MUST Journal of Research and Development (MJRD) Volume 7 Issue 4, October-December 2026
e ISSN 2683-6467 & p ISSN 2683-6475

Agro-Pest and Disease Prediction System: A Gradient Boosting Model for Improving Maize Productivity in the Southern Highlands of Tanzania

Simeo B. Kisanjara
Department of Computing Science Studies
Mzumbe University, P.O Box 87, Morogoro, Tanzania
DOI: https://doi.org/10.62277/mjrd2026v7i40003

ARTICLE INFORMATION

Article History
Received: 16th May 2026
Revised: 07th October 2026
Accepted: 29thSeptember 2026
Published: 15th November 2026

Keywords
Machine learning
Pests and Disease
Agriculture
Gradient Boosting model
Web based system

ABSTRACT

To enhance maize production and contribute to food sustainability in Tanzania, this study aimed to predict maize pests and diseases using a web-based system integrated with machine learning (ML) models. The study was conducted in the Ruvuma, Njombe, and Iringa regions, which are major maize-producing areas in Tanzania. We trained and validated six machine learning models, including Gradient Boosting (GB), Linear Regression (LR), Support Vector Regression (SVR), Long Short-Term Memory (LSTM), and K-Nearest Neighbours (KNN), to assess their effectiveness in predicting maize pests and diseases. The dataset was split into training and validation sets, with 80% for training and 20% for validation. The ML models were developed in Python using Pandas and TensorFlow and subsequently integrated into a web-based prediction system. The findings indicate that temperature, humidity, and rainfall had substantial effects on the model's ability to predict pest and disease occurrence, while wind speed, solar radiation, and historical data had moderate effects. Among the evaluated models, Gradient Boosting (GB) demonstrated strong predictive performance, achieving a Mean Absolute Error (MAE) of 102.03, a Root Mean Square Error (RMSE) of 0.057, and a coefficient of determination (R^2) of 0.70. The cross-validation results further indicated consistent performance, suggesting good model generalisation. The GB model also demonstrated suitability for capturing non-linear relationships among the predictor variables and performed effectively with the dataset used in this study. The integration of the model into a web-based system provides a potentially low-cost approach for pest and disease prediction in rural Tanzania, supporting early intervention and improved agricultural management. This study contributes to the existing literature by demonstrating the applicability of gradient boosting for regression-based pest and disease prediction using relatively small agricultural datasets.

*Corresponding author's e-mail address: simeokisanjara@gmail.com (Kisanjara, S.B.)

1.0 Introduction

According to the National Bureau of Statistics (NBS, 2022), agriculture is a major component of Tanzania's economy, employing over 65% of the country's workforce and accounting for about 27% of GDP. Smallholder farmers are the main suppliers in the sector and commonly rely on traditional farming methods. Pests and diseases significantly lower agricultural output and threaten food security.

Pests and diseases are major contributors to crop losses. The Food and Agriculture Organization (FAO, 2020) estimates that plant pests and diseases can account for substantial annual crop losses, while fall armyworm (Spodoptera frugiperda) is a major threat to maize production. Additionally, outbreaks of coffee leaf rust (Hemileiavastatrix) and banana Xanthomonas wilt have affected important crops in Tanzania.

The traditional methods of managing pests and diseases in Tanzania are mostly manual and reactive. Farmers primarily rely on local expertise and visual assessment. Limited access to agricultural extension and timely technical information can delay the identification and treatment of pests and diseases (Mpogole et al., 2020). These delays can contribute to infestations, crop damage, and lower yields.

The following reasons compound the problem: restricted access to resources, which means that many farmers in rural Tanzania cannot afford modern diagnostic tools or regular visits from agricultural extension experts. Climate change has disrupted weather patterns, contributing to changes in pest and disease occurrence (Skendžić et al., 2021). Additional concerns include the misuse and overuse of pesticides. Therefore, the combination of these factors can reduce agricultural output while also threatening the livelihoods of farming households. This situation calls for the use of emerging technologies like artificial intelligence (AI) and machine learning (ML).

The management of disease and pests could be drastically altered by AI/ML. AI/ML-powered systems may assess real-time data from satellite images, weather forecasts, and ground reports to forecast outbreaks and identify early signs of infestations. Tanzania has not yet embraced these technologies despite their successful adoption in wealthier countries due to their high costs, poor infrastructure, and low level of digital literacy among farmers. This historical context thus emphasises the critical need for a localised, cost-effective, and scalable AI/ML-driven pest and disease prediction system. By enabling early prediction and data-driven decision-making, such a system can improve food security, promote sustainable farming practices throughout Tanzania, and raise agricultural output.

1.1 Problem Statement

Over 70% of Tanzanians are employed in agriculture, which is a major part of the country's economy (Mpogole et al., 2020). However, the nation's agricultural sector is confronted with significant issues that restrict its potential for expansion and sustainability. The prevalence of crop pests and diseases, which result in crop losses, reduced yields, and increased costs for farmers, is one of the most significant problems (Arce and Caballero, 2025). Current approaches to managing pests and diseases may be inadequate or reactive, creating challenges for timely prediction and control (Demeke et al., 2016).

Although Tanzania has achieved progress in agricultural development, many smallholder farmers, particularly those in rural areas, lack timely and accurate information for disease and pest management. Conventional methods of controlling pests, including extensive pesticide use, can be costly and environmentally harmful. Furthermore, farmers may lack access to prediction models and real-time data that would enable them to take preventative measures before outbreaks occur. This knowledge gap in early diagnosis and forecasting can lead to inefficient farming practices that affect food security and the agricultural sector overall (Arce and Caballero, 2025). Therefore, this study developed a system integrating an ML model for pest and disease prediction under regression to provide farmers accurate, current information on potential dangers to their crops. Using machine learning models, farmers would be able to anticipate pest and disease outbreaks, take

preventative action, and optimise pesticide use. This will not only increase crop yields and contribute to food security but also reduce the adverse environmental effects of excessive chemical use and improve precision farming among Tanzanian maize farmers.

1.2 Study Objectives

1.2.1 General Objective

The general overall objective of this study is to develop a machine learning model (ML) for pest and disease prediction that uses meteorological and environmental data to provide Tanzanian farmers timely, useful insights.

1.2.2 Specific Objectives

i. To determine parameters (historical and real-time climatic data and regional patterns) required for prediction;
ii. To design a model capable of predicting pest and disease outbreaks based on integrated datasets;
iii. To develop an integrated web-based system that delivers region-specific predictions and actionable recommendations to farmers.

2.0 Related Literature Review

The literature review of related studies is presented in this section to highlight the gaps in knowledge regarding the use of developing technologies for pest and disease prediction to increase agricultural productivity, especially in rural areas.

2.1 Crops' Pest and Disease Prediction and Management

Pests and diseases have a considerable impact on agricultural production globally, accounting for 20–40% of crop losses each year (FAO, 2020). The effects are particularly important in Africa, including Tanzania, where access to modern farming equipment and expertise can be limited. Tanzania is affected by pests such as the coffee berry borer, stem borers, and fall armyworm (Spodoptera frugiperda), in addition to diseases including bacterial wilt and cassava mosaic virus. It is thought that these infestations disproportionately harm smallholder farmers, resulting in severe financial losses. Pesticide use, visual diagnosis, and physical inspection are all part of traditional pest and disease management techniques. These methods, which often rely on the expertise of extension agents or farmers, can be labour-intensive and subjective. Furthermore, overuse of chemical pesticides can be detrimental to human health and the environment (Demeke et al., 2016).

Rastogi et al. (2015) suggested a method for diagnosing diseases in leaves using computer vision and fuzzy logic. The study by Ghyar and Birajdar (2017) classified photos of rice leaves as either healthy or unhealthy using image-processing and classification techniques. For image segmentation, k-means clustering was employed in conjunction with preprocessing methods and classifiers, including ANN and SVM.

The work in (Francis and Deisy, 2019) discusses the use of k-means clustering and image-processing techniques for plant disease detection. Kuricheti and Supriya (2019) describe how to identify and categorise turmeric diseases, including leaf spot and leaf blotch, using K-means clustering and SVM. In order to improve image segmentation techniques used for plant disease classification, Rahman et al. (2019) used an improved segmentation approach that achieved higher accuracy than K-means clustering. As a result, many techniques are being used for feature extraction, image segmentation, and image enhancement.

Kulkarni (2018) provided a study on crop disease identification and prediction using the PlantVillage dataset. Transfer-learning models were used together with preprocessing techniques such as segmentation and scaling. However, research utilising AI/ML is still required to forecast maize pests and diseases in the context of Tanzania.

A Convolutional Neural Network (CNN) was used on a dataset of healthy and diseased plant leaves in the study by Francis and Deisy (2019). Several pre-processing techniques were applied before classification. A custom dataset of sugarcane disease images was used with a CNN in the study by Militante et al. (2019a), while Militante et al. (2019b) presented a deep-learning approach for plant leaf detection and disease recognition. These studies demonstrate the potential of CNN-based approaches for plant disease recognition and support the need for integrated AI and ML systems for agricultural applications.

To identify and predict various agricultural pests and diseases, the scientific community has developed a number of machine learning and deep learning algorithms. A review by Shruthi et al. (2019) shows the application of machine-learning classification techniques to plant disease detection. Guo et al. (2020a) developed a deep-learning-based model for plant disease identification and diagnosis. In complex environments, leaves were localised using a Region Proposal Network (RPN), followed by Chan–Vese image segmentation and disease identification. The reported accuracy of the method was 83.57%.

One of the traditional time-series models used in agricultural forecasting is the ARIMA model (Suresh and Priya, 2011). However, the complexity and dynamism of agricultural data can make it difficult for a single linear model to accommodate nonlinearity and short- and long-term interdependence. Researchers have therefore combined traditional statistical models with machine-learning and deep-learning approaches to improve prediction accuracy. For example, Grünig et al. (2021) demonstrated the application of deep neural networks to predict the incidence and phenology of plant pests and diseases, while Temraz et al. (2021) examined data-augmentation approaches for crop-growth prediction under climate uncertainty.

Numerous hybrid models have been applied to different agricultural prediction scenarios. Guo et al. (2020b) evaluated ARIMA, SVM, and LSTM models for forecasting hepatitis E and found that LSTM outperformed the other models. In a sugarcane disease prediction study, Dhawan et al. (2023) used a CNN–LSTM ensemble model for severity-level classification. More broadly, Wang and Li (2025) demonstrated the value of a hybrid ARIMA–LSTM model for sugarcane pest and disease prediction, showing how the combination can capture both linear trends and nonlinear fluctuations.

Additionally, external environmental elements such as weather and climate change must be taken into account when forecasting agricultural pests and diseases. Meteorological variables can strongly influence pest and disease occurrence. Xu et al. (2022) compared ARIMA, SVR, LS-SVR, and ARIMA–LSTM models for drought forecasting using the Standardised Precipitation Evapotranspiration Index (SPEI) in China and found that the hybrid models generally improved prediction accuracy. This provides methodological support for incorporating weather variables into agricultural prediction models.

In research on climate-related forecasting, combining ARIMA with models such as support vector regression (SVR) and least-squares support vector regression (LS-SVR) has also shown excellent prediction accuracy. Xu et al. (2022) found that ARIMA–SVR, LS-SVR, and ARIMA–LSTM models outperformed the corresponding single models for SPEI forecasting. A significant problem in current research is therefore choosing suitable models for particular agricultural contexts and purposes to increase the ability to anticipate complex pest and disease dynamics under changing environmental conditions.

2.2 The Role of AI/ML in Agriculture

The use of AI and ML in agriculture offers a way to address agricultural problems. Large datasets from sources such as satellite imagery, Internet of Things sensors, weather observations, and computer vision systems can be analysed using AI technologies (David, 2023). According to studies, AI/ML is effective at identifying plant diseases and pests. For example, convolutional neural networks trained on leaf images have been applied successfully to agricultural disease detection (Francis and Deisy, 2019). Similarly, machine-learning and deep-learning systems can be used to predict pest and disease occurrence by analysing historical data and meteorological variables.

Although AI/ML has enormous promise, its use in agriculture is uneven, with many advances occurring in settings with greater data and infrastructure availability. One of the difficulties specific to developing nations like Tanzania is limited data availability, where high-quality databases tailored to local crop types, pests, and diseases are often unavailable (Arce and Caballero, 2025). Another issue is that existing AI models may not adequately reflect region-specific factors such as pest life cycles, local weather patterns, and conventional farming practices. Initiatives to use AI solutions in developing countries, therefore, emphasise the

necessity of affordable technology, user-friendly interfaces, and regional data collection.

3.0 Materials and Methods

The key ideas of the research methodology and techniques used in this study are presented in this part. It covers the study area and data sources, data structure, data processing methods, feature selection, and the training and testing process. Others include data gathering and sampling, prediction model construction, and ethical issues.

3.1 Study Area and Data Sources

This study was carried out in Tanzania's southern highland areas of Ruvuma, Iringa, and Njombe. These areas were chosen because, in comparison to other Tanzanian regions, they produce the highest number of different types of crops, especially maize. The chosen areas are appropriate because this study is limited to maize production.

Data used in this study were collected from maize-producing areas in the aforementioned regions of Tanzania. The dataset consisted of agricultural and climatic observations associated with maize pests and disease occurrences. The climatic variables included temperature, relative humidity, rainfall, wind speed, and solar radiation, while historical pest and disease observations were used to represent previous occurrence patterns. The data were collected for the period of 10 years from the Tanzania Meteorological Agency (TMA) and the Tanzania Agriculture Research Institute (TARI). These data were used to develop and evaluate machine learning models for predicting maize pest and disease occurrence.

3.2 Data Collection and Sampling Techniques

Tools for interview guides and questionnaires were used to gather data. Only thirty extension agents and fifty seasoned farmers from each of the three regions were chosen. To identify the appropriate factors for pest and disease predictions, data were gathered using the questionnaire and interview guide. The sampling procedure made use of both simple random sampling and purposeful sampling. All seasoned farmers and agricultural extension agents were chosen through the use of purposeful sampling. Lastly, samples were chosen from the previously formed subgroup of extension agents and farmers using simple random selection. The list of farmers and extension agents with extensive experience is widely known in all regions, which is why simple random sampling is used.

3.3 Data Structure

The dataset consisted of 5000 observations and 20 predictor variables. The predictor variables included temperature (°C), relative humidity, rainfall (mm), wind speed (m/s), solar radiation (W/m^2), and historical pest and disease observations. The target variable represented the observed level of maize pest and disease occurrence.

3.4 Data Pre-Processing

Before model development, the collected data were inspected and cleaned to ensure consistency and completeness. Duplicate records were removed, and missing values were identified and handled by removal. Numerical variables were checked for abnormal values and outliers. Where necessary, numerical features were normalised using min-max, Z-score, and log transformation methods to ensure that variables with different measurement scales did not disproportionately influence model training. The processed dataset was then formatted into predictor and target variables for machine learning.

3.5 Feature Selection

Predictor variables were selected based on their relevance to maize pest and disease occurrence, supported by agricultural and climatic considerations and exploratory analysis of the dataset. The selected features included temperature, relative humidity, rainfall, wind speed, solar radiation, and historical pest and disease observations. These variables were considered because environmental conditions can influence pest and disease development and transmission. The contribution of the selected features was further examined during model development to determine their relative influence on prediction performance.

3.6 Model Training and Testing

The pre-processed dataset was divided into training and testing subsets using an 80:20 ratio. Eighty percent of the observations were used to train the machine learning models, while the remaining 20% were reserved for independent testing of model performance. The models evaluated included Gradient Boosting (GB), Linear Regression (LR), Support Vector Regression (SVR), Long Short-Term Memory (LSTM), and K-Nearest Neighbours (KNN). During training, the models learned the relationship between the selected environmental and historical variables and maize pest and disease occurrence. After training, the models were applied to the testing dataset, which was not used during model fitting. The aforementioned models were selected to be used in this study because the study has both classification and regression outcomes, which necessitate the complementary varieties of the six models. For instance, Logistic Regression, SVM, KNN, RF, and XGBoost can be used for classification, while SVR, Random Forest Regression, and XGBoost Regression can be used for continuous outcomes. The development was performed stage by stage as follows:

Figure 1

Model Development Process

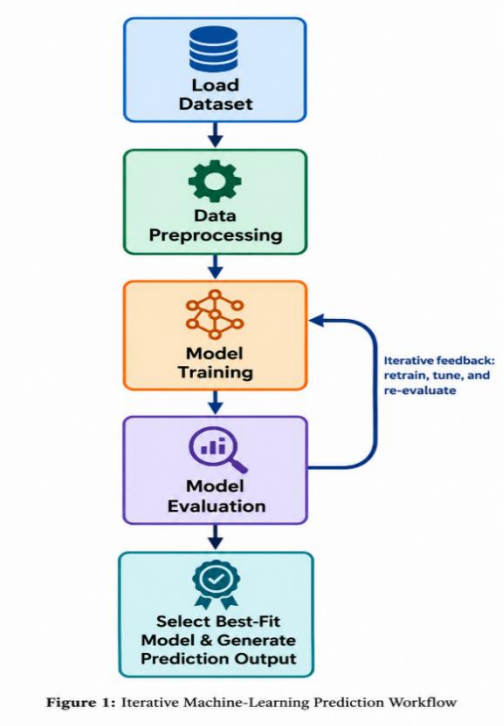


Figure 1: Iterative Machine-Learning Prediction Workflow

Figure 1 illustrates the iterative workflow for developing, executing, evaluating, and selecting a machine-learning model for prediction. The process begins with loading the dataset, followed by data preprocessing to prepare the data for model training. The trained model is then used to generate predictions, and its performance is assessed during the model evaluation stage using appropriate evaluation metrics. Based on the evaluation results, the best-performing model is selected for generating the final prediction output. The feedback loop from model evaluation to model training represents the iterative nature of machine learning, whereby the model can be retrained, its parameters tuned, and its performance re-evaluated when the initial results are unsatisfactory. This iterative process helped improve predictive performance and supported the selection of a suitable model for the agro-pest and disease prediction task.

The following six machine learning models were developed using Python, Pandas, Scikit-learn, and TensorFlow:

3.6.1 Linear Regression (LR)

The baseline model, which captured linear connections between input and target variables, was linear regression. The model's development and training algorithms are described as follows:

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

Where

$\hat{y}$: predicted pest and disease,

β_0: intercept,

β_i: regression coefficients,

x_i: independent variables

The model minimises the Mean Squared Error (MSE) generated by the model. Mathematically, MSE is expressed as:

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

Where,

y_j represents the actual Maize disease,

$\hat{y}_j$ is the predicted disease, and

N is the total number of observations.

The MSE ensures that the trained model produces accurate predictions.

3.6.2 Support Vector Regression (SVR)

Non-linear correlations were captured using SVR in an effort to construct a regression function that allows for minor deviations within an ϵ margin. The following is the SVR regression function:

$$f(x) = w.x + b$$

The optimization problem in SVR is defined as:

$$min\frac{1}{2}\|w\|^2\ s.t|y_i - f(x_i) \leq \epsilon|$$

Where

w is the weight vector, y_i the actual value, and $f(x_i)$ the predicted disease. Predictions that fall within the ϵ margin are accepted as correct, while deviations beyond this margin are penalised.

The training setup used$c = 100$,

$ramma = 0.1,\ \epsilon = 0.1, and\ kernel = "rbf"$

balancing model complexity with prediction accuracy.

3.6.3 K-Nearest Neighbours (KNN)

KNN regression predictspest and disease based on the similarity of new cases to historical records. For any input x, the model identifies the $nearest$ data points and averages their corresponding price values:

$$\hat{y}(x) = \frac{1}{k}\sum_{i\epsilon N_k(x)} y_i$$

Where

N_k(x) represents the set of the closest training points to x.

KNN regression uses the Euclidean distance metric to quantify the similarity between data points in order to predict values. Euclidean distance was used as the similarity metric when the model was set up using k=5.

3.6.4 Random Forest (RF)

Predictions from several decision trees are averaged by Random Forest. To minimize overfitting and variance, each tree was trained on a bootstrapped sample of the data using a random subset of features:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t$$

where

$\hat{y}_t$: prediction from tree t, T: total number of trees

The Random Forest model was configured with 200 trees and a maximum depth of 42. These settings enabled the model to capture non-linear relationships in the data and minimise overfitting.

3.6.5 Gradient Boosting (GB)

Gradient Boosting sequentially builds decision trees, with each new tree correcting the residual errors of the previous one. The prediction at this stage m is given as:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

where

$F_m(x)$: is the boosted model after the mth iteration, $F_{m-1}(x)$: is the model from the previous stage, η: is the learning rate that controls the contribution of each new learner, and $h_m(x)$: is the weak learner trained on the residuals of the previous model.

3.6.6 Long Short-Term Memory (LSTM)

Sequential dependencies in cashew nut prices over time are taken into account using LSTM. Gated memory cells are used to eliminate noise and preserve essential historical data. Three gates forget (f_t), input (i_t), and output (o_t)are included in the key equations to control the information flow.

Forget Gate$f_f = \sigma[w_f(h_{t-1}, x_t) + b_f]$

Where w_f is the weight, h_{t-1} is the output from the previous time step, x_t is the new input, and b_f is the bias? The new information that will be stored in the cell state is determined and updated by the sigmoid layer called the input gate layer. Next, a $tanh$ layer creates a vector of new candidate values that could be added to the state.

Input Gate:$i_t = \sigma[w_i(h_{t-1}, x_t) + b_i]$

Candidate Memory Update:

$$\hat{c}_i = tanh[w_c(h_{t-1}, x_t) + b_i]$$

The old cell state is updated.c_{t-1} into the new state, c_t The old state is multiplied by f_t, forgetting the things that one had decided to forget earlier. Then, $i_t * c_t$ is added. This gives a new candidate value scaled by how updated each value is.

Cell State Update:

$$c_t = f_{t-1} * c_t + i_t * \hat{c}_t$$

A sigmoid layer runs to decide what parts of the state are going to be output. Then, the cell state is put through tanh (to push the values between -1 and 1), and this is multiplied by the output of the sigmoid gate so that only the required parts are output.

Output Gate:

$$o_t = \sigma[w_o(h_{t-1}, x_t) + b_i]$$

Hidden State Update:

$$h_t = o_t * \tanh(c_t)$$

The final output layer matches the hidden states (h_t) with the target variable (y_t) to convert learned temporal patterns into a single prediction value for local prices. The equation is given below:

$$\hat{y}_t = W_{out} * h_t + b_{out}$$

The historical pest and disease data were transformed into a three-dimensional array (samples, timesteps, and features) during the LSTM model's training process. This allowed the model to process input as sequential patterns rather than discrete observations. Key hyperparameters, such as a batch size of 16 and 200 training epochs, were used to establish the model. These parameters balanced predicted accuracy and computational economy while guaranteeing consistent convergence.

3.7 Model Evaluation Metrics

The purpose of the validation step was to evaluate the trained models' predictive performance in order to ascertain the accuracy of the forecasts. As this study used both classification and regression models, standard regression metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2), were used to assess the performance of the model. Further, accuracy, precision, and F1-score were used for classification models.

MAE measures the average absolute difference between the observed and predicted values and provides an easily interpretable measure of the typical prediction error in the same units as the target variable. RMSE measures the square root of the average squared prediction errors and places greater emphasis on larger errors; therefore, it is useful for determining whether a model produces substantial prediction errors. R^2 measures the proportion of variability in the dependent variable that is explained by the model and provides an indication of the model's overall goodness of fit. MAE and RMSE were interpreted such that lower values indicate better predictive performance, whereas higher R^2 values indicate better model performance. The use of these three complementary metrics provides a more comprehensive assessment of predictive accuracy and model fit.

For classification models, accuracy, precision, and F1-score were used to assess predictive performance. Accuracy measures the proportion of correctly classified observations among all observations. Precision measures the proportion of observations predicted as positive that are actually positive, thereby indicating the model's ability to minimise false-positive predictions. The F1-score represents the harmonic mean of precision and recall and provides a balanced measure of classification performance, particularly when the classes are imbalanced. Therefore, higher values of accuracy, precision, and F1-score indicate better classification performance.

3.8 Web-Based Prediction System

The best-performing machine learning model identified during the experimental evaluation has to be integrated into the developed web application to provide real-time predictions to end users. After model training, hyperparameter tuning, and validation, the selected model would be saved together with the required pre-processing components, such as feature scaling or encoding procedures. These components would then be incorporated into the application's prediction backend.

3.9 Ethical Consideration

Research ethics were considered in this study. First, the Tanzania Agriculture Study Institute, which oversees agricultural affairs, gave the author permission to conduct the research. Mzumbe University wrote an introductory letter to the participating agricultural agencies because this is an academic project. According to the information in the questionnaire, the study guaranteed each participant's privacy and

anonymity. Respondents could stop participating in the study at any time and were not required to provide any information. Additionally, participants were informed of the study's purpose, which encouraged them to confidently provide information because it is necessary for academic objectives.

4.0 Results

The experiment's findings for each particular goal are shown in this section. The section shows the outcomes of the factors needed for prediction, created a model that can forecast pest and disease outbreaks using integrated datasets, and created an intuitive web interface that provides farmers with region-specific forecasts and practical suggestions.

4.1 Parameters Influencing Model Prediction

The 5-point Likert scale results of the parameters that were found to be most important in predictive analytics for disease and pest outbreaks are shown in Table 1. The factors that affect pest and disease forecasts were taken into account and measured, including historical data, current weather information, rainfall, temperature, humidity, sun radiation, wind speed, and crop type. The findings show that rainfall, temperature, and humidity significantly influence the prediction of pests and diseases in maize. Additionally, sun radiation, historical data, and wind speed agreed to significantly influence the prediction of pests and diseases. On the other hand, as seen in Table 1, crop type was found to strongly influence the prediction of pests and diseases.

Table 1
Parameters for Agriculture Pest and Disease Prediction

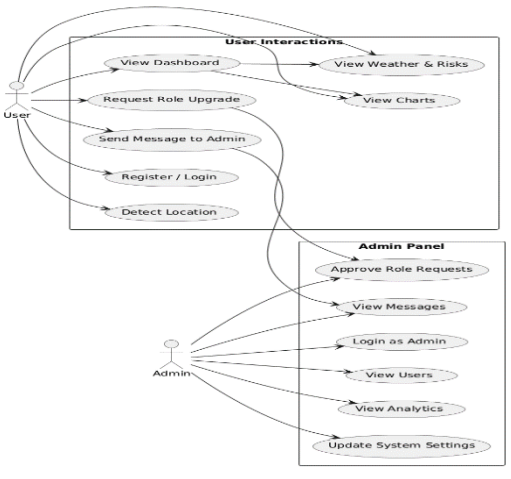
Measured Parameters	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Historical Data				V	
Rainfall					V
Temperature					V
Humidity					V
Solar radiation				V	
Wind speed				V	
Crop types	V				

4.2 Web-Based System Development Predicting Pest-Disease Outbreak

Several features of the system development outcomes in relation to the pest and disease prediction model are presented in this section. The user needs, both functional and non-functional, are first captured in the use case diagram. Second, the ER diagram displays the outcome of how data is saved and retrieved during the prediction process. Thirdly, Fig. 2 also shows the outcome of how the flow chart functions with respect to the prediction model. Lastly, the pest and disease prediction model is introduced.

Use case diagram: The effects of one or more actors' interactions inside the system are described in the use case diagram shown in Fig. 2. Use cases essentially outline how the system responds to user inputs by fulfilling its functional and non-functional needs. Any actor has access privileges to certain functions; an actor is a representation of a person who interacts with the system. We create the users: administrators, farmers, agricultural officers, and researchers—based on the access rights.

Figure 2
A Use Case Diagram



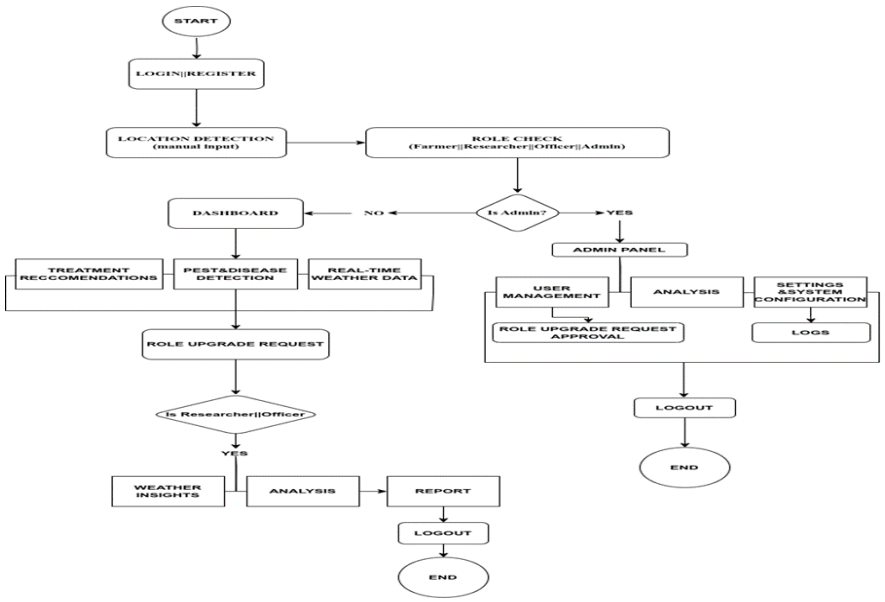
ER-Diagram: The database was created as the findings in Fig. 3 demonstrate after identifying the functional and non-functional requirements indicated in the results in Fig. 2. The ER-diagram illustrates the storage of connected items and their characteristics. The findings in Fig. 3 demonstrate that users, user settings, pest-disease prediction, role requests, weather logs, treatment guides, and contact messages are the most crucial entities that must be kept in the database for the process of pest and disease prediction. In the database for pest and disease prediction, these entities and their characteristics were often highly significant.

Figure 3
E-R Diagram



Flowchart diagram: The process of understanding, evaluating, and improving the pest and disease prediction model is made easier by Fig. 4, which provides a clear and concise explanation of a system or process. The flowchart that illustrates how users of the system log in and carry out specific functions which differ depending on their status is shown in Fig. 4. Farmers, researchers, agricultural extension officers, and administrators are the four categories of users, each having a distinct set of skills.

Figure 4
Flow Chart Diagram Indicating Model Prediction Process



4.3 Models Training and Validation

The results of model construction, training, and validation in connection with disease and pest prediction are shown in this section. Figure 5 displays the outcomes of the machine learning model training. Second, the results of the model validation are also shown in Table 2, which shows which model is the best and most potent.

4.3.1 Models Training Results

Using real-time weather data, humidity, rainfall, wind speed, solar radiation, and temperature, six machine learning models were created and tested to find the best model for forecasting pests and disease in Tanzanian agriculture. Although there were significant departures from actual values, the results show that Linear Regression was able to capture overall patterns in pests and diseases. While Gradient Boosting gave extremely accurate findings with only minor variations as shown in Figure 5, Random Forest produced predictions that were closely linked with actual pests and diseases, minimizing noticeable inaccuracies. SVR, on the other hand, did a poor job of capturing the timing and number of changes. Additionally, KNN generated overly smoothed predictions that missed the prevalence of disease by underestimating abrupt rises and overestimating falls. However, LSTM demonstrated great promise by identifying temporal and sequential pest and disease occurrence inconsistencies, indicating the need for additional tuning and larger datasets. Gradient Boosting demonstrated generally extremely good performance outcomes among the trained models due to its very good performance (accuracy ~70%) and resilience in handling non-linear data, as shown in Figure 5.

4.3.2 Modes Validation Results

Regression metrics MAE, MSE, RMSE, R^2, and their cross-validation equivalents were used in the validation findings to evaluate the prediction accuracy of six models as shown in Table 2. With MAE = 288.425, RMSE = 346.471, and R^2 = 0.338, the results of linear regression demonstrated a moderate level of competence. Stable cross-validation results verified Random Forest's nearly flawless accuracy with almost zero errors and R^2 = 0.522. Additionally, with consistent cross-validation metrics (MAE = 102.0003, RMSE = 0.057, R^2 = 0.7), Gradient Boosting showed exceptional generalization.

SVR, on the other hand, had unstable cross-validation results and large error rates (MAE = 461.577, RMSE = 590.801, R^2 = 0.013). KNN was likewise weak, with MAE = 431.998, RMSE = 450.231, and R^2 = 0.181. Poor generalization was confirmed by cross-validation. LSTM yielded conflicting results: cross-validation showed great potential (CVR^2 = 0.906), while normal validation showed minimal errors but a negative R^2 (−1.362).

Table 2
Model Validation Results Comparing Residue Metrics

Models/Residue Metrics Used	MAE	RMSE	R2
Linear Regression (LN)	288.425	346.471	0.338
Random Forest (RF)	234.224	437.260	0.522
Gradient Boosting (GB)	102.003	0.057 *	0.700
Support Vector Regression (SVR)	461.577	590.801	0.013
K-Nearest Neighbors (KNN)	431.998	450.231	0.181
LongShort-Term Memory (LSTM)	621.010	760.310	0.0024

Overall, the results in Table 2 demonstrate that Gradient Boosting was shown to be an excellent model for forecasting illnesses and pests. Pest population density, expressed as individuals per M2, serves as the study's measurement unit. The target variable ranged from 1 to 1500 M2 per person. The mistake in the real unit is represented by the MAE of 102.03, which is significant for the study. The updated metrics (MAE = 102.03) were shown on the original measurement scale, providing a clearer picture of the physical inaccuracy of the model. The experiment was conducted utilizing the approach of (leave-one group-out), employing data from six (6) distinct farms, five (5) for training and one (1) for testing. The data leakage was confirmed by demonstrating the validity of R2=0.7. In order to give farmers comprehensive information, the goal variable is continuous rather than binary, based

on the selection of Gradient Boosting Regression over categorization.

Figure 5 shows the gradient boosting experiment's results, displaying a distinct linear association with a realistic "cloud" of dots surrounding the identity line (y=x). The 30% of variance that the model is unable to account for is most likely due to variability or unmeasured environmental factors

Figure 5

Gradient Boosting Model Training Visualization

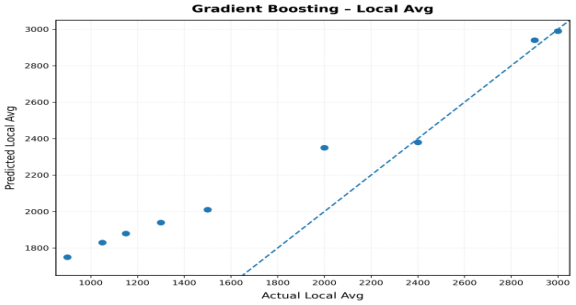


Table 3

Comparison of feature importance Analysis results

Feature / Parameter	LN	RF	GB	SVR	KNN	LSTM
Historical data	0.26	0.30	0.32	0.27	0.25	0.31
Temperature	0.22	0.24	0.25	0.23	0.21	0.24
Humidity	0.19	0.18	0.17	0.20	0.19	0.18
Solar radiation	0.14	0.12	0.11	0.13	0.15	0.12
Wind speed	0.10	0.09	0.08	0.10	0.11	0.08
Crop type	0.09	0.07	0.07	0.07	0.09	0.07
Total	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 3 presents the feature-importance analysis for the six evaluated models: Linear Network (LN), Random Forest (RF), Gradient Boosting (GB), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Long Short-Term Memory (LSTM). The results demonstrate a consistent pattern across all models, with historical data identified as the most influential predictor, followed by temperature, humidity, solar radiation, wind speed, and crop type. Historical data recorded the highest importance across all models, ranging from 0.25 for KNN to 0.32 for Gradient Boosting. The relatively high importance assigned to historical data indicates that previous pest and disease observations provide substantial predictive information for estimating future occurrence. Gradient Boosting assigned the highest importance to historical data (0.32), followed by LSTM (0.31) and Random Forest (0.30). Temperature was consistently ranked as the second most important predictor, with importance values ranging from 0.21 to 0.25. Gradient Boosting recorded the highest temperature importance (0.25), while KNN recorded the lowest (0.21). This consistent ranking across models suggests that temperature is an important environmental factor in the prediction of agro-pest and disease occurrence.

Humidity was the third most influential feature, with values ranging from 0.17 to 0.20. SVR assigned the highest importance to humidity (0.20), whereas Gradient Boosting assigned 0.17. Solar radiation showed moderate importance, ranging from 0.11 to 0.15, while wind speed contributed between 0.08 and 0.11. Crop type recorded the lowest importance in most models, with values ranging from 0.07 to 0.09. Overall, the consistency of the feature rankings across the six models suggests that the prediction task is strongly influenced by historical patterns and climatic variables, particularly historical data and temperature. The results also indicate that the Gradient Boosting model captures the same dominant predictors identified by the other approaches, while assigning comparatively higher importance to historical data and temperature. This provides useful evidence for interpreting the predictive behaviour of the proposed agro-pest and disease prediction system.

4.4 Overall System Evaluation Results

The evaluation findings of the system created for disease and pest prediction are shown in this section. The inputs to be tested, the anticipated system output, and a description of what is being tested were all specified in the test cases. The test scenario for the overall system results is shown in Table 4.

Table 4
Overall System Evaluation Results

ACTIVITY	EXPECTED RESULTs	TEST RESULT
Register as a new user.	User is successfully registered, assigned the default "farmer" role, and redirected to the dashboard.	Successfully
Submit role upgrade request.	Role upgrade request is saved in the database and becomes visible to the admin for review.	Successfully
Admin approves role	User's role is updated accordingly, and the user gains access to new permissions tied to the role.	Successfully
Submit contact message	Contact message is saved in the database and accessible for admin viewing in the message panel.	Successfully
Weather fetched and stored.	Real-time weather data is successfully fetched from the OpenWeather API and saved into the system.	Successfully
Pest prediction triggered	Flask API processes weather data and returns valid, scientifically accurate pest/disease.	Successfully
View analytics dashboard	Dashboard charts render correctly with up-to-date, accurate data reflecting system statistics.	Successfully

4.5 User Evaluation Results of the Agro-Pest and Disease Prediction System

Table 5 presents the results of the user evaluation of the proposed Agro-Pest and Disease Prediction System. The evaluation was conducted using a five-point Likert scale, where 1 represented strongly disagree and 5 represented strongly agree. The evaluation focused on usability, ease of navigation, clarity and relevance of prediction results, usefulness for pest and disease management, system responsiveness, and overall user satisfaction. The illustrative results show an overall mean score of 4.40, indicating a very positive user perception of the proposed system. The highest-rated criterion was the usefulness of the system for pest and disease management (mean = 4.50), followed by willingness to recommend the system to other maize farmers and agricultural stakeholders (mean = 4.47). The lowest mean score was observed for system responsiveness (mean = 4.28), although this still indicates a positive user perception. These results in Table 4 suggest that, in addition to its machine-learning performance, the proposed system has the potential to provide a usable and practical decision-support tool for maize pest and disease management.

Table 5
User Evaluation Results of the Agro-Pest and Disease Prediction System

S/N	Evaluation criterion	Mean	SD	Interpretation
1	The system is easy to use	4.45	0.63	Very satisfactory
2	The system is easy to navigate	4.45	0.63	Very satisfactory
3	The pest and disease predictions are easy to understand	4.42	0.61	Very satisfactory
4	The prediction results are relevant to maize production	4.36	0.72	Very satisfactory
5	The system provides useful information for pest and disease management	4.50	0.60	Very satisfactory
6	The system helps users make timely farming decisions	4.31	0.70	Very satisfactory
7	The system responds quickly to user requests	4.28	0.75	Very satisfactory
8	The information provided by the system is clearly presented	4.40	0.65	Very satisfactory
9	I am satisfied with the overall performance of the system	4.43	0.62	Very satisfactory
10	I would recommend this system to other maize farmers/agricultural stakeholders	4.47	0.64	Very satisfactory
	Overall	**4.40**	**0.66**	**Very satisfactory**

5.0 Discussion

5.1 Parameters Required for Pest and Crop Disease Prediction

The study's findings, as shown in Table 1, indicate a number of factors that were looked into as inputs for agricultural pest and disease prediction. Temperature, humidity, solar radiation, rainfall, wind speed, historical data, and current weather are all included in this parameter set. According to the study's findings, temperature, humidity, and rainfall have the most significant effects on

the process of predicting pests and diseases, which is corroborated by earlier research. Chen et al. (2020b), Lu et al. (2020), and Manjari et al. (2022) examined the influence of climatic variables on pest and disease prediction. Wang and Li (2025) confirmed the importance of humidity, temperature, and rainfall characteristics for sugarcane pest and disease forecasting. David (2023) similarly reviews the use of meteorological variables in weather-based disease and pest prediction. As a result, meteorological characteristics are considered important indicators for forecasting disease or pest outbreaks.

Wang and Li (2025) confirmed that humidity, temperature, and rainfall are important to the accuracy of the model for forecasting pests and diseases. Pest and disease outbreaks can be influenced by variables such as minimum and maximum temperature, temperature difference, humidity measures, and rainfall measures. David (2023) likewise emphasises the importance of weather variables in weather-based disease and pest predictions. Since weather parameters affect the life cycles of pathogens and pests under suitable climatic conditions, real-time weather conditions can be important predictors of disease or pest occurrence.

5.3 Pest and Crop Disease Prediction Model

Using various machine-learning models, numerous studies have been carried out globally to forecast pests and agricultural diseases in different crop varieties (David, 2023; Kittakorn, 2022; Manjari et al., 2022; Wang and Li, 2025). Six machine-learning models—linear regression, random forest, gradient boosting, SVR, KNN, and LSTM—were trained and evaluated separately for this work. The results showed that Gradient Boosting produced the best performance among the trained models. In contrast to the results of this study, Wang and Li (2025) demonstrated that a hybrid ARIMA–LSTM model can improve pest and disease prediction by combining linear and nonlinear components. Guo et al. (2020b), in a different application area, found that LSTM outperformed ARIMA and SVM for hepatitis E prediction. These studies illustrate that model performance is context-dependent and that hybrid approaches can be useful where both linear and nonlinear temporal patterns are present.

5.0 Performance Evaluation Metric Comparison

The effectiveness of linear regression, random forest, gradient boosting, SVR, LSTM, and KNN models in pest and disease prediction—particularly in maize pest and disease prediction—was compared in this work. Gradient boosting demonstrated strong generalisation and produced excellent results (MAE = 102.03, RMSE = 0.057, R^2 = 0.70) for the study dataset. The results can be compared with Chen et al. (2020a), who found that gradient boosting performed well in predicting grape downy mildew incidence and severity. Therefore, the gradient boosting model shows promise for forecasting maize pest and disease based on the validation comparisons in this study.

6.0 Conclusion

The principal conclusion of this research is the development and deployment of a localised, climate-aware gradient boosting model integrated into a web-based prediction system for forecasting maize pests and diseases in the Southern Highlands of Tanzania. The study therefore contributes both a context-specific machine-learning approach and a practical decision-support platform aimed at enabling timely pest and disease management and improving maize productivity. Specifically, the conclusion made in this study relies on its contributions in several ways: First, using localised Tanzanian datasets, such as pest/disease and environmental data relevant to maize production in the Southern Highlands of Tanzania, addresses the limitations of models developed from non-local datasets. Second, this study incorporates a climate-based prediction component through climatic/environmental factors to predict conditions associated with maize pests and diseases. Third, the application of machine learning models, like gradient boosting, has made a substantial contribution to learning relationships between environmental/agricultural factors and pest or disease occurrence. Fourth, this study has further integrated the prediction model into a web-based agro-pest and disease prediction system, which moves beyond a

standalone ML experiment and can provide actionable information to farmers or agricultural stakeholders. Finally, the findings of this study ultimately contribute to supporting early detection and preventive decision-making, which helps reduce pest and disease-related maize losses while improving productivity in the Southern Highlands.

6.1 Implication of the Study

6.1.1 Practical Implication

The study's findings practically provide a foundation for the nation's agriculture sector and important stakeholders to appropriately adopt and use machine learning prediction systems with suitable prediction parameters, as the study examined the prediction of pest and disease among farmers. Historical data, humidity, temperature, solar radiation, wind speed, real-time weather data, and rainfall are among the pest and disease prediction parameters that have been proven to have a substantial impact on the process. The Ministry of Agriculture is also informed by the study's findings to revise the National ICT policy to incorporate the use of ICTs, including machine learning and related technologies, as determined by the study's conclusion.

6.1.2 Theoretical Implication

In the context of Tanzanian agriculture, there hasn't been much study done on the use of machine learning technologies, namely gradient boosting, for pest and crop disease prediction. The findings of this study contribute to the literature on the use of the gradient boosting model, which is deemed suitable for studies with small datasets. The findings of this study provide Tanzanian agricultural organisations and farmers with the information they need to use machine learning systems to increase productivity. Additionally, the research fills in the knowledge gap in the literature about the potential applications of gradient boosting as a machine learning model for crop disease and pest prediction to increase agricultural output in Tanzania.

6.2 Research Limitation and Recommendation

Despite the promising predictive performance of the Gradient Boosting model, the possibility of overfitting cannot be completely excluded. The model may learn location-specific, seasonal, or noise-related patterns in the training data, particularly if the dataset is limited in size or diversity. Furthermore, the dataset may not fully represent the spatial and temporal variability of maize production and pest and disease occurrence across the Southern Highlands of Tanzania. Limitations may also arise from class imbalance, missing or inaccurately recorded observations, and the exclusion of potentially important factors such as soil properties, maize varieties, farming practices, and pesticide application. Consequently, the reported model performance should be interpreted with caution, and its generalisability to other seasons, districts, and maize-growing regions should be confirmed through independent external validation. Future research should expand the dataset across multiple locations and growing seasons and evaluate the model using spatial and temporal cross-validation to provide a more robust assessment of its predictive capability.

7.0 Conflict-of-Interest Statement

I declare that no conflict of interest was revealed in this study.

8.0 Acknowledgement and Funding

I acknowledge the Tanzania Agriculture Research Institute for their permission to conduct the research. I also acknowledge Mzumbe University for writing an introductory letter to the participating agricultural agencies. This research did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

9.0 References

Arce, C.E., & Caballero, J. (2025). Tanzania Agricultural Risk Assessment: World Bank Group Report Number 94883-TZ.

Chen, M., Brun, F., Raynal, M., & Makowski, D. (2020a). Forecasting severe grape downy mildew attacks using machine learning. *PLoS ONE*, *15(3)*, e0230254. doi:10.1371/journal.pone.0230254.

Chen, P., Xiao, Q., Zhang, J., Xie, C., & Wang, B. (2020b). Occurrence prediction of pests and diseases by bidirectional long short-term memory networks with climate and atmospheric circulation. *Computers and Electronics in Agriculture*, *176*, 105612.

David, D. (2023). Weather-Based Prediction Models for Disease and Pest Using Machine Learning: A Review. *Asian Journal of Agricultural Extension, Economics & Sociology*, *41(11)*, 334–345. Article no. AJAEES.103172.

Demeke, M., Kiermeier, M., Sow, M., & Antonaci, L. (2016). *Agriculture and Food Insecurity Risk Management in Africa*.

Dhawan, N., Kukreja, V., Sharma, R., Vats, S., & Verma, A. (2023). Deep learning-based sugarcane downy mildew disease prediction using CNN-LSTM ensemble model for severity level classification. In Proceedings of the *14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 6–8 July 2023.

Food and *Agriculture* Organization of the United Nations (FAO). (2020). *International Year of Plant Health 2020*. https://www.fao.org/plant-health-2020/.

Francis, M., &Deisy, C. (2019). Disease detection and classification in agricultural plants using convolutional neural networks—A visual understanding. In *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, 1063-1068. doi:10.1109/SPIN.2019.8711701.

Ghyar, B.S., &Birajdar, G.K. (2017). Computer vision-based approach to detect rice leaf diseases using texture and colour descriptors. In *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 1074–1078. doi:10.1109/ICICI.2017.8365305.

Grünig, M., Razavi, E., Calanca, P., Mazzi, D., Wegner, J.D., & Pellissier, L. (2021). Applying deep neural networks to predict incidence and phenology of plant pests and diseases. *Ecosphere*, *12*, e03791.

Guo, Y., Feng, Y., Qu, F., Zhang, L., Yan, B., &Lv, J. (2020b). Prediction of hepatitis E using machine learning models. *PLoS ONE*, *15(9)*, e0237750. doi:10.1371/journal.pone.0237750.

Guo, Y., Zhang, J., Yin, C., Hu, X., Zou, Y., Xue, Z., & Wang, W. (2020a). Plant disease identification based on a deep learning algorithm in smart farming. *Discrete Dynamics in Nature and Society*, *2020*, 2479172. doi:10.1155/2020/2479172.

Kittakorn, S. (2022). Weather-based rice blast disease forecasting. *Computers and Electronics in Agriculture*, *193*, 106685.

Kulkarni, O. (2018). Crop disease detection using deep learning. In 2018 4th International Conference on Computing Communication Control and Automation (ICCUBEA), 1–4. doi:10.1109/ICCUBEA.2018.8697390.

Kuricheti, G., &Supriya, P. (2019). Computer vision-based turmeric leaf disease detection and classification: A step to smart agriculture. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 545–549. doi:10.1109/ICOEI.2019.8862706.

Lu, W., Newlands, N.K., Carisse, O., Atkinson, D.E., & Cannon, A.J. (2020). Disease risk forecasting with Bayesian learning networks: Application to grape powdery mildew (Erysiphe necator) in vineyards. *Agronomy*, *10(5)*.

Manjari Singh, ParulSetiya, Anand Kumar Tewari, & Ajeet Singh Nain. (2022). Comparison of machine learning and regression approaches to forecasting Alternaria blight epidemic of Indian mustard. *The Pharma Innovation Journal*, *11(6)*, 2974–2984.

Militante, S.V., Gerardo, B.D., & Medina, R.P. (2019a). Sugarcane Disease Recognition using Deep Learning. In *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, 575 578.doi:10.1109/ECICE47484.2019.8942690.

Militante, S.V., Gerardo, B.D., & Dionisio, N.V. (2019b). Plant leaf detection and disease recognition using deep learning. In *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, 579-582. doi:10.1109/ECICE47484.2019.8942686.

Mpogole, H., Dimoso, P., &Mayaya, H. (2020). *Agriculture* for Rural Development in Tanzania. In *The Game Changer: President Magufuli's First Term in Office*. TEMA Publishers Co. Limited.

National Bureau of Statistics (NBS). (2022). The Economic Survey 2022. Dodoma, Tanzania.

Rahman, M.A., Islam, M.M., Shahir Mahdee, G.M., & Ul Kabir, M.W. (2019). Improved segmentation approach for plant disease detection. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT 2019), 1-5. doi:10.1109/ICASERT.2019.8934895.

Rastogi, A., Arora, R., & Sharma, S. (2015). Leaf disease detection and grading using computer vision technology & fuzzy logic. In *2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, 500-505. doi:10.1109/SPIN.2015.7095350.

Shruthi, U., Nagaveni, V., & Raghavendra, B.K. (2019). A review on machine learning classification techniques for plant disease detection. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*,-281-284. doi:10.1109/ICACCS.2019.8728415.

Skendžić, S., Zovko, M., Živković, I., Lešić, V., &Lemić, D. (2021). The Impact of Climate Change on Agricultural Insect Pests. *Insects*, *12*, 440.

Suresh, K.K., & Priya, S.R.K. (2011). Forecasting sugarcane yield of Tamil Nadu using ARIMA models. *Sugar Tech*, *13*, 23–26.

Temraz, M., Kenny, E., Ruelle, E., Shalloo, L., Smyth, B., & Keane, M. (2021). Handling Climate Change Using Counterfactuals: Using Counterfactuals in Data Augmentation to Predict Crop Growth in an Uncertain Climate Future. In *Case-Based Reasoning Research and Development: 29th International Conference, ICCBR 2021*; Springer.

Wang, M., & Li, T. (2025). Pest and Disease Prediction and Management for Sugarcane Using a Hybrid Autoregressive Integrated Moving Average—A Long Short-Term Memory Model. *Agriculture*, *15*, 500. https://doi.org/10.3390/agriculture15050500.

Xu, D., Zhang, Q., Ding, Y., & Zhang, D. (2022). Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. *Environmental Science and Pollution Research*, *29*, 4128–4144.