



## AI-Driven Big Data Analytics for Crop Disease Prediction in Smart Agriculture

**Elena Petrova**

Asst Professor

### Abstract

The conclusion highlights that crop disease prediction involves the fusion of different data sources, including field monitoring data, satellite imagery, and historical disease information. A Web-based, end-to-end solution is proposed to facilitate easy access and utilization of crop disease prediction models by researchers, farmers, and organizations. Future trends include monitoring and assessing the effects of climate change on crop disease, formulating preventive measures, and integrating remotely sensed data to enable foresighted, AI-enhanced analyses.

Early and accurate prediction of crop diseases can greatly minimize loss of yield, effort, and resources. Disease prediction considers multiple data sources, such as monitoring systems that capture temperature, humidity, soil moisture, light, rainfall, and disease episodes, and satellite images that help in monitoring disease outbreaks. Machine learning and deep learning methods are widely used for disease prediction; however, the models require periodic retraining for long-term effectiveness. Key steps include feature engineering, feature selection, training and validation of different ML and DL models, confusion matrix analysis, and performance assessment using metrics such as accuracy, precision, recall, F1 score, and AUC. An end-to-end system architecture addresses data pipelines, storage, and model-building and monitoring.

**Keywords :** Crop diseases, artificial intelligence, big data analytics, smart agriculture, remote sensing, Internet of Things.

### 1. Introduction

Agriculture has reached a tipping point necessitating a shift from traditional methods to “smart agriculture” that leverages big data ecosystems with embedded Artificial Intelligence (AI) capabilities. The creation of knowledge-based decision-making and advisory systems for different sectors of agriculture requires a large volume of spatiotemporal data collected in real-time from diverse sources to analyse complex multi-factor problems. An important potential application area is enhancing crop productivity. Government institutions and stakeholders can benefit from agriculture analytics based on predictive modelling to identify climatic conditions most conducive for triggering a particular disease and reporting alerts accordingly.

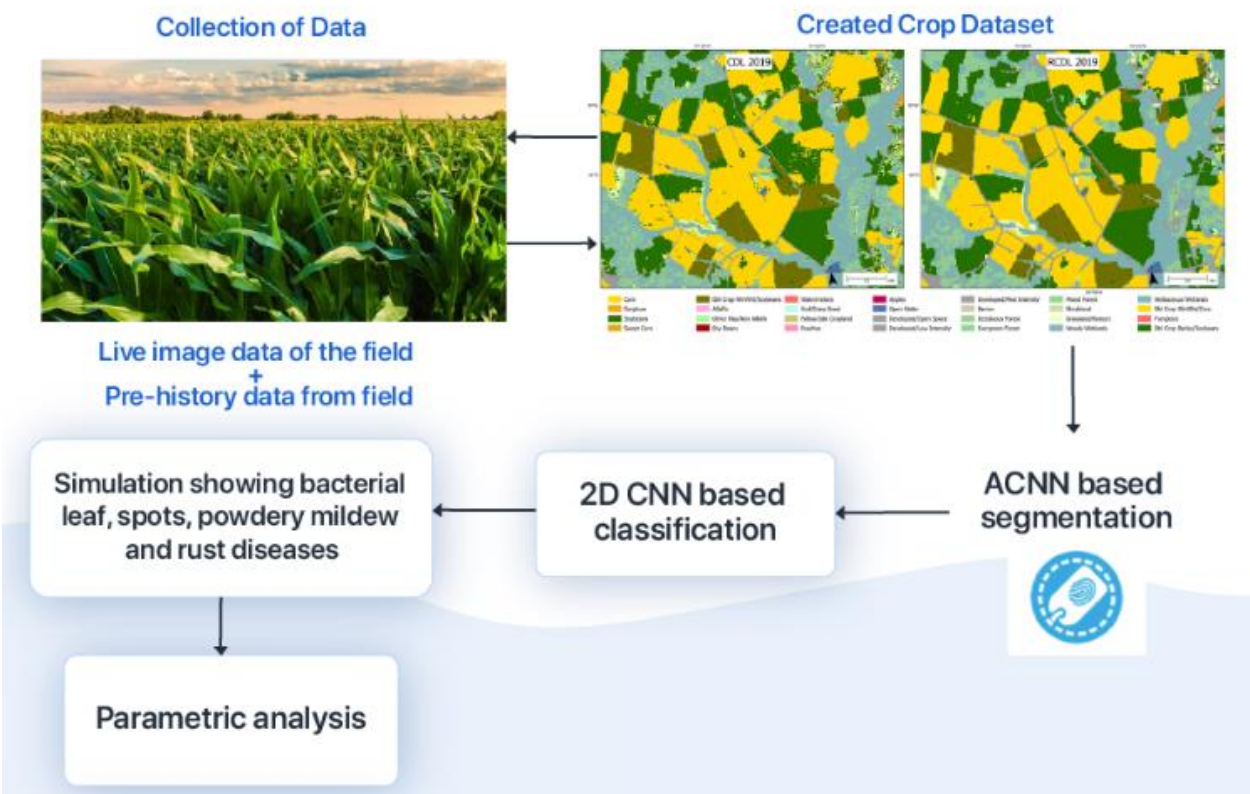
Decision-making in agriculture depends crucially on time-critical information from remote sensing and field sensors about the unanticipated emergence of crop diseases, which affect productivity and quality. Furthermore, spatiotemporal data from multiple sources, including weather conditions, soil moisture and temperature, shock temperature, leaf wetness, and moisture index, are useful for predicting crop disease outbreaks. As business processes fundamentally change with big data, the challenge is not only managing the large volume of data generated at high velocity but also converting it into useful information. Predictive modelling based on machine learning or deep learning techniques enables such big data analytics by employing feature engineering and model selection to improve predictive accuracy.

#### 1.1. Background and Significance

Recent advances in miniaturization and cost reduction of sensor technologies have made it feasible to instrument agricultural

fields with low-cost sensors. These devices enable monitoring of fine-grained environmental features such as soil moisture levels, and remote-sensing initiatives led by organizations like NASA offer multi-spectral satellite imagery with unprecedented temporal frequency. Alongside these data sources, a multitude of other types of data has emerged that captures all aspects of agricultural production, ranging from weather forecasts to commodity prices. AI-augmented big-data infrastructure forms the foundation on which modern-day data-driven business processes are constructed.

Crop diseases are amongst the most important factors that impact global food security. Traditional methods to manage crop diseases primarily rely on growers’ experiences and empirical knowledge, resulting in a "when I see it, I spray it" approach. A big-data-analytics-based system that predicts crop diseases before their actual occurrence would allow growers to apply fungicides only when diseases are imminent, resulting in significant savings in fungicide usage and reduced environmental impact. Recent research has shown that several key factors contributing to crop diseases can be captured by data obtained from field sensors, weather forecasts, satellite images, and disease incident reports. Machine learning and deep learning classifiers trained on such data sets have the potential to predict crop diseases with satisfactory accuracy.



**Fig 1: Early Crop Disease Detection with AI**



## 1.2. Research design

For the analysis of the impact of augmented big data on agriculture production, and especially for smart disease prediction in crop plants, the research design incorporates theoretical foundations, data sources, and methodology. The emphasis is on the role of sensors as well as remote sensing systems, and the intelligence-component draws on fusion processing between conventional machine learning and deep learning methods. The final development of a system architecture for the deployment and monitoring of data pipelines for these analytics is based on public cloud services.

The exploration of big data predictive analytics in smart agriculture encompasses the use of different machine learning and deep learning models for disease prediction in three groups of crops from diverse climatic zones: infection diseases of grapes and orange, and downy mildew of cucumber. Such predictive models are designed, assessed on selected metric evaluation measures, and cross-validated. The insights gathered during this exploration can enable improved living lab utility for model prediction and can facilitate ease of generalization across spatial and temporal zones. The currently emerging future directions of augmented big data predictive analytics in smart agriculture aim at incorporating development of novel crop physiology-based agronomic practices, enhancing input-energy-water-emission efficiencies for greater economy-ecology harmony, inter-plant constituency dynamics, closed-loop eco-system closeness in balance yields, and development of farm-internal and external eco-system modules.

### Equation 1: Dew point from temperature and relative humidity

Step 1:  $e_s(T) = 6.112 * \exp((aT)/(b+T))$   
 Step 2:  $e = (RH/100) * e_s(T)$   
 Step 3:  $e = e_s(T_d)$   
 Step 4:  $(RH/100) * 6.112 * \exp((aT)/(b+T)) = 6.112 * \exp((aT_d)/(b+T_d))$   
 Step 5:  $(RH/100) * \exp((aT)/(b+T)) = \exp((aT_d)/(b+T_d))$   
 Step 6:  $\ln(RH/100) + (aT)/(b+T) = (aT_d)/(b+T_d)$   
 Step 7: Let  $\gamma = \ln(RH/100) + (aT)/(b+T)$   
 Step 8:  $\gamma = (aT_d)/(b+T_d)$   
 Step 9:  $\gamma(b+T_d) = aT_d$   
 Step 10:  $\gamma*b + \gamma*T_d = aT_d$   
 Step 11:  $\gamma*b = T_d(a - \gamma)$   
 Step 12:  $T_d = (b*\gamma)/(a - \gamma)$

## 2. Theoretical Foundations

The utilities of big data and artificial intelligence (AI) in agriculture are outlined. The enormity of data quantities, along with the speed, variety, and veracity of modern data sources, are likely changing what is possible in agriculture beyond mere task acceleration to envisaging new tasks altogether. AI serves as a stimulating factor for the widespread introduction of Big Data, providing effective processing and analysing techniques, detecting complex patterns and using intelligent models to support decision-making and forecasting.



Agriculture has long been making use of advanced technologies. Synthetic fertilizers, herbicides, fungicides, insecticides, vaccines and genetically modified organisms (GMO) have all become common in modern agriculture. But modern agriculture invariably leads to pollution of water and soil, green-house-gas emission, climate change, lowering of soil organic matter and biodiversity, unbalance in natural ecosystems and so on, and these effects are now evident. Crop production has been typically heavy on resource consumption and poorly middle-in, leading to a relationship of increased production per unit area as connected to increased consumption per unit area. The Green Revolution had its impact at the national level and factored in both top down developments along with supporting technologies alike use of tractorization and tube wells. Devastating losses due to biotic agents such as diseases, insect pests, and weeds can affect large areas in a short time. AI can present and integrates Cloud Computing, The Internet of Things, and Deep Learning Technologies in order to improve the quality and profitability of farming product.

#### Equation 2: Core classification models mentioned in the article

Linear score:  $z = w^T x + b$

Probability:  $p(y=1|x) = 1/(1 + e^{(-z)})$

Decision:  $y_{\text{hat}} = 1$  if  $p(y=1|x) \geq \tau$ , else 0

Hyperplane:  $w^T x + b = 0$

Prediction:  $y_{\text{hat}} = \text{sign}(w^T x + b)$

Margin:  $i = y_i(w^T x_i + b) / \|w\|$

Optimization: minimize  $(1/2)\|w\|^2$  subject to  $y_i(w^T x_i + b) \geq 1$

Soft margin: minimize  $(1/2)\|w\|^2 + C \sum \xi_i$  with  $y_i(w^T x_i + b) \geq 1 - \xi_i$ ,  $\xi_i \geq 0$

Gini =  $1 - \sum p_c^2$

Binary case: Gini =  $1 - p^2 - (1-p)^2 = 2p(1-p)$

Gain = Gini(parent) -  $[(n_L/n)\text{Gini(left)} + (n_R/n)\text{Gini(right)}]$

Majority vote:  $y_{\text{hat}} = \text{mode}\{h_t(x)\}$

Mean probability:  $p_{\text{RF}}(y=1|x) = (1/T) \sum p_t(y=1|x)$

$Y(i,j) = \sum_m \sum_n X(i+m, j+n) K(m,n) + b$

$A(i,j) = \max(0, Y(i,j))$

$P(i,j) = \max \text{ over window of } A(u,v)$

Basic RNN:  $h_t = \tanh(W_x x_t + W_h h_{t-1} + b_h)$

Output:  $y_{\text{hat}_t} = g(W_y h_t + b_y)$

Forget gate:  $f_t = \sigma(W_f [h_{t-1}, x_t] + b_f)$

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

Candidate:  $c_{\text{hat}_t} = \tanh(W_c [h_{t-1}, x_t] + b_c)$

Cell state:  $c_t = f_t * c_{t-1} + i_t * c_{\text{hat}_t}$

Output gate:  $o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$

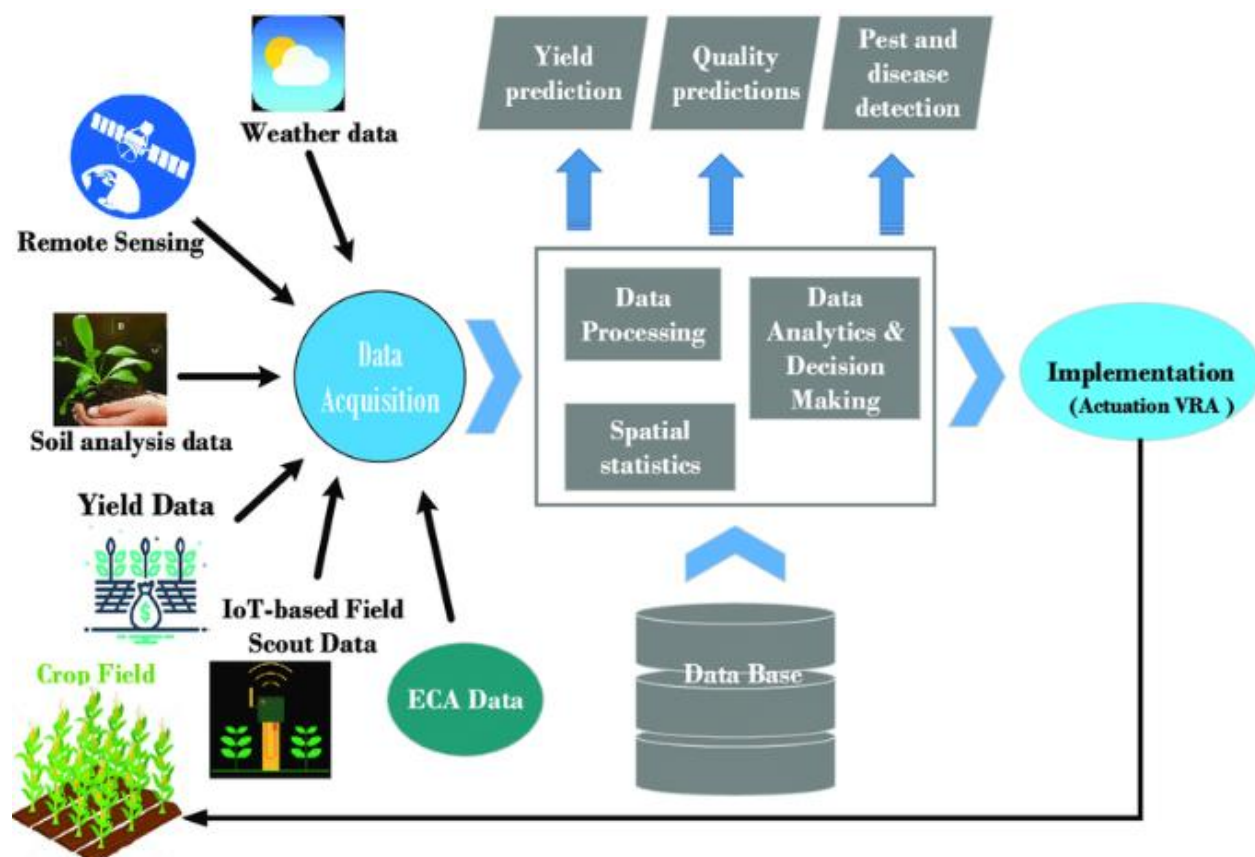
Hidden state:  $h_t = o_t * \tanh(c_t)$

#### 2.1. Big Data in Agriculture

Big data refers to large volumes of data, often unstructured, collected from a wide variety of sources that are constantly generating novel information. Big data is characterized by the three-volume dimensions of variety, velocity, and volume. The

multifaceted dimensions of agriculture generate huge volumes of data from different sources, including sensors, satellites, and social media. These multidimensional large datasets have the potential to solve complex problems, enabling timely monitoring of crops. The challenges of finding data and extracting useful knowledge from it are well-known.

In Pollution Technologies–Drastic Pollution Reduction in Renewable Energy for Smart Cities makes an important contribution by highlighting research on big-data generation technology in agriculture. It also describes a novel dictionary-learning analytical method for reducing pollution in corresponding energy facilities. This approach integrates big-data technologies into agriculture while using the idea of matching theory.



**Fig 2: Big data-based precision agriculture system representation**

## 2.2. Artificial Intelligence in Disease Prediction

Machine learning (ML) is a promising approach to address prediction problems as predictive models can be trained on past data with input features indicating the environmental conditions (temperature, rainfall, humidity, moisture, etc.) as well as the timing



of cropping activities (e.g., fertilizer application, sowing, and harvesting). Supervised ML formulations require a historical record of plant disease occurrences and an indication of corresponding diseases, which can be collected from governmental recommendations, agricultural universities, or farmers' records and then labeled in the training sets. In India, several smart phone applications have emerged that use such databases as sources for crop disease prediction. These are however limited to certain crops and geographic regions, and in many cases provide limited accuracy levels (sensitivity  $\leq 0.68$ ), likely due to the narrow training sets used.

Crop disease prediction can also be approached as a computer vision categorization problem, where the labels indicate the observed infected leaves in the training sets. Convolutional Neural Networks (CNNs) trained on such datasets have shown success in leaf disease recognition and been deployed in mobile applications for farmer assistance. Although such models predict the possible disease based only on the leaf features, practical crop monitoring must monitor the overall field and any disease attack detection can be further leveraged for prediction. Integrating the capabilities of such CNN and the supervised ML algorithms, an integrated staged approach is proposed for disease prediction.

### 3. Data Sources and Infrastructure

Although weather, soil, and other environmental data are essential for crop development, they are not sufficient to predict diseases. Software and hardware are needed for efficient disease prediction. Proper detection of crown rot in wheat improves productivity, yields, and quality while reducing management costs.

Data from field-fitted sensors, in conjunction with remote sensing data from satellite and ground verification, constitute the raw data set. The sensors are connected over the Internet of Things (IoT) network and acquire soil temperature, humidity, pH, moisture, and moisture content; weather parameters at one-minute intervals through a retrofitted microcontroller; and leaf nitrogen concentration through a leaf-nitrogen-analyzer. Thermal and multispectral images are acquired from drones to create surfaces for vegetation indices (VIs); S, G, and GNDVI are calculated. Optical rainfall data obtained by the laser technique are collected with a radar receiver. Optical soil-microclimatic data are recorded with a multichannel microcontroller using a protocol board at intervals. Optical imagery from the Indian remote sensing satellite series on a 25-m spatial resolution and multispectral images from Landsat and Sentinel are being acquired during the disease initiation period.

Crown rot disease creates yield loss in wheat crops and subsequent cereal crops due to elevated inoculum levels and less nitrous-oxide release during nitrous-oxide production. Early disease detection avoids repeated costs of postdetection management using forecasting models. The system has an architecture of a data pipeline from distinct sources that defines the target disease prediction and real-time monitoring of associated parameters.

#### Equation 3: Confusion matrix and metric derivations

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$F1 = 2 / (1/P + 1/R)$$



$$F1 = 2PR/(P + R)$$

$$\text{Specificity} = TN/(TN + FP)$$

$$\text{Balanced accuracy} = (\text{Recall} + \text{Specificity})/2$$

$$MCC = (TP*TN - FP*FN) / \sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}$$

Using  $TN=42$ ,  $FP=8$ ,  $FN=5$ ,  $TP=45$ :

$$\text{Accuracy} = (45 + 42)/100 = 0.87$$

$$\text{Precision} = 45/53 = 0.8491$$

$$\text{Recall} = 45/50 = 0.90$$

$$F1 = 2*0.8491*0.90/(0.8491 + 0.90) = 0.8738$$

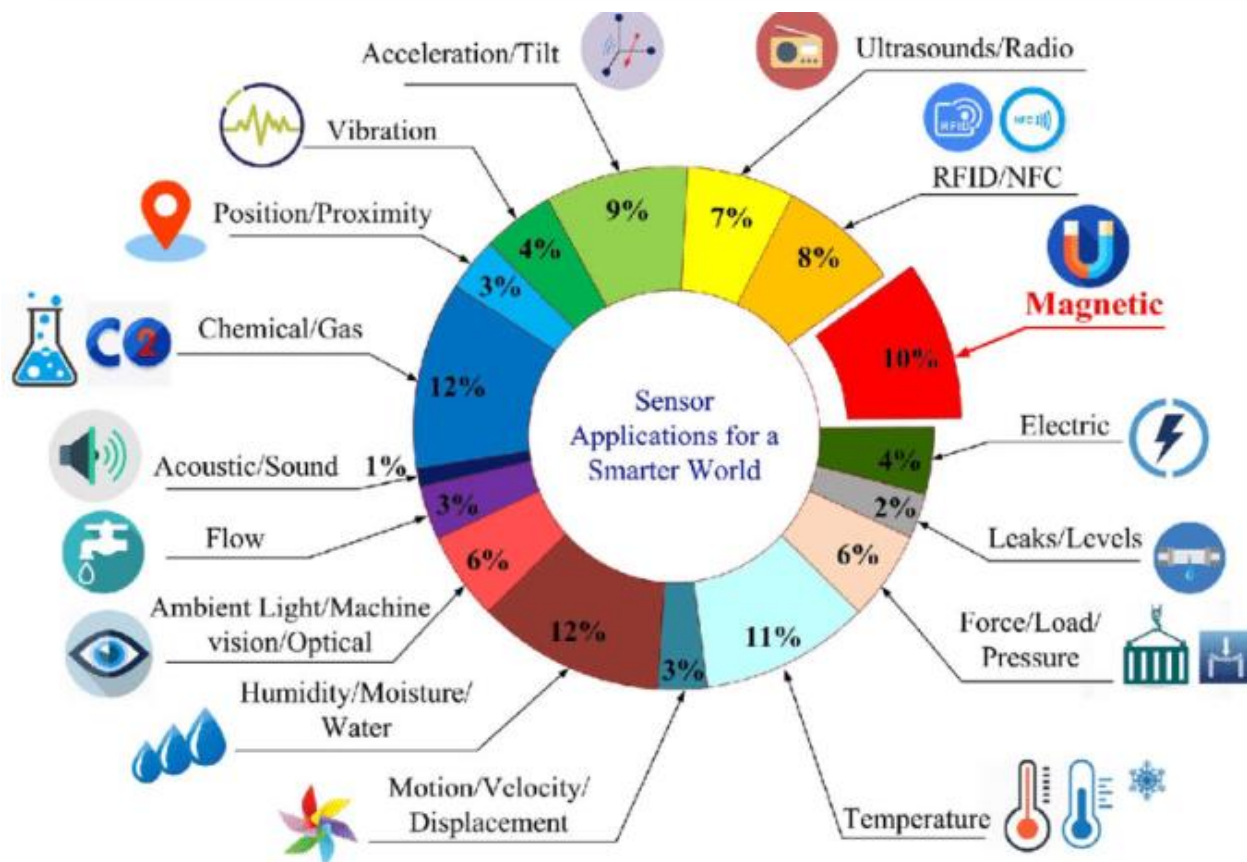
$$\text{Specificity} = 42/50 = 0.84$$

$$\text{Balanced accuracy} = (0.90 + 0.84)/2 = 0.87$$

### 3.1. Field Sensors and IoT

Agricultural systems are often enlarged by enhanced mechanization; however, in many cases, human resources are complementary if not essential. Unfortunately, such change passes many localities in development, and even the use of automated telemetry is either not possible or impractical for some parameters. Therefore, the objective is to facilitate a decentralized approach by designing a cheap data-collection unit, using common off-the-shelf components, and simple programming. This is valid for a drive station receiver transmitting to groups of local users dozens of microclimatic and floristic parameters. The technology uses a geo-referenced ZIG-Bee like network with very low consume in cars, and at the end, the similar wireless camera are also power by the Tags.

In the last decades, climate change increased the interest on phenological observation and prediction because of its strong relationship with e.g. insect or fungi abundance and distribution, as well as the modeling of crop simulation models. While studies that explore crop development at country or regional scales often relay on data-gathering organization like the US National Phenology Network, to have an insight on such patterns in a local environment is often more tricky. Treeboard is an original company located in Switzerland and it collects this type of information for the Helvetian territory. Data released are simultaneously transmitted on various platform like Global Unique Identifiers (G.U.I.D.), Yahoo! or Twitter pages. However, even in need or diffculted region, phenological user experiment with the ZIG-Bee technology idea. Thus, if all tree must be observed, the cost for each single tree can allow thousands of cameras and a large variety of users generation tasks.



**Fig 3: Sensor categories including magnetic sensors in IoT applications**

### 3.2. Remote Sensing and Satellite Imagery

In addition to IoT sensors at the farm level, remote sensing data serve as a key component of the augmented disease analytics engine. Regional observations at medium to coarse spatial resolution support the prediction of localized outbreaks, while finer-resolution satellite images enable precision forecasting applications. Open-access satellite datasets from Europe’s Copernicus Earth monitoring programme, such as Sentinel-2 and Copernicus Global Land Service, provide weekly multispectral imagery with spatial resolutions up to 10 m. Weather data from the European Centre for Medium-Range Weather Forecasts, the National Oceanic and Atmospheric Administration, and India Meteorological Department, along with soil moisture information from the Advanced Microwave Scanning Radiometer (Radiometer) on NASA’s EOS Aqua satellite and Normalized Difference Vegetation Index estimates, represent additional important sources of qualitative information.

Long-range forecasts for seasonal climate prediction can be obtained from the experimental seasonal weather forecast system developed at the Indian Institute of Tropical Meteorology, Pune. The resulting products include similar to 15-day forecasts of the



meteorological parameters required for agriculture. Weather conditions during the current season can also be predicted using various dynamic and statistical forecasting initiatives based on the monsoon ensemble. Space-based capabilities are further expanded by higher-frequency operating systems, such as the Unmanned Aerial Vehicle (UAV), which can be positioned close to the crops. The on-board sensors capture high-resolution spatio-temporal information for timely crop disease warnings and site-specific decision-making. The integration of UAV-based data and location-specific weather parameters enhances prediction performance.

#### Equation 3: ROC, PR, and AUC derivation

$$TPR(\tau) = TP(\tau) / (TP(\tau) + FN(\tau))$$

$$FPR(\tau) = FP(\tau) / (FP(\tau) + TN(\tau))$$

The ROC curve is the set of points  $(FPR(\tau), TPR(\tau))$  as  $\tau$  varies.

AUC = integral of TPR with respect to FPR

Discrete trapezoidal approximation:  $AUC \approx \sum (x_{(i+1)} - x_i) * (y_{(i+1)} + y_i) / 2$

## 4. Methodologies for Disease Prediction

Disease prediction requires careful selection of features and appropriate modeling methods. The vast array of input data creates opportunities to derive new features that capture complex relationships, and additional feature selection techniques can help identify the most influential drivers of disease occurrence. Machine learning methods, including boosted decision trees, random forests, and bagged ensemble classifiers, are particularly suited to tabular data but may lack the depth necessary to capture highly nonlinear relationships. Convolutional neural networks can process imagery directly, but require a large quantity of labelled input-output pairs. Transfer learning can reduce the size of the training set by leveraging established patterns in a pretrained model. Deep recurrent networks provide additional perspectives for timeseries prediction.

### 4.1. Feature Engineering and Selection

Feature engineering involves deriving new predictors based on existing data that better represent the underlying factors affecting the response variable. The initial set of predictors obtained from field sensors and remote sensing may not capture all the complex interactions that contribute to disease, but such patterns can be detected from historical data and expert knowledge. For instance, in addition to air temperature and humidity, substituting dew point for relative humidity more accurately represents the moisture content of the air as an initiating factor for the infection process of many foliar diseases. Deficient nutrient levels can predispose plants to diseases, and synthetic indices, especially on large domains, have been shown to provide better correlations with crop disease incidence than the individual components.

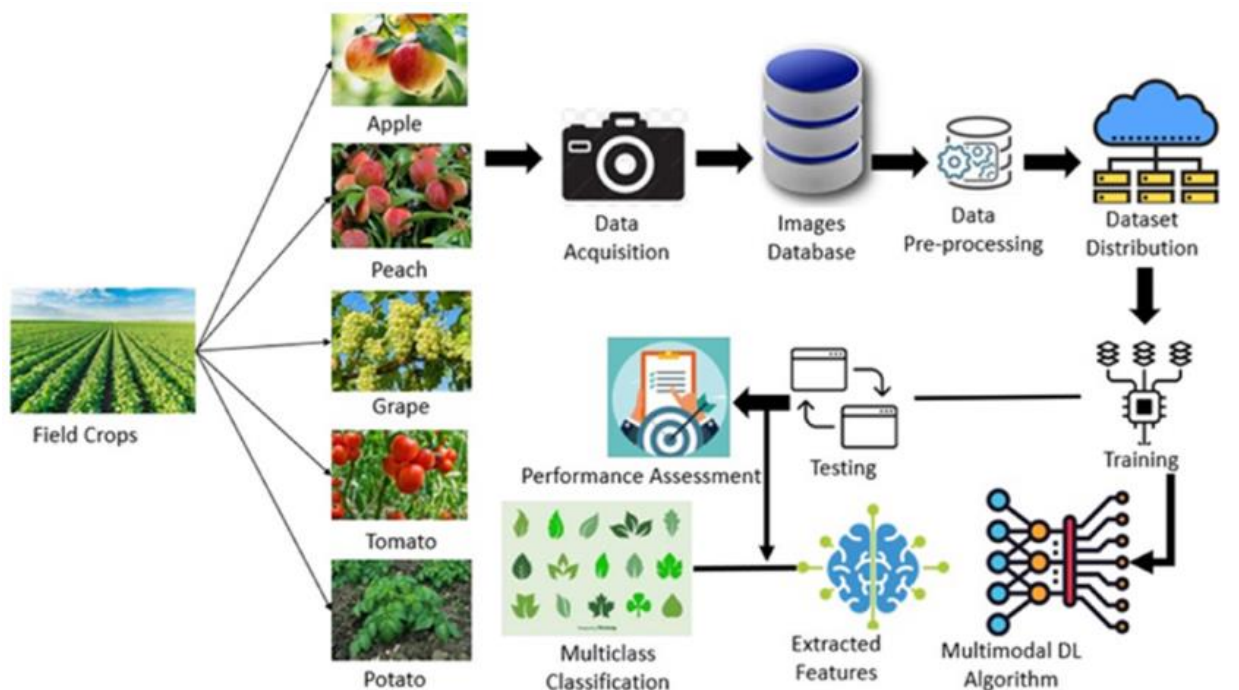
Feature selection methods can then be applied to identify the predictors that have a significant influence on disease occurrence. Standard approaches include univariate tests for continuous and categorical predictors, multivariate adaptive regression splines, decision trees, and LASSO regression. These techniques can reduce the number of predictors retained for machine learning models and also assist with feature sourcing and engineering by identifying suitable predictors for subsets of input data.

### 4.2. Machine Learning and Deep Learning Models

Boosted decision trees, random forest classifiers, and ensemble classifiers built using bagging techniques are popular choices for tabular data, particularly when predictors present a mixture of types. These methods incorporate set partitioning within the learning process and are relatively robust against overfitting when the dataset is sufficiently large. Achieving low prediction errors of these base classifiers prevents their weaknesses from degrading the performance of the final ensemble. However, they are less effective when trained on highly correlated predictors.

Convolutional neural networks have established new benchmarks in classification tasks involving image data, although for crop disease timing only limited labelled datasets exist. Transfer learning techniques that take advantage of the deep feature representations learned by previously trained models improve performance on small datasets by using a pretrained model as a starting point.

Recurrent architectures with long short-term memory cells are hierarchically designed for learning from interpretable structured data and combined with convolutional modules for processing visual information. Drought stress, weather data, and vegetation indices provide direct influences on diseased area prediction, while remotely sensed Red Edge and near-infrared bands serve as ancillary inputs.



**Fig 4: Enhancing Agricultural Disease Detection**



#### 4.1. Feature Engineering and Selection

Predictive modelling for disease incidence relies on a variety of features that influence outbreaks. Designing suitable and robust features for the display and diagnosis of disease prediction models is crucial, given that a small subset contributes to a predictive model while the remainder act as noise. Various data mining, machine learning and natural language processing techniques are generally employed in machine-generated data analysis. Natural language processing techniques, applied to millions of articles in crop disease archives, extract key words and phrases that highlight diseases in a geolocation and time context. However, such techniques are not easily transferrable to disease prediction modelling because predictive modelling analyses such as support vector machines and neural networks do not treat redundant features well, creating predictions that do not reflect features' relative importance. Consequently, model performance can be downgraded unless proper methods are used to delineate the appropriate features used for prediction.

Feature selection and feature extraction are therefore major issues in modelling and predicting epidemiological change. Feature selection methods are generally subdivided into three classes: wrapper approaches; filter methods; and embedded approaches. Wrapper approaches apply a greedy algorithm to a predictive model that assigns a score to individual feature subsets based on error rate or prediction accuracy. The wrapper-based method remains one of the best performers because it successfully identifies and removes irrelevant features in a noisy environment. Filter methods rank features according to a defined statistical measure, evaluating the individual relevance of features but ignoring feature dependencies. An embedded method is built into a learner (e.g. recursive feature elimination in convolution networks, regularization methods for regression problems, and tree ensembles using a node's split score).

#### Equation 4: Regression metrics for disease severity or yield-loss estimation

$$e_i = y_i - y_{\hat{i}}$$

$$MAE = (1/N) \sum |y_i - y_{\hat{i}}|$$

$$RMSE = \sqrt{(1/N) \sum (y_i - y_{\hat{i}})^2}$$

$$r = \frac{\sum (y_i - y_{\bar{y}})(y_{\hat{i}} - y_{\hat{\bar{y}}})}{\sqrt{\sum (y_i - y_{\bar{y}})^2 * \sum (y_{\hat{i}} - y_{\hat{\bar{y}}})^2}}$$

$$SST = \sum (y_i - y_{\bar{y}})^2$$

$$SSE = \sum (y_i - y_{\hat{i}})^2$$

$$R^2 = 1 - SSE/SST$$

#### 4.2. Machine Learning and Deep Learning Models

The primary mechanism for mapping feature vectors to predicted labels is a classification model. Models are trained using features from the previous season along with the labelled status of the crop at the end of the planting season. Classical machine learning methods, such as Support Vector Machines (SVM), the Random Forest Classifier (RFC), and k-Nearest Neighbours (kNN), are often complemented with Deep Learning algorithms. These complex models show promise in identifying increasingly abstract representations and handling modern datasets involving extremely large and possibly unstructured datasets.

Consequently, several variations of Multi-Layer Perceptrons (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) architectures have been adapted to model predictive relationships within Agriculture Big Data. These modifications have enabled food crop yield prediction, sensor-based soil quality measurement, disease detection in plant leaves, and pest activity prediction.



MLP architectures classify static datasets, RNN frameworks usually focus on dynamic datasets involving time-series features, and novelty-based CNN methods extract information from highly unstructured image datasets. Their inherent capability to readily extract salient features from real agricultural images has made them especially popular in identifying plague-infected crops and deriving pest activity levels from satellite and aerial imagery. Recently, the inherent sequential nature of these expert-labeled images has combined them with LSTM neural networks to detect symptoms of plant disease using image sequences. These methodologies have successfully assisted farmers in pest detection and symptom-specific pesticide recommendations. However, despite recent progress, a consensus classification model for disease prediction in food crops is still lacking.

## 5. System Architecture for AI-Enhanced Analytics

Analytics tools can be developed by a data scientist to monitor and visualize weather changes that affect crop yield. However, traditional machine learning models for disease prediction are not efficient for real-time, farm-level disease forecasting because they are usually not deployable models. They are based on limited training data and trained only once. Model drift often occurs due to unmonitored changes in the data—data changes that occur due to climatic variables, changes in farming practices, and genetic and varietal differentials in the cropping system.

The architecture presented in this research provides a foundation for a new approach to near-real-time, farm-level crop disease prediction driven by an AI-augmented big data infrastructure that automatically collects, processes, and curates massive amounts of data for machine learning model training and maintenance. Data pipelines ingest a wide range of environment, satellite, and sensor data related to crops, ultimately delivering feature sets to machine learning models hosted on cloud platforms. Automated procedures for model training and performance monitoring are in place, enabling a responsive model-operations ecosystem.

### Equation 5: Cross-validation derivations

Partition the N samples into k folds  $F_1, \dots, F_k$ .

Train on k-1 folds and validate on the held-out fold.

$CV_k = (1/k) \sum M_j$

Repeated stratified k-fold:  $RSKF = (1/R) \sum_r [ (1/k) \sum_j M_{(r,j)} ]$

Sliding-window F1 at time t:  $F1_t(w) = 2 P_t(w) R_t(w) / (P_t(w) + R_t(w))$

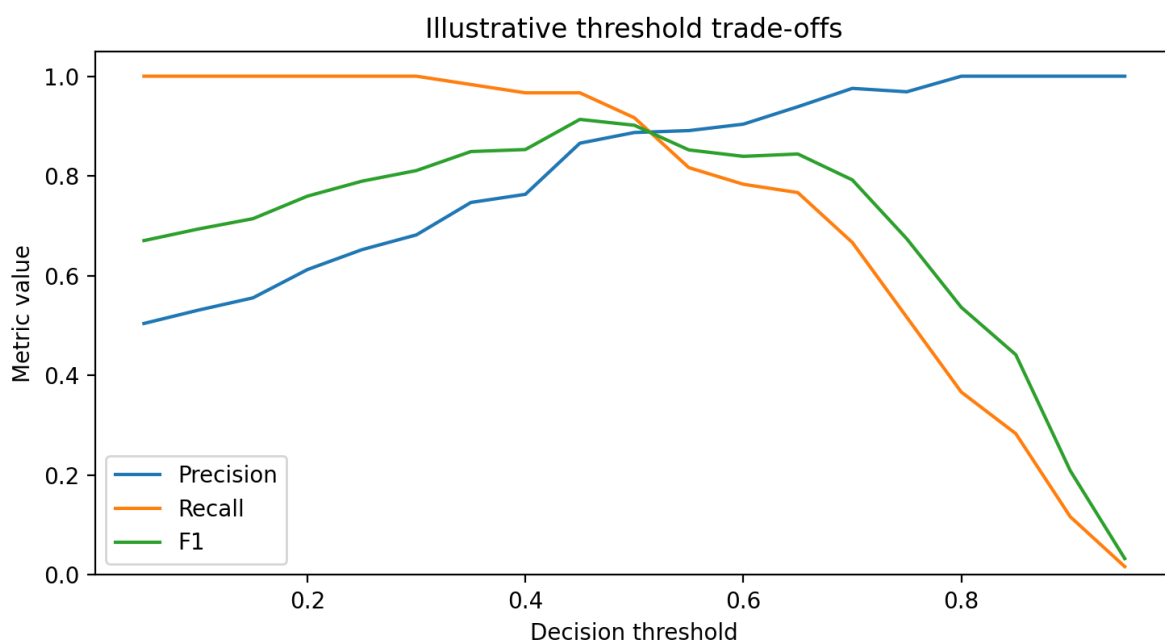
Drift alert rule:  $baseline\_F1 - F1_t(w) > \delta$

### 5.1. Data Pipelines and Storage

While the mainstream cloud service providers support sophisticated data pipelines that can accommodate big data applications, similar solutions can also be developed using open-source components, hosted on an internal datacentre or in a single cloud. With the continuous evolution of IaaS, PaaS and Big Data components, it is straightforward to set up pipelines for data ingestion, storage and processing. The proposed components can be integrated into each data-pipeline.

Data from field sensors, weather stations and IoT applications can be ingested and stored in a Big Data store using tools such as Apache Flume, Apache Kafka or Apache Nifi. This storage is usually required for a limited time period. A long-term repository for sensor data, such as an Apache HBase database with a time-series structure, can also be created. Such a solution will enable

faster access to the sensor data during the prediction process. Weather forecasting and satellite-based weather data can be made available as service-based repositories.



## 5.2. Model Deployment and Monitoring

With pipelines, storage, and joint models defined, the next step is to prepare the predictive models for operational use by extending their functionality for seamless integration into the analytics ecosystem, enabling proactive monitoring, and establishing alerting mechanisms. The first function to implement is prediction, a common requirement for almost all deployed models. For trivial models, such as those that select just one of the two possible labels for a binary classification task, the prediction can be defined explicitly, but it is more typical to utilize the base model's `.predict` method to perform the actual prediction. As uncertainty estimation is less common for standard machine learning models, this capability is made optional.

As background, these models need to be augmented with the capability to make prediction requests via HTTP, so that they can be called from anywhere with a simple REST API call. When a prediction request is received, the model checks for the presence of raw pixel data in the request, and if present, the data is used to make a prediction. This prediction is either returned to the caller or used to perform some other action. Multi-class and multi-label classification models also receive a translation table mapping model labels to their human-readable names as an input to the prediction function, enabling the translation of model outputs into a more useful form. For models that establish an alerting trigger, the rationale for the prediction is evaluated against the established severity thresholds, and an alert is sent when appropriate.



## 6. Evaluation and Validation

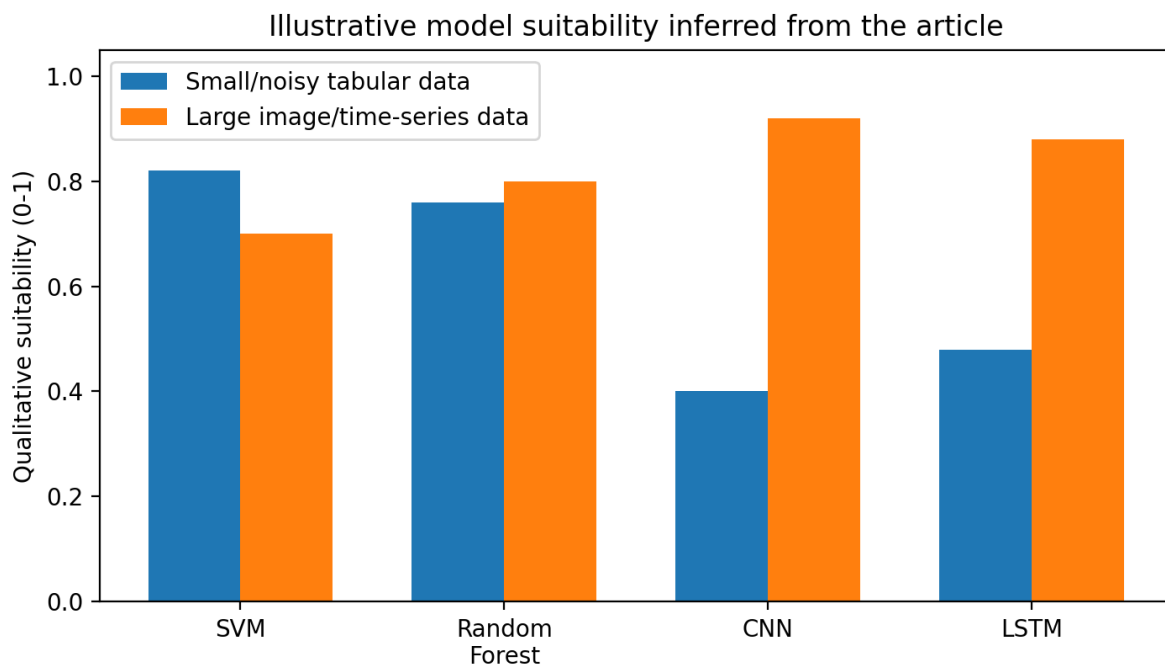
Evaluating predictive performance calls for appropriate metrics linked with the objective and the manner of predictive output. Most commonly, key-performance measures for supervised classification include accuracy, precision, recall, and F1 score. In cases where imbalance occurs between the number of samples in the categories, such as disease-versus-no-disease distinction, evaluation should also employ area under the receiver operating characteristic curve (AUC-ROC). The use of reliable cross-validation techniques provides additional insurance, especially in the case of quantity-limited datasets. A trained model may perform well on the training set but poorly generalize to unseen data due to complete or partial overfitting. Control sets may be kept aside to allow the covert validation of model generalization. An independent test set created from new samples is often a requirement at deployment. Such precautions are especially important in high-stakes applications such as disease prediction; it is better to err in predicting false negatives than false positives.

The deployment of real-time models also requires ongoing performance monitoring and maintenance. Concept-drift detection strives to capture a gradual deterioration in the predictive power of machine-learning models caused by factors such as climate changes and population dynamics in disease-causing agents. Such drift can be spotted by monitoring online changes in performance using, for example, F1 score computed in a sliding-window manner. The need to retrain the model can be notified through alerts. Alternatively, concepts can be incrementally updated online through a dedicated framework that automatically and continuously retains, discards, and retrains concept models as more data are received.

### 6.1. Metrics for Predictive Performance

To evaluate the predictive performance of machine learning and deep learning disease models, quantitative metrics from the areas of classification and regression can be employed. For classification models, metrics from confusion matrices, such as accuracy (Acc), Matthews correlation coefficient (MCC, Matthews 1975), and F1 score (F1) are well established for assessing predictive performance based on true/false positives/negatives. Nevertheless, as these classification metrics do not consider the relative severity of different classes, alternative scoring systems (Jiang et al. 2016) can provide different perspectives on model functionality. Area under the receiver operating characteristic (ROC) curve (AUC-ROC) and area under the precision-recall curve (AUC-PR) quantify the performance of classification models over threshold values, while balanced accuracy (B-Acc) and McNemar's test provide insights into class imbalance. For regression models, the more intuitive Pearson correlation coefficient ( $r$ ) in addition to root mean-square error (RMSE), mean absolute error (MAE), and coefficient of determination ( $R^2$ ) are commonly reported.

To better evaluate model generalization and robustness, repeated stratified k-fold (RSKF) cross-validation of TSS and B-Acc for classification or  $r$  for regression becomes standard procedure. In RSKF, the dataset is divided into k-folds, with strategy explicit for class-imbalanced problems (Dewan et al. 2019). Rather than hold out the same fold(s) for each model training and valuation, repeated iterations rotate the identity of the hold-out fold(s) k times, such that every sample undergoes testing at one point, before being trained upon by the remainder of the data. For RSKF with N samples, the indices of data allocated to the hold-out fold(s) are randomly selected from N, and replotted for a user-defined number of repetitions (Nrep). Final scoring aggregates the observations over the RSKF iterations. RSKF, along with model selection, hyper-parameter tuning, and test data partitioning, remains important for disease models owing to the sensitivity of TSS performance to individual samples, even in large datasets.



## 6.2. Cross-Validation and Generalization

The risk of overfitting when training machine learning models needs to be considered when making long-term predictions in smart agriculture applications. The goal is to generalize well to unseen data. One approach to estimating generalization performance is k-fold cross-validation. In k-fold cross-validation, the dataset is randomly partitioned into k subsets. A model is trained using  $k - 1$  of the subsets as training data and validated on the remaining subset. This process is then repeated k times, with each subset used exactly once as the validation data. The k results are then averaged to produce an overall performance metric. The advantage of k-fold cross-validation is its ability to optimize model parameters while avoiding overfitting.

Cross-validation is particularly useful for time-series modeling since conventional overfitting checks cannot be performed by randomizing the training and test set division. In time-series data, the prediction at a given point depends on earlier estimations, so the model should not have access to future data when training predicts a target one time unit forward. Therefore, split-sample validation with an unseen test set is not a rigorous check for overfitting; cross-validation is a more reliable technique of estimating predictive performance.

## 7. Conclusion



Research description considers climate change, soil health, population growth, increasing food demand, etc., which will increase threats. A combination of intelligent world sensors for large-scale data collection and Artificial Intelligence-Augmented Data Analytics, AI-ApDA, will make crop disease prediction efficient. Different Field-computer coupling path analysis of crop diseases is explored. Big Data State of the Art focuses on natural sources using Agro IoT (A-IoT) sensors, Satellite Sensors, and Remote Sensors. A-IoT proposes multiple sensor setup with minimum cost and defence mechanism for data integrity and privacy. Unique technical and conceptual frameworks define satellite imagery data depth, meaning, relationship, redundancy, and uniqueness. Predictive features Definition is based on optical and near-infrared sensing principles. Multi-level prediction Capabilities of Machine Learning and Deep Learning techniques for crop diseases are emphasised. For feature suitability, scenario-wise suitability, domain suitability, and dataset size suitability of different feature engineering methods DL models are preferred except for very small datasets, where other ML techniques are used. SVM is the most generalised technique for basic ML methods. Prediction capability of A-IoT sensors is established for different spatial-temporal levels of depth.

The critical requirements for model design, deployment, and monitoring of farm, state, and country-level models in a Cloud-Outlook environment are discussed, and some solutions are proposed. Box-model-type predictive capability may generate demand at district and farmer levels. Value-added services for model owners are also suggested. A multi-location and multi-path dataset generation is increasing demand. New prediction dimensions, such as pest prediction and deeper inner-level disease prediction in plants, are explored. Technological integration and country- and state-level crop insurance prediction models are highlighted. The emerging trend is fully automatic pipeline-model generation capable of supplying data to different models owned by different providers at various spatial-temporal levels.

### 7.1. Emerging Trends

Emerging trends in crop production, pest control, and fertilizer usage are transforming agriculture globally. Data-intensive smart agriculture combines massive data collection through advanced field, satellite, and drone sensors with predictive analytics. Despite its appeal, the deluge of environmental, agronomic, and remote sensing data makes intelligent decision support challenging. Predictive models, especially in agriculture, are often under-validated, hampered by overfitting, domain adaptation, and spatial-temporal generalization problems. Augmenting big data analytics with inherent knowledge can address some of these limitations. Tapping into the explosion of cloud computing and big data technologies, industry and academic interest in deploying production-ready models is surging. An integrated cloud-based infrastructure enables the development, deployment, and life-cycle management of predictive models using available data. Datadriven pipelines extract, transform, train, publish, and continuously assess models while monitoring incoming data. This approach is exemplified by a novel big data analytics system for predicting plant disease outbreaks in smart agriculture using machine learning and deep learning.

Predicting crop diseases months in advance aids timely protective measures. Large weather, soil, historical disease, and satellite data can be harnessed for intelligent prediction. Many predictive models are simply showcased, followed by piecemeal attempts at life-cycle management. Validating all publicly available machine-learning models using a folding scheme uncovers weaknesses and high prediction uncertainty. The world's first vegetation index-based deep-convolutional-network model for sugarcane disease prediction validates well. However, the rice blast prediction model performs poorly because the large and noisy dataset used overwhelms the model's learning ability.



## 8. References

1. Zhou, J., Li, J., Wang, C., Wu, H., Zhao, C., & Teng, G. (2021). Crop disease identification and interpretation method based on multimodal deep learning. *Computers and Electronics in Agriculture*, 189, 106408.
2. Abbas, A., Jain, S., Gour, M., & Vankudothu, S. (2021). Tomato plant disease detection using transfer learning with C-GAN synthetic images. *Computers and Electronics in Agriculture*, 187, 106279.
3. Yandamuri, U. S. (2024). AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology. *Metallurgical and Materials Engineering*.
4. Gui, P., Dang, W., Zhu, F., & Zhao, Q. (2021). Towards automatic field plant disease recognition. *Computers and Electronics in Agriculture*, 191, 106523.
5. Canals, R., et al. (2021). Computer vision, IoT and data fusion for crop disease detection using machine learning: A survey and ongoing research. *Remote Sensing*, 13(13), 2486.
6. Mattaparthi, R. (2024). Transformer-Based Fault Diagnosis for Large-Scale Standby Power Generators: Partial Discharge Pattern Recognition at Hyperscale Data Center Installations. *Journal of Computational Analysis and Applications (JoCAAA)*, 33(08), 8781-8799.
7. Albattah, W., Nawaz, M., Javed, A., Masood, M., & Albahli, S. (2022). A novel deep learning method for detection and classification of plant diseases. *Complex & Intelligent Systems*, 8, 507–524.
8. Borhani, Y., Khoramdel, J., & Najafi, E. (2022). A deep learning based approach for automated plant disease classification using vision transformer. *Scientific Reports*, 12, 11554.
9. Kolla, S. H., & Peddi, R. K. (2024). Designing Governance-Aligned GenAI Pipelines Using Small Language Models for Enterprise Workflow Intelligence. *International Journal of Science, Research and Technology*, 7(6), 13256-13268.
10. Nawaz, M., Nazir, T., Javed, A., Masood, M., Rashid, J., Kim, J., & Hussain, A. (2022). A robust deep learning approach for tomato plant leaf disease localization and classification. *Scientific Reports*, 12, 18568.



11. Mangalampalli, B. M. (2024). Transparent Intelligence Explainability Frameworks for AI-Driven Clinical Decision Support in Healthcare Business Intelligence. *International Journal of Research Publications in Engineering, Technology and Management (IJPETM)*, 7(3), 10566-10579.
12. Chen, W., Chen, J., Duan, R., Fang, Y., Ruan, Q., & Zhang, D. (2022). MS-DNet: A mobile neural network for plant disease identification. *Computers and Electronics in Agriculture*, 199, 107175.
13. Janarthan, S., Thuseethan, S., Rajasegarar, S., & Yearwood, J. (2022). P2OP—Plant Pathology on Palms: A deep learning-based mobile solution for in-field plant disease detection. *Computers and Electronics in Agriculture*, 202, 107371.
14. Inala, R. (2021). A New Paradigm in Retirement Solution Platforms: Leveraging Data Governance to Build AI-Ready Data Products. *Journal of International Crisis and Risk Communication Research*, 286-310.
15. Dey, B., Haque, M. M. U., Khatun, R., & Ahmed, R. (2022). Comparative performance of four CNN-based deep learning variants in detecting Hispa pest, two fungal diseases, and NPK deficiency symptoms of rice (*Oryza sativa*). *Computers and Electronics in Agriculture*, 202, 107340.
16. Inala, R. (2022). Engineering Data Products for Investment Analytics: The Role of Product Master Data and Scalable Big Data Solutions. *International Journal of Scientific Research and Modern Technology*, 155-171.
17. Saleem, M. H., Potgieter, J., & Arif, K. M. (2022). Plant disease detection and classification by deep learning. *Plants*, 11(6), 1–24.
18. Li, Z., Yang, J., Wu, H., & Zhang, Y. (2022). Deep learning for plant disease recognition: A comprehensive review. *Agriculture*, 12, 1350.
19. Aravind, R., Deon, E., & Surabhi, S. N. R. D. (2024). Developing cost-effective solutions for autonomous vehicle software testing using simulated environments using AI techniques. *Educational Administration: Theory and Practice*, 30(6), 4135-4147.
20. Alzahrani, M. S., Alotaibi, M. A., & Alharbi, A. (2022). Machine learning for detection and prediction of crop diseases and pests: A comprehensive survey. *Agriculture*, 12(9), 1350.



21. Jung, M., Song, J. S., Shin, A. Y., Choi, B., Go, S., Kwon, S.-Y., Park, J., Park, S. G., & Kim, Y.-M. (2023). Construction of deep learning-based disease detection model in plants. *Scientific Reports*, 13, 7331.
22. Reddy, V. A. R. (2024). Generative Intelligence for Healthcare Claims Processing and Personalized Benefits Management. *International Journal of Advanced Engineering Science and Information Technology (IJAESIT)*, 7(3), 14099.
23. Shi, T., Liu, Y., Zheng, X., Hu, K., Huang, H., Liu, H., & Huang, H. (2023). Recent advances in plant disease severity assessment using convolutional neural networks. *Scientific Reports*, 13, 2336.
24. Salman, Z., Muhammad, A., Piran, M. J., & Han, D. (2023). Crop-saving with AI: Latest trends in deep learning techniques for plant pathology. *Frontiers in Plant Science*, 14, 1224709.
25. Aravind, R., & Surabhi, S. N. R. D. (2024). Smart charging: AI solutions for efficient battery power management in automotive applications. *Educational Administration: Theory and Practice*, 30(5), 14257-1467.
26. Shahoveisi, F., Taheri Gorji, H., Shahabi, S., Hosseini-rad, S., Markell, S., & Vasefi, F. (2023). Application of image processing and transfer learning for the detection of rust disease. *Scientific Reports*, 13, 5133.
27. Luo, Y., Cai, X., Qi, J., Guo, D., & Che, W. (2023). FPGA-accelerated CNN for real-time plant disease identification. *Computers and Electronics in Agriculture*, 207, 107715.
28. Dey, N., et al. (2023). Automation of crop disease detection through conventional machine learning and deep transfer learning approaches. *Agriculture*, 13(2), 352.
29. Aitha, A. R. (2024). Generative AI-Powered Fraud Detection in Workers' Compensation: A DevOps-Based Multi-Cloud Architecture Leveraging, Deep Learning, and Explainable AI. *Computer Fraud and Security*.
30. Hassan, S. M., et al. (2023). DeepCrop: Deep learning-based crop disease prediction with web application. *Journal of Agriculture and Food Research*, 14, 100764.
31. 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.



32. Natarajan, S., Chakrabarti, P., & Margala, M. (2024). Robust diagnosis and meta visualizations of plant diseases through deep neural architecture with explainable AI. *Scientific Reports*, 14, 13695.
33. Aravind, R. (2024). Integrating controller area network (CAN) with cloud-based data storage solutions for improved vehicle diagnostics using AI. *Educational Administration: Theory and Practice*, 30(1), 992-1005.
34. Yang, B., Li, M., Li, F., Wang, Y., Liang, Q., Zhao, R., Li, C., & Wang, J. (2024). A novel plant type, leaf disease and severity identification framework using CNN and transformer with multi-label method. *Scientific Reports*, 14, 11664.
35. Rezaei, M., Diepeveen, D., Laga, H., Jones, M. G. K., & Sohel, F. (2024). Plant disease recognition in a low data scenario using few-shot learning. *Computers and Electronics in Agriculture*, 219, 108812.
36. Wang, H., Nguyen, T. H., Nguyen, T. N., & Dang, M. (2024). PD-TR: End-to-end plant diseases detection using a transformer. *Computers and Electronics in Agriculture*, 224, 109123.