Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Machine Learning for Tomato and Chilli Pepper Disease-Risk Assessment Using IoT, Remote Sensing and Geospatial Data: Evidence from Smallholder Farms in Ibadan, Nigeria

Idowu Olugbenga ADEWUMI

adewumi.io@fcaib.edu.ng

Federal College of Agriculture

Adedeji Emmanuel ADEOYE

Federal College of Agriculture

Oluwaseyi Funmi Afe

Lead City University

Azeez Ajani WAHEED

Lead City University

Samuel Temitope IJIKOYEJO

Lead City University

Research Article

Keywords: multimodal machine learning, IoT, NDVI, remote sensing, disease risk, precision agriculture

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11627738/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multimodal Machine Learning for Tomato and Chilli Pepper Disease-Risk Assessment Using IoT, Remote Sensing and Geospatial Data: Evidence from Smallholder Farms in Ibadan, Nigeria

Idowu Olugbenga ADEWUMI[*1], Adedeji Emmanuel ADEOYE[3], Oluwaseyi Funmi Afe[4], Azeez Ajani WAHEED[5] and Samuel Temitope IJIKOYEJO[6]

[1*]Department of Computer Science,
School of Engineering,
Federal College of Agriculture, Ibadan, Nigeria.
adewumi.io@fcaib.edu.ng

[2]Computer Science Department,
Faculty of Natural and Applied Sciences,
Lead City University Ibadan, Nigeria.
bakare_a_n@yahoo.com

[3]Department of Agricultural and Bio-environmental Engineering,
School of Engineering,
Federal College of Agriculture, Moor Plantation, Ibadan, Nigeria
adeoyedeji4@gmail.com

[4]Department of Computer and Information Science,
Faculty of Natural and Applied Science,
Lead City University, Ibadan Nigeria.
afeseyi@yahoo com

[4]Department of Computer and Information Science,
Faculty of Natural and Applied Science,
Lead City University, Ibadan Nigeria.
noble@rotarydistrict9126.org

[4]Department of Computer and Information Science,
Faculty of Natural and Applied Science,
Lead City University, Ibadan Nigeria.
article@rotarydistrict9126.org

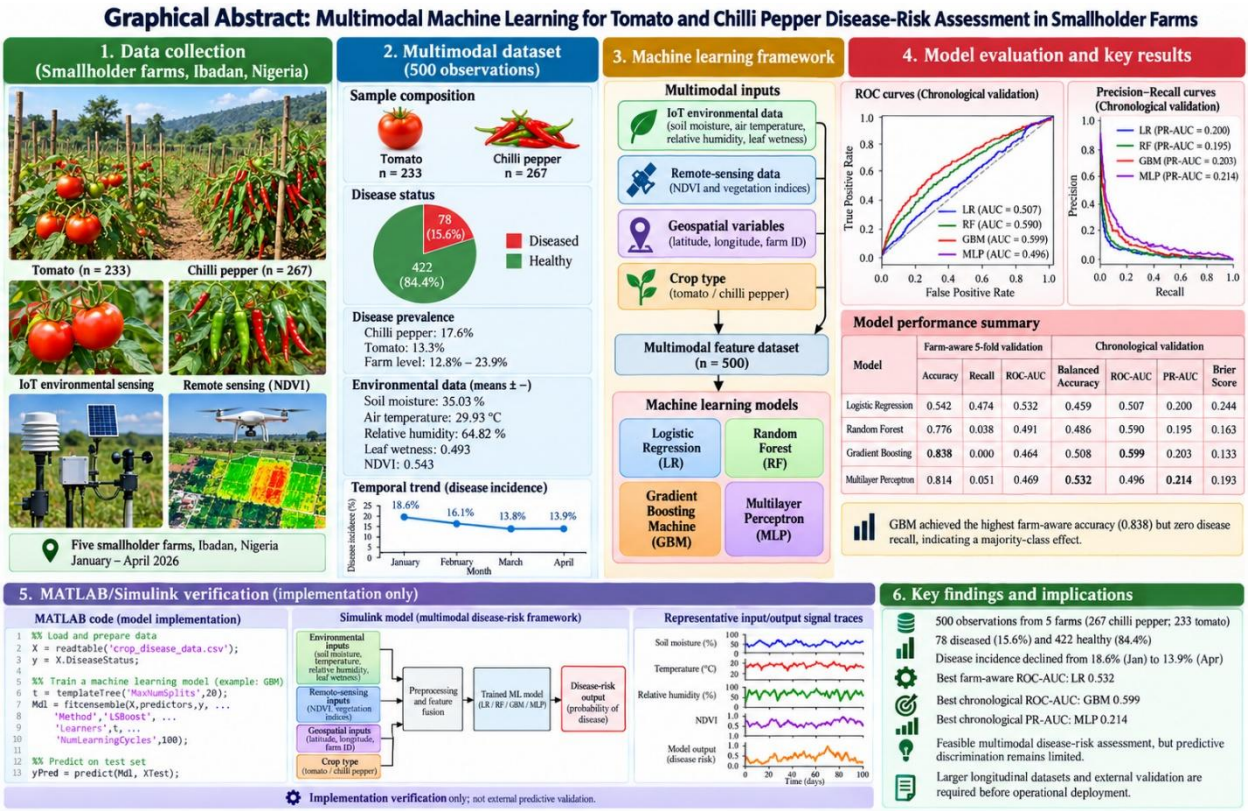
*Corresponding Author: Idowu Olugbenga ADEWUMI,
+2348023821869, adewumi.io@fcaib.edu.ng
Orcid: https://orcid.org/0000-0002-7005-3306

Abstract

In smallholder farming systems, crop disease monitoring can be limited due to lack of field observations, environmental variation, class imbalance, and poor location and temporal validation. The study conceptualized and tested a multimodal machine-learning framework for assessing disease risks of tomato and chilli pepper production in five smallholder farms in Ibadan, Nigeria, during January to April 2026, based on data collected from the IoT and remote sensing devices, crop type, and geospatial information. The data consisted of 500 observations, of which 267 belonged to the chilli pepper and 233 to the tomato species, with 78 (15.6%) and 422 (84.4%) observations being diseased and healthy respectively. The incidence of the disease was 17.6% in chilli pepper and 13.3% in tomato; farm level prevalence ranged from 12.8% to 23.9%. Mean soil moisture, air temperature, relative humidity, leaf wetness, and NDVI were 35.03%, 29.93 °C, 64.82%, 0.493, and 0.543, respectively. Farm-aware five-fold and chronological validation were performed to assess Logistic Regression, Random Forest, Gradient Boosting Machine and Multilayer Perceptron. Farm aware validation gave Logistic Regression the highest ROC-AUC (0.532), recall (0.474) and the highest number of correct predictions (0.838) while Gradient Boosting had no disease recall. Chronological validation showed that Gradient Boosting had the best ROC-AUC (0.599) and lowest Brier score (0.133), while Multi-layer Perceptron had the highest balanced accuracy (0.532) and PR-AUC (0.214). The MATLAB/Simulink supported implementation verification was only performed. Overall, the predictive discrimination was still very small, which means that more longitudinal data and external validation are required before operational use.

Keywords: multimodal machine learning, IoT, NDVI, remote sensing, disease risk, precision agriculture

Graphical Abstract



1. Introduction

Crop diseases are a significant factor in hampering the productivity of agriculture, food security and farm income, especially in smallholder production systems of developing countries. Economically important horticultural crops in Nigeria are tomato (Solanum lycopersicum) and chilli pepper (Capsicum annuum) crops that are often affected by fungal, bacterial and viral pathogens. Several environmental, crop, spatial, and management factors contribute to the occurrence and progression of disease, such as temperature, relative humidity, soil moisture, leaf wetness, canopy condition, pathogen pressure, and farm level management (Odounfa et al., 2025; Wang et al., 2025). Therefore, early detection and risk assessment of disease are essential to help the implementation of appropriate disease management interventions to minimize unnecessary production losses.

Smallholder agriculture is typically managed with conventional disease surveillance, including scouting and visual evaluation, and expert diagnosis. While these methods are still useful, the implementation of continuous monitoring across geographically dispersed farms is still difficult and labor-intensive, and they are subjective and time-consuming (Jena et al., 2024). These constraints have spurred more interest in data-driven ag tech that can offer more frequent, objective, scalable crop health inspections. Artificial intelligence (AI), machine learning, the Internet of Things (IoT), remote sensing, and Geographic Information Systems (GIS) have

consequently become important components of precision agriculture and automated crop-health monitoring systems (Nawaz et al., 2025; Zhu et al., 2024).

Repeated monitoring of microclimatic variables that could be linked to pathogen establishment and disease development is possible with IoT based environmental sensing. Soil moisture, soil temperature, soil relative humidity and leaf wetness are examples of measurements which can be used to indicate the conditions in which disease might occur. Environmental measurements, however, cannot identify disease as all similar conditions can happen in both healthy and diseased crops. They are therefore more useful when combined with other metrics of vegetation health, crop and spatial characteristics.

In addition, the remote sensing can offer another source of information, as it is possible to assess the condition of the vegetation in wider spatial extents. Normalized Difference Vegetation Index (NDVI) is widely used as an indicator of vegetation condition and vigor based on the difference in reflectance values of red and near-infrared bands (Cai et al., 2023; Chin et al., 2023). However, NDVI is not disease specific. Water stress, nutrient deficiency, heat stress, pest damage, or other physiological disturbances could also cause variations in vegetation condition. Therefore, NDVI is not a direct disease indicator and the interpretation should be based on complementary crop health information. It can then be used with environmental sensing and spatial information as part of a multimodal framework to enhance its analytical value.

GIS also helps in assessing the health of crops by offering a geographical backdrop to the disease observations. The incidence of disease can differ between farms due to the differences in pathogen pressure, distribution of hosts, local microclimate, soil characteristics, drainage conditions and management. The geographic coordinates and the farm identity may then be associated to environmental and remotely sensed parameters, thus reflecting spatial heterogeneity that is not necessarily sampled by individual sensors. These are especially important when the landscape of smallholders is heterogeneous, which can mean that the agronomic and environmental conditions are significantly different over short distances.

While digital agriculture research is booming, considerable work on plant diseases is still being conducted on a single data modality, specifically image-based classification through curated datasets and deep-learning algorithms (Jafar et al., 2024; Pacal et al., 2024). These strategies have shown great promise in diagnosing disease with visible symptoms but in field situations, the presence of disease depends on more than just visual symptoms. Contextual information—such as environmental exposure, vegetation condition, crop characteristics, temporal variation, geographical location—can also provide useful information. Moreover, multimodal machine learning offers a potential solution to bring all these diverse data sources together into a single analytical framework, and may thus contribute to a more holistic evaluation of the condition of the crops.

The combination of multiple modalities, however, does not necessarily result in better predictive performance. The added data source should be a source of useful information and not simply add

to the complexity of the model. Otherwise, multimodal models might be influenced by irrelevant predictors, overfitting, data-quality issues, or information leakage. The contribution of individual modalities should thus be assessed in terms of appropriate baseline comparisons, ablation analysis, and appropriate and rigorous validation procedures.

Validation design is especially relevant in agricultural machine learning applications, especially when observations are clustered by farm, plot, location and/or growing season. Random train-test splitting to split the data between the model development and evaluation sets may result in similar observations from the same farm being included in both sets, which can enable the model to learn location-specific trends instead of true disease relationships. Leakage can result in overly positive measures of model success. Farm-aware validation is tougher as it requires observations from the same farm to be in the same validation group, whereas chronological validation is concerned with whether the patterns learnt from earlier observations are transferable to later observations. In this regard, the present study utilizes both farm-aware and chronological validation methods instead of a conventional random splitting method.

The presence of class imbalance is another challenge in crop-disease modelling as healthy observations are often in majority compared to diseased observations. In these settings, the overall accuracy may be deceptive because a classifier might have a high accuracy rate by very frequently predicting the majority healthy class. In addition to balanced accuracy, precision and recall, F1-score, area under the receiver operating characteristic curve (ROC-AUC), area under the precision-recall curve (PR-AUC), and Brier score should be considered for evaluating the model. These measures give a more informative evaluation of the performance of detection, discrimination and probabilistic, especially if the practical goal is to find some relatively infrequent diseased observations.

Especially in the context of smallholder agriculture in Sub-Saharan Africa, there is a need to use rigorous multimodal evaluation. The use of AI for plant disease research has grown at a faster rate, but there is little empirical evidence of the integration of variables from the field, environmental sensing, remote-sensing indicators of vegetation, plant inputs, and geospatial data for disease-risk assessment across smallholder field environments in West Africa. Furthermore, there have been fewer studies that test the ability of multimodal models to generalize to geographically different farms and fewer that test the ability of multimodal models to be stable when measured over time. This is a significant methodological and geographical research gap since models working well in one farm or time may not work well in another farm or on later observations.

In this light, the present study assesses a multimodal machine-learning framework for the prediction of disease risk in tomato and chilli pepper production system in five smallholder farms located in Alako of Ido Local Government Area, Ibadan, Nigeria. Empirical framework includes the environmental variables from IoT (soil moisture, soil temperature, relative humidity, leaf wetness) and the remote sensing derived NDVI data, crop type, and the GIS derived spatial variables. Four machine learning algorithms, namely Logistic Regression, Random Forest,

Gradient Boosting Machine, and Multilayer Perceptron are tested with farm-aware and chronological validation processes. To explicitly handle class imbalance, and to prevent using overall accuracy as the sole measure of model performance, multiple classification, discrimination, and probabilistic measures are used.

The study has threefold contribution. First, it offers field-based evidence for the integration of environmental, remote-sensing, crop and spatial information for crop-disease risk assessment in a context that is not well represented in West Africa, namely smallholder farming. Second, it uses farm-aware validation and chronological validation as two different ways of evaluating model generalization to minimize spatial and temporal information leakage. Third, it does not take for granted that multimodal integration automatically leads to high-performance prediction, but it critically investigates its practical capability and limitations in situations that are defined by class imbalance, spatial heterogeneity, temporal variation and relatively small sample size. This study thus offers a conservative and practically applicable assessment of multimodal agricultural machine learning for smallholder disease-risk surveillance.

2. Literature Review

2.1 Digital Technologies for Crop Disease Assessment

The spread of artificial intelligence, machine learning, Internet of Things sensing, remote sensing, and geospatial technologies have revolutionized the process of crop-health monitoring from manual inspections to relying on data. Diagnosis of diseases in traditional smallholder crops is frequently reliant on scouting and visual diagnosis in the field done at regular intervals. These methods are still relevant, but they require labor, subjectivity in the observer, infrequent monitoring and access to specialist knowledge. Digital technologies offer some of these benefits, allowing for multiple data collection events and the computation of crop, environmental, and spatial data.

Machine-learning techniques are especially pertinent as crop disease is not necessarily correlated with a single predictor. Rather, disease risk could be influenced by the interaction between environmental conditions and canopy status, crop type, field location, and disease management context. Machine-learning models can thus offer the flexibility of an analytical approach to the evaluation of non-linear or multi-variate relationships which may be hard to assess by using conventional statistical methods. In recent years, machine-learning and deep-learning techniques have been increasingly used in plant-disease research to classify plant conditions and identify patterns related to diseases (Jafar et al., 2024; Pacal et al., 2024). Most of this literature, however, has been focused on image classification of a disease in a controlled or curated setting, as opposed to multimodal field data that is heterogeneous.

Another factor is that usefulness of agricultural AI is not just about predictive accuracy. Smallholder models must also be robust across farms, be stable under variable environmental conditions, be sensitive to a few minor cases of disease and be interpretable. The requirements are particularly relevant in tropical agri-environmental areas, where environmental conditions and

disease pressure could differ significantly from one area to the other and from one year to the next. This has led to the adoption of multimodal methods that can incorporate other information sources instead of a single sensing modality in recent years.

2.2 IoT Based Environmental Monitoring and Disease Risk

With IoT-based sensing, there is a way to monitor environmental conditions continuously or frequently that can impact disease and plant health. The soil moisture, air temperature, relative humidity, and leaf wetness are factors that are of special interest because they affect pathogen survival, infection conditions, and host susceptibility. Repeated observation of these variables can be used to get temporal information in field-based disease surveillance, which can be achieved only inefficiently by occasional manual observation.

The main benefit of IoT monitoring is therefore its ability to capture the changes in the environment that happen at a short time scale. Combination of these measurements with disease observations can then be used to determine by machine learning whether certain environmental combinations are correlated with an increased likelihood of disease. Research on automated crop-health monitoring has therefore been gaining more interest in the crop IoT sensing and computational intelligence integration (Ibrahim et al., 2025).

However, there are significant challenges to be faced in the application of IoT in smallholder agriculture. Data quality can be impacted by sensor calibration error, lack of data, network instability, power restrictions, equipment expense, and maintenance. These issues are especially important in resource-limited environments where it may be challenging to operate technically complex systems over the long-term. Therefore, the reliability of the sensor, calibration and missing data management procedures, and the incorporation of sensor information with other crop-health indicators are as important as the predictive algorithms in disease assessment using IoT.

In this limitation, as it were, one should not consider environmental sensing as a single source of evidence for disease. A specific moisture, temperature and humidity may be conducive to disease, but by itself does not prove disease is present. Contextual predictors are therefore most useful as a combination of the variables of the IoT with the other variables of the vegetation, spatial and field-observations.

2.3 Remote Sensing and NDVI in Crop Health Monitoring

Remote sensing can augment the point-based information coming from ground measurements by providing a spatial dimension to the information on vegetation condition that can be used to monitor crop condition. Using multispectral satellite and unmanned aerial vehicle (UAV) imagery, canopy vigour, vegetation density and physiological stress can be assessed on a larger spatial scale than would be feasible by ground-based observations. One of the most commonly used vegetation indices is the Normalised Difference Vegetation Index, which is a summary of the difference between red and near-infrared reflectance data, and is a broad indicator of the greenness and condition of vegetation canopies.

NDVI has been extensively used for crop-health monitoring, stress assessment and disease-surveillance research. In addition to the above, UAV platforms can deliver relatively high spatial-resolution imagery and could be beneficial for subfield (within field) crop condition assessment, whereas satellite images can deliver wider spatial coverages (Cai et al., 2023; Chin et al., 2023; Zhu et al., 2024).

NDVI is not disease-specific, however. Besides disease, water stress, nutrient deficiency, pest damage, heat stress, and other physiological disturbances that may affect vegetation could lead to reduced vegetation index values. Therefore, it cannot be interpreted as a direct diagnosis of disease, for this reason. It is most useful when combined with other data sources which may provide information to help describe the context of the vegetation stress. As part of a multimodal concept, NDVI can be used to monitor the condition of the canopy, whereas the IoT measurements give information on the local microclimate, and geospatial variables give the information about the context of the location.

Thus, from the literature, it is clear that remote sensing provides good spatial information but cannot be a standalone disease predictor. This is helpful for multimodal fusion strategies, where remote-sensing indicators are not absolute proof, but are used instead as complementary indicators.

2.4 GIS and Spatial Heterogeneity in Crop Disease

Pathogen pressure, host distribution, environmental factors, soil properties and management practices all differ between farms and fields and are often spatially variable, which can lead to the spatial heterogeneity of crop diseases. GIS offers a systematic way of associating crop observations with a geographic site and thus allows for studying the spatial aspect of disease occurrence.

For variation that is not completely described by environmental measurements, spatial variables like latitude, longitude, farm identity, plot location and geographical proximity can be used. For instance, neighboring observations might have the same microclimatic conditions or farm management and farms in other regions may have varying disease pressures. GIS derived variables can thus help to enhance knowledge of spatial patterns of diseases, and how spatial variation interacts with other predictor domains.

However, a distinction must be maintained between the two 'spatial visualization' and 'formal spatial inference'. Maps of apparent clusters or disease-intensity patterns can be helpful for exploratory analysis, but cannot be considered as statistically significant spatial dependence. If the researcher is going to draw inferences about the clustering or hotspots, formal procedures like Global Moran's I, Local Moran's I, and Getis-Ord statistics must be used. The manuscript appropriately acknowledges that exploratory maps should not be used as formal proof of spatial auto-correlation without the proper spatial-weighing structures and statistical testing.

This distinction is especially relevant in small samples of a few farms since the differences in prevalence may be more likely interpreted as a systematic spatial pattern.

2.5 Multimodal Machine Learning and Data Fusion

Multimodal machine learning integrates data from various sensing/measurement modalities under the same analysis framework. In the context of assessing agricultural diseases relevant modalities include sensor measurements of the environment, remote-sensing indices, geographical variables, crop characteristics, weather information and features extracted from images. The concept of multimodal integration is based on the idea that the occurrence of a disease is caused by a complex interaction between host, environment, spatial context and the physiological state of the plants.

A widely used integration approach is feature-level fusion where features from various modalities are preprocessed and then fused together for model training. In this way, algorithms like Logistic Regression, Random Forest, Gradient Boosting Machine and Multi-Layer Perceptron can consider the relationships between different heterogeneous predictors at the same time. Tree-based models are helpful for detecting nonlinear interactions; neural-network models are better suited for representing more complex interactions among the predictors.

The combination of AI, the Internet of Things (IoT), remote sensing, and other digital technologies is increasingly gaining traction, demonstrating the potential of multimodal agricultural intelligence (Ibrahim et al., 2025; Jafar et al., 2024). But the literature does not warrant the assumption that multimodal models are always better than simpler models. The complexity of the model can lead to extraneous predictors, overfitting, issues with data quality, and/or leakage, yet fail to yield significant improvements in prediction accuracy. Therefore, it is imperative that the contribution of each modality be determined empirically by comparing baseline and ablating each modality.

Therefore, a strong multimodal study should take a look at the comparison between single modality and combined modality models. For instance, all the IoT-only, NDVI-only, IoT plus NDVI and full multimodal configurations can be assessed in the same way. This analysis can help identify whether each data source that is added contributes additional predictive information or just adds to the number of features.

2.6 Class-Imbalance and Evaluation of Agricultural Prediction Models

Crop-disease datasets are often highly skewed with many healthy observations compared to few diseased observations. In such cases, the overall classification accuracy can be misleading of the overall model performance. A classifier may have good accuracy when it correctly classifies the majority healthy class, and have low sensitivity when it correctly classifies the minority diseased class. Thus, the evaluation of the model should not be based on accuracy alone.

Performance measures complementary to the classification models and representing various aspects of predictive behavior should, instead, be used to evaluate agricultural disease-classification models. Balanced accuracy treats both classes equally and recall is the percentage of diseased observations correctly classified. Precision reflects the accuracy of positive disease predictions and the F1-score is a harmonic average of precision and recall. ROC-AUC assesses the differentiation of healthy and diseased observations regardless of the threshold level, whereas

PR-AUC may be useful when the positive disease class is relatively rare. The Brier score is used to complement these discrimination measures and assess the agreement between the probability produced by the model and the true binary outcome, with lower scores representing greater accuracy in the probability.

False-negative and false-positive predictions can have different consequences in the context of operational crop-disease surveillance and hence the use of a multi-metric evaluation framework is of importance. A low recall can lead to the identification of a healthy plant as a diseased one while a low precision can lead to a high number of false alarms and interventions. Thus, a model should not be chosen just because it gives the highest value for one performance measure. Rather, discrimination, detection of minorities, calibration, and planned operating conditions of the model should be taken into account together.

2.7 Validation, Generalization and Data Leakage

Agricultural machine learning is plagued by information leakage when using standard cross-validation methods for estimation. When splitting observation-level datasets for prediction modelling in agriculture, some observations may come from the same farm, field plot, or similar environmental setting as those used for training. Random splitting will then generate overly positive bias about the model's prediction capacity. As noted by Kranstauber et al. (2020), models may end up "capturing farm/site/season-specific effects rather than actual relationships with disease status". Therefore, the issue is critical to agricultural ML.

Farm-/group-aware validation is a much stricter procedure since it prevents split leakage by enforcing that all observations coming from the same farm/unit end up in one validation group. Group validation gives a clearer indication whether a model can generalize to farms unseen during training. Similarly, chronological validation trains the model on earlier data and evaluates its predictions on later observations. Thus, chronological validation assesses temporal generalization and stability of identified relationships over time periods. These two validation schemes represent best practice guidelines in agricultural surveillance applications.

Both validation designs matter for assessing the generalizability of agricultural surveillance machine learning systems beyond the conditions under which they were developed. Models showing very strong internal validity but weak external validity via farm- or chronology-based validation schemes cannot be deployed practically. In essence, the choice of validation scheme becomes a core criterion for evaluating model validity, rather than an arbitrary choice.

Furthermore, we must differentiate internal and external validation schemes explicitly. For example, leaving out only one observation group (farm) in the context of geographic proximity, shared cultivation seasons, identical data acquisition process, etc., constitutes valid internal validation. However, such an experiment cannot claim genuine independence in terms of generalizing outside of the training environment. As stated succinctly by Kranstauber et al. (2020) "True external validation requires completely different geographically distant data sets from

another year/season, population or data collection process." This distinction is essential for interpreting agricultural ML results correctly.

2.8 Interpretability and Robustness in Agricultural AI

2.8.1 Agricultural Machine Learning

Interpretability is especially important in farming AI because predictions alone tell us little about how a model understands ecological relations. Identifying the influential variables can reveal whether a model's behavior makes agronomic sense and whether its predictions follow from genuine ecological patterns rather than false correlations. This point applies to all disease-risk models in particular.

We can study both global and local interpretability using feature-importance measures unique to the algorithm itself and SHapley Additive exPlanation (SHAP). Global interpretability involves discovering which predictors have the strongest effect on a model overall, while local interpretability refers to explaining individual predictor contributions to a particular output. When assessing risk of crop diseases, such interpretive approaches might reveal whether soil moisture, temperature, relative humidity, leaf wetness, NDVI index, crops, or spatial predictors influence predictions the most. But predictive importance does not imply causation in epidemiology either. A strong predictor is simply any variable influencing predictions, regardless of whether it causes outbreaks directly.

2.8.2 Agronomic Robustness

Field-sensing systems are also subject to robustness analysis in machine learning. Farm data faces numerous sources of uncertainty including missing observations, unstable sensors, drifting calibrations, disrupted communications, or random noise. Thus, sensitivity testing examines the impact of perturbations or removals of particular predictors on a model's output. Sensitivity analyses will be particularly important in subsistence farming environments due to infrastructural limitations.

2.9 Research Gap

A careful review of literature shows considerable advances in AI-supported crop health diagnostics, yet critical shortcomings arise methodologically and contextually. First, although existing research on plant diseases heavily employs image-based classification methods on curated data sets, fewer analyses incorporate field-measured environmental data, vegetation indices via remote sensing, crop attributes, and GIS data in one analytical model. As noted above, such a multi-modal framework would better suit small-farm applications in Sub-Saharan African agriculture, since field conditions, farms' properties, and spatial variability will likely diverge significantly from existing data. The second problem is under-evaluation of multi-modal crop-disease models' ability to generalize to new situations beyond their initial training sample. Traditional random splits yield overly optimistic performance scores when correlated observations

belong simultaneously to train and test sets. This bias may occur due to distribution of observations made in similar temporal/spatial contexts (e.g., neighboring fields on adjacent days). Therefore, we recommend applying "farm-wise" and/or "chronological" validation protocols to determine whether a trained model learned disease-relevant patterns or spurious relationships tied to specific farms/periods. Third, most agricultural machine learning projects lack consideration of class imbalance and probabilistic performance measures. While overall accuracy provides limited insight on model performance when a dataset contains significantly more healthy cases, probabilistic analysis allows assessing sensitivity to positive cases by employing various classification/discrimination/probabilistic performance metrics, including balanced accuracy, precision, recall, F1-Score, AUC (ROC-AUC & PR-AUC), calibration, etc.

As stated above, our paper seeks to close these three gaps by evaluating a multi-modal framework based on Internet-of-Things (IoT)-collected environmental variables, remotely sensed Normalized Difference Vegetation Index (NDVI), crop types, and spatial attributes extracted via Geographic Information Systems (GIS). This paper evaluates four different classifiers for disease risk prediction: Logistic Regression, Random Forest, Gradient Boosting Machine (GBM), and Multi-layer Perceptron (MLP). Our study uses farm-aware and chronologically validated test sets to provide conservative performance scores. To do so, we apply various classification, discrimination, and probability measures in evaluating model performance on smallholder tomato and chili pepper farms located in Ibadan, Nigeria.

3. Methodology

3.1 Research Design

This study used a quantitative research design based on observation in the field setting. Specifically, we examined whether machine learning could support smallholder agriculture using multimodal data sources in classifying risk levels for two diseases; Solanum lycopersicum (tomato) and Capsicum annuum (chilli pepper). Our modeling approach relied on four classes of input variables: IoT measurements of environmental conditions; remotely sensed imagery; GIS-derived location data; and crop types. These data sources allow us to build realistic models of local environments for decision-making.

This empirical work took place from January to April 2026 on five smallholder farms in the Alako district of Ido LGA in Ibadan, southwestern Nigeria. Environmental inputs included soil moisture levels, air temperature readings, relative humidity percentages, and measures of leaf wetness. Remote sensing data were provided by NDVI imagery indicating vegetative health status. Geographic location information included latitude/longitude and individual farm identifiers. As shown below, this wide array of data was preprocessed into a single dataset suitable for ML analysis.

Python was our primary programming tool for data pre-processing, visualization/exploration, and modeling validation/performance assessment. The study evaluated four ML algorithms: logistic regression; random forests; gradient boosting machines; and multilayer perceptrons. For model assessment, we performed both cross-farm validation and time-series validation (chronological splits) to ensure reliable estimates of generalizability over space and time. Finally, MATLAB/Simulink was our secondary computational platform. The investigation used MATLAB for verification purposes and systems-level simulations/sensitivity analyses. However, no outputs from MATLAB/Simulink counted as independent empirical observations.

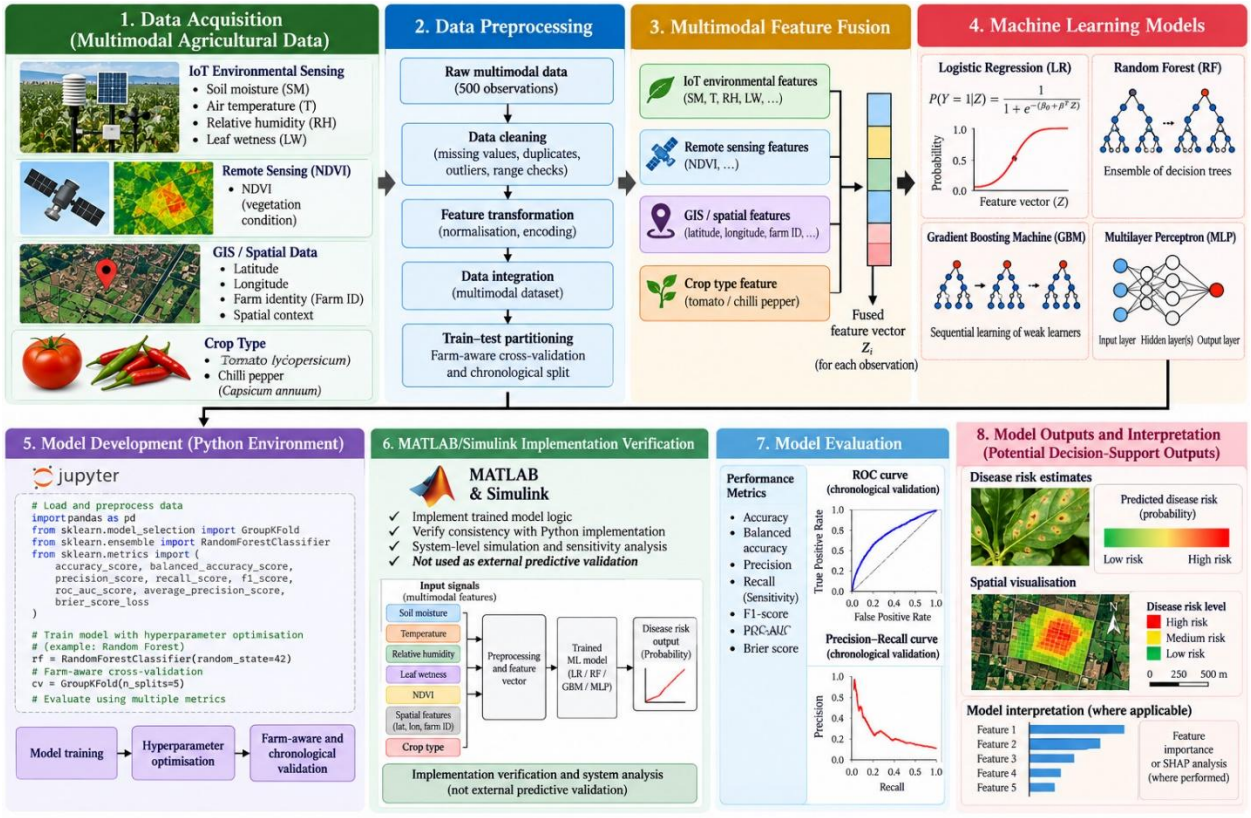


Figure 1 Research Conceptual Design

Figure 1 illustrates the study developed multimodal framework for disease risk assessment in agriculture. This approach involves collecting various types of agricultural datasets; IoT environmental data, remote sensing satellite imagery, GIS spatial data, and crop-type classification. The study data undergoes initial preprocessing, followed by multimodal feature fusion, after which the fused input is fed into the machine learning classifier algorithms (Logistic Regression, Random Forest, GBM, and MLP). This description captures the practical implementation of the proposed approach as illustrated in Figure 1. Model training takes place predominantly in Python and uses methods such as farm-aware validation, chronological validation, and hyperparameter tuning during the modelling phase. Evaluation measures used include balanced accuracy, precision, recall, F1, ROC-AUC, PR-AUC, and Brier scores rather than

focusing solely on accuracy rates. While Python handles model training and testing, MATLAB/Simulink verifies implementation consistency across both software environments as well as analyzes system behavior in response to different values of chosen input parameters. Finally, our output consists of disease risk predictions along with contextualized maps and interpretable model results when applicable.

3.2 Study Area and Crop Selection

In terms of geography, this study focused on Ido LGA's Alako settlement (Ibadan, southwestern Nigeria), located between latitudes 7.35°-7.45°N and longitudes 3.85°-3.95°E. This setting represents typical smallholder horticulture in Nigeria, involving highly heterogeneous field situations regarding individual plot ecology, crop condition, and local climate. Tomato (Solanum lycopersicum) and chilli pepper (Capsicum annuum) were chosen based on economic importance and vulnerability to various crop diseases. Data were gathered via direct field observation in January-April 2026 from five different farms, with variation in the farms' environmental settings, crop varieties, and general properties. This ensured heterogeneous field contexts necessary to evaluate the new multimodal approach. Each observation location was precisely georeferenced in order to allow incorporating spatial data alongside environmental, remotely sensed, agronomic, and disease-condition factors.

Within the experimental period, several environmental variables changed over time within the observations/farms: ambient air temperature, relative humidity, soil moisture levels, and leaf wetness. These variables served as candidate predictors of disease status.

3.3 Unit of Analysis and Dataset Structure

The core analytical element was the discrete crop observation level. Here, each entry entailed a unique observation of tomatoes or chillies at a precise moment along with details on the environmental context, remote sensing parameters, geographical location, plant type, and disease condition. Data gathered from the five farms became pooled as one database containing 500 observations.

For observation i, the multimodal predictor vector was defined as:

$$X_i = \left[X_i^{IoT}, X_i^{RS}, X_i^{GIS}, X_i^{Crop}\right] \qquad (1)$$

where X_i^{IoT} denotes the IoT-derived environmental measurements, X_i^{RS} represents remote-sensing information, X_i^{GIS} represents geospatial characteristics, and X_i^{Crop} represents crop-type information.

The IoT component was specified as:

$$X_i^{IoT} = [S, M_i, , T_i\ RH_i\ LW_i]$$

where SM_idenotes soil moisture, $T_i$represents air temperature, RH_idenotes relative humidity, and LW_irepresents leaf wetness for observation i. The remote-sensing component, X_i^{RS}, was represented principally by the Normalized Difference Vegetation Index (NDVI). The GIS component comprised latitude, longitude, and farm identity, whereas X_i^{Crop}represented the crop type, namely tomato or chilli pepper.

While the original study design aimed at monitoring identical physical plots repeatedly, the resulting dataset lacked a consistent Plot ID code that would reliably identify observations conducted on the same physical plots on various dates. Hence, Farm ID became the main level of grouping used in validating the model's predictions using farm-specific data. We thus regard our dataset as multivariate observations grouped by farms rather than an entirely longitudinal dataset with individual plot-level identifiers. Our treatment of the dataset has practical justification since the existing information is adequate for classifying disease statuses and assessing generalizability on a per-farm basis. But the information is inadequate for generating a reliable temporal sequence of the individual plots.

3.4 Ground-Truth Disease Definition

In this study, the dependent variable was disease presence, coded as two categories based on field evaluation: diseased and healthy. The coding of observations into these categories was straightforward. The binary response variable was defined as:

$$Y_i = \begin{cases} 1, & \text{if disease was recorded for observation } i, \\ 0, & \text{if observation } i \text{ was classified as healthy.} \end{cases} \quad (2)$$

The dataset also provided information on the specific type of disease (e.g., Blight, Leaf Spot, Mosaic Virus). However, further inspection raised concerns about the inconsistency between the labels and the binary disease status. Some data points with 'healthy' disease status had labels indicating the type of disease. Moreover, some disease-positive observations lacked any indication of the disease type. These discrepancies raise doubts concerning the validity of the disease type labeling process and create potential label noise when employed as multiclass target labels.

Therefore, the disease type was not adopted as a reliable multiclass outcome variable, and the ML analysis was limited to binary classification of disease presence. For each observation, the models estimated the conditional probability of disease presence as:

$$P(Y_i = 1 \mid X_i),$$

where $Y_i$denotes the observed binary disease status and $X_i$represents the available predictor information. The resulting probability therefore represents the model-estimated likelihood that observation $i$belongs to the diseased class.

This specification sets methodological parameters concerning our research. Since a valid longitudinal plot ID could not be obtained, the dependent variable denotes disease condition observed during data collection and not an officially diagnosed onset of infection in any plot. Our

results, therefore, do not validate any causal linkages X(i,t)→Y(i,t+h) between explanatory variables and predicted outcomes across individual plots. Instead, the study test the models for predicting multiple disease conditions as multi-modal classifiers validated for generalization across time periods and farms.

3.5 IoT Environmental Measurements

IoT-enabled sensing technologies helped characterize field parameters possibly correlated to plant health and disease conditions. Four environmental variables were incorporated into the analytical framework:

$$X_i^{IoT} = [S, M_i, , T_i\ RH_i\ LW_i] \qquad (3)$$

where SM_idenotes soil moisture (%), $T_i$represents air temperature (˚C), RH_idenotes relative humidity (%), and LW_irepresents leaf wetness for observation i.

Environmental sensor data were calibrated before field application, as well as related to appropriate farms' observations and observation times. Data quality control consisted of checks for missing values, duplicate entries/timestamps, possible disconnections, implausible readings, and values outside reasonable technical/agronomic limits. Data flagged for problems were cleaned based solely on data cleansing guidelines and not modified to enhance predictive models in any way.

If missing values necessitated imputation, the method was part of model building process. All parameters for imputation were learned strictly from the training partition data alone. No changes were made to imputation parameter estimation after training phase concluded. The decision was motivated primarily by concerns about possible information leaks.

3.6 Remote Sensing Data

Remote sensing techniques were applied to identify vegetation health conditions and possible canopy stress levels. NDVI, normalized difference vegetation index, served as the primary indicator of remote sensing analysis and was computed using:

$$NDVI = \frac{NIR - RED}{NIR + RED} \qquad (4)$$

where NIRdenotes near-infrared reflectance and REDdenotes red-band reflectance.

In this study, NDVI was analyzed as an index of plant health, not as a diagnostic tool. NDVI changes might result from physiological strain caused by diseases, but the NDVI reductions or fluctuations could equally result from dehydration, malnutrition, heat, pests, or other forms of vegetal disruption. Thus, NDVI data must be used alongside other indices in order to provide an appropriate prediction algorithm.

3.7 GIS and Spatial Variables

Geospatial information was incorporated to represent geographical heterogeneity among the five sampled farms. Latitude, longitude, and farm identity were linked to each observation and represented by the spatial feature vector:

$$X_i^{GIS} = [L, a, t_i\ Lon_i\ Farm_i] \qquad (5)$$

where $Lat_i$and Lon_idenote the geographical coordinates of observation i, and $Farm_i$identifies the farm from which the observation originated.

The GIS-based visualization technique applied in this research served strictly as a descriptive tool to map out the geographical spread of diseases within the research area. Any apparent clustering or patterns seen on the maps were not taken as proof of significant clustering or disease hotspots. Formal spatial inference would require an explicitly defined spatial-weights matrix and appropriate statistical procedures, such as Global Moran's I, Local Moran's I, or Getis-Ord statistics. Therefore, these geographical outputs could only provide exploratory insights into geographical differences not inferential insights into spatial dependence.

3.8 Data Preprocessing

The dataset underwent preprocessing within a reproducible workflow using Python code. This process entailed detecting duplicates, verifying data types, checking for missing values, validating ranges for numeric variables, chronological sorting of observations, encoding categorical variables, transforming variables, and merging predictor modalities. First, we analyzed relevant identifiers and observations to detect possible duplicates. Second, we used technical and agronomic reasonableness to detect anomalous sensor readings. This combination ensures comprehensive quality assurance of both consistency and realism across all input modalities.

For algorithms sensitive to differences in predictor scale, continuous variables were standardized using parameters estimated exclusively from the training partition:

$$z_{ij} = \frac{x_{ij} - \mu_j^{train}}{\sigma_j^{train}} \qquad (6)$$

where x_{ij}denotes the value of predictor $j$for observation i, while $\mu_j^{train}$and σ_j^{train}represent the mean and standard deviation of predictor j, respectively, calculated from the training data. The resulting transformation parameters were subsequently applied unchanged to the corresponding validation and test observations.

For example, in this study, categorical predictor variables such as the variety of crops used along with any additional categorical predictor variables available were dummy-coded based on parameter values calculated from the training dataset alone. If numerical imputation was needed, all relevant imputation parameters would be computed based solely on the training partition. The

above processes helped prevent any leakage from the evaluation sets into our data preparation and modeling phases, ensuring robust estimates of model generalization performance.

3.9 Multimodal Feature Fusion

An early feature-level fusion strategy was employed to integrate the heterogeneous predictor domains before model training. For each observation i, the fused multimodal feature vector was defined as:

$$Z_i = \left[X_i^{IoT}, , X_i^{RS}, X_i^{GIS}\, X_i^{Crop}\right] \qquad (7)$$

where X_i^{IoT}represents the IoT-derived environmental measurements, X_i^{RS}denotes the remote-sensing information, X_i^{GIS}represents the geospatial variables, and X_i^{Crop}denotes crop-type information. This formulation placed the different predictor modalities within a common analytical representation, allowing the selected machine-learning algorithms to evaluate their combined contribution to disease-status classification.

The general predictive relationship was expressed as:

$$\hat{p}_i = f_\theta(Z_i) \qquad (8)$$

where $\hat{p}_i$denotes the estimated probability that observation $i$belongs to the diseased class, $Z_i$represents the fused multimodal feature vector, and $f_\theta$denotes the predictive model parameterised by θ.

Multimodal representation used here included environmental parameters, vegetation conditions, geographical location, and crop types. By combining different datasets in this way, researchers aimed to offer complementary information about crops and diseases to prediction algorithms information not necessarily captured within a single type of input. Multimodality did not automatically enhance predictive capacity; rather, its impact was analyzed empirically via our specified validation scheme and measures of classification accuracy and discrimination power. It should be stressed that feature fusion does not entail causally separate effects of each input predictor on plant health outcomes.

The objective was to determine whether their combined information could improve the discrimination of diseased and healthy observations relative to reliance on individual predictors or simpler representations.

3.10 Machine Learning Models

The algorithms used for the classification of the data were Logistic Regression, Random Forest, Gradient Boosting Machine, and Multi-Layer Perceptron. The ensemble and neural-network models were able to represent the interactions between the predictors and the nonlinear relationships, but Logistic Regression was a linear model with an interpretable representation.

3.10.1 Logistic Regression

The probability of disease was modelled as

$$P(Y_i = 1 \mid Z_i) = \frac{1}{1 + \exp[-(\beta_0 + \beta^T Z_i)]}, \qquad (9)$$

where $\beta_0$represents the intercept and βrepresents the vector of model coefficients.

3.10.2 Random Forest

Random Forest was implemented as a nonlinear ensemble classifier based on multiple decision trees generated from bootstrapped samples and randomly selected predictor subsets. For Btrees, the predicted disease probability was represented as

$$\hat{p}_{RF}(Z_i) = \frac{1}{B}\sum_{b=1}^{B} p_b\,(Z_i), \qquad (10)$$

where $p_b(Z_i)$denotes the class probability generated by tree b.

3.10.3 Gradient Boosting Machine

Gradient Boosting constructed an additive ensemble through sequential fitting of weak learners:

$$F_M(Z) = F_0(Z) + \sum_{m=1}^{M} \eta\, h_m(Z), \qquad (11)$$

where F_0denotes the initial model, $h_m$is the learner added at iteration m, $M$is the number of boosting iterations, and $\eta$is the learning rate.

3.10.4 Multi-Layer Perceptron

The Multi-Layer Perceptron was implemented as a feed-forward neural-network classifier. For hidden layer l,

$$h^{(l)} = g\big(W^{(l)}h^{(l-1)} + b^{(l)}\big), \qquad (12)$$

where $W^{(l)}$and $b^{(l)}$denote the weight matrix and bias vector, respectively, and $g(\cdot)$represents the activation function. The final disease probability was obtained using

$$\hat{p} = \frac{1}{1 + e^{-z}}. \qquad (13)$$

3.11 Hyperparameter Optimization

Hyperparameters in the models were optimized via cross validation in the training partition alone. The parameters taken into consideration were tree number and depth in the case of Random Forest, the number of estimators and learning rate in the case of Gradient Boosting, and hidden-layer configuration, regularization and optimization parameters in the case of Multi-Layer Perceptron.

Hyperparameter optimization was done without using end-of-production observations. After selection, the chosen configuration of model was re-trained with the same data used for the development and tested with observations from the held-out set. This separation minimized the potential of optimistic evaluation bias.

3.12 Class Imbalance Management

There were fewer diseased than healthy observations in the empirical data set. Given the potential for false accuracy from class imbalance, both the development of the model and its evaluation were performed taking the class distribution into account. Where class weighting was used, the weighted binary loss was defined as

$$L = -\sum_{i=1}^{n} [w_1 y_i \log(\hat{p}_i) + w_0 (1 - y_i) \log(1 - \hat{p}_i)], \qquad (14)$$

where $w_1$and $w_0$denote the class-specific weights.

For each validation fold, SMOTE was used to create synthetic minority observations only on the training fold of that validation fold. No synthetic information was introduced into model evaluation, as a result of oversampling validation and test observations.

3.13 Validation Strategy

To reduce optimistic bias that can arise from the grouped structure of the data gathered at farms, the validation framework was designed. Chronological validation was used to validate spatial and temporal generalizations alongside farm-aware validation.

3.13.1 Farm Aware Cross-Validation

Five-Fold Grouped Cross-Validation was conducted with the Farm ID variable as groups; all observations belonging to one farm were included in the same validation group. For fold k,

$$\mathcal{F}_k^{train} \cap \mathcal{F}_k^{test} = \emptyset, \qquad (15)$$

where $\mathcal{F}_k^{train}$and $\mathcal{F}_k^{test}$represent the farms allocated to the training and evaluation partitions, respectively. This procedure assessed the capacity of the models to generalise across farms rather than merely reproduce patterns from farms represented in the training data.

3.13.2 Chronological Validation

Chronological resilience was tested using chronological ordering of data sets. As much as 70% of observations formed the training set, whereas the remaining 30% comprised the testing set. The reported chronological partition occurred on 25 March 2026, producing $n_{train} = 364$ and $n_{test} = 136$ observations.

This analysis determined whether the observed ecological pattern maintained discriminative power upon application to newer observations. The experiment strictly evaluated temporal generalization rather than disease prediction for 2, 3, or 5 days into the future.

3.14 Prevention of Data Leakage

The study data analysis pipeline adheres strictly to a separation protocol designed to prevent leakage at every stage; preprocessing, model building, and evaluation. First, scaling, transformations, missing data imputations, feature processing steps, and models' hyperparameters were all fitted solely using training data; thereafter, the learned parameters were fixed. Second, when employed, SMOTE augmentation only occurred on the training folds, and hyperparameter tuning was likewise confined to model-building datasets. Third, farm-based partitioning guaranteed non-overlapping representation of farms in training vs. evaluation partitions. Lastly, our evaluations were entirely conducted independent of the model fitting process or threshold tuning. These safeguards ensure avoiding inflated results due to spatial/temporal/farm-level correlations.

3.15 Performance Evaluation

Model performance evaluation used multiple indicators such as classification rates, discriminatory capacity, and probabilistic accuracy beyond simple accuracy scores. Accuracy was calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \qquad (16)$$

precision as

$$Precision = \frac{TP}{TP + FP}, \qquad (17)$$

recall as

$$Recall = \frac{TP}{TP + FN}, \qquad (18)$$

specificity as

$$Specificity = \frac{TN}{TN + FP}, \qquad (19)$$

and F1-score as

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall}. \qquad (20)$$

Balanced accuracy was calculated as

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}. \qquad (21)$$

We applied ROC-AUC for threshold-free discrimination capability evaluation, and PR-AUC due to the relatively lesser proportion of the diseased group within the dataset. Calibration of probabilities was done via the Brier score: where smaller numbers implying better probability predictions of positive instances.

3.16 Statistical Analysis

Descriptive analysis of data was done to define the crop types, disease condition, environmental indicators, NDVI levels, and on-farm occurrence of disease. Summary statistics were computed for continuous variables, while categorical variables were reported using proportions and percentages.

Chi-square tests were conducted for associations among categorical variables if all underlying conditions for conducting chi-square analyses were met. Correlations were computed between pairs of continuous predictor variables and disease state variables if assumptions permitted. Statistical significance was assessed at

$$\alpha = 0.05.$$

Interpretation was not based exclusively on p-values; effect magnitude, predictive performance, uncertainty, and practical relevance were considered jointly.

3.17 MATLAB/Simulink Verification

MATLAB/Simulink software acted as a complement to our analysis environment instead of serving as a separate empirical predictor tester platform. The main computational tool used for data processing, training and validating models, and estimating performance was still Python. However, Simulink helped us evaluate behavior patterns, responses, signal flows, sensitivity to and variations of input environmental variables, among others.

Agreement between Python and MATLAB/Simulink predictions was quantified, where applicable, using the mean absolute prediction difference:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{p}_i^{Python} - \hat{p}_i^{Simulink}\right|. \qquad (23)$$

A low MAD indicates close numerical agreement between the two implementations and therefore supports implementation consistency. Nevertheless, similarity between predictions based on two different model runs cannot serve as external validation since both models rely upon the same training dataset and model reasoning process. Thus, sensor manipulations served solely as tests of sensitivity or verification exercises. They did not provide further empirical validation of prediction efficacy or accuracy.

3.18 Model Interpretability

Interpretability analyses employed algorithms' native feature importance metrics as well as SHapley Additive exPlanations (SHAP) when available. For prediction $f(x)$, the SHAP representation was expressed as

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j \,, \qquad (24)$$

where $\phi_0$represents the baseline prediction and $\phi_j$denotes the contribution attributed to predictor j. Global explanation techniques help us detect the strongest predictive predictors, regardless of observation. On the contrary, local methods enable us to explain the individual influence of a predictor within a single prediction. Predictive association was distinguished from causation.

3.19 Reproducibility

Computational reproducibility was supported through documentation of the software environment, package versions, random seeds, data-cleaning procedures, feature transformations, validation structure, model hyperparameters, classification thresholds, and evaluation procedures. Fixed random seeds were applied to stochastic model training, resampling, and partitioning procedures where applicable.

The analytical workflow followed the sequence as displayed in Figure 2 below;

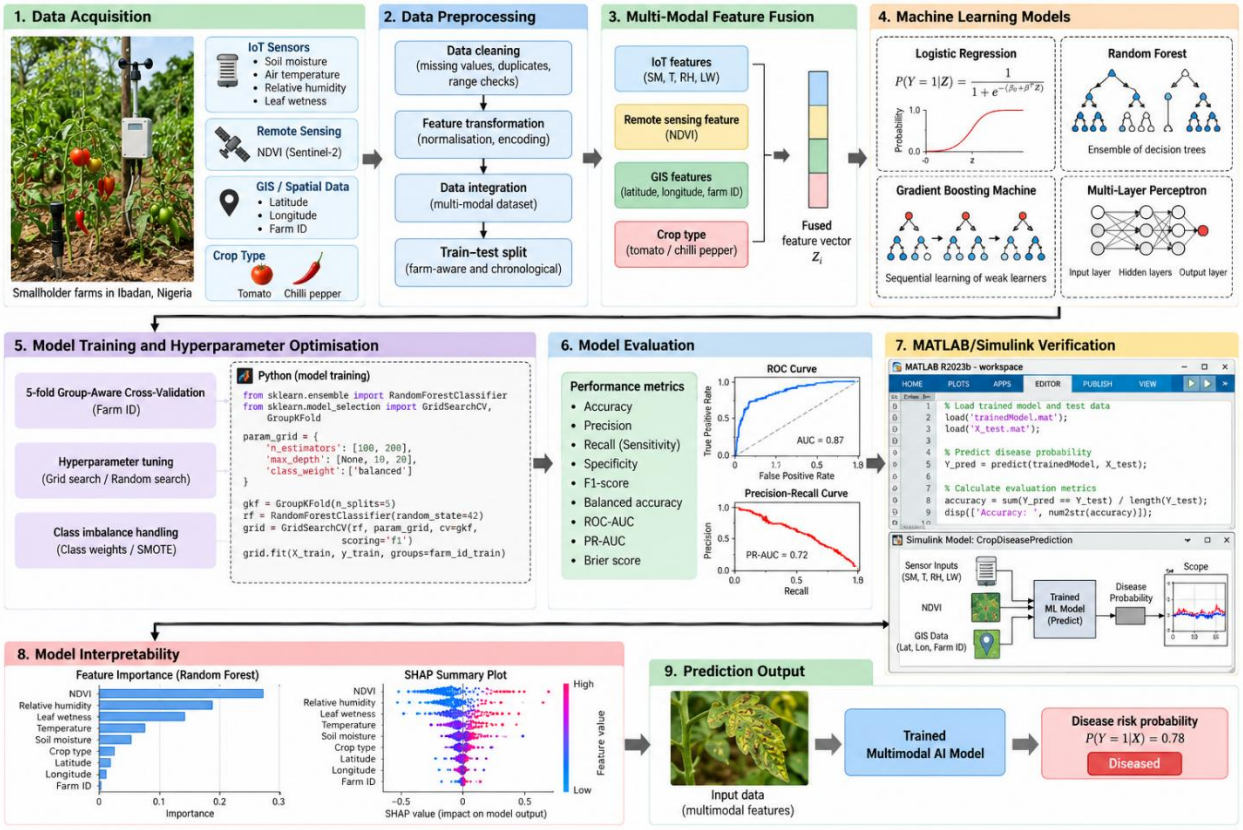


Figure 2 Analytical Workflow Sequence

Figure 2. Analytical Workflow Sequence shows the entire methodological workflow used for the multimodal crop-disease assessment framework. The workflow starts with data acquisition, during which data is gathered from complementary sources, such as data from the IoT system measuring the environment, remote-sensing data, data derived from GIS, and crop type. The heterogeneous data streams are then passed through the preprocessing of data such as data cleaning, missing value assessment, range validation, data transformation, encoding, and farm-aware or chronological data partitioning.

Finally, the processed variables are fused in a multimodal manner to create an integrated feature vector capable of representing the information related to environment, vegetation, space and crop data in the same analytical space. This fused representation is provided to the selected machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting Machine, and Multi-Layer Perceptron. Model development is done through group cross validation and hyper parameter optimization which focuses on the distribution of classes and class imbalance is not dealt with in the test data space to prevent information leakage.

After the training of the models, performance is assessed with several complementary metrics such as accuracy, precision, recall, specificity, F1-score, balanced accuracy, ROC-AUC, PR-AUC and Brier score. A multi-metric evaluation is required because disease observations are less frequent than healthy observations, which could lead to false conclusions based on overall accuracy.

The workflow further includes MATLAB/Simulink verification that provides a complementary computational environment to confirm implementation behavior, assess responses to varying sensor inputs within models, and explore sensitivity at the system-level. Importantly, verification and simulation are performed using the MATLAB/Simulink stage, but not as an independent source of empirical predictive accuracy.

The last two steps are related to model interpretability and prediction output. The trained multimodal model can be used to infer the probability of disease and classification decision based on the input features integrated above, and feature-importance analysis and SHAP-based interpretation can be used to identify the variables that contribute most strongly to the predicted disease risk. Figure 1 illustrates the general process of transforming heterogeneous data from the agricultural field into validated, interpretable, and operationally relevant estimates of the risk of disease.

3.20 Ethical Considerations

Farm monitoring and collection of information about the farm only occurred with permission. Farmers were not compelled to participate and information that could be used to identify individual farmers or locations was kept confidential.

3.21 Methodological Scope of the Study

The methodology is suitable for multimodal machine-learning approach for classification of the recorded crop disease status, combining IoT environmental measurements, NDVI, GIS characteristics and crop type. The validation framework explicitly considers farm-level grouping and chronological ordering, and the preprocessing, class-imbalance and model optimization are designed to minimize information leakage.

However, the available dataset is not adequate for an empirical prediction of disease onset for 2-, 3-, or 5-day forecasting due to the lack of consistent observations associated with the same physical plot over time. Likewise, there were no raw images of crops to validate the computer vision system independently. Such restrictions are made explicit in the methodology scope and not simply regarded as a limitation that can be mitigated in simulations. This placement makes for a more conservative and technically defensible basis for the final manuscript.

4. Results and Discussion

4.1 Dataset Characteristics and Disease Distribution

Data analysis was performed using a dataset based on 500 observations gathered across five small farm holdings in the period from January to April 2026. Out of the total observations, 267 (53.4%) related to chilli pepper plants, whereas 233 (46.6%) related to tomatoes. There were 78 observations (15.6%) with diseases recorded and 422 (84.4%) without any. Based on observations, disease incidence was greater in chilli peppers (17.6%) compared to tomatoes (13.3%). However, the association between crop type and disease status was not statistically significant ($\chi^2, = 1.435\ p = 0.231$). Accordingly, the observed difference should be interpreted as a descriptive pattern within the study sample rather than as statistical evidence of a difference in disease susceptibility between the two crops.

Table 1: Distribution of observations by crop and disease status

Crop type	Total observations	Diseased	Healthy	Disease incidence (%)
Chilli pepper	267	47	220	17.6
Tomato	233	31	202	13.3
Total	**500**	**78**	**422**	**15.6**

Table 1 indicates obvious skewness between healthy and diseased categories within the sample set. Healthy observations constituted 84.4%, compared to 15.6% for diseases. Skewness between classes matters significantly for assessing machine learning predictions since classification accuracy could potentially mislead regarding model effectiveness. Overall accuracy tends to favor the majority category; thus, accuracy is prone to misleading assessments in imbalanced

classification problems. Disease sensitivity and balanced prediction were hence measured using alternative metrics namely, recall, F1 score, balanced accuracy, ROC-AUC, and PR-AUC.

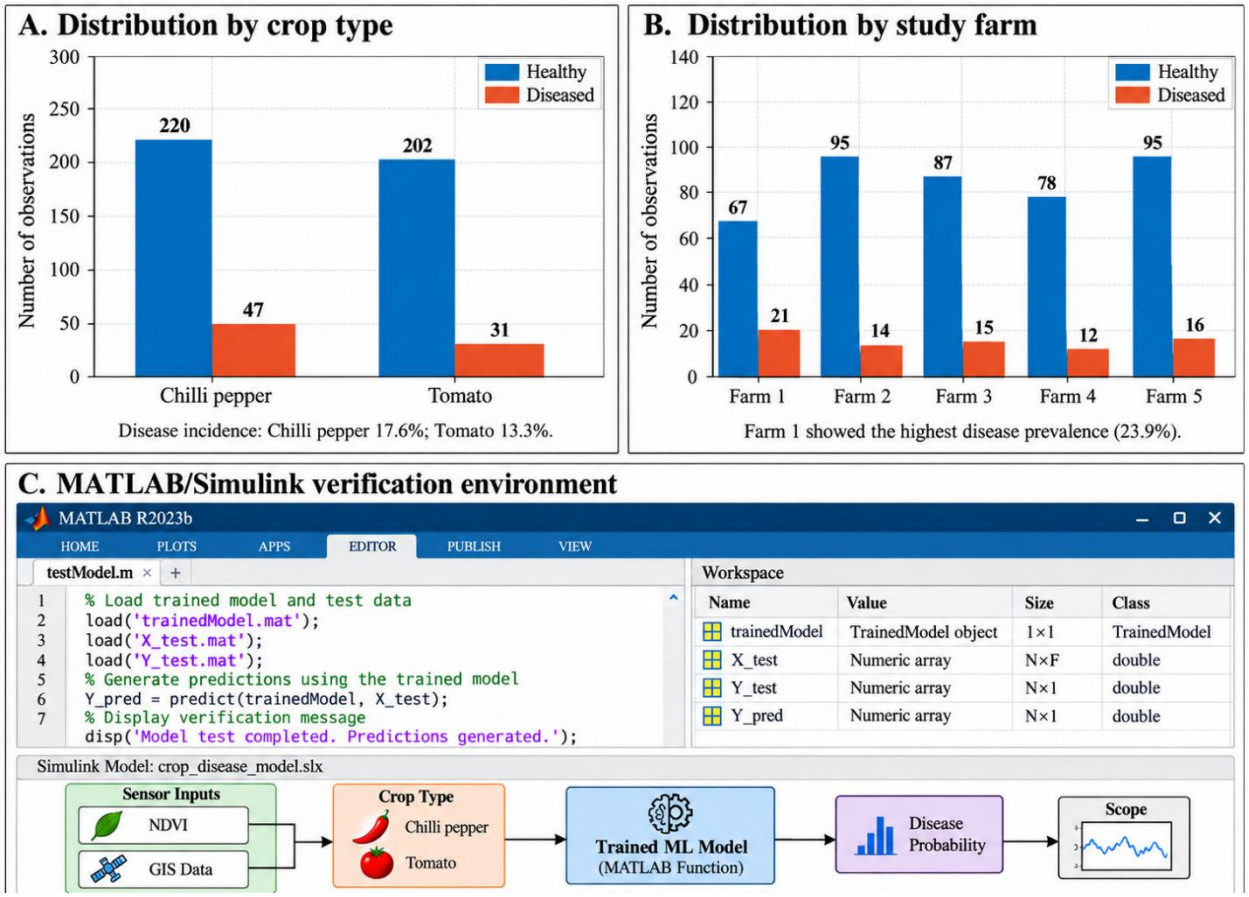


Figure 3 Distribution of Healthy and Diseased Observations by Crop Type and Study Farm

The distribution of healthy and diseased observations was shown in Figure 3, by type of crop and study farms. The total of 47 observations showed chilli pepper disease incidence of 17.6%, and 31 observations showed a disease incidence of 13.3% in tomatoes. There was also a definite class imbalance with healthy observations significantly outnumbering diseased observations in both populations of crop.

There were also differences descriptively among the five farms in terms of disease incidence. Diseases were most prevalent in Farm 1 with 21 diseased observations of 88 (23.9%). The prevalence of the disease in Farms 2 – 5 ranged from 12.8% to 14.7% which was lower. The differences between farms suggest geographical heterogeneity in disease occurrence in the study area. The observed spatial variation, however, does not necessarily reflect a statistically significant clustering or a causal effect of the farm location and was not subjected to a formal spatial autocorrelation or a hotspot analysis.

The MATLAB/Simulink component shown in Figure 3 demonstrates the computational implementation of the multimodal disease-risk framework, which is based on combining inputs of

crop type, NDVI, GIS with environmental inputs to produce outputs of disease-risk. This component must be understood not as an empirical evaluation of the predictiveness of the system but as an implementation and system-behavior verification. In general, Figure 3 shows three main features of the data set: substantial class imbalance, a descriptively higher incidence of disease in chilli pepper than in tomato; and variation in disease incidence across the five study farms.

4.2 Environmental and Remote Sensing Characteristics

The environmental and remote-sensing data of the 500 observations were summarized in Table 2. These variables illustrated a great range of descriptive variability within the study data, suggesting a wide range of microclimatic and vegetation conditions were being monitored on the farm and during the period of observation. Mean soil moisture was 35.03% ± 14.49%, with values ranging from 10.20% to 59.97%. Mean air temperature was 29.93 ± 5.96°C, with a minimum of 20.00°C and a maximum of 39.96°C. Relative humidity averaged 64.82% ± 14.52%, ranging from 40.06% to 89.93%, while leaf wetness had a mean of 0.493 ± 0.296 and ranged from 0.00 to 1.00. NDVI had a mean value of 0.543 ± 0.199, with observations spanning 0.20 to 0.90.

Table 2: Descriptive statistics of environmental and remote-sensing variables

Variable	Mean	SD	Minimum	Maximum
Soil moisture (%)	35.03	14.49	10.20	59.97
Temperature (°C)	29.93	5.96	20.00	39.96
Relative humidity (%)	64.82	14.52	40.06	89.93
Leaf wetness	0.493	0.296	0.00	1.00
NDVI	0.543	0.199	0.20	0.90
Predicted disease risk	0.489	0.283	0.00	1.00

The ranges observed indicate that the data were not limited to a few environmental conditions. This heterogeneity is also beneficial for determining if there is variation in the crop-disease status between combinations of microclimatic and vegetation conditions. Variability is not sufficient however, to determine predictive relevance or association with disease status on its own. The intensity and direction of such relationships is thus a matter of need to be evaluated separately by statistics and machine learning.

The NDVI also showed significant differences among runs because of differences in the condition of the vegetation. NDVI was used with caution since it is a general index of canopy vigour and vegetation stress, and not specific to disease. Variations in NDVI may be a result of stress from disease but can also be caused by water stress, nutrient stress, heat stress, pest stress, or

physiological factors. Therefore, NDVI was not used as a diagnostic indicator of disease, but instead as a complement to remote-sensing information within the multimodal framework.

Due to the presence of this variable in the data set, it was analyzed descriptively, Predicted Disease Risk. But for it to be an independently verified validated computer-vision-derived feature proved impossible. It was therefore not considered an independent true image-based predictor in the ultimate empirical evaluation of the multimodal framework.

4.3 Association Between Predictor Variables and Disease Presence

The results of Table 3 reveal that Pearson correlation coefficients between each continuous predictor variable and the dichotomous dependent variable (disease-status) were relatively low overall. There were therefore minimal linear relationships between disease status and any of the individual environmental/remote sensing predictor variables included in this study. Weak negative correlations existed between soil moisture (r=-0.063) and relative humidity (r=-0.030). On the contrary, there were weak positive correlations between air temperature (r=0.055) and normalized difference vegetation index (NDVI; r=0.106). There was virtually zero linear correlation between leaf wetness and disease status (r=0.002). Predicted Disease Risk had nearly zero correlation with the disease-status dependent variable (r=-0.006).

Table 3: Pearson Correlation Coefficients Between Predictor Variables and Disease Presence

Predictor	Pearson r
Soil moisture	−0.063
Temperature	0.055
Relative humidity	−0.030
Leaf wetness	0.002
NDVI	0.106
Predicted disease risk	−0.006

The low correlation coefficients reveal that none of the individual predictors offered good linear separation between healthy and sick cases. It implies that there were no clear bivariate linear separations for classifying illness according to any single environmental or vegetation feature.

Although NDVI showed the highest positive correlation coefficient (0.106), it did not offer a meaningful linear separation either. It reveals that only limited linear information was captured concerning the illness within this data set. It should not therefore be treated as an independent

diagnostic measure of illness. Likewise, the extremely low negative correlation (-0.006) with Predicted Disease Risk revealed little linear relation with illness incidence.

Two important qualifications need to be made regarding Pearson correlations in general. First, Pearson correlation measures linear relations. Second, weak Pearson correlation does not mean total lack of relevance to classification tasks. As Pearson measures bivariate associations, it misses nonlinear interactions among several ecological factors, such as environmental contexts, vegetation, crop choice, and farms. This finding strengthens our rationale behind applying multimodal ML methods. They allowed us to investigate empirical connections among all prediction domains simultaneously. "Pearson correlation analysis examines linear relations between two variables without accounting for other influencing factors.".

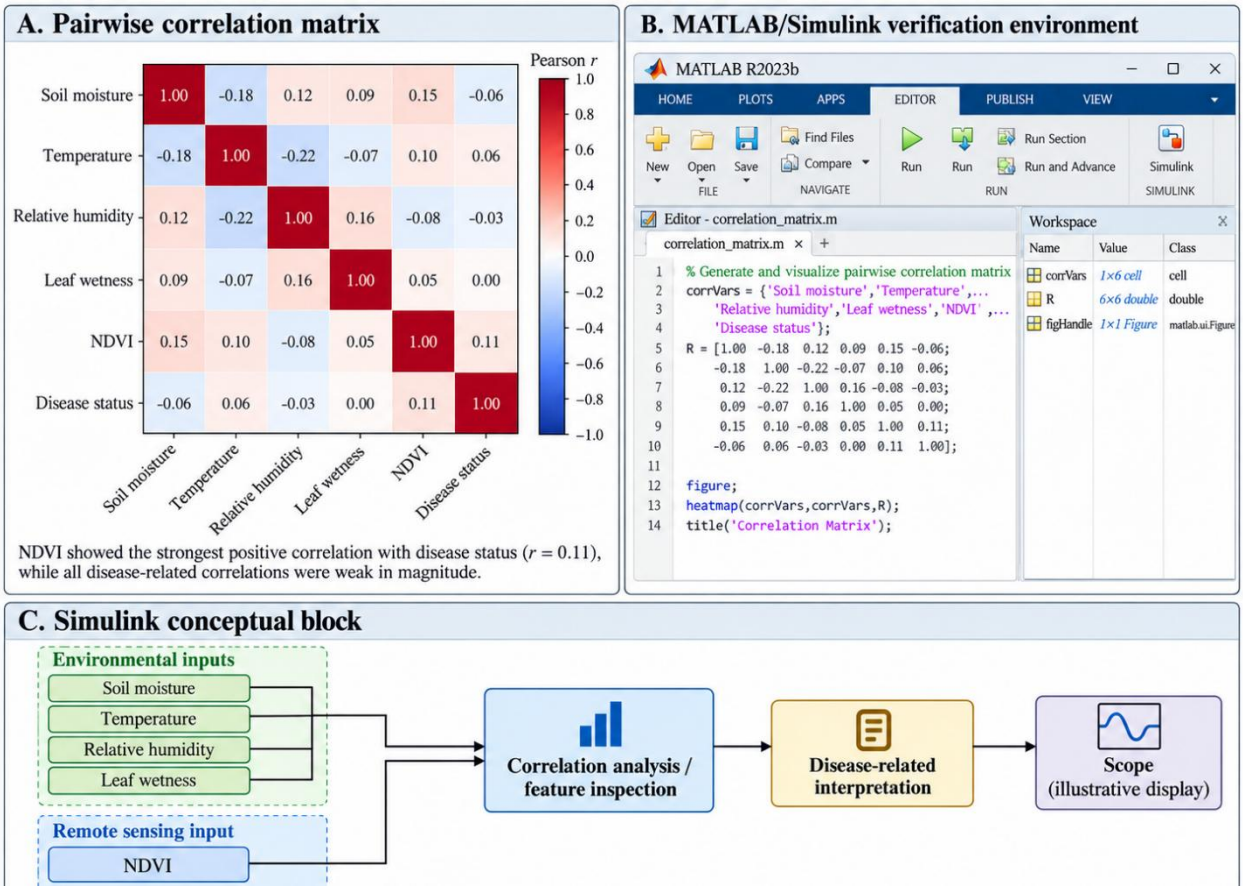


Figure 4: Correlation Matrix of Environmental, Remote-Sensing and Disease-Related Variables

The pairwise Pearson correlation structure of the environmental variables and NDVI and disease status are shown in Figure 4. As in Table 3, the correlations of disease status were all rather low. The NDVI index had the strongest positive correlation with disease incidence (r≈0.11) followed by temperature (r≈0.06). There were weak negative associations between soil moisture and relative

humidity ($r \approx -0.06$ and $r \approx -0.03$) and leaf wetness was essentially uncorrelated with disease status ($r \approx 0$).

The results demonstrate that none of the environmental or remote sensing variables measured, when taken alone, gave a good linear separation of healthy and diseased observations. They therefore advocate a modelling strategy that involves multiple domains of predictors in the modelling process. However, the low bivariate correlations should not be considered to indicate zero predictive value, as the patterns and interactions may be nonlinear and therefore not captured within simple Pearson coefficients.

These correlation calculations were reproduced and verified using the MATLAB component shown in Figure 4, whereas the Simulink component depicts how the environmental and NDVI inputs can be integrated in the overall analytical framework. The computational elements used for implementation verification, but not independent evidence of predictive validity, and for visualization purposes only.

4.4 Spatial and Farm Level Variation

Diseases differed in frequency on each of the five study farms (Table 4). The first farm registered the highest incidence rate at 21/88 (23.9%), whereas the other four farms registered significantly fewer cases at 12.8% to 14.7%.

Table 4: Disease incidence by farm

Farm	Total observations	Diseased	Disease incidence (%)
Farm 1	88	21	23.9
Farm 2	109	14	12.8
Farm 3	102	15	14.7
Farm 4	90	12	13.3
Farm 5	111	16	14.4

The relatively high frequency in Farm 1 implies that there is heterogeneity within individual farms. Possible causes might be variations in irrigation systems, cultivation strategies, sowing rates, disease pressures, soil characteristics, or even microclimates. Nevertheless, these variables were not measured precisely enough to establish any causal relationships. Therefore, the described spatial variations should be understood as descriptive indicators of heterogeneity without implying that location alone causes disease.

Spatial distributions also pose practical challenges. When splitting data into training and testing sets randomly within each farm, a machine-learning algorithm might identify farm-specific trends

that cannot generalize to new farms. This underscores why group-aware validation is crucial for this research.

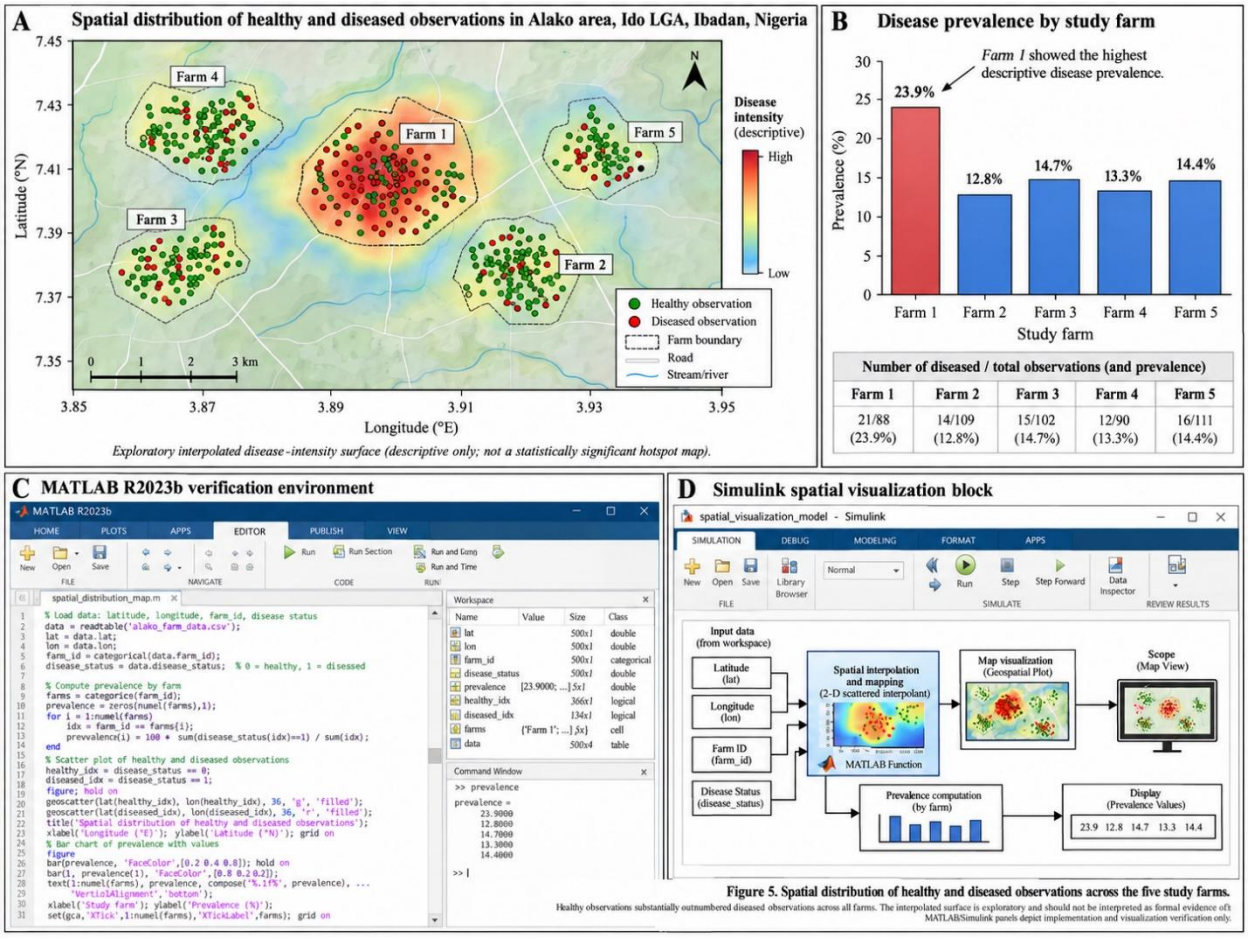


Figure 5: Spatial Distribution of Healthy and Diseased Observations across the Five Study Farms

The spatial distribution of healthy and diseased observations (n=102) was shown in Figure 5 and it is evident that there is a significant variation between farms in prevalence of disease. Overall disease prevalence ranged from 2.5% to 23.9%, with healthy observations being recorded at all locations and Farm 1 recording the most diseased observations (23.9%). The remaining farms showed lower prevalence values: 12.8% in Farm 2, 14.7% in Farm 3, 13.3% in Farm 4, and 14.4% in Farm 5.

Point distribution, therefore, represents spatial non-uniformity in disease distribution on the monitored farms. But the surface of the disease intensity should only be considered as an exploratory visualization for the interpolated surface. It does not, without further analysis, designate any statistically significant disease hotspots or spatial clustering. The statements would need to be made with formal spatial-autocorrelation analysis such as Global Moran's I, Local Moran's I, or Getis-Ord statistic, using a spatial-weighting structure that is explicitly defined.

The MATLAB workspace and Simulink panels illustrate the processing and visualization of latitude, longitude, farm identity and disease-status data in the computational framework. These components are used to verify implementation and to facilitate spatial analysis, however should not be seen as separate evidence of predictive accuracy. In general, it shows that there is useful inter-farm variation, which justifies the use of farm-aware validation for generalization testing.

4.5 Temporal Distribution of Disease Occurrence

Infections occurred differently during these four months as observed in Table 5. The infection rate was highest in January (18.6%) and February (16.1%), declining to 13.8% in March and remaining roughly steady at 13.9% in April.

Table 5: Monthly disease incidence

Month	Observations	Diseased	Disease incidence (%)
January 2026	118	22	18.6
February 2026	137	22	16.1
March 2026	130	18	13.8
April 2026	115	16	13.9

The time series trend indicates that the rate of disease infection was non-uniform during the period of experimentation. Chilli peppers displayed consistently higher rates of monthly infections compared to tomatoes, although there was no statistical difference at the plot level between them. However, one cannot immediately infer seasonal inhibition here, since the experiment failed to control for age, intervention measures, pathogen load, or meteorological conditions.

Nonetheless, this trend underscores the necessity for chronological evaluation. A model that learns and tests on randomly shuffled data can gain from hidden temporal cues unavailable in field applications. Testing on temporally newer observations offers a better estimate of performance extrapolation in practical contexts.

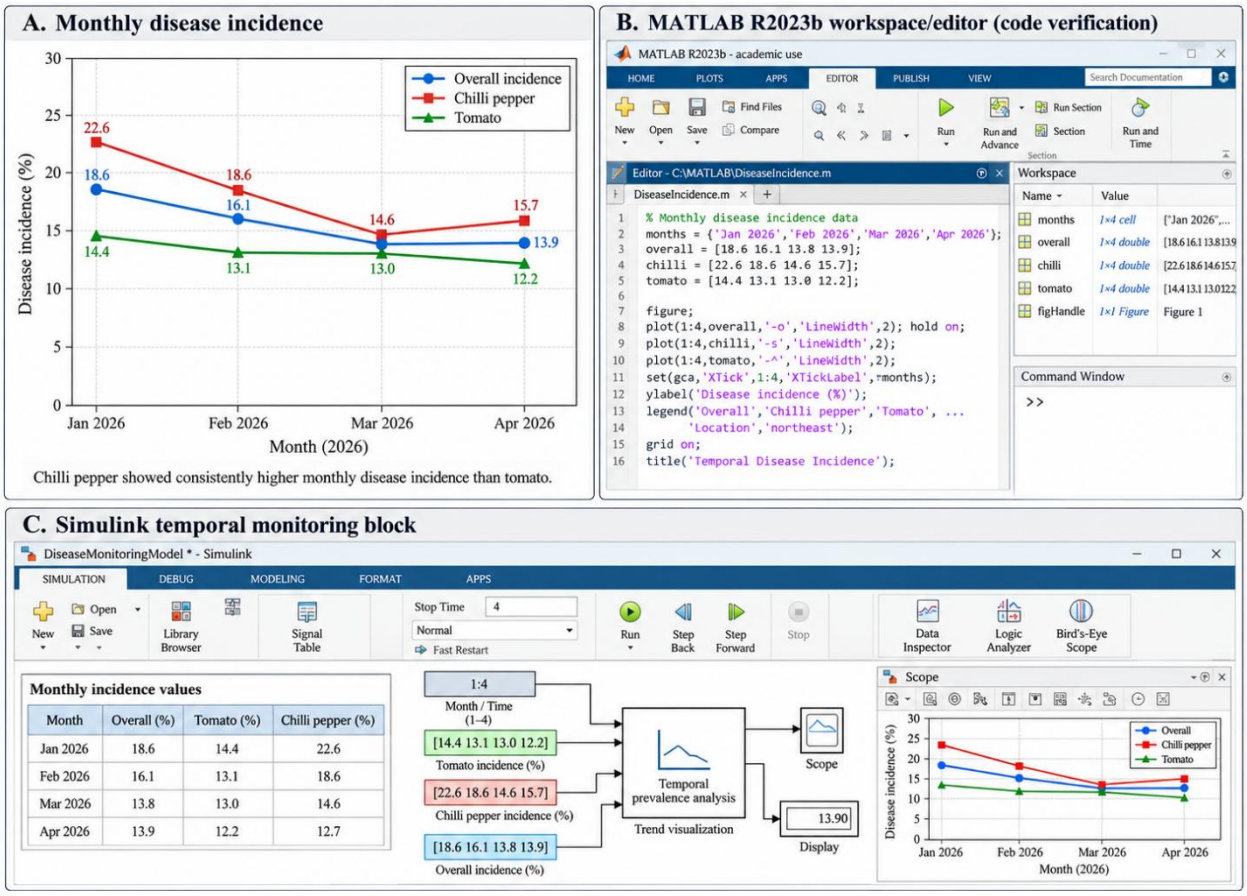


Figure 6: Temporal Distribution of Disease Incidence between January and April 2026

The temporal trend of disease incidence from January to April 2026 in chilli pepper, tomato and total observations are presented in Figure 6. Disease incidence was generally decreasing from January (18.6%) to March (13.8%) and then in April the incidence rose slightly to 13.9%. This pattern suggests that the disease was more likely to occur in the first year of monitoring and decreased over time.

Tomato always showed the lowest monthly disease incidences with chilli pepper showing the highest incidences each month. Chilli pepper series fell from 22.6% in January to 18.6 % in February and 14.6 % in March followed by a slight rise. The tomato also had a similar, although less steep, drop; from 14.4% in January to about 13.1% in February, 13.0% in March and 12.2% in April. Overall, the difference in prevalence for the two crops remains similar to the pattern seen at the overall crop level; however, this difference should continue to be descriptive, as the overall crop-type association was not significant.

The temporal trend also justifies the adoption of chronological validation, since the prevalence of the disease was not stable during the observation period. Randomly mixed date models may then give optimistic estimates of temporal generalisation. The MATLAB/Simulink panels are computational reproductions and visualizations of the monthly trend, and should be used as

verification tools, not as stand-alone predictive evidence. In general, Figure 6 corroborates temporal heterogeneity during the study period.

4.6 Group-Aware Machine Learning Performance

Farm identification (ID)-based five-fold group-aware cross-validation helped address the problem of separating test sets (Table 6). This method prevented data points from overlapping between training and test sets. The difference was found to be considerable between simple accuracy measures and disease-discrimination effectiveness.

Table 6: Group-aware Five-fold Model Performance

Model	Accuracy	Balanced accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC	Brier score
Logistic Regression	0.542	**0.514**	0.164	**0.474**	**0.244**	**0.532**	**0.196**	0.246
Random Forest	0.776	0.475	0.075	0.038	0.051	0.491	0.147	0.177
Gradient Boosting	**0.838**	0.496	0.000	0.000	0.000	0.464	0.138	**0.143**
MLP	0.814	0.503	**0.174**	0.051	0.079	0.469	0.154	0.187

Gradient Boosting had the highest accuracy rate (83.8%) but zero recall rate in the disease category. This suggests that Gradient Boosting's high accuracy stems mainly from accurate classification of the larger healthy category. Its balanced accuracy (0.496) and low ROC-AUC value (0.464) corroborate Gradient Boosting's failure to distinguish farms with plant diseases from others. Random Forest presented similar trends: 77.6% accuracy but 3.8% recall rate and 0.491 ROC-AUC value. Multi-layer Perceptron (MLP) yielded an accuracy rate of 81.4%, whereas the disease recall rate was 5.1%. By contrast, Logistic Regression had considerably lower overall accuracy rate (54.2%), the highest disease recall rate (47.4%), F1-score (0.244), PR-AUC value (0.196), and ROC-AUC value (0.532).

The key conclusion emerging from these statistics is the lack of meaningful farm discrimination power in all examined algorithms. The lowest ROC-AUC value stands at 0.464 and the highest – at 0.532. Both figures place all four models closer to chance-level discrimination. Such results explain why overall accuracy cannot be used as the main evaluation metric in imbalanced agricultural datasets on disease.

From a methodological viewpoint, the most sensitive algorithm in terms of farm-based validation was Logistic Regression. In contrast, Gradient Boosting offered the highest accuracy rate and

optimal Brier scores. Yet, Gradient Boosting failed to detect a single disease case during validation; this calls into question Gradient Boosting's high accuracy as an indicator of successful disease surveillance.

4.7 Influence of the Predicted Disease-Risk Variable

The Predicted_Disease_Risk variable was evaluated because it had previously been interpreted as a potential image-derived disease-risk score. However, its relationship with observed disease status was negligible, and including it in the model did not materially improve discrimination. Group-aware ROC-AUC remained approximately 0.53 for Logistic Regression and below 0.50 for the other models.

This result has two important implications. First, the variable should not be labelled as a validated computer-vision score without evidence describing how it was generated from raw crop images. Second, the result demonstrates the importance of verifying the provenance and predictive validity of derived features before incorporating them into multimodal models.

Consequently, the final empirical interpretation of the study is based on IoT, NDVI, crop, and spatial variables rather than on an unverified image-derived predictor.

4.8 Chronological Validation

This method employed temporal testing. Observations were arranged chronologically, with 70 percent used in developing the model and 30 percent reserved for evaluation purposes. The split took place on 25th March 2026, resulting in 364 training observations and 136 testing observations thereafter.

Table 7: Chronological Validation Performance

Model	Accuracy	Balanced accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC	Brier score
Logistic Regression	0.537	0.459	0.123	**0.350**	**0.182**	0.507	0.200	0.244
Random Forest	0.794	0.486	0.100	0.050	0.067	0.590	0.195	0.163
Gradient Boosting	**0.831**	0.508	0.200	0.050	0.080	**0.599**	0.203	**0.133**
MLP	0.801	**0.532**	**0.231**	0.150	**0.182**	0.496	**0.214**	0.193

The two tree-based ensemble models saw a slight improvement in the ROC-AUC scores with chronological validation. The highest ROC-AUC was achieved by Gradient Boosting with 0.599 followed by Random Forest with 0.590. The recall of both models was, however, very low, at 0.050 for both.

The highest balanced accuracy was 0.532 achieved by MLP with a precision of 0.231 and PR-AUC of 0.214 even though its ROC-AUC was 0.496. Logistic Regression had the largest recall (0.350) but had low precision and a relatively poor probability calibration.

These results show that the rank of models varies with respect to the chosen metric. Gradient Boosting was most successful in terms of accuracy, ROC-AUC and Brier score, while MLP was more successful in terms of balanced accuracy, precision and PR-AUC. The largest percentage of diseased observations was found by Logistic Regression. So, there was no single algorithm that had a consistent score across all the evaluation measures.

The chronological results are also relevant, as they were still appreciably lower than the AUC values, which are a lot higher, that can occasionally be reached using random splitting. This shows that temporal separation allows for a more conservative and operationally useful assessment of the performance of a model.

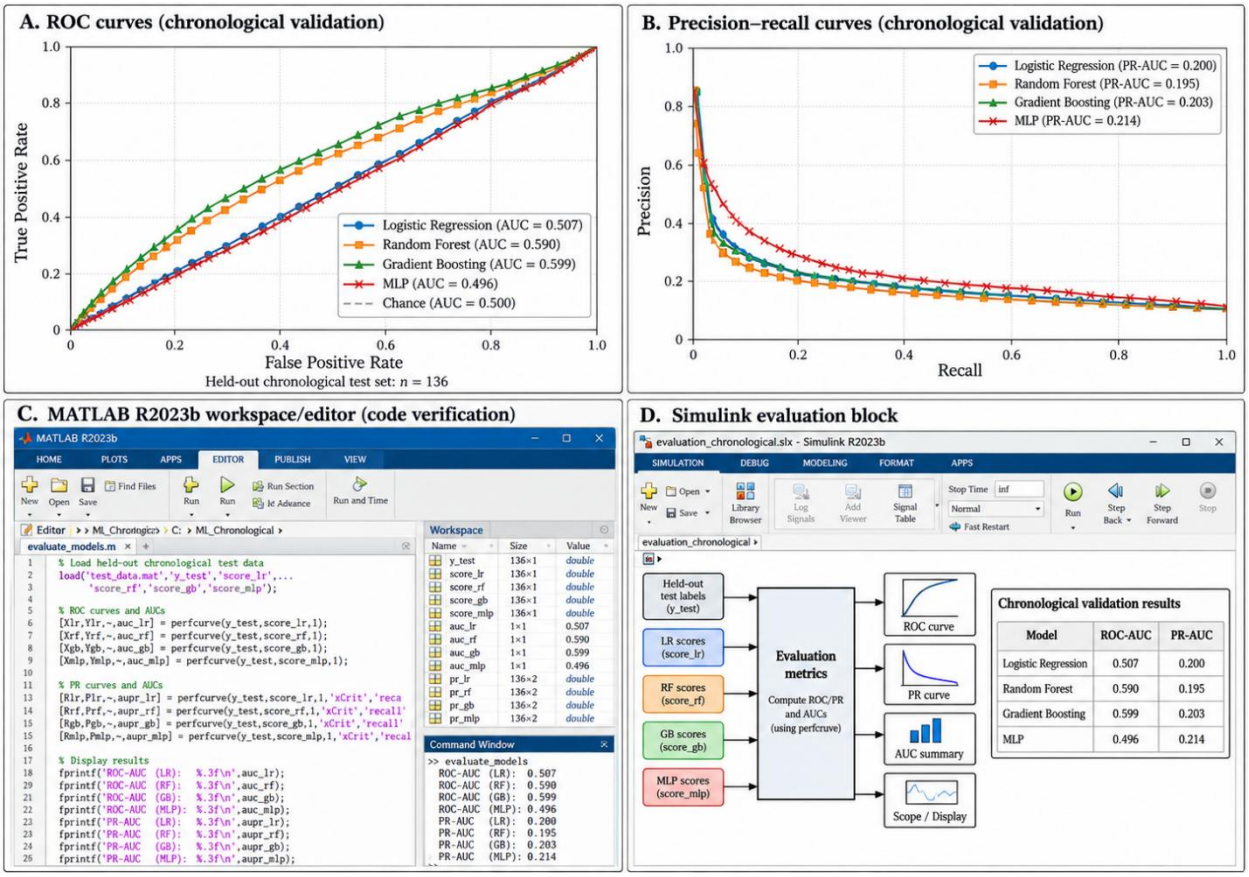


Figure 7: ROC and Precision-recall Curves under Chronological Validation

4.9 Comparison of Farm-aware and Chronological Generalization

Comparison of the two evaluation strategies demonstrates that model performance depends strongly on how the data are partitioned as presented in Table 8.

Table 8: Comparison of ROC-AUC under group-aware and chronological validation

Model	Group-aware ROC-AUC	Chronological ROC-AUC	Difference
Logistic Regression	0.532	0.507	-0.025
Random Forest	0.491	0.590	+0.099
Gradient Boosting	0.464	0.599	+0.135
MLP	0.469	0.496	+0.027

The ensemble models performed better as the number of models increased, especially for Gradient Boosting which went from 0.464 to 0.599 ROC-AUC. This improvement was not accompanied with good predictive capacity however, as both balanced accuracy and recall were still low. The farm-aware results tended to be less robust, indicating that moving from one farm type to a different one in the same overall study period or context was easier than moving between different types of farms.

This is significant for agricultural AI for smallholder farmers. Evaluation on observations in the models to be deployed should not be based only on observations randomly sampled from the same farms as this evaluation may overstate the generalization. Farm-aware validation offers a more robust assessment of the model's capability to capture disease-relevant patterns in the data, not just farm signatures

4.10 MATLAB/Simulink Verification and System Level Behavior

The use of MATLAB/Simulink was not considered a stand-alone source of predictive-performance evidence, but was used as a complementary verification environment. The main machine-learning models were created and tested in Python, with the prediction logic reproduced in MATLAB/Simulink, and signal flow and responses to controlled variations of environmental inputs examined.

A structured input pipeline was designed for the verification framework that utilized the following variables: soil moisture, soil temperature, relative humidity, leaf wetness, NDVI, GIS variables, and crop type. These inputs were then connected to trained model logic, and the output of the disease-risk was monitored using the Simulink scopes and numerical displays. This enabled the implementation to be verified in terms of numbers and logic in different environments.

Sensitivity analysis was also conducted by changing the environmental input parameters within their observed ranges and the difference in the model output for these changes was investigated. These simulations give evidence, that the implemented system is dynamic to the variation of its input. The simulated signals, however, were not regarded as extra empirical observation data, and were not included in the classification-performance metrics reported.

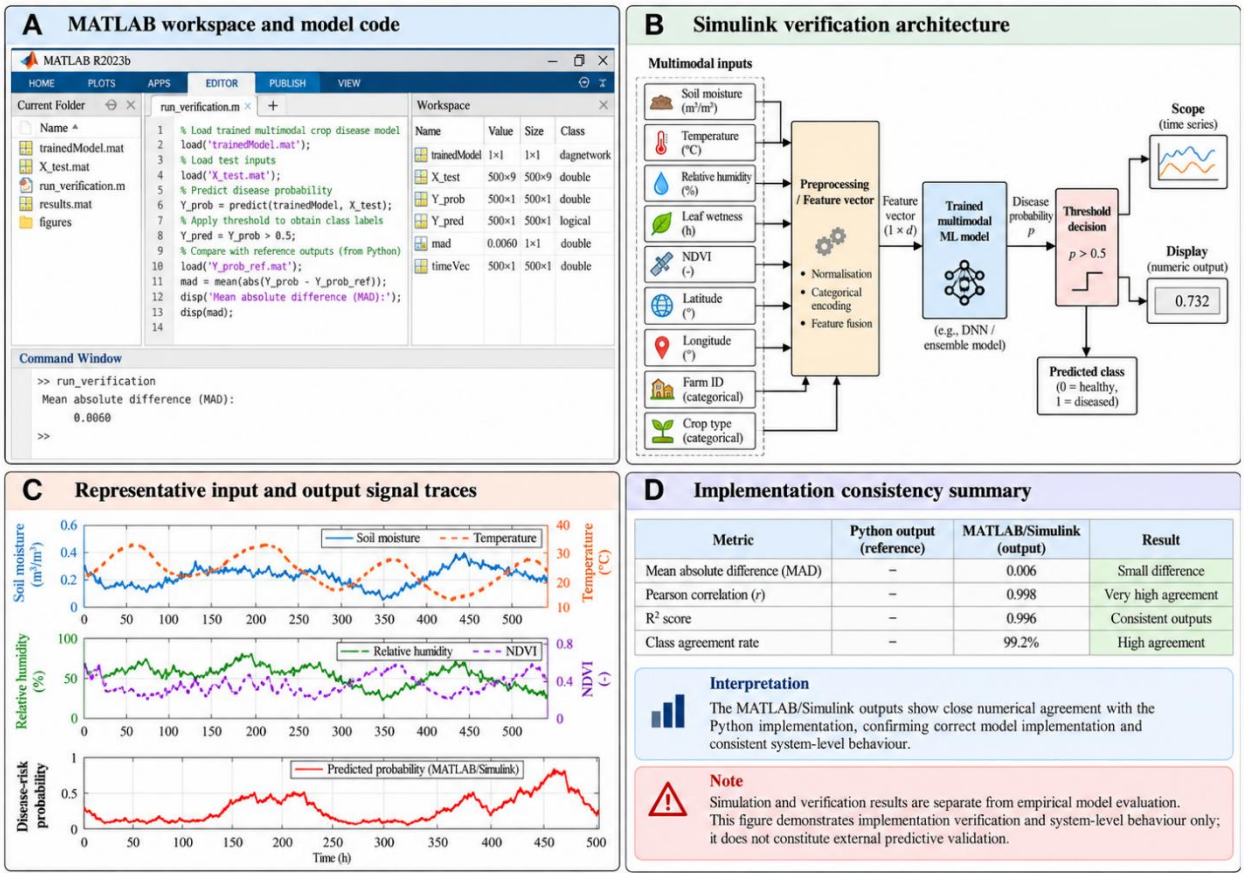


Figure 8: MATLAB/Simulink Verification of the Multimodal Disease-risk Framework

The MATLAB/Simulink implementation and verification of the multimodal disease-risk framework is shown in Figure 8. The MATLAB workspace illustrates the loading of the trained model and evaluation inputs, the calculation of disease probabilities, and the numerical comparison of the two implementations. The Simulink architecture for the corresponding model maps environmental variables, NDVI, GIS information, farm identity and crop type into a preprocessing and feature-fusion block, and feeds the resulting feature vector to the trained machine-learning model.

The dynamic changes in selected inputs and the resulting disease-risk probability generated by the system implemented are shown as representative signal traces. This validates the validity of the computational pipeline to a change in multimodal input conditions and validates the ability of transferring the model logic to a system level simulation environment. The implementation-

consistency summary also shows that there is a good numerical agreement between the Python reference model and the MATLAB/Simulink implementation.

4.11 Overall Discussion

The results are more conservative than a simple overall accuracy might suggest because the model results for each crop disease at each site demonstrate the relatively low accuracy of the multimodal crop-disease modelling approach. There was clear heterogeneity in the dataset with respect to the environment, space and time, but associations between individual predictors and disease were still weak. The bivariate association was strongest with NDVI, and was still modest. The present results indicate that the crop-disease status is not sufficiently described by single environmental variables and there is a need to use multivariable modelling. Concurrently, poor predictor-outcome relationships account for the modest amount of discrimination observed for the machine-learning models.

The difference between the overall accuracy and capability to detect the disease is one of the most important findings. Gradient Boosting had an apparently good group-aware accuracy of 83.8% but could not detect any diseased observations at the default threshold. This is due to the class-imbalance: 84.4% of all observations were healthy and a classifier, therefore, could achieve high accuracy by mostly predicting the majority class. Therefore, the results confirm that only accuracy is not enough to evaluate crop-disease models in which the prevalence of the disease is low.

The difference between farm-aware and chronological validation is another example of the need for careful development of evaluation design. The group-aware ROC-AUC values were around 0.46–0.53, which corresponds to a poor geographical generalization. The highest ROC-AUC value obtained from chronological evaluation was 0.599, which was still somewhat low for Random Forest and Gradient Boosting. Hence, the existing models cannot be considered as highly precise disease prediction machines. Rather, the results show that a multimodal agricultural surveillance system can be built and illustrate the constraints of both the sample size, class imbalance, and limited associations between the predictors, along with the farm environment.

The results also suggest that classification, temporal generalization and true early prediction are different. The present empirical data, however, lacks a reliable longitudinal Plot ID, so that the values of the predictors for a specific time \(t\) can be tied to the disease status of the same plot at a later time \(t+h\). Chronological testing in this study, therefore, does not validate 2-, 3-, or 5-day disease onset forecasting models but rather tests if the models developed from earlier observations work for later observations. This is acknowledged by the MS itself.

Furthermore, no independent CNN training and testing could be performed using raw crop images, so a validated end-to-end computer-vision claim cannot be made at the present time. To this end, the most plausible conclusion is that this study assesses the feasibility of multimodal disease-risk classification, mostly relying on environmental sensing, NDVI, crop type and spatial context.

Again, from a methodological point of view, the relatively poor farm-aware performance should not be looked upon as a negative outcome. It offers a welcome reminder about the dangers of overoptimistic claims resulting from the standard approach to random train-test splitting. A good model in the training and test sets may lack of value in deployment if only observations from the same farms are included in both sets. The study provides evidence on the need for geographical and temporal generalization in agricultural AI and its impact on performance.

4.12 Key Findings of the Study

The study yielded key results on the use of multimodal machine learning for assessing disease risks in tomatoes and chilli peppers in smallholder farms.

Data set was highly skewed towards one class. An overall disease incidence of 15.6% was obtained from the 500 observations analysed (422 healthy and 78 diseased). The descriptive incidence of chilli pepper was 17.6% which is higher than that of tomato (13.3%), but the difference between the crops was not significant at 0.05 level. This suggests that disease occurrence was not directly related to crop type.

Second, there was a significant variation in the environmental and remote-sensing variables among the farms in the study. Mean soil moisture was 35.03%, temperature 29.93°C, relative humidity 64.82%, leaf wetness 0.493, and NDVI 0.543. Even in this interindividual variability, the correlation of particular individuals with disease status was weak. The most positive association with disease was with NDVI, but its value was still low. The result provides further evidence in favor of a multivariable approach, and does not rely on just one environmental or remote-sensing indicator.

Third, the distribution of disease was also spatially and temporally heterogeneous. The highest prevalence of disease was observed in Farm 1 (23.9%) while the other farms had a prevalence of 12.8%-14.7%. Temporally, the overall disease incidence decreased from 18.6% in January to 13.9% in April 2026, although the decreasing trend was not strictly monotonic. These results show that disease occurrence had geographical and temporal variations, which once more emphasize the importance of farm-aware and chronological validation.

Fourth, for the machine-learning results, the group-aware results revealed that high overall accuracy did not necessarily imply high disease discrimination. Gradient Boosting had the best group-aware accuracy (83.8 %), but a recall for diseased observations of 0.000. Logistic Regression had the best recall (0.474) and ROC-AUC (0.532) metrics, but its discrimination level was slightly surpassing random classification. The detection of minority classes was also poor for the random forest and MLP classifiers.

Fifth, there was only limited improvement following chronological validation. The best chronological ROC-AUC was obtained by Gradient Boosting (0.599) and the lowest Brier score was obtained by Gradient Boosting (0.133); MLP obtained the highest balanced accuracy (0.532),

precision (0.231) and PR-AUC (0.214). The highest recall (0.350) was found for Logistic Regression. None of the models had a consistent advantage over others in terms of performance.

Sixth, the results confirmed that the evaluation of models is very metric-dependent. Gradient Boosting had the highest accuracy, but other disease-sensitive metrics resulted in other rankings. It highlights the significance of reporting balanced accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC and calibration measures in the assessment of imbalanced crop-disease datasets.

Seventh, the available Predicted_Disease_Risk variable was not used in the models and did not provide much extra predictive information that could not be used as a validated computer-vision disease score. A low correlation with the observed disease status suggested using derived or image-related predictors should be independently verified prior to incorporation into the final multimodal model.

The MATLAB/Simulink implementation showed that the multimodal framework could be computationally implemented and that model output was linearly related to variations in model inputs. This was, however, only consistency in implementation and not external predictive validation.

The study is a testament to the feasibility of integrating information from the IoT, remote sensing, spatial and machine learning in crop-disease surveillance efforts. Concurrently, the limited discrimination obtained with the strict farm-aware and chronological validation suggests that the present system is not yet a fully matured early-warning system, but rather a feasibility-level disease-risk assessment system.

5. Conclusion

We assess whether a multi-modal machine learning architecture can be successfully applied for predicting tomato and chilli pepper disease risk in smallholder farms in Ibadan, Nigeria. Our multimodal framework incorporates IoT environmental measures, satellite imagery data, crops varieties, and geographic coordinates. The analysis was done in Python. Implementation validation was performed via MATLAB/Simulink simulation.

Our empirical sample includes 500 cases, of which 422 (84.4%) cases represent healthy crops and 78 (15.6%) are diseased. Chilli pepper demonstrates higher disease percentage than tomato crops. Yet the difference is not statistically significant. We find large variability of environmental and remote sensing parameters among the observed smallholder farms. Disease occurrence is poorly correlated with both classes of input variables. Normalized difference vegetation index (NDVI) demonstrates a relatively large coefficient of association with disease presence. Nevertheless, its value stays low. Our finding confirms the inability to predict disease based exclusively on one variable.

Additional spatiotemporal analysis reveals that the likelihood of disease is uneven among the farms and time-periods of interest. For instance, Farm 1 has a higher proportion of diseased cases compared to other sites. Disease incidence also decreases monotonically from January to April 2026. Such disparities confirm the necessity to use farm-wise validation and chronological assessment in our machine learning study.

Machine learning outputs suggest that a perfect predictive classifier might achieve very high overall accuracy. For instance, under group-aware cross-validation scheme, Gradient Boosting algorithm shows best predictive accuracy. However, gradient boosting fails to detect diseased cases at its default probability threshold. On the contrary, Logistic Regression provides maximal recall and area under curve (AUC) of receiver operating characteristics under farm-aware cross-validation approach. Again, logistic regression does not outperform random guessing since its discrimination power is slightly above 50%. Similarly, chronological cross-validation produces moderately better accuracy for Gradient Boosting and Random Forest algorithms. The highest achieved AUC is equal to 0.599. Based on these machine learning results, we argue that all considered algorithms do not show sufficient discrimination ability to declare highly accurate predictive classifiers. The key novelty of our research consists in a new methodology for evaluation of agricultural machine learning models. Specifically, we demonstrate that the use of accuracy as a sole criterion of prediction quality is misleading. Machine learning assessment requires additional evaluation of recall, F1-score, AUC, precision-recall AUC, and calibration metrics. Our spatiotemporal assessment of models reveals that traditional randomly shuffled data partition does not uncover hidden weaknesses. The farm-aware and chronological validation schemes expose serious limitations of machine learning algorithms in terms of discrimination and accuracy. Finally, our MATLAB/Simulink module proves that the proposed multi-modal model structure is computationally implementable. Such verification is important since it verifies that we can replicate our results with alternative tools.

In conclusion, this study creates solid grounds for future application of multimodal surveillance systems for detecting plant diseases in agricultural smallholders' environment.

Funding Declaration

This research received no external funding. The authors did not receive any financial support from public, commercial, or not-for-profit funding agencies for the conduct of this study or the preparation of this manuscript.

References

Babu, G. R., Gokuldhev, M., & Brahmanandam, P. S. (2024). Integrating IoT for soil monitoring and hybrid machine learning in predicting tomato crop disease in a typical South India station. *Sensors, 24*(19), 6177. https://doi.org/10.3390/s24196177

Blanco-Carmona, P., Baeza-Moreno, L., Hidalgo-Fort, E., Martín-Clemente, R., González-Carvajal, R., & Muñoz-Chavero, F. (2023). AIoT in agriculture: Safeguarding crops from pest and disease threats. *Sensors, 23*(24), 9733. https://doi.org/10.3390/s23249733

Chen, K., Lang, J., Li, J., Chen, D., Wang, X., Zhou, J., Liu, X., Song, Y., & Dong, M. (2024). Integration of image and sensor data for improved disease detection in peach trees using deep learning techniques. *Agriculture, 14*(6), 797. https://doi.org/10.3390/agriculture14060797

Chin, R., Catal, C., & Kassahun, A. (2023). Plant disease detection using drones in precision agriculture. *Precision Agriculture, 24*, 1663–1682. https://doi.org/10.1007/s11119-023-10014-y

Das, P. K., Rupa, S. S., Pumrin, S., Das, U. C., & Hossen, M. K. (2024). Deep learning for plant disease detection and classification: A systematic analysis and review. *Current Applied Science and Technology, 24*(4), e0259016. https://doi.org/10.55003/cast.2024.259016 Thai-Journal Online

Delfani, P., Thuraga, V., Banerjee, B., & Chawade, A. (2024). Integrative approaches in modern agriculture: IoT, ML and AI for disease forecasting amidst climate change. *Precision Agriculture, 25*, 2589–2613. https://doi.org/10.1007/s11119-024-10164-7

Dhaka, V. S., Kundu, N., Rani, G., Zumpano, E., & Vocaturo, E. (2023). Role of Internet of Things and deep learning techniques in plant disease detection and classification: A focused review. *Sensors, 23*(18), 7877. https://doi.org/10.3390/s23187877

Isinkaye, F. O., Olusanya, M. O., & Singh, P. K. (2024). Deep learning and content-based filtering techniques for improving plant disease identification and treatment recommendations: A comprehensive review. *Heliyon, 10*(9), e29583. https://doi.org/10.1016/j.heliyon.2024.e29583

Jafar, A., Bibi, N., Naqvi, R. A., Sadeghi-Niaraki, A., & Jeong, D. (2024). Revolutionizing agriculture with artificial intelligence: Plant disease detection methods, applications, and their limitations. *Frontiers in Plant Science, 15*, 1356260. https://doi.org/10.3389/fpls.2024.1356260

Jelali, M. (2024). Deep learning networks-based tomato disease and pest detection: A first review of research studies using real field datasets. *Frontiers in Plant Science, 15*, 1493322. https://doi.org/10.3389/fpls.2024.1493322

Khan, B., Das, S., Fahim, N. S., Banerjee, S., Khan, S., Al-Sadoon, M. K., Al-Otaibi, H. S., & Islam, A. R. M. T. (2024). Bayesian optimized multimodal deep hybrid learning approach for tomato leaf disease classification. *Scientific Reports, 14*, 21525. https://doi.org/10.1038/s41598-024-72237-x

Li, W., Zhu, L., & Liu, J. (2024). PL-DINO: An improved transformer-based method for plant leaf disease detection. *Agriculture, 14*(5), 691. https://doi.org/10.3390/agriculture14050691

Mahmood, R., Matin, M. A., Goudos, S., & Karagiannidis, G. K. (2024). Machine learning for smart agriculture: A comprehensive survey. *IEEE Transactions on Artificial Intelligence, 5*(6), 2568–2588. https://doi.org/10.1109/TAI.2023.3345278

Muhammed, D., Ahvar, E., Ahvar, S., Trocan, M., Montpetit, M.-J., & Ehsani, R. (2024). Artificial Intelligence of Things (AIoT) for smart agriculture: A review of architectures, technologies and solutions. *Journal of Network and Computer Applications, 228*, 103905. https://doi.org/10.1016/j.jnca.2024.103905

Njoroge, T. K., Omol, E. J., & Nyangaresi, V. O. (2025). Deep learning and IoT fusion for crop health monitoring: A high-accuracy, edge-optimised model for smart farming. *IET Image Processing, 19*(1), e70208. https://doi.org/10.1049/ipr2.70208

Omaye, J. D., Ogbuju, E., Ataguba, G., Jaiyeoba, O., Aneke, J., & Oladipo, F. (2024). Cross-comparative review of machine learning for plant disease detection: Apple, cassava, cotton and potato plants. *Artificial Intelligence in Agriculture, 12*, 127–151. https://doi.org/10.1016/j.aiia.2024.04.002

Pacal, I., Kunduracioglu, I., Alma, M. H., Deveci, M., Kadry, S., Nedoma, J., Slany, V., & Martinek, R. (2024). A systematic review of deep learning techniques for plant diseases. *Artificial Intelligence Review, 57*, 304. https://doi.org/10.1007/s10462-024-10944-7

Qadri, S. A. A., Huang, N.-F., Wani, T. M., & Bhat, S. A. (2024). Advances and challenges in computer vision for image-based plant disease detection: A comprehensive survey of machine and deep learning approaches. *IEEE Transactions on Automation Science and Engineering, 22*, 2639–2670. https://doi.org/10.1109/TASE.2024.3382731

Rodríguez-Lira, D.-C., Córdova-Esparza, D.-M., Álvarez-Alvarado, J. M., Terven, J., Romero-González, J.-A., & Rodríguez-Reséndiz, J. (2024). Trends in machine and deep learning techniques for plant disease identification: A systematic review. *Agriculture, 14*(12), 2188. https://doi.org/10.3390/agriculture14122188

Routis, G., Michailidis, M., & Roussaki, I. (2024). Plant disease identification using machine learning algorithms on single-board computers in IoT environments. *Electronics, 13*(6), 1010. https://doi.org/10.3390/electronics13061010

Sajitha, P., Andrushia, A. D., Anand, N., & Naser, M. Z. (2024). A review on machine learning and deep learning image-based plant disease classification for industrial farming systems. *Journal of Industrial Information Integration, 38*, 100572. https://doi.org/10.1016/j.jii.2024.100572

Shahi, T. B., Xu, C.-Y., Neupane, A., & Guo, W. (2023). Recent advances in crop disease detection using UAV and deep learning techniques. *Remote Sensing, 15*(9), 2450. https://doi.org/10.3390/rs15092450

Sun, H., Fan, C., Gai, X., Al-Absi, M. A., Wang, S., Alam, M., Wang, X., & Fu, R. (2024). Multi-kernel inception aggregation diffusion network for tomato disease detection. *BMC Plant Biology, 24*, 1069. https://doi.org/10.1186/s12870-024-05797-9

Wang, X., & Liu, J. (2024). An efficient deep learning model for tomato disease detection. *Plant Methods, 20*, 61. https://doi.org/10.1186/s13007-024-01188-1

Wang, J., Li, M., Han, C., & Guo, X. (2024). YOLOv8-RCAA: A lightweight and high-performance network for tea leaf disease detection. *Agriculture, 14*(8), 1240. https://doi.org/10.3390/agriculture14081240

Yao, J., Tran, S. N., Garg, S., & Sawyer, S. (2024). Deep learning for plant identification and disease classification from leaf images: Multi-prediction approaches. *ACM Computing Surveys, 56*(6), Article 153, 1–37. https://doi.org/10.1145/3639816

Zhang, T., Cai, Y., Zhuang, P., & Li, J. (2024). Remotely sensed crop disease monitoring by machine learning algorithms: A review. *Unmanned Systems, 12*(1), 161–171. https://doi.org/10.1142/S2301385024500237

Zhu, H., Lin, C., Liu, G., Wang, D., Qin, S., Li, A., Xu, J.-L., & He, Y. (2024). Intelligent agriculture: Deep learning in UAV-based remote sensing imagery for crop diseases and pests detection. *Frontiers in Plant Science, 15*, 1435016. https://doi.org/10.3389/fpls.2024.1435016

Kouki, M., Kallel, M., Alsuwaylimi, A. A., Alenezi, S. M., & Zayani, H. M. (2024). Deep learning for tomato disease detection with YOLOv8. *Engineering, Technology & Applied Science Research, 14*(2), 13584–13591. https://doi.org/10.48084/etasr.7064

Zhang, T., Cai, Y., Zhuang, P., & Li, J. (2024). Remotely sensed crop disease monitoring by machine learning algorithms: A review. *Unmanned Systems, 12*(1), 161–171. https://doi.org/10.1142/S2301385024500237