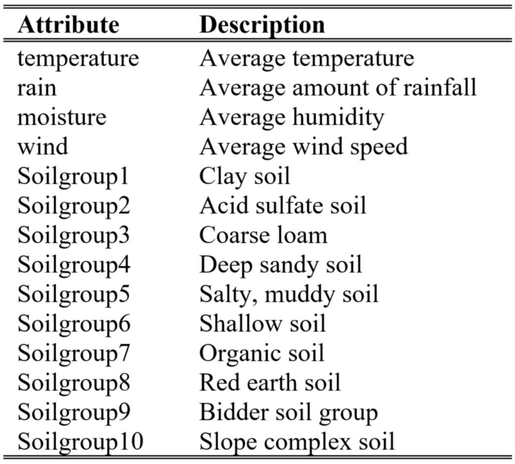
4.1 Experimental Settings

The decision tree model, ANN, and Naïve Bayes model were employed for predictions, each utilizing distinct techniques and configurations. The models were evaluated using different percentage splits (60%, 80%, and 90%) and cross-validation folds (5-fold and 10-fold). Additionally, batch sizes of 8 or 16 were set during model evaluation experiments. The decision tree model (J48) was tuned using C values of 0.1, 0.25, and 0.5. For the ANN, learning rates were adjusted to 0.01, 0.02, 0.03, 0.1, and 0.2, with epoch values ranging from 50 to 300 with a fixed momentum of 0.2. Data preprocessing involved normalization using a Z-score during the training phase. These configurations were systematically explored to evaluate model performance and optimize prediction capabilities in alignment with the research objectives.

4.2 Experimental Evaluation Metrics

Data Description: The dataset used in this study was structured into clusters to enhance the prediction accuracy. It comprised preprocessed attributes, including temperature, rainfall, moisture content, wind speed, and soil classification, covering 93 districts in the Lower Northern region of Thailand.

Table 2: *Attribute Description.*



Attribute	Description
temperature	Average temperature
rain	Average amount of rainfall
moisture	Average humidity
wind	Average wind speed
Soilgroup1	Clay soil
Soilgroup2	Acid sulfate soil
Soilgroup3	Coarse loam
Soilgroup4	Deep sandy soil
Soilgroup5	Salty, muddy soil
Soilgroup6	Shallow soil
Soilgroup7	Organic soil
Soilgroup8	Red earth soil
Soilgroup9	Bidder soil group
Soilgroup10	Slope complex soil

This clustering provided a comprehensive and structured representation of the environmental and geographical factors identified for the study area, facilitating robust prediction modeling and analysis. This study considered key variables affecting crop production, such as rainfall, humidity, temperature, and wind speed. Additionally, variations in soil types across different regions contributed to a precise assessment of crop cultivation suitability. Soil classification data were obtained from the Soil Research Survey and Research Division. Table 2 presents the attributes used in the prediction model.

Model Evaluation: The model was evaluated using MAE, RMSE, precision, recall, and F-measure. The MAE, given by Equation (2), quantifies the average absolute deviation between the predicted and actual values of the target variable:

$$\text{MAE} = \frac{\sum |y_i - p_i|}{n}, \tag{2}$$

where y represents the actual value of the target variable, p represents its predicted value, and n is the number of cases.

The RMSE, defined in Equation (3), measures the average squared difference between predicted and actual values of a variable:

$$\text{RMSE} = \sqrt{\frac{\sum (y_i - p_i)^2}{n}}. \tag{3}$$

Precision, computed using Equation (4), quantifies the proportion of correctly predicted positive instances among all predicted positive instances:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{4}$$

Recall, defined in Equation (5), assesses a model's ability to identify positive instances:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{5}$$

Here, TP (true positive) denotes correctly identified positive instances, FP (false positive) represents instances incorrectly classified as positive, and FN (false negative) refers to instances incorrectly classified as negative.

The F-measure, calculated using Equation (6), evaluates the balance between precision and recall:

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision+Recall}}. \tag{6}$$

4.3 Experimental Result

Three distinct prediction models—the decision tree model, ANN, and Naïve Bayes model—were tested for their ability to identify the suitability of alternative-crop cultivation in the given region. The models were evaluated using different percentage splits and cross-validation folds to optimize the training and testing datasets.

Table 3: *Performance Evaluation Metrics of the Training Models.*

Performance Evaluation Metrics										
Model	***Evaluation Metrics***	***Parameter***	***Accuracy (%)***	***TP***	***FP***	***Precision***	***Recall***	***F-Measure***	***MAE***	***RMSE***
Rubber										
Decision tree	80%	C = 0.1	70.83	0.708	0.118	0.751	0.708	0.688	0.1904	0.3359
ANN	80%	Hidden = 6, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	70.83	0.708	0.092	0.733	0.708	0.695	0.2341	0.3681
Naïve Bayes	60%	N/A	54.16	0.542	0.182	0.572	0.542	0.523	0.2474	0.4374
Coconut										
Decision tree	5-fold	C = 0.25	71.62	0.716	0.095	0.716	0.716	0.707	0.1711	0.3569
ANN	90%	Hidden = 4, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	85.71	0.857	0.024	0.929	0.857	0.867	0.1374	0.2792
Naïve Bayes	80%	N/A	60.00	0.600	0.127	0.597	0.600	0.593	0.2105	0.3657
Longan										
Decision tree	80%	C = 0.5	90.90	0.909	0.027	0.924	0.909	0.904	0.0647	0.2019
ANN	90%	Hidden = 3, LR = 0.1, Momentum = 0.2, Seed = 1, Epochs = 250	90.90	0.909	0.034	0.932	0.909	0.900	0.1554	0.2502
Naïve Bayes	60%	N/A	75.00	0.750	0.089	0.738	0.750	0.716	0.1518	0.324
Durian										
Decision tree	80%	C = 0.5	73.33	0.733	0.097	0.800	0.733	0.749	0.1583	0.3736
ANN	5-fold	Hidden = 6, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	50.64	0.506	0.165	0.504	0.506	0.504	0.2829	0.438
Naïve Bayes	5-fold	N/A	48.05	0.481	0.174	0.573	0.481	0.471	0.2945	0.4387
Rambutan										
Decision tree	90%	C = 0.1, C = 0.25, C = 0.5	85.71	0.857	0.057	0.905	0.857	0.848	0.1508	0.2627
ANN	5-fold	Hidden = 4, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	57.74	0.577	0.140	0.574	0.577	0.570	0.2515	0.4211
Naïve Bayes	5-fold	N/A	53.52	0.535	0.155	0.525	0.535	0.517	0.2535	0.4053
Mangosteen										
Decision tree	5-fold	C = 0.1	60.97	0.610	0.129	0.611	0.610	0.610	0.2163	0.3792
ANN	5-fold	Hidden = 6, LR = 0.1, Momentum = 0.2, Seed = 1, Epochs = 300	56.09	0.560	0.145	0.562	0.561	0.560	0.2529	0.3933
Naïve Bayes	5-fold	N/A	57.31	0.573	0.142	0.571	0.573	0.554	0.2253	0.4175

Using the percentage split method, the dataset was divided into two subsets: a training set and a testing (or validation) set. The first set was used to train the prediction model, while the testing set was used to evaluate model performance. To determine the optimal prediction model, the dataset was partitioned into training and testing sets using different percentages

For cross-validation, the dataset was divided into K partitions, commonly referred to as folds. The model was trained on all partitions except one, which was reserved as the testing set. This process was repeated "K" times, ensuring that each fold served as the testing set once. The mean performance of all K iterations was then computed. The experiment aimed to identify the optimal tuning parameter "K" to enhance the prediction accuracy. The model was trained using the K-fold forward chaining cross-validation procedure.

During the training stage, the three prediction models were evaluated by varying the percentage splits (60%, 80%, and 90%) and number of cross-validation folds (5-fold and 10-fold). The designated alternative crops were analyzed under these experimental settings.

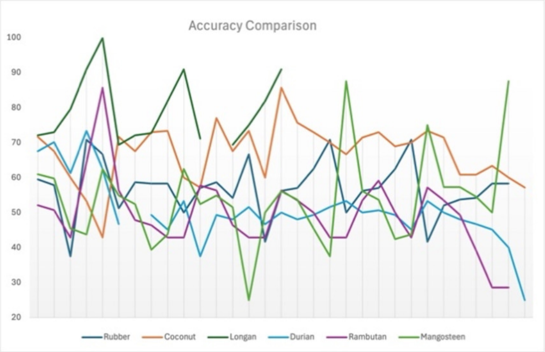


Fig.4: *Comparison of Model Accuracy Across Various Hyperparameter Tuning Settings.*

Fig. 4 presents a comparative analysis of various hyperparameter tuning settings across different models, facilitating the identification of optimal hyperparameters for each model. Through extensive experimentation, the most effective configurations were determined to maximize model accuracy in predicting cultivation suitability. This analysis highlighted the crucial role of hyperparameter tuning in enhancing model performance and ensuring the selection of the most suitable parameters for achieving the highest prediction accuracy.

Table 3 shows a comprehensive comparison of three ML models, highlighting the optimal hyperparameters identified for each model. It also presents a detailed evaluation of the performance metrics, offering valuable insights into each model's prediction capabilities and effectiveness. The experimental results demonstrate the effectiveness of the various ML models in forecasting crop suitability. The decision tree model was optimized by fine-tuning the C value and cross-validation folds. Similarly, the ANN model, in an iterative process, was refined through experiments involving multiple parameters, including the number of hidden layers (Hidden), learning rate (LR), momentum (Momentum), seed (Seed), and epochs (Epochs). This iterative optimization process was tailored for each crops to enhance model prediction performance. The choice of hyperparameters significantly impacted model effectiveness, as reflected in the variations in accuracy and performance metrics across different parameter configurations.

The decision tree model exhibited satisfactory performance for most crops, with accuracy ranging from 60.97% to 90.90%. Notably, it achieved high precision and recall scores, indicating its ability to make accurate predictions and effectively identify the most suitable crops.

The performance of the ANN varied across different crops. For crops such as coconut and longan, precision scores were exceptional, matching or even surpassing those obtained when using the decision tree model. However, for durian and rambutan, accuracy of the ANN was lower, ranging from 50.64% to 57.74%.

Overall, the Naïve Bayes model demonstrated lower accuracy than the decision tree and ANN models. Prediction performance varied across crops, with accuracy ranging from 48.05% to 60.97%. While Naïve Bayes performed well for certain crops, it generally failed to achieve accuracy levels comparable to those of the other models.

The comparative analysis revealed that the decision tree model consistently produced reliable and accurate results across different crops. Its widespread applicability stemmed from its simplicity and adaptability, allowing it to effectively process quantitative and qualitative data. While the ANN demonstrated high prediction potential for certain crops, its performance was inconsistent, requiring extensive parameter tuning and higher computational resources than the decision tree model. Although Naïve Bayes is computationally efficient and easy to implement, it struggled to achieve the same level of accuracy as the decision tree and ANN models. This highlighted the limitations of Naïve Bayes in capturing complex data relationships.

The most effective model was subsequently evaluated using a series of testing methods, with the results presented in Table 4. These findings provide critical insights into the model's performance and effectiveness in practical applications, contributing significantly to advancements in this research domain.

The experimental findings (Table 4) demonstrated the efficacy of various ML models in predicting the suitability of locations for crop cultivation. The decision tree model consistently exhibited high accuracy, precision, recall, and F-measure for rubber, longan, durian, rambutan, and mangosteen crops. The decision tree model achieved accuracy values ranging from 85.00% to 94.74%, indicating its effectiveness in classifying the location suitability level for crop cultivation across different datasets. Its high precision scores (84.80%–95.60%) reflected its ability to minimize false positives. Similarly, recall values (lying between 85.00% and 94.70%) demonstrated the model's capability to correctly identify a substantial proportion of suitable alternative crops, reducing incorrect rejections. The F-measure values, which integrates precision and recall, ranged from 84.50% to 94.70%, further confirming the model's reliability in predicting suitable cultivation areas for these crops.

Table 4: *Comparison of ML Model Performance in Predicting Crop Suitability.*

Crop	Model	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Rubber	Decision tree	85.00	84.80	85.00	84.50
Coconut	ANN	81.82	84.80	81.80	82.10
Longan	Decision tree	94.74	95.60	94.70	94.70
Durian	Decision tree	94.11	95.30	94.10	94.20
Rambutan	Decision tree	86.67	91.10	86.70	86.30
Mangosteen	Decision tree	89.47	90.40	89.50	89.30

The ANN was the most effective model for estimating the suitability of a location for coconut cultivation. The evaluation yielded an accuracy of 81.82%, with a precision of 84.80%, a recall of 81.80%, and an F-measure of 82.10%.

The decision tree model's consistent performance across multiple crops demonstrated its resilience and reliability in forecasting suitable cultivation areas in diverse agricultural settings. This model exhibited strong adaptability and effectively captured the relationships between crop attributes and suitability levels across various datasets.

4.4 Comparison Results Based on a Benchmark Dataset

This research evaluated the proposed optimal model using a benchmark dataset to assess its effectiveness in predicting the suitability of locations for the cultivation of specific crops. The experimental analysis focused on rubber and coconut, employing the optimal model with precisely tuned hyperparameters. The benchmark dataset for coconut, available at DOI:10.21227/12nr-fe03, categorizes locations as being either suitable or unsuitable without specifying the degree of appropriateness. Meanwhile, our study identified specific levels of suitability. To enable comparison between our model and previously published research, on the same dataset, we transformed our output predictions into a binary classification of "suitable" or "unsuitable" to align with the benchmark dataset. Similarly, the rubber dataset (available at `https://datasets.omdena.com/dataset/crop-yield-prediction`) provides crop-yield forecasts rather than location suitability for crop culti-

Table 5(a): *Performance Evaluation Metrics When Using the Testing Dataset*

Performance Evaluation Metrics										
Model	***Evaluation Metrics***	***Parameter***	***Accuracy (%)***	***TP***	***FP***	***Precision***	***Recall***	***F-Measure***	***MAE***	***RMSE***
Coconut										
Decision tree	5-fold	C = 0.25	71.62	0.716	0.095	0.716	0.716	0.707	0.1711	0.3569
ANN	90%	Hidden = 4, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	85.71	0.857	0.024	0.929	0.857	0.867	0.1374	0.2792
Naïve Bayes	80%	N/A	60.00	0.600	0.127	0.597	0.600	0.593	0.2105	0.3657
Rubber										
Decision tree	80%	C = 0.1	70.83	0.708	0.118	0.751	0.708	0.688	0.1904	0.3359
ANN	80%	Hidden = 6, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	70.83	0.708	0.092	0.733	0.708	0.695	0.2341	0.3681
Naïve Bayes	60%	N/A	54.16	0.542	0.182	0.572	0.542	0.523	0.2474	0.4374

Table 5(b): *Performance Evaluation Metrics When Using the Testing Dataset*

Performance Evaluation Metrics										
Model	***Evaluation Metrics***	***Parameter***	***Accuracy (%)***	***TP***	***FP***	***Precision***	***Recall***	***F-Measure***	***MAE***	***RMSE***
Coconut										
Decision tree	5-fold	C = 0.25	87.50	0.875	0.125	0.900	0.875	0.873	0.131	0.338
ANN	90%	Hidden = 4, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	100.00	1.000	1.000	1.000	1.000	1.000	0.017	0.019
Naïve Bayes	80%	N/A	100.00	1.000	1.000	1.000	1.000	1.000	0.0001	0.0002
Rubber										
Decision tree	5-fold	C = 0.25	96.77	1.000	0.067	0.941	1.000	0.970	0.032	0.179
ANN	90%	Hidden = 4, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	100.00	1.000	1.000	1.000	1.000	1.000	0.030	0.036
Naïve Bayes	80%	N/A	100.00	1.000	1.000	1.000	1.000	1.000	0.006	0.015

Table 6: *Performance Evaluation Metrics When Using the Three Crop Datasets*

Performance Evaluation Metrics											
Model	***Evaluation Metrics***	***Parameter***	***Dataset***	***Accuracy (%)***	***TP***	***FP***	***Precision***	***Recall***	***F-Measure***	***MAE***	***RMSE***
Decision tree	5-fold	C = 0.25	Crop Recommendation using Soil Properties and Weather Prediction Dataset (1)	100.00	1.000	0.000	1.000	1.000	1.000	0.000	0.000
			Crop Recommendation Dataset (2)	100.00	1.000	0.000	1.000	1.000	1.000	0.000	0.000
			Agriculture Crop Yield Dataset (3)	99.99	1.000	0.000	1.000	1.000	1.000	0.000	0.002
ANN	90%	Hidden = 4, LR = 0.2, Momentum = 0.2, Seed = 1, Epochs = 300	Crop Recommendation using Soil Properties and Weather Prediction Dataset (1)	100.00	1.000	0.000	1.000	1.000	1.000	0.007	0.007
			Crop Recommendation Dataset (2)	100.00	1.000	0.000	1.000	1.000	1.000	0.010	0.011
			Agriculture Crop Yield Dataset (3)	99.41	0.994	0.006	0.994	0.994	0.994	0.007	0.067
Naïve Bayes	80%	N/A	Crop Recommendation using Soil Properties and Weather Prediction Dataset (1)	80.57	0.806	0.220	0.810	0.806	0.803	0.189	0.391
			Crop Recommendation Dataset (2)	100.00	1.000	0.000	1.000	1.000	1.000	0.000	0.000
			Agriculture Crop Yield Dataset (3)	95.42	0.954	0.046	0.954	0.954	0.954	0.078	0.181

Dataset accessible at:
(1) https://data.mendeley.com/datasets/8v757rr4st/1
(2) https://ieee-dataport.org/documents/crop-recommendation-dataset
(3) https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield

vation. We converted crop-yield predictions into binary classifications, designating locations with above-average yields as “suitable” and those below average as “unsuitable.” This preprocessing step enabled a standardized evaluation of the proposed model across datasets. The results presented in Table 5(a) for our testing dataset and Table 5(b) for the benchmark dataset demonstrated that the proposed model performed effectively across diverse datasets, highlighting its reliability and high prediction accuracy. A comprehensive validation was conducted using multiple crop datasets to assess the model’s robustness and generalizability. The methodological validation approach incorporated three datasets to rigorously evaluate the model’s interoperability and prediction efficacy across various crop types. Each dataset was systematically assessed using optimal model hyperparameters derived from preliminary model development. Binary classification results were obtained, indicating whether locations were “suitable” or “unsuitable” for agricultural intervention.

The experimental approach standardized the hyperparameter settings across all three datasets to ensure a uniform basis for comparison. The performance metrics of the prediction model, which included the key statistical indicators, are presented in Table 6, providing a quantitative assessment of the model’s transferability and prediction capabilities across diverse agricultural settings.

This methodology presents a systematic framework for evaluating ML models in agricultural predictions, emphasizing the critical role of cross-dataset validation for ensuring the reliability and generalizability of agricultural models used for predictions.

5. CONCLUSION AND FUTURE WORK

This study examined the effectiveness of various ML models in predicting the suitability of different locations for crop cultivation. Suitability was assessed on a spectrum, ranging from "high" to "low." Building on previous research [13], which focused on three primary crops and identified the most suitable crops for a given location, this study expanded those findings by providing a more detailed assessment of location suitability for crop cultivation. The proposed approach classified location suitability on a continuum while incorporating additional critical factors influencing crop yield. By systematically evaluating multiple ML models, this research identified the most effective model for crop selection, contributing to a precise and data-driven approach toward agricultural decision-making.

Comprehensive tests were conducted to evaluate the effectiveness of the decision tree model, ANN, and Naïve Bayes model in predicting the suitability of cultivation locations for coconut, longan, durian, rambutan, and mangosteen crops. The models were assessed using accuracy, precision, recall, and F-measure.

The experimental results indicated that the ANN model performed exceptionally well in predicting coconut crop suitability but exhibited lower accuracy for other crops. This underscores the importance of selecting models tailored to specific agricultural contexts. Naïve Bayes—with simplicity and low computational requirements—demonstrated lower accuracy than the decision tree and ANN models, highlighting its limited ability to capture complex relationships within data. The empirical findings demonstrated the decision tree model's high efficacy in accurately predicting suitable cultivation locations across various crops. This underscores its role as a critical tool for informed decision-making and strategic planning in agriculture. By providing precise predictions, this model aids the decision-making process regarding crop selection and land use, ultimately improving agricultural productivity and sustainability. Accurate predictions of crop suitability facilitate the efficient allocation of resources such as land, water, and labor, leading to improved agricultural outcomes and economic growth.

This research significantly contributes to advancing agricultural decision-making by demonstrating the utility of ML models in predicting crop suitability. Future studies can explore advanced optimization techniques, ensemble learning methodologies, or integrating additional features to enhance the prediction accuracy and better address real-world agricultural challenges.

The proposed model underwent a thorough evaluation, incorporating various of relevant factors. These factors included various crop types, distinct soil compositions, as well as seasonal weather conditions such as temperature, moisture content, and rainfall patterns. This comprehensive approach provided a structured and robust foundation for evaluation. Multiple interrelated factors influence crop productivity and soil health. Understanding the interactions between crop varieties and specific soil characteristics is essential for optimizing agricultural practices. Comparative analysis with existing models ensured that the proposed model was rigorously evaluated against established benchmarks. This comparative analysis not only confirmed the reliability of the model but also highlighted its potential advantages and limitations in different agricultural contexts. By employing a comprehensive and multifaceted evaluation methodology, this study provided valuable insights into agricultural research, contributing to developing resilient and adaptable farming systems.

Additional investigation is needed to examine the factors influencing variations in model performance across different crops. Analyzing dataset characteristics, feature importance, and model interpretability can provide valuable insights for improving model prediction accuracy. While decision tree models generally yield satisfactory results, there remains room for improvement by developing more advanced models to improve prediction accuracy and reliability. Our findings contribute to a deeper understanding of ML applications in agriculture, supporting informed decision-making for crop management and optimization strategies.

ACKNOWLEDGEMENT

This research was supported by Naresuan University (NU) and the National Science, Research, and Innovation Fund (NSRF) under Grant No. R2566B034. Neither NU nor NSRF had involvement in the study design, data collection, analysis, decision to publish, or manuscript preparation.

AUTHOR CONTRIBUTIONS

Conceptualization, S.M. and A.S.; software, S.M; methodology, S.M. and A.S.; investigation, S.M. and A.S.; data curation, S.M.; visualization, S.M. and A.S.; writing - original draft, S.M; validation, A.S.; supervision, A.S.; writing - review & editing, A.S.; funding acquisition, A.S. The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper. All authors have read and agreed to the published version of the manuscript.

References

[1] P. Appiahene, A. Kofi, K. Santo and R. Bannor, "Application of Machine Learning in Crop Cultivation and Production: Systematic Review,"

Journal of Energy and Natural Resource Management (JENRM), vol. 8, no. 2, pp. 51-62, 2022.

[2] T. van Klompenburg, A. Kassahun and C. Catal, "Crop Yield Prediction using Machine Learning: A Systematic Literature Review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, Oct. 2020.

[3] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar and A. Kumar, "A Systematic Literature Review on Machine Learning Applications for Sustainable Agriculture Supply Chain Performance," *Computers & Operations Research*, vol. 119, p. 104926, Jul. 2020.

[4] M. A. Riaño, A. O. R. Rodriguez, J. B. Velandia, P. A. G. García and C. E. M. Marín, "Design and Application of An Ontology to Identify Crop Areas and Improve Land Use," *Acta Geophys.*, vol. 71, No. 3, pp. 1409–1426, Jun. 2023.

[5] J. Lacasta, F. J. Lopez-Pellicer, B. Espejo-García, J. Nogueras-Iso and F. J. Zarazaga-Soria, "Agricultural Recommendation System for Crop Protection," *Computers and Electronics in Agriculture*, vol. 152, pp. 82–89, Sep. 2018.

[6] K. Patel and H. B. Patel, "A State-of-the-Art Survey on Recommendation System and Prospective Extensions," *Computers and Electronics in Agriculture*, vol. 178, p. 105779, Nov. 2020.

[7] P. Kaur, J. K. Chahal and T. Sharma, "A Data Mining Approach for Crop Yield Prediction in Agriculture Sector," *Advances in Mathematics: Scientific Journal*, vol. 10, no. 3, pp. 1425–1430, Mar. 2021.

[8] C. N. Vanitha, N. Archana and R. Sowmiya, "Agriculture Analysis using Data Mining and Machine Learning Techniques," in *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, pp. 984–990, Mar. 2019.

[9] T. Deshmukh, A. Rajawat, S. B. Goyal, J. Kumar and A. Potgantwar, "Analysis of Machine Learning Technique for Crop Selection and Prediction of Crop Cultivation," in *2023 International Conference on Inventive Computation Technologies (ICICT)*,pp. 298–311, Apr. 2023.

[10] S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika and J. Nisha, "Crop Recommendation System for Precision Agriculture," in *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pp. 32–36, Jan. 2017.

[11] U. P. Jakarbet, S. S P, S. R M and A. B P, "An Empirical Analysis of Machine Learning Algorithms for Agricultural Crop Prediction," *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Kirtipur, Nepal, pp. 1782-1790, 2024.

[12] G. Mariammal, A. Suruliandi, S. P. Raja and E. Poongothai, "Prediction of Land Suitability for Crop Cultivation Based on Soil and Environmental Characteristics Using Modified Recursive Feature Elimination Technique With Various Classifiers," in *IEEE Transactions on Computational Social Systems*, vol. 8, no. 5, pp. 1132-1142, Oct. 2021

[13] R. Kammoonrat and A. Salaiwarakul, "Machine Learning Models for Crop Prediction in Agriculture: A Comparative Study," *ICIC Express Letters Part B: Applications*, vol. 16, no. 1, pp. 1–8, 2025.

[14] S. S. B, Anusha, A. Shetty, R. R. Shetty, B. A. D. Alva and A. D. Shetty, "Machine Learning Techniques in Crop Recommendation based on Soil and Crop Yield Prediction System – Review," in *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp. 230–235, Dec. 2022.

[15] P. Parameswari and C. Tharani, "Crop Specific Cultivation Recommendation System using Deep Learning," in *Information and Communication Technology for Competitive Strategies (ICTCS 2022)*, Singapore: Springer Nature, pp. 781–787, 2023.

[16] S. Rani, A. K. Mishra, A. Kataria, S. Mallik and H. Qin, "Machine Learning-Based Optimal Crop Selection System in Smart Agriculture," *Scientific Reports*, vol. 13, no. 1, p. 15997, Sep. 2023.

Sujitranan Mungklachaiya received her B.S. degree in applied computer science from King Mongkut's Institute of Technology, North Bangkok, Thailand, in 2001 and her M.Sc. degree in information technology from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, in 2004. She worked as a lecturer in the Computer and Information Technology Department, Faculty of Science and Technology, Loei Rajabaht University, Loei, Thailand, since 2006. She is currently pursuing her Ph.D. degree in information technology at the Naresuan University, Phitsanuloke, Thailand. Her research interests include web technology, machine learning, and image processing.

Anongporn Salaiwarakul eceived her B.Sc. degree in computer science from the Assumption University in 1997, her M.Sc. degree in computer science from the Chulaongkorn university in 2004, and her Ph.D. degree in computer science from the University of Birmingham, UK, in 2009. She worked as a lecturer in the Department of Computer Science and Information Technology, Faculty of Science, Naresuan University. University, Phitsanuloke, Thailand, since 1998, where she is currently working as an assistant professor. Her research interests include computer security, semantic webs, ontologies, project management, and image processing. She has led more than 10 research projects as the principal investigator, receiving funding support through collaborative grants from the university and the National Science Research and Innovation Fund.