

AI-Powered Intrusion Detection Framework For Smart Agriculture In Remote And Disaster-Prone Area

Kowshika D K, PG Scholar, Department of Electronics and Communication Engineering, Government College of Engineering, Salem, India.

Manuscript Received: July 23, 2026; Revised: July 28, 2026; Published: July 28, 2026

Abstract: Smart agriculture utilizes IoT devices, sensors, and wireless communication to enhance crop yield and promote sustainable farming methods. In India, the impacts of climate change have notably influenced key crops, leading to inconsistent yields over the past few decades. Precise crop yield forecasting prior to harvest is crucial for effective planning and resource allocation. This study introduces an AI-powered Intrusion Detection System (IDS) aimed at securing smart agriculture networks while facilitating intelligent crop yield forecasting. The system integrates IoT-driven agricultural data with machine learning approaches. A Random Forest algorithm is utilized to predict crop yields with a high level of accuracy. An interactive web-based platform has been created to offer a user-friendly experience for farmers and stakeholders. Standard datasets like NSL-KDD are employed to train and assess the IDS module. The IDS consistently supervises the IoT network. The proposed system improves overall agricultural productivity. It also strengthens cybersecurity in smart farming environments. The approach promotes sustainable and data-driven agriculture.

Keywords: Intrusion Detection System (IDS), Artificial Intelligence (AI), Machine Learning (ML), Internet of Things (IoT).

1. Introduction

Smart agriculture has gained significant attention as an effective solution to improve crop productivity, optimize resource utilization and promote sustainable farming practices through the use of Internet of Things (IoT) technologies[1]. In India, climate change and environmental variability over the past two decades have adversely affected the yield of major crops, creating uncertainty for farmers and policymakers[1][2][3]. Early and accurate crop yield prediction plays a vital role in supporting informed decisions related to resource management, storage and market planning.[4][5] However, the increasing reliance on IoT devices in agriculture exposes smart farming systems to cybersecurity threats, which can compromise data integrity and system reliability[5][6][7]. To address these challenges, this paper proposes an AI-driven Intrusion Detection System integrated with a crop yield prediction framework[7]. By employing machine learning techniques such as the Random Forest algorithm and utilizing standard datasets, the proposed system aims to enhance both agricultural productivity and network security in smart agriculture environments.

An Intrusion Detection System (IDS) is a security mechanism that monitors network traffic and system activities to identify malicious behaviour and unauthorized access. In smart agriculture environments, where IoT devices and sensors are widely deployed, an IDS is essential for protecting sensitive agricultural data and ensuring uninterrupted system operation. The IDS analyses network patterns to detect cyber threats[7][8]. Modern IDS solutions utilize machine learning and artificial intelligence to distinguish normal and abnormal network behaviour and adapt to emerging threats[9]. By providing real-time alerts, the IDS enables timely response to security incidents [10]. Overall, integrating an IDS enhances the reliability, integrity and security of smart agriculture networks

2. Need For Intrusion Detection System In Smart Agriculture

Smart agriculture relies heavily on Internet of Things (IoT) devices, wireless sensors, and cloud-based data systems to monitor and control various farming activities such as irrigation, soil moisture analysis, weather monitoring and crop yield prediction. While these technologies enhance productivity and efficiency, they also expose agricultural networks to numerous cybersecurity threats. Cyberattacks such as Denial of Service (DoS), Distributed DoS (DDoS), data

tampering and unauthorized access can disrupt communication between IoT devices, corrupt critical data and even cause system downtime, leading to severe economic losses and crop damage.

An Intrusion Detection System (IDS) is therefore essential to monitor network traffic, detect malicious activities, and alert users in real time. It helps in maintaining data integrity, confidentiality and availability within smart farming systems. By integrating AI and machine learning techniques, an IDS can intelligently identify abnormal network behavior and adapt to new or evolving cyber threats. This proactive security mechanism ensures that smart agricultural operations remain uninterrupted, reliable and secure, especially in remote and disaster-prone regions where manual monitoring is difficult. Ultimately, implementing an IDS strengthens the overall resilience of smart farming infrastructures and safeguards the future of data-driven agriculture.

3. Implementation Methodology

The proposed work focuses on developing a machine learning-based system for crop prediction. The system's workflow begins with collecting a "Crop Dataset," which consists of relevant crop-related images or data. This raw data undergoes a pre-processing phase, where unnecessary noise is removed, and the dataset is enhanced for better accuracy in the subsequent steps. The pre-processing phase ensures that the data is clean and standardized, improving the model's performance.

3.1 Workflow Of Crop Prediction System

Parallel to the crop dataset, the trained data, which has been previously processed and labelled, is also pre-processed. This trained data is essential for teaching the classifier how to make accurate predictions by learning from historical crop data. Both the crop dataset and trained data undergo "Feature Extraction," a process where relevant features such as colour, shape, texture, or spectral data are extracted to represent the data effectively. Feature extraction is crucial for reducing the complexity of the data while maintaining the most significant attributes for prediction.

Once the features are extracted, the data is passed to the "Classifier Algorithm." The classifier algorithm is a machine learning model, trained using the processed data. It uses the extracted features to classify the crop accurately. The classifier's role is to analyse the input data and predict the corresponding crop type based on patterns learned from the trained data.

In addition to generating crop predictions, the system evaluates the "Accuracy" of the classifier's performance. This step involves comparing the predicted results with actual outcomes to measure the system's reliability and performance. The accuracy metric ensures that the classifier can make reliable predictions and provides feedback for improving the model.

The goal of this proposed work is to create an efficient, automated system for crop prediction that can aid farmers, agricultural researchers and crop management professionals in making data-driven decisions. By leveraging machine learning techniques, the system aims to accurately predict crops based on input data, improving agricultural productivity and planning. Furthermore, the system can be adapted to various environments and crop types, making it a versatile tool in the agricultural sector.

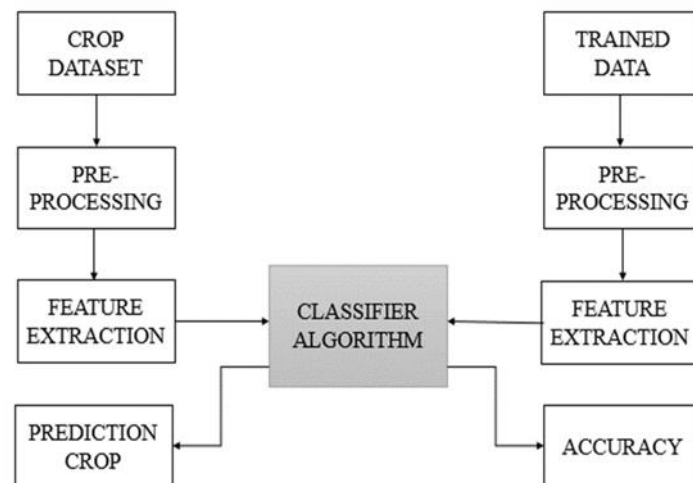


Fig 3.1: Block Diagram of Crop Prediction Model

3.3 Workflow

The block diagram Fig:3.1 demonstrates a systematic approach to intelligent crop prediction using machine learning. By integrating efficient data pre-processing, robust feature extraction and an accurate classifier, the proposed system enhances prediction accuracy, decision-making efficiency, and sustainable agricultural planning.

3.3.1 Data pre-processing

Data pre-processing is an essential stage in any data-driven project. It involves transforming the raw dataset into a structured, clean, and consistent format suitable for analysis and model training. The quality of data significantly affects the performance of machine learning models. In this study, several pre-processing steps were applied to the crop dataset to ensure high data quality, reliability, and readiness for feature extraction and classification. The steps are explained in detail below.

3.3.1.1 Importing Required Libraries

The Python libraries NumPy, Pandas, Matplotlib, Seaborn, and Plotly Express were imported to facilitate data handling, computation, and visualization. NumPy provides mathematical operations on arrays, while Pandas offers high-level data structures and tools for data manipulation. Matplotlib and Seaborn are used to generate static, publication-quality plots, and Plotly Express supports interactive visualizations. These libraries together provide a complete framework for data pre-processing and exploratory analysis, allowing for both numerical and graphical understanding of the dataset.

3.3.1.2 Loading the Dataset

The dataset was loaded into the Python environment using the Pandas function `read.csv()`. This command reads the comma-separated values file and converts it into a Data-Frame for easy data manipulation and inspection. The functions `head()` and `shape` were used to view the first few records and check the number of rows and columns respectively. This helps confirm that the dataset has been imported correctly without truncation or formatting errors. It also provides an initial understanding of the dataset's dimensions and structure, which is essential before proceeding to cleaning and analysis.

3.3.1.3 Handling Missing Values

The command `df.isnull().sum()` was used to check for missing or null values in each column. Missing data can occur due to human error during data entry or sensor malfunction during data collection. Such values can lead to inaccurate statistical measures and negatively impact model accuracy. Identifying and handling missing values ensures data

consistency and integrity. In this project, the output showed that there were no missing values, indicating that the dataset was complete and ready for further analysis without the need for imputation or record removal.

3.3.1.4 Descriptive Statistical Analysis

Descriptive statistics were computed using the `df.describe()` function, which provides information such as mean, standard deviation, minimum, maximum, and quartile values for each numerical feature. This step gives a general overview of the data distribution and central tendencies of attributes like Nitrogen, Phosphorus, Temperature, and Rainfall. It also helps to detect skewness and potential outliers that may affect model performance. Understanding these summary statistics aids in determining whether normalization, transformation, or scaling of features might be necessary in later stages.

3.3.1.5 Data Type and Structure Verification

The function `df.info()` was employed to examine the data types of each attribute and verify that all values were stored in appropriate formats. This function also reports the number of non-null entries, which is essential for detecting incomplete or incorrectly formatted columns. Correct data types ensure that numerical features can be used in mathematical operations and categorical features are processed properly during encoding. For instance, attributes like Nitrogen or Temperature should be numeric, whereas the "Label" column, representing crop type, should be of string or categorical type. Ensuring correct data types prevents computational errors and model mis-interpretation later.

3.3.1.6 Removing Duplicate Records

Duplicate entries in a dataset can introduce bias and affect the performance of predictive models. The command `df.duplicated().sum()` was used to check for any repeated rows. Duplicate records often occur due to repeated data collection or merging of multiple datasets without proper indexing. If present, they can cause overrepresentation of certain samples, leading to incorrect learning by the algorithm. In this dataset, the output confirmed that there were no duplicate entries, indicating that the dataset was unique and reliable. This verification step ensures the dataset's integrity and prevents redundancy.

3.3.1.7 Renaming Columns

Renaming columns improves readability and helps in maintaining consistency throughout the analysis. It provides clear, meaningful names for each feature, making it easier to interpret during visualization and modelling. This step also eliminates issues that may arise due to special characters or inconsistent naming conventions in the original dataset. Using uniform names ensures that the attributes can be accessed programmatically without confusion or errors during the model-building phase.

3.3.1.8 Exploratory Data Visualization

To understand the distribution and variability of each numerical feature, several visualizations were plotted using Seaborn and Matplotlib. Histograms with kernel density estimation (KDE) curves were used to visualize the frequency distribution and identify whether the data was normally distributed or skewed. Violin plots displayed both the distribution and density of the data, offering deeper insight into variability. Box plots were generated to detect the presence of outliers and visualize the median and quartile range. These visualizations help determine whether feature scaling or outlier removal is required before model training.

3.3.1.9 Grouping Data by Crop Type

The dataset was grouped based on the Label (crop type) column using the Pandas `groupby()` function, and the mean values of all numerical attributes were calculated for each crop. This grouping enables the comparison of average environmental and soil requirements for various crops. For instance, certain crops may require higher nitrogen levels, while others may thrive under specific temperature or rainfall conditions. This analysis provides valuable insights into the relationship between crop type and its environmental factors, which assists in developing a robust classification model.

3.3.1.10 Comparative Analysis Using Bar Plots

Finally, bar plots were generated to visually compare the mean values of all features for different crop categories. Each subplot represented one feature (e.g., Nitrogen, pH, or Rainfall) across all crop types. These visualizations provide a clear, comparative understanding of how different crops respond to varying soil nutrients and environmental parameters. Such comparisons are crucial in feature selection and in identifying which factors contribute most to crop differentiation. This step completes the exploratory phase and ensures that the data is well understood before feeding it into the classification algorithm.

4. Software Requirements

4.1 Python Programming Language

Python is selected as the primary programming language for this project due to its simplicity, flexibility and strong support for data science and machine learning applications. It offers a rich set of libraries for data preprocessing, visualization, model development and performance evaluation. The clear syntax and open-source nature of Python make it well suited for academic and research use. Additionally, libraries such as Scikit-learn, TensorFlow and Keras enable efficient implementation of machine learning algorithms with reduced code complexity.

4.2 Jupyter Notebook

Jupyter Notebook is used as the development environment for this project due to its interactive and user-friendly nature. It enables step-by-step execution of Python code with immediate visualization of results. By combining code, outputs, and documentation in a single interface, Jupyter Notebook enhances workflow clarity and reproducibility.

4.2.1 Libraries and Packages Used

Several Python libraries were utilized to perform data handling, visualization and pre-processing operations efficiently. These include:

- NumPy: Used for performing numerical and mathematical operations on large datasets efficiently using multidimensional arrays.
- Pandas: A powerful data analysis library used for loading, cleaning, manipulating and organizing structured data into Data-Frames.
- Matplotlib: A plotting library used for creating static visualizations such as line charts, bar graphs, histograms, and scatter plots.
- Seaborn: Built on top of Matplotlib, Seaborn provides advanced statistical visualization tools and attractive default styles to represent data distributions and correlations.
- Plotly Express: Used to create interactive and dynamic visualizations for better data interpretation and presentation.

These libraries collectively provide a complete toolkit for handling the entire data science pipeline — from loading and pre-processing the dataset to building and evaluating machine learning models.

8. Conclusion

This study explored the possibility of using machine learning techniques to identify mental health risk based on Instagram usage behaviour. The primary objective of this study was to determine whether behavioural usage patterns on Instagram could be used to identify potential mental health risks through machine learning techniques. Five behavioural features were used as input: age, daily Instagram usage hours, posting frequency, average likes per post, and follower count. Since the dataset had no pre-existing mental health labels, K-Means clustering was applied first to group users into Low Risk and High-Risk categories. Four classification models were then trained and tested on these labels.

Among all models, Random Forest performed the strongest with 99% accuracy, 100% precision, 99% recall, and an F1-score of 99%. Decision Tree matched these results closely, while Support Vector Machine achieved 89% accuracy and Logistic Regression reached 84%. These results confirm that tree-based and ensemble models handle this type of behavioural data better than linear approaches.

That said, one limitation must be acknowledged honestly. The risk labels in this study came from a clustering algorithm, not from actual clinical assessments. This means the models learned to replicate data groupings rather than real mental health diagnoses. Any conclusions drawn from this study should be understood within that context.

Going forward, future work should aim to collect datasets with verified clinical labels, explore deep learning approaches, and test models across broader and more diverse user groups. These steps would bring this type of system closer to practical real-world use.

Despite its limitations, this study shows that analysing social media behaviour through machine learning is a promising direction for early mental health risk detection. With the right data and ethical safeguards in place, such approaches could eventually support mental health professionals in identifying individuals who need help before their condition worsens.

5. Results And Discussion

This presents a crop-wise analysis by grouping the dataset according to crop labels and computing the mean values of key soil and environmental features. This analysis highlights variations in nutrient availability and climatic conditions across different crops, providing essential insights that support accurate crop prediction and recommendation models.

5.1 Pre-Processing

5.1.1 Missing Value Detection

This step validates the integrity and reliability of the dataset, ensuring that the analytical results derived from it will be accurate and dependable.

```
In [4]: df.isnull().sum()
Out[4]: N          0
        P          0
        K          0
        temperature  0
        humidity     0
        ph           0
        rainfall     0
        label        0
        dtype: int64
```

Fig 5.1 Output for missing values

The output shows zero (0) missing values across all attributes such as N, P, K, Temperature etc., as shown in Fig 5.1. Since no data is missing, no imputation or correction is required, and the dataset can be directly used for further processing and analysis.

5.1.2 Descriptive Statistical Summary

This step provides an essential overview of dataset characteristics, ensuring that further data analysis and modelling steps are based on a clear understanding of value distribution and variation

```
In [5]: df.describe()
Out[5]:
```

	N	P	K	temperature	humidity	ph	rainfall
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	50.551818	53.362727	48.149091	25.616244	71.481779	6.469480	103.463655
std	36.917334	32.985883	50.647931	5.063749	22.263812	0.773938	54.958389
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	20.211267
25%	21.000000	28.000000	20.000000	22.769375	60.261953	5.971693	64.551686
50%	37.000000	51.000000	32.000000	25.598693	80.473146	6.425045	94.867624
75%	84.250000	68.000000	49.000000	28.561654	89.948771	6.923643	124.267508
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	296.560117

There are outliers present in the dataset

Fig 5.2 Output of Descriptive Statistical Summary

A large difference between the minimum and maximum values for attributes like N, P and K suggests the presence of outliers or extreme values in the dataset as shown in the Fig 6.2. The statistical summary aids in understanding data spread and identifying anomalies.

5.1.3 Duplicate Check

```
In [7]: # To check for duplicates
df.duplicated().sum()
#No duplicates present

Out[7]: 0
```

Fig 5.3 Output of Overall Interpretation

The output obtained is “0”, which indicates that there are no duplicate entries present in the dataset as shown in the Fig 6.3.

5.1.4 Renaming Columns

Renaming columns is an important data pre-processing step to make the dataset more readable, organized and meaningful.

```
In [8]: #Renaming columns
df.columns = ['Nitrogen', 'Phosphorus', 'Potassium', 'Temperature', 'Humidity', 'pH', 'Rainfall', 'Label']
```

Fig 5.4 Output of Renaming Columns

As shown in Fig 6.5 command `df.columns = [...]` is used to assign new names to all column headers in the dataset at once. Renaming the columns helps make each feature name **self-explanatory**, thereby improving the dataset’s clarity.

5.1.5 Exploratory Data Visualization

The Exploratory Data Visualization which helps in understanding the underlying patterns, trends, and relationships within the dataset.

```
In [9]: plt.style.use('dark_background')
sns.set_palette("Set2")
for i in df.columns[:-1]:
    fig, ax = plt.subplots(1, 3, figsize=(18, 4))
    sns.histplot(data=df, x=i, kde=True, bins=20, ax=ax[0])
    sns.violinplot(data=df, x=i, ax=ax[1])
    sns.boxplot(data=df, x=i, ax=ax[2])
    plt.suptitle(f'Visualizing {i}', size=20)
```

Fig 5.5 Code for Exploratory Data Visualization

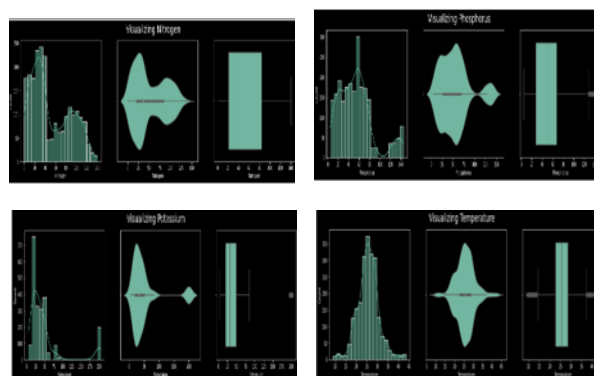


Fig 5.6 Visualization of Nitrogen Distribution across Crops

Histogram with KDE, Violin Plot and Box Plot are used to visualize feature distributions, central tendency, variability and skewness as shown in Fig 5.6. Together, these plots help identify data patterns and outliers, supporting informed pre-processing decisions such as normalization and feature scaling for improved model performance.

5.1.6 Grouping Data by Crop Type

Grouping helps in analysing trends and relationships between crops and their associated soil or environmental parameters.

In [11]: grouped

Out[11]:

	Label	Nitrogen	Phosphorus	Potassium	Temperature	Humidity	pH	Rainfall
0	apple	20.80	134.22	100.89	22.630942	92.333383	5.920983	112.654779
1	banana	100.23	82.01	50.05	27.378708	60.356123	5.983893	104.626980
2	blackgram	40.02	67.47	19.24	29.973340	65.118426	7.133952	87.884151
3	chickpea	40.09	67.79	79.92	18.872847	16.860439	7.336957	80.058977
4	coconut	21.98	16.93	30.59	27.409802	94.844272	5.976582	175.688646
5	coffee	101.20	28.74	29.94	25.540477	58.888846	6.790386	158.096295
6	cotton	117.77	46.24	19.56	23.968058	79.843474	6.912675	80.398043
7	grapes	23.18	132.53	200.11	23.849575	81.875228	6.025937	69.611829
8	jute	78.40	46.86	39.99	24.958376	79.635884	6.732778	174.792798
9	kidneybeans	20.75	67.54	20.05	20.115085	21.665357	5.749411	105.919778
10	lentil	18.77	68.36	19.41	24.609052	64.894785	6.927932	45.689454
11	maize	77.76	48.44	19.79	22.589204	65.092249	6.245190	84.795988
12	mango	20.07	27.18	29.92	31.208770	50.156673	5.766373	94.704515
13	mothbeans	21.44	48.01	20.23	28.194920	53.160418	6.831174	51.198487
14	mungbean	20.99	47.28	19.87	28.525775	85.499975	6.723957	46.403601
15	muskmelon	100.32	17.72	50.08	28.663066	92.342092	6.358806	24.689952
16	orange	19.58	16.55	10.01	22.785725	92.170289	7.016957	110.474089
17	papaya	49.88	59.05	50.04	33.723859	92.483386	6.741442	142.627839

Fig 5.7 Output of Grouping Data by Crop Type

The grouping process is performed to organize the dataset based on a common feature here, the Label column, which indicates the type of crop is shown in Fig 5.7

5.1.7 Comparative Analysis of Environmental Parameters for Different Crops

The comparative analysis of environmental parameters for different crops evaluates how factors like temperature, rainfall and humidity influence crop growth.

```
In [12]: fig,ax=plt.subplots(7,1,figsize=(25,25))
for index,i in enumerate(grouped.columns[1:]):
    sns.barplot(data=grouped,x='Label',y=i,ax=ax[index])
plt.suptitle("Comparision of Mean Attributes of various classes",size=25)
plt.xlabel("")
```

Fig 5.8 Code for Comparative Analysis

This analysis supports precision agriculture, helping farmers choose crops best suited to their soil and climate. The summarized data also aids in crop prediction and recommendation systems for sustainable farming.

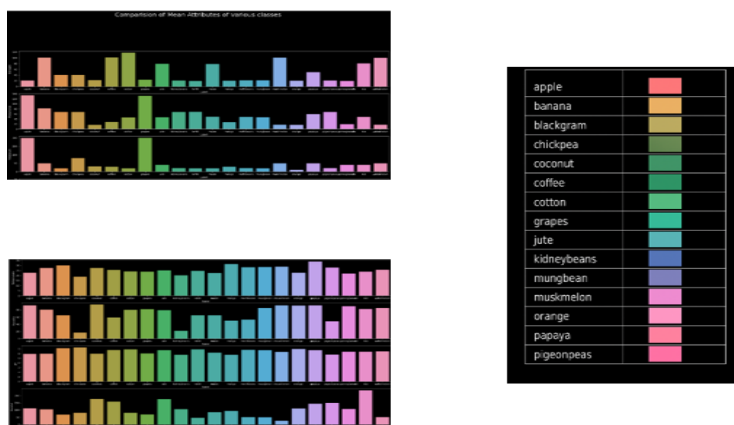


Fig 5.9 Comparison of Mean Soil Attributes for Different Crop Classes

Fig. 5.9 illustrates how average soil parameters vary among various crop types. The Fig 5.9 shows the mean soil and environmental parameters (Nitrogen, Phosphorus, Potassium, Temperature, Humidity, pH and Rainfall) across different crops.

5.2 Classifier Algorithm

A classifier algorithm is a machine learning method used to classify data into different categories and evaluate model performance. It is applied to predict the correct class label and measure performance using metrics such as accuracy, precision, recall and F1-score. These metrics help determine how well the model performs in predicting the correct outcomes.

5.2.1 Logistic Regression

The Logistic Regression classifier was applied to the crop prediction dataset to evaluate its performance in classifying crops based on soil nutrients and environmental parameters.

Table 5.1 Performance Evaluation of Logistic Regression Classifier

Model	Accuracy in %	Precision in %	Recall in %	F1-Score in %
Logistic Regression	91	88	87.6	90.2

The Logistic Regression model achieved 91% accuracy, demonstrating good performance in crop prediction. It correctly classified most crop classes with only a few minor misclassification errors. Some classes showed slightly lower precision and recall due to overlapping feature patterns. However, the model is less effective in handling complex non-linear relationships in the dataset.

5.2.2 Support Vector Classifier

The Support Vector Classifier (SVC) was applied to the crop prediction dataset to classify crops based on soil nutrients and environmental parameters. The performance of the model was evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score.

Table 5.2 Performance Evaluation of Support Vector Classifier

Model	Accuracy in %	Precision in %	Recall in %	F1-Score in %
Support Vector Classifier	93	86.2	86.2	91.3

The Support Vector Classifier (SVC) achieved an accuracy of 93% by identifying the optimal parameter combination for classification. The model correctly classified most crop classes, demonstrating strong predictive performance. Only a few classes showed slightly lower precision due to similar soil and environmental characteristics among certain crops. Overall, the SVC model provides reliable and accurate crop prediction.

5.2.3 Gradient Boosting Classifier

The Gradient Boosting Classifier was applied to the crop prediction dataset to classify different crop types based on soil nutrient levels and environmental conditions. The performance of the model was evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score.

Model	Accuracy in %	Precision in %	Recall in %	F1-Score in %
Gradient Boosting Classifier	93	91.4	90.6	93

Table 5.3 Performance Evaluation of Gradient Boosting Classifier

The Gradient Boosting Classifier achieved an accuracy of approximately 93%, indicating strong classification performance. Although the model correctly predicted most crop classes, a few classes exhibited slight prediction errors.

due to similar feature patterns. These minor misclassifications had a limited impact on the overall performance. However, the model is comparatively more complex and less interpretable than simpler classification algorithms.

5.2.4 Random Forest Classifier

The Random Forest Classifier was applied to the crop prediction dataset to classify different crop types based on soil nutrients and environmental parameters. The model performance was evaluated using standard classification metrics including accuracy, precision, recall, and F1-score.

Table 5.4 Performance Evaluation of Random Forest Classifier

Model	Accuracy in %	Precision in %	Recall in %	F1-Score in %
Random Forest Classifier	96	90	91	94

The Random Forest classifier achieved a high accuracy of 96%, demonstrating excellent performance in crop prediction. It constructs multiple decision trees and combines their predictions through an ensemble approach, which improves classification accuracy and reduces overfitting. Using soil nutrients and environmental parameters, the model successfully predicts the most suitable crop with high reliability and robustness.

5.2 Comparison Of Performance Metrics For Various Classifier Algorithm

Table 5.5 Performance Evaluation of various Classifier algorithms

Model	Accuracy in %	Precision in %	Recall in %	F1-Score in %
Logistic Regression	91	88	87.6	90.2
Support Vector Classifier	93	86.2	86.2	91.3
Gradient Boosting Classifier	93	91.4	90.6	93
Random Forest Classifier	96	90	91	94

The performance of four classification algorithms was evaluated using accuracy, precision, recall, and F1-score. Among the evaluated models, the Random Forest classifier achieved the highest performance with 96% accuracy, 90% precision, 91% recall, and an F1-score of 94%, demonstrating its superior classification capability. Both Support Vector Classifier and Gradient Boosting Classifier achieved 93% accuracy, while Logistic Regression obtained 91% accuracy. Based on the experimental results, the Random Forest classifier outperformed the other models, making it the most suitable algorithm for accurate crop prediction.

6. Conclusion

This work presents a machine learning-based crop prediction system that recommends suitable crops using soil nutrients and environmental parameters, including Nitrogen, Phosphorus, Potassium, Temperature, Humidity, pH, and Rainfall. Data pre-processing and feature extraction techniques such as standardization, correlation analysis, and label encoding were performed to improve data quality and prepare the dataset for model training. Four classification algorithms, namely Logistic Regression, Support Vector Classifier, Gradient Boosting, and Random Forest, were evaluated using accuracy, precision, recall, and F1-score. Among these models, the Random Forest classifier achieved the highest accuracy of 96%, demonstrating superior performance in crop prediction. The proposed system provides reliable crop recommendations, helping farmers make informed decisions, improve agricultural productivity, and promote sustainable smart farming practices.

7. Future Work

Although the proposed system achieved high accuracy in crop prediction, there is scope for further improvement. Future work can focus on using larger and more diverse datasets collected from different geographical regions to improve the robustness and generalization of the model. Additional parameters such as soil moisture, sunlight intensity, fertilizer usage, and seasonal variations can be incorporated to enhance prediction accuracy. Furthermore,

advanced machine learning and deep learning algorithms can be explored to better capture complex relationships within agricultural data. The system can also be deployed as a real-time web or mobile application to provide farmers with instant crop recommendations. In addition, integrating crop prediction with yield prediction and fertilizer recommendation can help develop a comprehensive smart agriculture decision-support system.

8. References

- [1] Vansiya, J., Chandi, A., & Khan, R. A. (2025). AI-based intrusion detection & prevention models for smart home IoT systems: A literature review. *Journal of Computer Science and Technology Studies*, 982–996.
- [2] Paul, J. (2025). *Integration of AI and ML in intrusion detection and prevention systems*. ResearchGate.
- [3] Ferreira, R., Bispo, I., Rabadão, C., Santos, L., & Costa, R. L. de C. (2025). Farm-flow dataset: Intrusion detection in smart agriculture based on network flows. *Elsevier*.
- [4] Raghuvanshi, A., Pallathadka, H., Umesh, & Phasinam, K. (2022). Intrusion detection using machine learning for risk mitigation in IoT-enabled smart irrigation in smart farming. *IEEE*.
- [5] Subrahmanyam, K., Pavani, B., Kishore, M., Kumar, M., & Suresh, K. (2025). AI-driven animal intrusion detection system for agriculture: Real-time monitoring and automated response using deep learning. *International Journal of Innovative Research in Technology (IJIRT)*.
- [6] Gandhi, L. J., Armstrong, O., Petkar, P., & Tripathy, A. K. (2016). Rice crop yield prediction in India using support vector machines. In *Proceedings of the 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 1–5).
- [7] Khan, M., & Noor, S. (2019). Irrigation runoff volume prediction using machine learning algorithms. *European International Journal of Science and Technology*, 8*, 1–22.
- [8] National Informatics Centre. (2023). *Annual report 2020–21*. <https://agricoop.nic.in/Documents/annual-report2020-21>
- [9] Government of India. (2023). *Crop production statistics information system*. <https://aps.dac.gov.in/APY/Index.htm>
- [10] Khan, M., & Noor, S. (2019). Performance analysis of regression machine learning algorithms for prediction of runoff time. *Agrotechnology*, 8*(1), 1–12.
- [11] India Brand Equity Foundation. (2023). *Agriculture in India: Industry overview, market size, role in development*. <https://www.india.gov.in/>
- [12] Abdinasir, H. A., Lukman, A., Adeb, S. A., Rasheed, H., Ahmed, S., & Tahir, A. (2024). Security control and data planes of SDN: A comprehensive review of traditional, AI and MTD approaches to security solutions. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2024.3393548>
- [13] Akmalbek, A., Dismurod, K., Rashid, N., Ilkhom, R., & Cho, Y. I. (2024). Optimizing smart home intrusion detection with harmony-enhanced extra trees. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2024.3422999>
- [14] Albasir, A., Naik, K., & Manzano, R. (2022). Towards improving the security of IoT and CPS devices: An AI approach. *Digital Threats: Research and Practice*. <https://doi.org/10.1145/3497862>
- [15] Almaraz-Rivera, J. G., Perez-Diaz, J. A., Cantoral-Ceballos, J. A., Botero, J. F., & Trejo, L. A. (2022). Toward the protection of IoT networks: Introducing the LATAM-DDoS-IoT dataset. *IEEE Access*, 10*, 106909–106920. <https://doi.org/10.1109/ACCESS.2022.3211513>
- [16] Alrayes, F. S., Wafa, M., Aljameel, S. S., Mashael, M., Rizwanullah, M., & Salama, A. S. (2023). Improved radial movement optimization with fuzzy neural network enabled anomaly detection for IoT-assisted smart cities. *IEEE Access*, 11*, 143060–143068. <https://doi.org/10.1109/ACCESS.2023.3342698>