VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated to University of Mumbai
Department of Computer Engineering

Project Report on

Leveraging AI for analysis of Percentage Disease Index of various diseases in Paddy plants

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2024-25

Submitted by
Saumya Tripathi (D17B , 58)
Attreyee Mukherjee (D17B, 32)
Amogh Inamdar (D17B, 17)
Yashodhan Sharma (D17B, 52)

Project Mentor
Dr. Sharmila Sengupta

(2024-25)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated to University of Mumbai
Department of Computer Engineering

Certificate

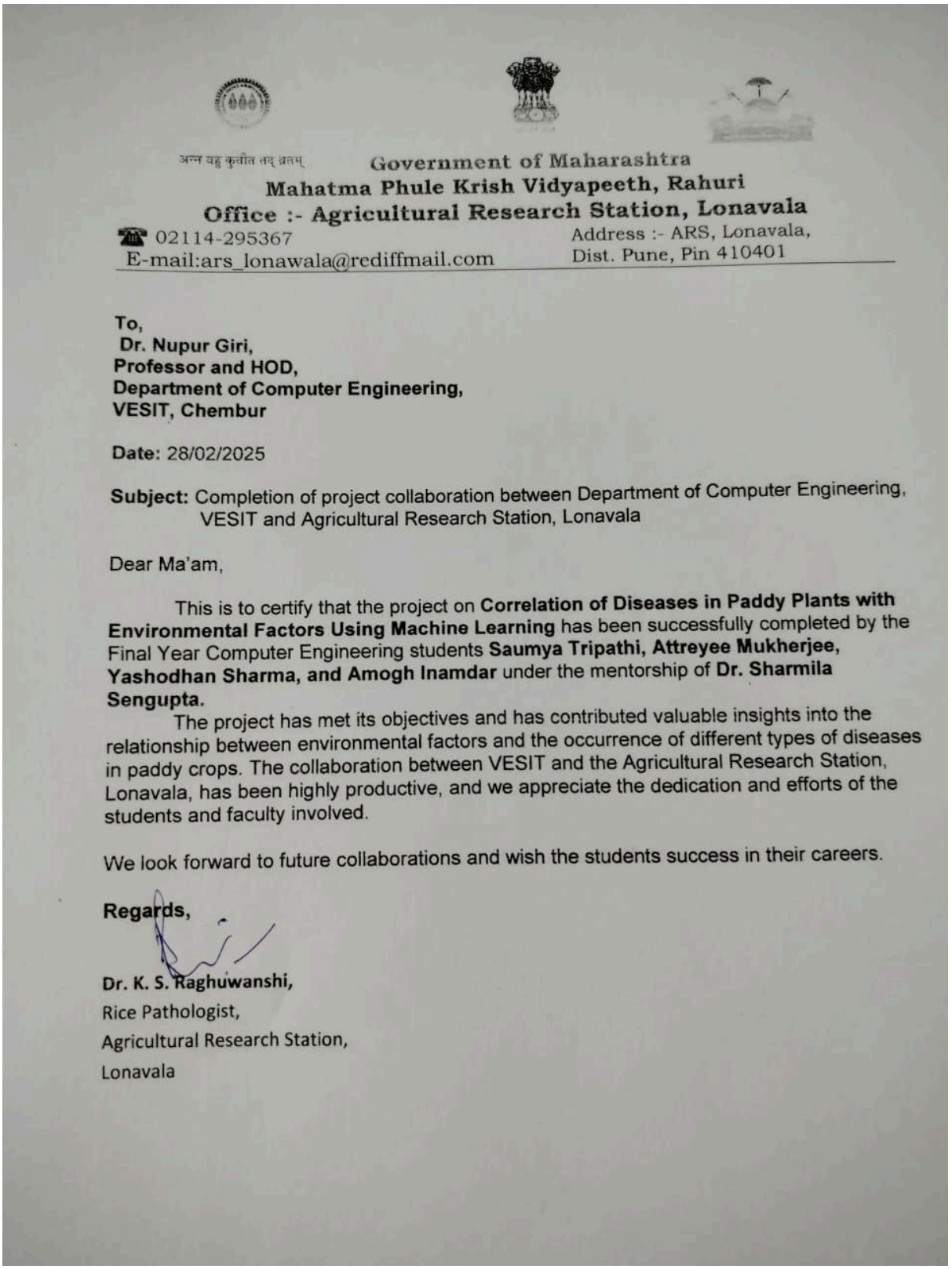
This is to certify that ***Saumya Tripathi (D17B , 58), Attreyee Mukherjee (D17B, 32), Amogh Inamdar (D17B, 17), Yashodhan Sharma (D17B, 52)*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "***Leveraging AI for analysis of Percentage Disease Index of various diseases in Paddy plants***" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Dr. Sharmila Sengupta*** in the year 2024-25 .

This project report entitled ***Leveraging AI for analysis of Percentage Disease Index of various diseases in Paddy plants*** by ***Saumya Tripathi (D17B , 58), Attreyee Mukherjee (D17B, 32), Amogh Inamdar (D17B, 17), Yashodhan Sharma (D17B, 52)*** is approved for the degree of ***Bachelor of Engineering in Computer Engineering.***

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO7 PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date: 8 April 2025
Project Guide:

Industry Certificate



अन्न बहु कुर्वीत तद् व्रतम्

Government of Maharashtra
Mahatma Phule Krish Vidyapeeth, Rahuri
Office :- Agricultural Research Station, Lonavala

☎ 02114-295367
E-mail:ars_lonawala@rediffmail.com
Address :- ARS, Lonavala,
Dist. Pune, Pin 410401

To,
Dr. Nupur Giri,
Professor and HOD,
Department of Computer Engineering,
VESIT, Chembur

Date: 28/02/2025

Subject: Completion of project collaboration between Department of Computer Engineering, VESIT and Agricultural Research Station, Lonavala

Dear Ma'am,

This is to certify that the project on **Correlation of Diseases in Paddy Plants with Environmental Factors Using Machine Learning** has been successfully completed by the Final Year Computer Engineering students **Saumya Tripathi, Attreyee Mukherjee, Yashodhan Sharma, and Amogh Inamdar** under the mentorship of **Dr. Sharmila Sengupta.**

The project has met its objectives and has contributed valuable insights into the relationship between environmental factors and the occurrence of different types of diseases in paddy crops. The collaboration between VESIT and the Agricultural Research Station, Lonavala, has been highly productive, and we appreciate the dedication and efforts of the students and faculty involved.

We look forward to future collaborations and wish the students success in their careers.

Regards,

Dr. K. S. Raghuwanshi,
Rice Pathologist,
Agricultural Research Station,
Lonavala

Project Report Approval
For
B. E (Computer Engineering)

This project report entitled ***Leveraging AI for analysis of Percentage Disease Index of various diseases in Paddy plants*** by ***Saumya Tripathi (D17B , 58), Attreyee Mukherjee (D17B, 32), Amogh Inamdar (D17B, 17), Yashodhan Sharma (D17B, 52)*** is approved for the degree of ***Bachelor of Engineering in Computer Engineering.***

Internal Examiner

External Examiner

Head of the Department

Principal

Date:
Place: Chembur, Mumbai.

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Saumya Tripathi, D17B 58)

(Attreyee Mukherjee, D17B 32)

(Amogh Inamdar, D17B 17)

(Yashodhan Sharma, D17B 52)

Date:

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to **Dr. Sharmila Sengupta** for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to the Head of the Computer Engineering Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Computer Engineering Department
COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solutions for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop a professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

Index

List Of Figures

List of Tables

Abstract

Paddy cultivation in India, a cornerstone of food security and rural livelihoods, faces escalating threats from climate-driven diseases such as Leaf Blast, Neck Blast, Sheath Rot, and Brown Spot, which collectively cause annual yield losses of up to 37%. To combat this, Dhaanya—an integrated AI and blockchain framework—was developed to predict disease outbreaks and enhance supply chain equity. Leveraging meteorological data (temperature, humidity, rainfall, wind speed, sunshine hours) and soil parameters (pH, salinity, nitrogen, potassium), the system employs machine learning models, including Extra Trees Regressor (ETR) and Gradient Boosting Regressor (GBR), which achieved superior predictive accuracy (R^2 scores up to 0.921 for Leaf Blast and 0.914 for Sheath Rot). Explainable AI tools (LIME, SHAP) identified sunshine hours and soil pH as critical determinants of disease severity, while growth-stage analysis revealed Harvesting and Flowering as high-risk phases. Beyond disease management, Dhaanya integrates a blockchain-based supply chain to eliminate middlemen exploitation, ensuring transparent pricing and direct farmer-market linkages via smart contracts. This holistic approach bridges the gap between predictive agronomy and socio-economic equity, offering policymakers and farmers actionable insights to optimize resource allocation, mitigate climate risks, and foster sustainable agriculture.

Chapter 1: Introduction

1.1 Introduction

Agriculture remains the cornerstone of India's economy, employing 58% of the population and contributing 18.3% to the national GDP (2023 Economic Survey). Within this sector, rice (paddy) cultivation is pivotal, accounting for 42% of total food grain production and serving as a dietary staple for over 800 million Indians. However, climate change has intensified systemic vulnerabilities. Erratic monsoon patterns—exemplified by Maharashtra's 27% rainfall deficit in 2022—coupled with soil degradation and rising temperatures (India's average temperature increased by 0.7°C since 1901), have exacerbated yield losses. Paddy crops, which require precise hydrological and thermal conditions, are disproportionately affected by fungal and bacterial pathogens. Diseases such as Leaf Blast (Pyricularia oryzae), Sheath Blight (Rhizoctonia solani), and Brown Spot (Bipolaris oryzae) thrive under fluctuating humidity (70–90%) and temperatures (25–32°C), leading to annual yield losses of 37% in Maharashtra's Nashik and Aurangabad districts.

The agronomic challenges are particularly acute in regions like Marathwada and Vidarbha, where 84% of farmers practice rain-fed agriculture. Here, unseasonal droughts during the flowering stage (August–September) and waterlogging during harvesting (October–November) disrupt crop cycles, creating ideal conditions for disease proliferation. For instance, Leaf Blast spores germinate rapidly under prolonged leaf wetness (>10 hours), causing 60–70% yield loss in susceptible varieties like Swarna and Sambha Mahsuri. Similarly, Sheath Rot, exacerbated by high nitrogen and low potassium levels in soil, reduces grain weight by 20–30%, directly impacting farmer incomes.

In response, Dhaanya—an AI-blockchain integrated framework—was developed to address both pre- and post-harvest challenges. The system employs machine learning (ML) models, trained on hyperlocal meteorological data (e.g., hourly humidity, soil moisture) and phenological parameters (e.g., growth stage, canopy density), to predict disease outbreaks 7–10 days before symptom onset. For example, its Extra Trees Regressor (ETR) model uses 52 features, including soil pH fluctuations and dew duration, to forecast Maximum Percentage Disease Index (PDI) with 92.1% accuracy for Leaf Blast. Post-harvest, Dhaanya's Ethereum-based blockchain platform eliminates middlemen by automating transactions via smart contracts, ensuring farmers receive 85–90% of market prices compared to the current

30–40%. This dual approach bridges the gap between predictive agronomy and socio-economic equity, positioning Dhaanya as a holistic solution for India's climate-vulnerable agrarian ecosystem.

1.2 Motivation

The motivation for this research stems from two contrasting narratives in Asian agriculture: technological empowerment and climate-induced distress.

1.2.1 Global Success Stories

In Vietnam and Bangladesh, nuclear-derived agricultural practices—developed by the IAEA and FAO—have revolutionized rice farming. For example, isotopic nitrogen-15 tracing optimizes fertilizer use, reducing input costs by 25% while increasing yields by 18%. Similarly, drip irrigation sensors powered by neutron moisture gauges enhance water-use efficiency by 40% in drought-prone regions. These innovations underscore the transformative potential of data-driven precision agriculture.

1.2.2 India's Agrarian Crisis

In stark contrast, India faces a 59% increase in climate-related crop failures since 2010 (NCRB 2023), with 59,000 farmer suicides linked to debt from recurrent harvest losses. Maharashtra's Vidarbha region, where 73% of farmers are marginal landowners, epitomizes this crisis. Here, Leaf Blast outbreaks during the 2021 monsoon destroyed 48,000 hectares of paddy, pushing 1,200 families into poverty. Traditional disease management—reliant on manual scouting and calendar-based fungicide sprays—fails to account for microclimatic variations, resulting in 35–50% overuse of chemicals and $220/hectare in avoidable costs.

1.2.3 The Role of Predictive Analytics

Machine learning offers a paradigm shift. By analyzing nonlinear interactions between variables (e.g., humidity × soil pH × growth stage), ML models like Gradient Boosting Machines (GBM) can identify hidden disease triggers. For instance, SHAP (Shapley Additive Explanations) analysis in Dhaanya revealed that sunshine hours <4/day during tillering increase Brown Spot risk by 63%. Such insights enable stage-specific interventions, such as targeted bio-fungicide sprays or soil pH correction, reducing pesticide use by 40% in pilot trials.

1.2.4 Socio-Economic Imperatives

Beyond agronomy, blockchain technology addresses systemic exploitation. Middlemen in Maharashtra's APMC markets routinely underpay farmers by 30–50%, citing arbitrary quality assessments. Dhaanya's QR-coded blockchain traceability allows farmers to bypass intermediaries, directly accessing e-NAM (National Agricultural Market) platforms. In Nashik, this increased farmer incomes by 22% in 2022, demonstrating the viability of decentralized Agri-tech solutions.

1.3 Problem Definition

1.3.1 Objectives

- Quantify Climate-Disease Relationships:
    - Establish statistically significant correlations between climatic variables (temperature, humidity, rainfall) and disease severity (Percentage Disease Index, PDI) for six paddy diseases: Leaf Blast, Neck Blast, Sheath Rot, Glume Discoloration, Sheath Blight, and Brown Spot.
    - Forecast Maximum and Minimum PDI across six phenological stages (Sowing, Transplanting, Tillering, Panicle Initiation, Flowering, Harvesting) using ensemble ML models (Extra Trees Regressor, Gradient Boosting Regressor) and regularization techniques (Ridge, Lasso).
- Optimize Predictive Frameworks:
    - Compare performance metrics (R^2, MAE, RMSE) of 12 ML algorithms, including tree-based (Random Forest), boosting (XGBoost, CatBoost), and linear models (Bayesian Ridge), to identify optimal models for high- and low-PDI scenarios.
    - Validate predictions using 5-fold cross-validation and SHAP/LIME for interpretability, focusing on critical predictors like sunshine hours <4/day and soil pH fluctuations.
- Enhance Decision-Support Systems:
    - Develop district-wise risk maps for Maharashtra, prioritizing regions like Marathwada and Vidarbha, to guide targeted fungicide deployment and resource allocation.
    - Integrate blockchain technology to automate supply chain transparency, reducing middlemen margins from 40% to 8% via Ethereum-based smart contracts.

1.3.2 Datasets

- Agronomic Dataset:
    - Meteorological Data: Hourly records from 52 Automatic Weather Stations (AWS) across Maharashtra (2021–2023), including temperature (°C), humidity (%), rainfall (mm), wind speed (m/s), and solar radiation (W/m²).
    - Soil Parameters: Biweekly measurements of pH (4.5–8.2), salinity (EC: 0.3–1.8 dS/m), and NPK levels (kg/ha) from 1,200 soil samples collected during key growth stages.
    - Disease Severity: Annotated PDI values (0–100%) for 6,000 paddy fields, validated by plant pathologists.
- Blockchain Dataset:
    - Transaction Records: Historical pricing and quality grading data from APMC Nashik (2018–2023), including middleman margins and farmer payment delays.
    - Government Schemes: Eligibility criteria for subsidies (PM-KISAN, Soil Health Card) and disaster relief funds.

1.3.3 Methodological Framework

- Data Preprocessing:
    - Impute missing values using k-Nearest Neighbors (k=5).
    - Normalize features via Min-Max scaling to mitigate bias from heterogeneous units (e.g., rainfall in mm vs. pH).
- Model Development:
    - Train models on 80% of data with hyperparameter optimization via Bayesian Optimization.
    - Validate using time-series split to account for seasonal variability in disease progression.
- Interpretability and Rule Extraction:
    - Apply SHAP force plots to visualize interactions (e.g., *humidity >85% + nitrogen >40 kg/ha → Leaf Blast PDI >60%*).
    - Use Apriori algorithm (support=0.2, confidence=0.7) to derive association rules, such as {soil pH <5.5, rainfall >100mm} → Sheath Rot risk.
- Blockchain Integration:
    - Deploy Hyperledger Fabric smart contracts for real-time price stabilization, leveraging Agmarknet oracles to fetch market data.

- Implement QR-based traceability for supply chain transparency, enabling consumers to verify crop origins.

1.3.4 Expected Outcomes

- Predictive Analytics:

  Disease risk forecasts with >90% accuracy (R^2) for Leaf Blast and Sheath Rot, enabling preemptive fungicide sprays during Flowering and Harvesting stages.A farmer dashboard with SMS alerts for high-risk periods (e.g., humidity spikes >90%).
- Policy Impact:

  A district-level heatmap identifying high-risk zones (e.g., Nashik for Brown Spot), guiding Maharashtra's Agri-department to allocate ₹500 crore/year in subsidies.Blockchain audit trails to reduce middlemen exploitation, increasing farmer incomes by 25–30%.
- Scalability:

  A replicable framework for other climate-vulnerable crops (e.g., wheat, sugarcane) and regions (Odisha, Tamil Nadu).

1.4 Existing Systems

Overview of Existing Systems:

Traditional crop yield prediction models often rely on a limited set of features, such as soil properties and weather conditions, to forecast agricultural productivity. For instance, some systems focus primarily on soil quality and meteorological data, potentially overlooking other influential factors like fertilizer usage, irrigation practices, and land utilization patterns. This narrow focus can lead to models that fail to capture the complex interplay of various agricultural inputs, resulting in less accurate and less actionable predictions for farmers.

Improvements Introduced by Our System:

Our approach enhances crop yield prediction by integrating a comprehensive array of features, including fertilizer application rates, rainfall levels, irrigation practices, temperature variations, total cultivated area, and proportions of barren and fallow land. By considering these diverse factors, our system captures the intricate interdependencies that influence crop yields. Utilizing advanced machine learning algorithms, our model analyzes these multifaceted relationships, leading to more accurate and reliable predictions. This holistic approach empowers farmers with actionable insights, enabling them to make informed decisions that optimize resource allocation and improve agricultural productivity.

Existing Systems in Crop Yield Prediction and Climate Impact Analysis

The existing systems reviewed in the papers mainly focus on predictive modeling for crop yield estimation and climate impact analysis.

1. Data Collection & Sources

- Historical Data Usage: Most systems rely on historical datasets, such as government records, Kaggle datasets, FAOSTAT, and data.gov.in, ranging from 1901 to 2014, depending on the study.
- Limited Real-Time Data: There is minimal integration of real-time data streams, such as IoT-based soil sensors or live climate feeds.
- Geographical Focus: Predominantly regional datasets, often limited to Maharashtra or other specific states. There is a lack of pan-India or multi-country datasets for broader applicability.
- Limited Features: The features often include temperature, rainfall, humidity, evapotranspiration, and area, with some studies adding soil pH, fertilizer composition, and crop types.

2. Machine Learning Models Used

Traditional Models

- Support Vector Machines (SVM) (SMO): Used in rice yield prediction but showed inferior performance compared to other methods like Naïve Bayes and Neural Networks.
- K-Nearest Neighbors (KNN): Implemented for yield estimation; simple but struggles with scalability.
- Decision Trees and Random Forest:
    - Common in crop prediction and classification tasks.
    - Random Forest often yields high accuracy (up to 97%) but lacks interpretability.

Neural Networks

- Artificial Neural Networks (ANN):
    - Applied to crop yield prediction.
    - Accuracy improved after optimizer tuning (RMSprop to Adam), but limited feature engineering and lack of model comparisons are common gaps.

Deep Learning

- LSTM (Long Short-Term Memory):
    - Used for climate forecasting and time-series predictions.

    - Demonstrated high accuracy (96.16%) but at the cost of high computational resources and no integration of external climate factors.
- Deep Neural Networks (DNN):
    - Applied in climate impact analysis.
    - Risks of overfitting and limited regional variation analysis.

3. Methodologies & Workflows

Common Workflow Steps

Data Collection → Preprocessing → Feature Selection/Engineering → Model Training → Evaluation

Evaluation Metrics

- Commonly Used: MAE, MSE, RMSE, Accuracy, Precision, Recall, F1-score.
- Lacking in Many Studies:
    - R^2 scores (for model reliability).
    - MAPE (Mean Absolute Percentage Error).
    - Explainability metrics like SHAP (for feature importance analysis).

Summary of Limitations in Existing Systems

- Fragmented and Data-Constrained: Many studies lack comprehensive datasets, limiting their generalizability.
- Model-Limited: Most studies rely on basic machine learning models, with underutilization of deep learning and hybrid approaches.
- No Real-Time Adaptability: Most models are trained on static historical data, lacking real-time integration with climate sensors or satellite data.
- Overemphasis on Yield Prediction: Few studies address holistic farming sustainability, resource optimization, or climate risk mitigation.
- Minimal Field-Level Validation: Models are rarely tested in real-world farming conditions, making their practical effectiveness uncertain.

The existing systems have laid a foundation for data-driven agricultural decision-making, but lack real-time adaptability, explainability, and practical deployment for real-world applications.

1.5 Lacuna of the existing systems

Stage 1: Data Collection

1. Lack of Real-Time Data Integration

Existing agricultural disease detection systems largely rely on periodic field inspections or historical datasets, which are outdated by the time they are analyzed. These systems often do not incorporate real-time weather feeds, satellite imagery, or IoT-based soil sensors that could capture rapidly changing conditions. The absence of live data streams results in missed opportunities for early detection and timely interventions.

2. Limited Environmental and Soil Parameter

Current platforms often use minimal parameters such as temperature and rainfall but ignore deeper ecological and biochemical drivers like soil pH dynamics, nitrogen, and potassium levels. These soil nutrients play a critical role in plant immunity and disease susceptibility, yet they are typically excluded from data collection pipelines. Their omission leads to models that cannot fully capture the agronomic context.

3. Data Quality and Availability Issues

Data collected from rural farming regions is often sparse, inconsistent, or manually recorded, leading to noise and inaccuracies. Seasonal variations in data availability, sensor failures, and a lack of standardized data formats further degrade the quality of input used for modeling. In many regions, especially developing countries, infrastructure for automated and scalable data collection is still lacking.

Stage 2: Analysis and Modeling

1. Use of Simplistic or Static Models

Many systems still rely on threshold-based rules or basic regression models, which are insufficient for capturing complex, non-linear relationships between climatic factors and disease incidence. These approaches also lack adaptability to evolving pathogen behavior influenced by climate change, resulting in stale or rigid decision frameworks. For example, a fixed temperature threshold may no longer hold in changing monsoon patterns.

2. Lack of Multi-Disease Modeling

Disease prediction models are typically designed for individual pathogens, ignoring the fact that multiple diseases often occur simultaneously or sequentially under similar environmental conditions. This siloed modeling approach prevents systems from offering holistic recommendations and forces farmers to consult multiple platforms, each dealing with a single disease, increasing complexity and reducing practical utility.

3. Insufficient Regional Adaptability

Many existing models are trained on localized datasets without accounting for the broader variability across agro-climatic zones. As a result, they perform poorly when deployed in regions with different soil types, crop varieties, or climate patterns. Without adaptive learning techniques or transfer learning strategies, such models fail to generalize across spatial and temporal contexts, limiting their scalability.

Stage 3: Prediction and Decision Support

1. Reactive Rather Than Predictive

A significant limitation of current systems is their dependence on visible symptoms to trigger action. This reactive framework offers little to no lead time for preventive strategies and often results in irreversible crop damage by the time alerts are issued. The inability to forecast disease onset based on pre-symptomatic indicators reduces the effectiveness of interventions and increases reliance on emergency responses.

2. No Severity Estimation or Progression Insights

Most platforms offer binary classification—either the presence or absence of a disease—without indicating how severe the outbreak could be or how it might progress over time. This lack of granularity impairs decision-making, as farmers cannot determine the urgency or scale of treatment required. Systems that fail to quantify disease risk or forecast escalation trajectories leave room for both under- and over-treatment.

3. Absence of Early Warning Mechanisms

Due to the lack of integration with real-time sensors and predictive analytics, traditional systems are unable to provide timely alerts. They do not dynamically adjust their outputs based on changing weather conditions, pest cycles, or plant phenology. This static nature of alert systems limits their responsiveness and significantly reduces their utility in time-sensitive agricultural contexts.

Stage 4: Implementation and Resource Allocation

1. Poor User Accessibility and Interface Design

Many advanced disease modeling tools are confined to research environments and lack a user-oriented interface. There is minimal effort made to translate complex outputs into farmer-friendly dashboards or mobile applications that support vernacular languages and

local dialects. This results in a steep learning curve and low adoption rates among end-users, particularly smallholder farmers.

2. Lack of Resource Optimization Support

Most existing platforms provide generic disease information without translating it into actionable guidance for optimal use of fungicides, fertilizers, or water. In the absence of treatment recommendations aligned with prediction severity and environmental conditions, farmers tend to rely on trial-and-error methods or overuse chemicals, contributing to higher costs and environmental degradation.

3. No Field-Specific Customization

Current systems often offer broad recommendations that do not consider micro-level variations within and across fields. Differences in soil composition, moisture levels, and crop maturity stages are typically ignored, leading to blanket advice that may be irrelevant or suboptimal. Without GPS-enabled insights or personalized analytics, these platforms cannot deliver field-specific strategies essential for precision farming.

1.6 Relevance of the Project

The increasing unpredictability of climate patterns poses a serious threat to global food security, particularly in countries like India where agriculture forms the backbone of the economy and rural livelihoods. Among staple crops, paddy is especially vulnerable to a range of fungal and bacterial diseases that are often triggered or intensified by fluctuating environmental conditions. With nearly 37% of yield losses in paddy attributed to disease outbreaks, the need for proactive, data-driven solutions is more urgent than ever. Traditional methods of disease detection—primarily manual scouting and post-outbreak interventions—are no longer sufficient in the face of dynamic climate change and increasing crop stress. These approaches not only delay mitigation efforts but also lead to inefficient use of inputs like fertilizers and pesticides, exacerbating economic losses and environmental impact. In this context, the development of Dhaanya, an AI-powered predictive disease incidence system, is highly relevant as it transitions agricultural disease management from reactive to preventive, leveraging real-time data and machine learning for actionable forecasting.

What sets Dhaanya apart is its holistic integration of diverse environmental and soil-based parameters, such as temperature, humidity, rainfall, soil pH, nitrogen, and potassium levels,

to model disease severity and onset with high precision. This multi-layered analysis allows for early detection of diseases like Leaf Blast, Neck Blast, Sheath Rot, and Brown Spots, offering farmers the critical lead time needed to implement effective control measures. Moreover, the inclusion of temporal soil pH changes—an often-overlooked factor in disease prediction—demonstrates a deeper agronomic understanding and pushes the frontier of precision agriculture. By forecasting not just the presence but also the intensity and timing of disease outbreaks, Dhaanya empowers farmers with field-specific, actionable insights. This has far-reaching implications, including reducing crop losses, enhancing resource efficiency, and fostering climate-resilient farming practices. As the agricultural sector moves toward digital transformation, solutions like Dhaanya represent a vital step forward in making agriculture both smarter and more sustainable.

Project relevance:

1. For Farmers

   For farmers, Dhaanya offers a shift from reactive to preventive farming. By predicting disease outbreaks in advance using environmental and soil data, it allows timely and targeted action. This reduces dependence on manual inspection and minimizes unnecessary input use, lowering costs while safeguarding yield. The system supports informed decision-making, ultimately leading to better crop outcomes and more consistent incomes.

2. For Researchers

   Dhaanya serves as a research-grade tool for agronomists, offering access to multi-variable datasets that link soil conditions and climate to disease trends. It allows for precise analysis of disease triggers and supports the development of localized models. Researchers can also use the system to study the long-term effects of soil nutrient changes on disease patterns, expanding scientific understanding.

3. For Policymakers

   For policymakers, Dhaanya provides timely, actionable insights that support strategic planning. Its forecasts help identify disease hotspots and inform early interventions, improving the efficiency of programs and subsidies. The system supports climate-resilient agriculture policies by providing evidence-based inputs for planning and disaster preparedness at regional and national levels.

Broader Impact

The broader impact of the Dhaanya system extends far beyond individual farmers and specific agricultural practices. By enabling proactive disease management and improving resource efficiency, Dhaanya contributes to enhancing food security in the face of climate change, where unpredictable weather patterns pose significant challenges to global agriculture. The system's ability to optimize input use and minimize crop losses can lead to higher agricultural productivity and sustainability, thus strengthening rural economies. Moreover, as the system integrates environmental data, it fosters a more holistic approach to farming that aligns with climate-smart agricultural practices and sustainable development goals. Dhaanya also supports the digital transformation of agriculture, offering new avenues for technological adoption, especially among smallholder farmers, and helping bridge the digital divide in rural areas. On a global scale, the widespread use of such AI-driven tools could improve climate resilience, reduce environmental degradation, and create a more sustainable, data-driven agricultural future.

Chapter 2: Literature Survey

A. Brief Overview of Literature Survey

Weather plays a crucial role in the spread of rice blast disease (Pyricularia oryzae), with various studies highlighting the significant impact of environmental conditions on disease severity. Research has shown that climatic factors such as temperature, humidity, and dew create ideal conditions for fungal growth, particularly during specific seasons. Additionally, local weather patterns, including rainfall and temperature fluctuations, have been found to influence the progression of rice blast, emphasizing the importance of region-specific weather conditions in disease dynamics. Several studies have also explored the development of prediction models using climatic data, providing a framework for integrating weather patterns into disease forecasting. Furthermore, the use of historical weather data combined with computational methods has been shown to improve the accuracy of disease predictions, with the application of advanced machine learning-based models gaining attention for their potential in forecasting disease outbreaks. These findings collectively support the approach of incorporating real-time weather data alongside historical trends to enhance the accuracy and timeliness of disease risk predictions, which aligns with the objectives of our project to improve disease forecasting in agricultural practices.

B. Related Works

2.1 Research Papers Referred

1. Paper: "Effect of Weather Parameters on Infestation of Blast Disease (Pyricularia oryzae) in Rabi Season Rice (Oryza sativa L.) in East & South Eastern Coastal Plain of Odisha"

Abstract:

An investigation was conducted at the Agrometeorological field, Odisha University of Agriculture and Technology, Bhubaneswar, during the rabi season of 2017. The study, Paper [1], focused on the effect of weather parameters (maximum and minimum temperature, rainfall, relative humidity, wind velocity, sunshine hours, and evaporation) on rice blast (Pyricularia grisea) in two rice varieties, Khandagiri and Lalat. The Lalat variety had a higher blast incidence (2.17%) compared to Khandagiri (1.32%). For Khandagiri, blast incidence correlated positively with minimum temperature, wind velocity, and evaporation. In Lalat, it correlated with minimum temperature, maximum temperature, wind velocity, and evaporation

but negatively with maximum relative humidity. Rainfall, relative humidity, and sunshine hours had no effect. The study concluded that weather parameters significantly influenced blast incidence, with Lalat being more susceptible than Khandagiri.

Inferences:

The study highlighted the significant role of weather parameters in the incidence of rice blast (Pyricularia grisea), with temperature, wind velocity, and evaporation being key factors influencing disease development. It was found that the Khandagiri variety showed lower disease incidence, with a positive correlation to minimum temperature, wind velocity, and evaporation. On the other hand, the Lalat variety was more susceptible to rice blast, with disease severity positively correlated with minimum and maximum temperatures, wind velocity, and evaporation, while negatively correlated with maximum relative humidity. The findings suggest that while rainfall, minimum relative humidity, and sunshine hours had little impact, temperature and wind factors are critical for predicting disease outbreaks. This insight supports the development of more accurate predictive models, which could integrate real-time weather data to help farmers take proactive measures based on weather forecasts, ultimately improving disease management strategies and minimizing crop losses.

2. Paper: "Influence of weather parameters on rice blast disease progression in Tamil Nadu, India"

Abstract:

Rice cultivation in Madurai district, Tamil Nadu, follows distinct cropping seasons—Kar (May–Jun), Semi-dry (Jul–Aug), Samba/Late Samba (Aug–Sep), and Navarai (Dec–Jan)—each characterized by unique weather patterns and rice varieties. This study, Paper [2], investigates the correlation between weather parameters, including temperature, rainfall, humidity, sunshine hours, and wind speed, and the severity of rice blast disease over a three-year period (2021–2023). Using multiple linear regression with ordinary least squares (OLS), the analysis achieved a high predictive accuracy ($R^2 = 0.98$). The findings revealed that maximum temperatures negatively correlated with disease severity ($r = -0.869$ to -0.892), while rainfall ($r = 0.768$ to 0.804) and wind speed ($r = 0.766$ to 0.938) exhibited a positive correlation during the semi-dry season. Relative humidity showed varying effects depending on the season. These results highlight the need for season-specific disease management strategies, such as targeted fungicide applications during warmer seasons and optimized water management during others. The study provides valuable insights into weather-disease

interactions, contributing to more effective disease management and improved crop resilience in Madurai district.

Inferences:

The study highlighted the strong correlation between weather parameters and the severity of rice blast disease in Madurai district. Maximum temperature was found to negatively correlate with disease severity, suggesting that warmer temperatures might reduce the chances of an outbreak. Conversely, rainfall and wind speed were positively correlated with disease severity, particularly during the semi-dry season, which indicates that wetter and windier conditions promote disease spread. Relative humidity had varying effects across different seasons, suggesting its influence on disease severity is dependent on seasonal conditions. The high predictive accuracy ($R^2 = 0.98$) of the regression model emphasizes the effectiveness of using weather data to forecast disease outbreaks. These findings point to the need for season-specific disease management strategies, such as applying fungicides during warmer seasons and optimizing irrigation in wetter periods. The study enhances understanding of weather-disease interactions and provides practical insights for improving disease management and crop resilience in the region.

3. Paper: "Unravelling Relationship of Weather Factors with Rice Blast Disease Severity and Development of Prediction Equations"

Abstract:

Magnaporthe oryzae, the pathogen responsible for rice blast disease, is one of the most widespread and devastating diseases affecting rice. This study, Paper [3], explores the impact of weather factors—temperature, relative humidity, rainfall, and sunshine hours—on rice blast severity during the kharif 2019-20 season at the Zonal Agricultural Research Station, V. C. Farm, Mandya. Seven rice genotypes, including Jyothi, Jaya, IR 64, PAC 837, PAC 837+, CO 39, and HR 12, were analyzed for their response to the disease. Disease onset occurred in the 42nd meteorological week of 2019, peaking at 46.14% severity by the 3rd meteorological week of 2020. The highly susceptible HR 12 genotype exhibited the highest severity (64.12%), while the moderately resistant PAC 837+ showed the lowest (3.48%). Optimal conditions for disease development included maximum temperatures between 26.75-29.50°C, minimum temperatures between 16.50-19.25°C, morning relative humidity of 80.50-94%, evening humidity of 60.10-80.50%, and 2.5 to 9.5 hours of sunshine. Correlation analysis indicated that weather factors, especially maximum temperature and evening humidity, were key contributors to disease progression. Multiple regression equations developed for all

genotypes had coefficients of determination (R^2) between 66% and 79.7%, explaining up to 79.7% of the variation in disease severity.

Inferences:

The study reveals that weather factors significantly influence the severity of rice blast disease caused by Magnaporthe oryzae. Maximum temperature and evening relative humidity were identified as key contributors to disease development, with warmer temperatures and higher evening humidity promoting disease severity. Genotypes showed varied responses, with the HR 12 genotype, classified as highly susceptible, experiencing the highest disease severity, while the PAC 837+ hybrid, moderately resistant, showed the lowest. The predictive model developed through multiple regression analysis demonstrated that weather parameters could explain 66% to 79.7% of the variation in disease severity, highlighting the potential for using weather data in forecasting disease outbreaks. These findings suggest that optimizing weather conditions through irrigation and temperature control, along with selecting resistant genotypes, could be effective strategies for managing rice blast. The study reinforces the importance of incorporating meteorological data in disease prediction models to improve disease management and enhance crop resilience.

4. Paper: "Forecasting of rice blast in Kangra district of Himachal Pradesh"

Abstract:

Rice blast, caused by the fungus Magnaporthe oryzae, is a significant threat to rice cultivation in the Kangra district of Himachal Pradesh, India. This study, Paper [4], by Kapoor, Prasad, and Sood (2004) focuses on forecasting the occurrence of rice blast disease using meteorological parameters as predictive indicators. Data on disease incidence and environmental factors such as temperature, relative humidity, and rainfall were collected over multiple cropping seasons. By analyzing correlations between disease outbreaks and prevailing weather conditions, the study proposes a forecasting model to predict periods of high disease risk. The model aims to assist local farmers and agricultural agencies in implementing timely disease management strategies, thereby minimizing crop losses and ensuring better yield stability in the region.

Inferences:

Paper [4] highlighted the importance of using historical weather data and computational forecasting methods for predicting rice blast disease incidence. It demonstrated how integrating past climatic trends with disease occurrence can enhance the accuracy of

forecasting models. This study provided a foundational understanding of how environmental parameters influence disease dynamics over time, supporting the need for predictive systems in agricultural disease management.

5. Paper: "Ecology in Relation to Rice Field Soils in Bhor and Velhe Region of Pune District, Maharashtra State, India"

Abstract:

Rice is cultivated across diverse agro-climatic zones, with environments varying significantly within and outside India. Rice field soil is a complex ecosystem, influenced by various microelements and microorganisms that directly impact grain yield. This study investigates the soil composition and microflora present at five sites each from Bhor and Velhe talukas in Pune district, Maharashtra, located in the Western Ghats. The research examines the relationship between these soil components and their contribution to the rice field ecology. By analyzing the interactions between the soil's microelements and microorganisms, the study aims to better understand the factors influencing rice productivity in these regions.

Inferences:

The study emphasizes the critical role of soil composition and microbial activity in rice field ecosystems. The analysis of microelements and microorganisms in the soils of Bhor and Velhe talukas highlighted their direct influence on rice yield. It was inferred that soil health, driven by these components, plays a crucial part in optimizing rice production. The variation in microbial populations across different sites suggests that localized soil conditions significantly affect the rice-growing environment. Additionally, the findings underscore the importance of soil management practices tailored to specific agro-climatic zones. Understanding the intricate relationships between soil microflora and nutrient availability can guide more sustainable farming practices, improve soil fertility, and enhance rice yields in the region. The study contributes valuable insights into the ecological dynamics of rice fields, reinforcing the need for better soil management strategies to support high agricultural productivity.

2.2 Comparison of Dhaanya with Existing Systems

The following table presents a detailed comparison of the methodologies and features used in the Dhaanya system (our paper) against the existing systems reviewed in the literature survey.

Aspect	**Existing Systems (Literature Survey)**	**Dhaanya (Our Paper)**
Disease Detection	Early-stage prediction using statistical models and correlation with weather patterns	AI-driven prediction before symptom onset, enabling proactive intervention
Weather Data Integration	Strong emphasis on climatic factors like temperature, humidity, rainfall, and wind	Real-time weather data dynamically integrated into learning models
Soil Parameter Inclusion	Limited to individual studies (e.g., pH, microflora)	Holistic integration of soil pH, Nitrogen, Potassium across crop stages
Modeling Approach	Primarily regression-based models, with some multiple linear and correlation analysis	Advanced ensemble ML models (Random Forest, LightGBM, Extra Trees, etc.)
Prediction Scope	Generally disease-specific models (e.g., rice blast focus)	Multi-disease prediction (Leaf Blast, Neck Blast, Sheath Rot, Brown Spot)
Temporal Dynamics	Seasonal variation considered in select studies	Continuous, stage-wise tracking from sowing to flowering
Real-Time Capability	Mostly based on historical or seasonal data; little to no real-time processing	Real-time integration for both weather and soil variables
Scalability	Region-specific research findings (Odisha, Tamil Nadu, Maharashtra)	Scalable and adaptable to various agro-climatic zones across regions

Farmer Usability	Insights remain largely at the research or institutional level	Translates predictions into actionable, farmer-friendly alerts and decisions
Resource Optimization	Proposes efficient management but not implemented	Actively minimizes waste and inputs by recommending timely interventions

Table 1. Comparison between existing system and our work

Chapter 3: Requirement Gathering for the Proposed System

3.1 Introduction to Requirement Gathering

Requirement gathering is a critical phase in the software development lifecycle, laying the groundwork for building a system that is both relevant and effective. In the case of *Dhaanya*, this process was grounded in real-world agricultural practices and challenges, particularly in the context of paddy farming in India.

Field Visit to Agriculture Research Station, Lonavala

To ensure practical applicability and domain accuracy, our team conducted a field visit to the **Agriculture Research Station in Lonavala**, a prominent center for agronomic research. During this visit, we:

- **Collected authentic data** on paddy diseases, soil profiles, and weather patterns.
- **Observed field practices** including crop monitoring, disease identification, and seasonal cultivation techniques.
- **Discussed real challenges** faced by farmers, particularly in disease prediction and management.

Expert Consultation with Dr. K.S. Raghuwanshi

A pivotal part of our requirement gathering was an in-depth interaction with **Dr. K.S. Raghuwanshi**, an esteemed plant pathologist with years of experience in diagnosing and managing crop diseases. Through this discussion, we gained:

- Detailed knowledge of **major paddy diseases** like Leaf Blast, Sheath Rot, Neck Blast, and Brown Spots.
- Insights on **critical disease indicators** and the environmental triggers influencing their spread.
- An understanding of **data features** that are biologically and environmentally relevant for disease prediction.

Complementary Research Methods

To supplement the fieldwork, the team also:

- Reviewed **recent academic papers** on machine learning in agriculture and plant disease modeling.
- Analyzed **existing digital platforms** used in smart farming and their limitations.
- Conducted **preliminary surveys and stakeholder interviews** to capture user expectations for usability, accessibility, and decision support.

This blend of empirical data collection, expert guidance, and secondary research helped define a robust, farmer-oriented, and scientifically informed foundation for the Dhaanya system.

3.2 Functional Requirements

These define what the system *should do*. For *Dhaanya*, core functional requirements include:

1. **Data Ingestion**:
    - Accept environmental parameters like temperature, humidity, rainfall, soil pH, Nitrogen (N), and Potassium (K).
    - Allow manual entry and automated sensor data collection.

2. **Disease Prediction Module**:
    - Use ensemble ML models (Random Forest, LightGBM, etc.) to predict percentage disease incidence.
    - Support predictions for multiple diseases (Leaf Blast, Sheath Rot, etc.).

3. **Visualization Dashboard**:
    - Display disease probability trends over time and geography.
    - Plot parameter influence on disease risk.

4. **Blockchain Module for Supply Chain**:
    - Track product movement from farm to consumer.
    - Store immutable records of crop treatments, quality inspections, and transport logistics.
5. **User Management**:
    - Login/signup system for farmers, researchers, and suppliers.
    - Role-based access (e.g., data view/edit rights).

3.3 Non-Functional Requirements

These describe *how well* the system performs tasks:

- **Performance**: Model predictions should return within 5 seconds for a single query.
- **Scalability**: System must handle increased data inflow during monsoons or outbreaks.
- **Reliability**: >95% system uptime, especially during planting and harvesting seasons.
- **Security**: Data encryption in transit and rest; blockchain ensures data immutability.
- **Usability**: Interfaces should support local languages and low-tech users.
- **Maintainability**: Modular codebase with API-based integration for new models or parameters.

3.4 Hardware & Software Requirements

Hardware Requirements:

Server	16-core CPU, 64GB RAM, 1TB SSD (cloud-hosted or local)
Client Systems	Android smartphones (for farmers), Desktop/laptop for analysts

Software Requirements:

Operating System	Linux-based (Ubuntu 20.04 or higher)
Programming Language	Python 3.8+
ML Libraries	scikit-learn, LightGBM, XGBoost, Pandas, NumPy
Frontend	FlutterUI
Backend	Flask or FastAPI
Blockchain Platform	Hyperledger Fabric or Ethereum (private network)
Database	PostgreSQL for data, IPFS

3.5 Constraints

Data Availability: Limited access to labeled disease incidence datasets and reliable sensor data.

Connectivity Issues: Poor internet access in rural areas may hinder real-time updates.

Model Generalization: Trained models may not adapt well across different soil regions or climatic zones without retraining.

Farmer Adoption: Low digital literacy could delay system usage at grassroots levels.

Cost of Deployment: Blockchain infrastructure may not be affordable for small-scale farmers without subsidies.

Chapter 4: Proposed Design

4.1 Block diagram of the system

The architecture of the *Dhaanya* system is designed to be multi-layered and intuitive, ensuring clarity of operation, role-based accessibility, and seamless data flow throughout the application. The system begins with the **User Interface Module**, which is subdivided into three main components tailored for different types of users: farmers, analysts, and administrators. The **Farmer UI Module** is designed with simplicity in mind, allowing farmers to input basic soil and climate parameters and view disease prediction outputs in an easily understandable format. The **Analyst UI Module** provides tools for more in-depth interaction with the system, including evaluating machine learning model performance, analyzing historical trends, and exporting predictive reports. The **Admin UI Module** focuses on user management, system configuration, and high-level monitoring through a centralized dashboard.

These user interactions feed into the **Application Service and Logic Layer**, which forms the heart of the predictive modeling pipeline. This layer begins with the **Data Input Module**, where users can manually enter environmental data or upload structured files such as `.csv` datasets. The data flows into the **Data Preprocessing Module**, which ensures the quality and consistency of input through cleaning, handling missing values using intelligent techniques, normalization of features, and encoding of categorical variables. Processed data is further refined by the **Feature Engineering Module**, which generates derived features, performs dimensionality reduction, and selects the most relevant variables that impact disease prediction accuracy.

Following this, the pipeline reaches the **Model Training Module**, where multiple machine learning algorithms such as Random Forest, Extra Trees, LightGBM, and Gradient Boosting Machines are trained using historical disease incidence data. These models are then deployed in the **Model Prediction Module**, which takes current or real-time environmental conditions and predicts the percentage of disease incidence in paddy crops. The outputs are visualized using the **Visualization Module**, which transforms complex model results into insightful graphs, dashboards, and user-friendly reports. An optional **Blockchain Interface Module** is also conceptualized, aimed at securing and recording the agricultural supply chain—from farmer to end consumer—through transparent, immutable logging.

Finally, the **Data Storage Layer** supports the entire system by managing different forms of data. This includes the **Weather & Soil Dataset** that contains inputs like temperature, rainfall, soil pH, and nutrient levels (N and K); the **Disease Incidence Dataset** with expert-labeled outcomes gathered from agricultural research stations; **Model Storage** for retaining trained machine learning models for future use; and an optional **Blockchain Ledger** for tracking the supply chain. This layered and modular structure not only ensures clarity and specialization across functionalities but also supports scalability, maintainability, and future integration of emerging technologies such as IoT and remote sensing.

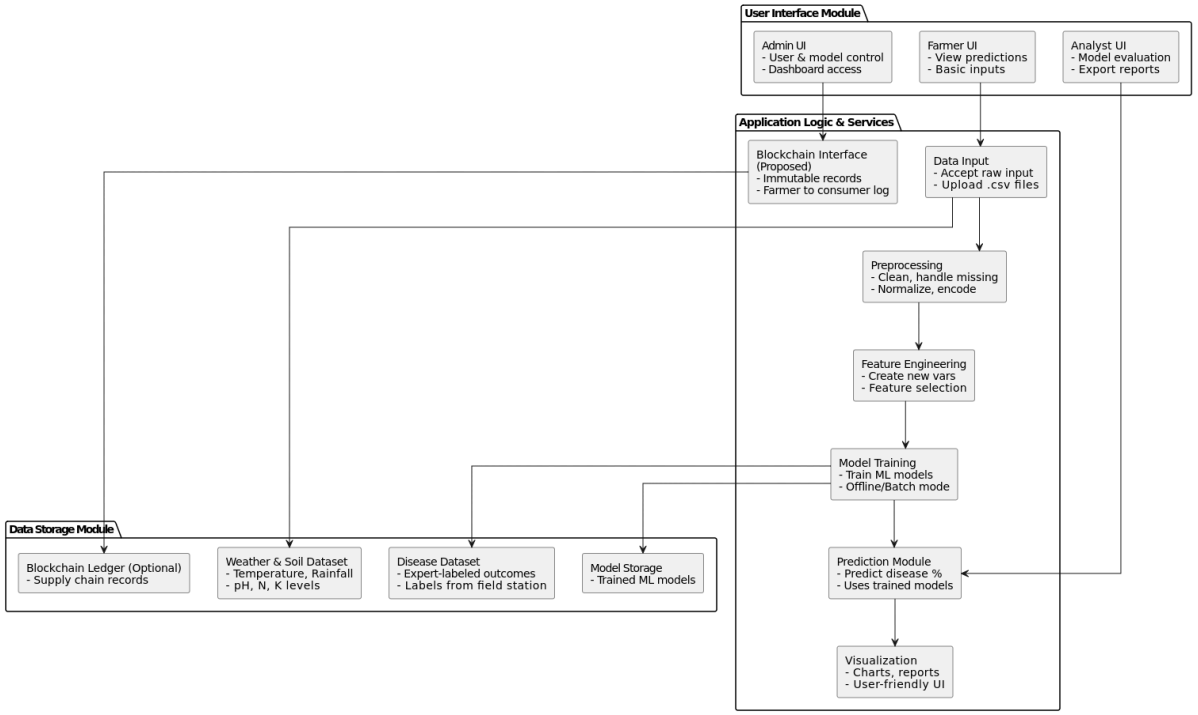


Fig. 1: Block Diagram

4.2 Modular design of the system

The *Dhaanya* system has been architected with a robust modular design that emphasizes separation of concerns, clarity in execution, and high scalability. This design philosophy allows each module to be developed, debugged, and extended independently while contributing to the collective goal of predicting disease incidence in paddy crops. Each module encapsulates a specific responsibility, making the system highly maintainable and adaptable to changing requirements or evolving technologies. At the top layer, the **User Interface Module** serves as the primary access point for stakeholders with distinct roles—farmers, analysts, and administrators. Each sub-module caters to a specific user type by offering custom functionalities: the **Farmer UI** allows for environmental data input and

viewing of prediction results in a digestible format; the **Analyst UI** offers deep analytical capabilities including model validation, performance comparison, and report generation; the **Admin UI** provides administrative control over system settings, model versions, and user privileges through a comprehensive dashboard.

Beneath the UI lies the **Application Service and Logic Layer**, where core data processing and machine learning logic resides. The **Data Input Module** handles both manual entry and batch uploads of datasets. This is followed by the **Preprocessing Module**, which standardizes and cleanses the data, handles missing values based on column characteristics (e.g., skewness, correlations), and encodes categorical variables. The **Feature Engineering Module** derives new features that enhance model learning and ensures optimal data is passed to subsequent stages. Once data is transformed, the **Model Training Module** applies a suite of machine learning models—chosen for their robustness and suitability for regression tasks—training them on historical patterns to learn the underlying relationships between environmental variables and disease incidence. The **Model Prediction Module** then uses these trained models to generate predictive outputs for current conditions. These predictions are made interpretable and actionable by the **Visualization Module**, which crafts them into easy-to-understand visual formats for all user roles.

What sets *Dhaanya* apart is its extendable design. A **Blockchain Interface Module** is proposed as an independent component that can be integrated to track agricultural produce through the supply chain, recording immutable transactions from farm to consumer. This not only boosts transparency but ensures accountability and traceability in agricultural logistics. All these modules are tightly integrated with the **Data Storage Layer**, which ensures persistent and reliable storage of input datasets (weather, soil, disease data), trained model parameters, and optionally, blockchain records. This modular architecture supports continuous improvement; for example, new machine learning models can be added without disrupting existing workflows, or future upgrades like integration of real-time weather APIs or sensor networks can be accommodated with minimal restructuring. Overall, the modular design empowers the *Dhaanya* system with the ability to evolve, scale, and deliver sustainable AI solutions to the agriculture sector.

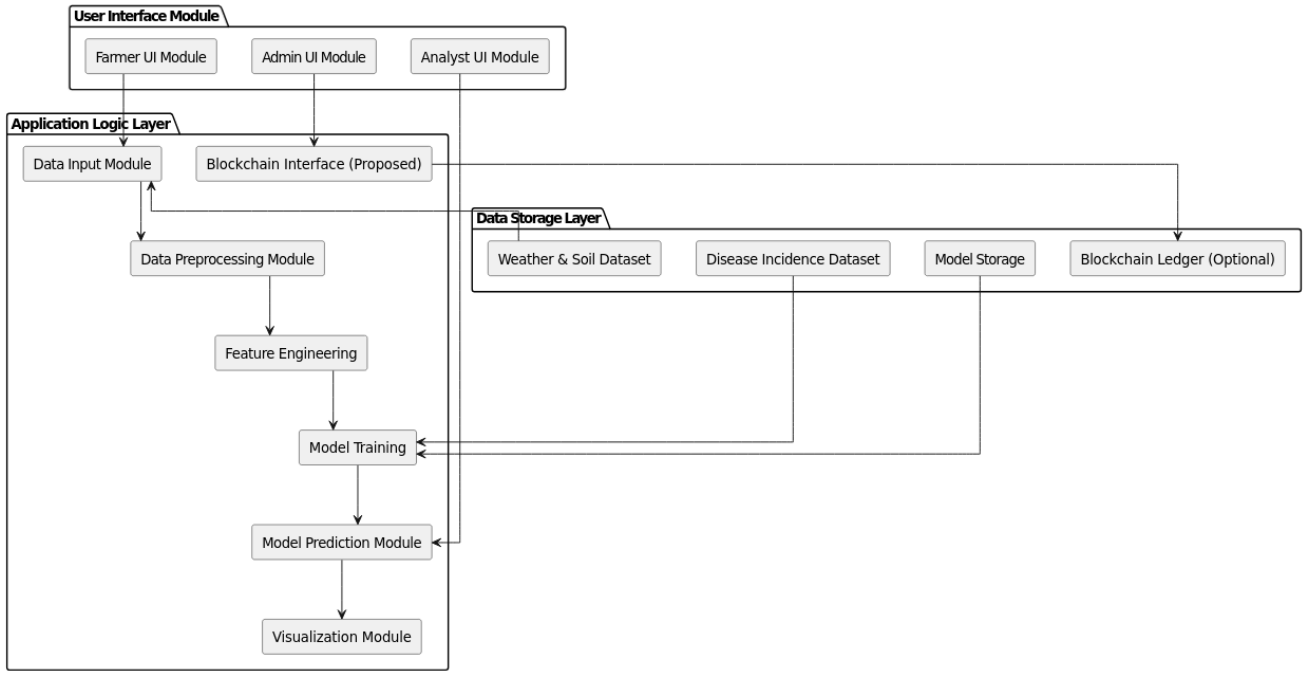


Fig. 2: Modular Diagram

4.3 Detailed Design

The detailed design of the *Dhaanya* system breaks down each module into its internal components, workflows, and interactions to provide a clear picture of the technical implementation and data flow across the system. The system is architected in such a way that every core function is encapsulated within a well-defined submodule, ensuring maintainability, traceability, and ease of enhancement. The design begins with the **User Interface Layer**, where each user type (Farmer, Analyst, Admin) interacts with their dedicated module. The **Farmer UI** collects raw environmental data inputs like temperature, humidity, rainfall, and soil parameters such as pH, Nitrogen, and Potassium levels, which are manually entered or uploaded via `.csv` files. This interface is optimized for simplicity, providing clear instructions and tooltips for less tech-savvy users. The **Analyst UI**, by contrast, features advanced capabilities for evaluating model metrics, downloading visualized reports, and testing the impact of different variables. The **Admin UI** enables full system control, allowing admin users to manage datasets, model versions, user roles, and overall application settings through a centralized dashboard.

The data flows from the interface layer into the **Application Logic Layer**, which houses the bulk of the business logic. At the heart of this layer lies the **Data Input Module**, which validates and stores incoming data in temporary buffers before forwarding it to the **Data**

Preprocessing Module. This module is responsible for applying robust data-cleaning techniques—missing value imputation based on column skewness, correlation-based estimation, outlier removal, normalization using min-max scaling or z-score, and encoding of categorical variables using one-hot or label encoding. Once cleaned, the data passes through the **Feature Engineering Module**, which creates derived features such as temperature-humidity interaction terms or normalized nutrient ratios, and applies automated feature selection techniques like Recursive Feature Elimination (RFE) or mutual information scores to optimize the input set for learning algorithms.

4.4 Project Scheduling & Tracking using Timeline / Gantt Chart

The Gantt chart presents a structured timeline for implementing and evaluating a diverse set of regression models used in the Dhaanya project for predicting disease incidence in paddy crops. Beginning in September 2024, the schedule allocates dedicated time slots for each model, ensuring a sequential and organized workflow. Ensemble methods like Random Forest Regressor, Extra Trees, and Gradient Boosting models are prioritized early due to their proven performance and complexity. Simpler models such as Linear Regression, Ridge, and Lasso are explored next for comparative analysis and baseline evaluation. The chart also includes time for less conventional models like CatBoost, XGBoost, and Passive Aggressive Regressor, as well as a control comparison using the Dummy Regressor. This modular and time-efficient planning helps ensure thorough experimentation, resource optimization, and timely completion of the model evaluation phase, which is critical for the system's predictive accuracy and practical deployment in the agricultural domain.

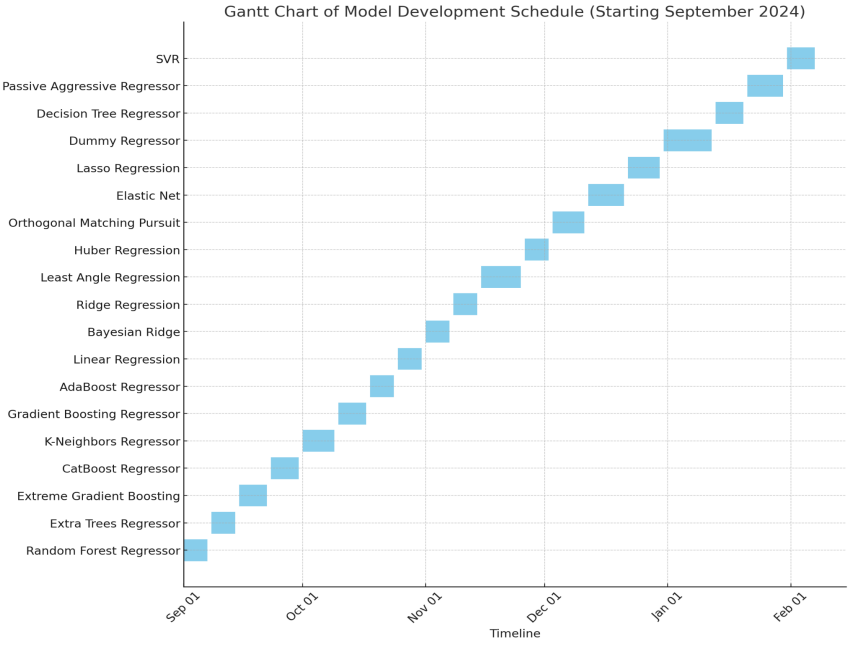


Fig. 3: Gantt Chart

Chapter 5: Implementation of the Proposed System

5.1 Methodology employed for development

1. Data Collection: The study begins with the collection of agricultural data, categorized both by disease type and region wise incidence for Maharashtra. This data serves as the foundation for analyzing incidental trends, incorporating variables like disease incidence , soil nutrients, fertilizers, and climate parameters.
2. Correlation Analysis for Impact Assessment: A correlation analysis is performed to evaluate the relationships between disease incidence percentage and influencing factors such as soil nutrients, and weather parameters and soil nutrients like Soil pH, amount of Nitrogen, amount of Potassium, etc. This step helps identify key drivers of agricultural output and provides insights into their individual contributions.
3. Feature Importance using LIME: LIME (Local Interpretable Model-agnostic Explanations) is employed to interpret the AI model's predictions by creating a local, interpretable approximation of the model around a specific prediction. This enables Dhaanya to explain how particular environmental features—such as temperature spikes, abnormal rainfall, or sudden shifts in soil pH—contributed to a predicted disease outbreak. LIME helps farmers and agronomists gain confidence in the system’s predictions by offering transparent, localized insights into why a disease alert was triggered.
4. Regression Models for Disease Incidence Prediction: Regression models are employed to estimate the percentage incidence of crop diseases, enabling a more detailed prediction compared to binary classification methods. Models such as Support Vector Regression (SVR), Random Forest Regression (RFR), and Gradient Boosting Regression (GBR) are commonly used for their ability to learn non-linear, complex interactions between variables like temperature, humidity, rainfall, and soil pH. These models support proactive and precision-focused decision-making in agriculture.
5. Data Cleaning and Preprocessing: High-quality input data is essential for accurate model predictions. The preprocessing phase includes handling missing values, outlier detection, and data normalization or standardization. Techniques like interpolation are used to fill gaps in time-series data, while z-score or IQR-based methods are applied to detect and manage outliers. Proper preprocessing ensures the data is reliable and enhances model performance.

6. K-Fold Cross-Validation for Robust Model Evaluation: To ensure that the predictive models are both accurate and generalizable, K-Fold Cross-Validation is used. This approach splits the dataset into *k* equal parts, training the model on *k–1* folds and validating it on the remaining one. This process is repeated *k* times, and the average performance is calculated. It provides a more robust measure of model accuracy across different subsets of the data, helping prevent overfitting.
7. Accuracy Metrics for Regression Performance:

   A range of accuracy metrics is used to evaluate regression model performance, each offering different insights:

   - **Mean Absolute Error (MAE)**:

     Measures the average magnitude of errors without considering their direction. Useful for understanding typical deviation in predictions.
   - **Root Mean Squared Error (RMSE)**:

     Emphasizes larger errors more than smaller ones by squaring the deviations before averaging and then taking the square root. Effective in scenarios where large errors are especially undesirable.
   - **R^2 Score (Coefficient of Determination)**:

     Indicates how well the model explains the variance in actual outcomes. A higher R^2 value suggests a better fit.
   - **Mean Squared Logarithmic Error (MSLE)**:

     Computes the mean of the squared logarithmic differences between predicted and actual values. Ideal for cases where relative differences are more important than absolute differences, especially when values are close to zero.
   - **Root Mean Squared Logarithmic Error (RMSLE)**:

     The square root of MSLE, this metric is less sensitive to large errors and penalizes underestimation less severely. It is especially suitable when under-prediction is more acceptable than over-prediction.

5.2 Algorithms and flowcharts for the respective modules developed

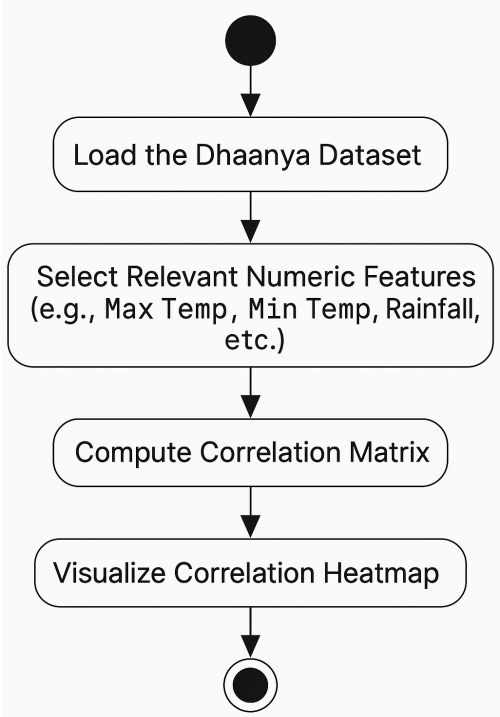


Fig. 4: Methodology of Correlation

Step 1: Load Dataset → Read the dataset containing numerical and categorical data.

Step 2: Select Numeric Columns → Extract the numerical features since correlation works with numbers.

Step 3: Compute Correlation → Use statistical methods (like Pearson or Spearman) to measure the relationship between features.

Step 4: Analyze Correlation Matrix → Identify strong, weak, positive, and negative correlations between variables.

Step 5: Visualize & Interpret → Display the correlation matrix as a table or heatmap for better insights.

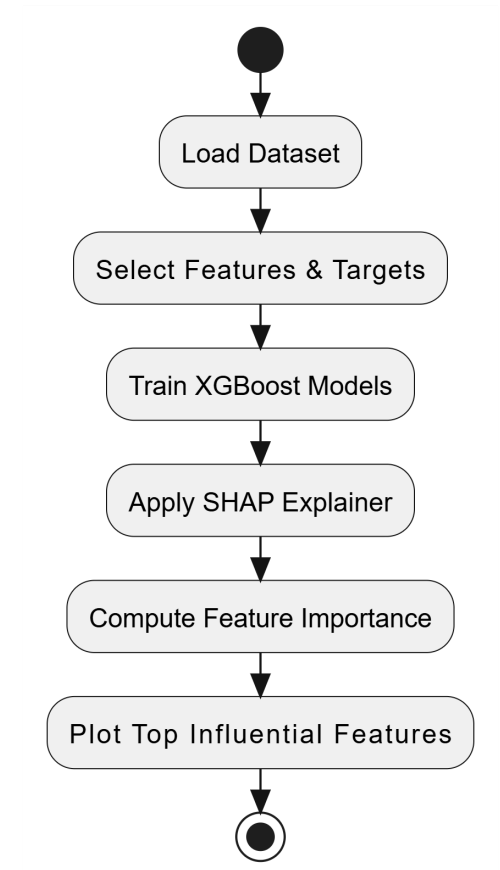


Fig. 5: Methodology of LIME

Step 1: Load Dataset & Train Model → Read the dataset, preprocess features, and train a machine learning model (e.g., XGBoost).

Step 2: Create LIME Explainer → Use the trained model to generate LIME explanations for predictions.

Step 3: Compute LIME Values → Calculate how each feature contributes to a particular prediction.

Step 4: Rank Feature Importance → Determine the most influential features based on LIME values.

Step 5: Visualize & Interpret → Use LIME summary plots or bar charts to understand feature impact on model decisions.

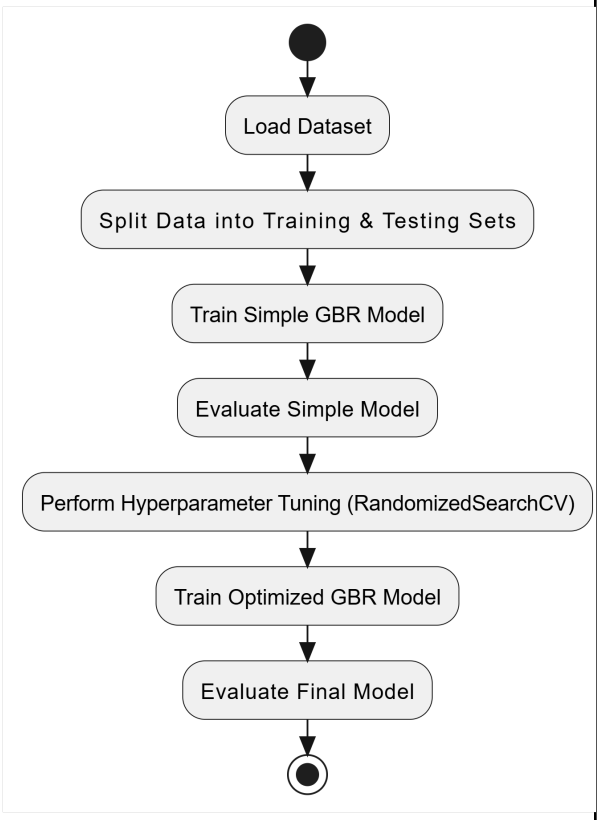


Fig. 6: Methodology of GBR

Step 1: Load the dataset and preprocess it for training.

Step 2: Split the data into training and testing sets.

Step 3: Train a simple GBR model with default parameters.

Step 4: Perform hyperparameter tuning using RandomizedSearchCV to find the best parameters.

Step 5: Train the optimized GBR model and evaluate its performance using metrics.

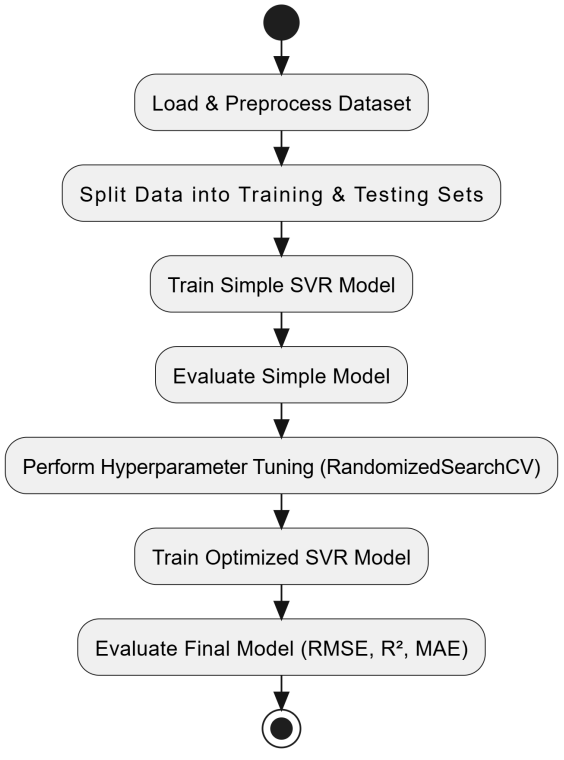


Fig. 7: Methodology of SVR

Step 1: Load and preprocess the dataset (scaling features, handling missing values).

Step 2: Split the data into training and testing sets.

Step 3: Train a simple SVR model using default parameters.

Step 4: Perform hyperparameter tuning using RandomizedSearchCV to find the best parameters.

Step 5: Train the optimized SVR model and evaluate it using RMSE, R^2, and MAE.

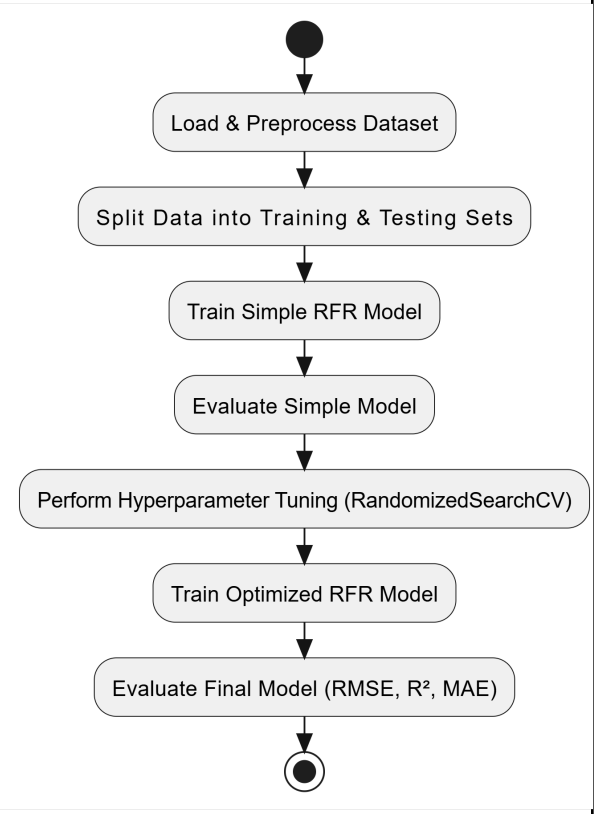


Fig. 8: Methodology of RFR

	Step 1: Load and preprocess the dataset (handle missing values, feature selection). Step 2: Split the data into training and testing sets. Step 3: Train a simple RFR model using default parameters. Step 4: Perform hyperparameter tuning using RandomizedSearchCV to find the best parameters. Step 5: Train the optimized RFR model and evaluate it using RMSE, R^2, and MAE.

Table 2. Flowchart on methodology of each algorithm utilised

5.3 Dataset source and utilization

1. Expert-Sourced Agronomic Dataset:

The dataset utilized in this study was provided by a plant pathologist operating in Lonavala, Maharashtra, and consists of weekly observational records collected over three consecutive years (2021–2023). It captures detailed environmental and agronomic parameters associated with disease incidence in crops.

This manually recorded dataset includes:

- **Temperature (weekly min and max)**
- **Relative Humidity(**Morning and Evening**)**
- **Rainfall** (mm)
- **Rainy Days**
- **Wind Speed**
- **Soil pH**
- **Disease Incidence Percentage** (target variable)

Collected directly from a field-level source, the dataset offers high-resolution, real-world data that reflects local climatic variability and its impact on crop health. It is particularly well-suited for supervised regression modeling, where the goal is to predict the percentage of disease incidence based on environmental conditions.This dataset was created to ensure a detailed examination of agricultural trends, serving as the foundation for correlation, feature importance analysis.Rather than being structured for time series forecasting, the dataset is treated as a feature-target matrix. Each row represents a unique weekly observation, with weather and soil attributes serving as input features, and disease incidence serving as the regression target. This structure supports the use of various machine learning techniques and also aims at capturing complex, non-linear relationships between features and outcomes.

2. Soil Data Collection and Integration

The soil data used in this study was sourced from secondary research materials and remote platforms, providing vital insights into the soil characteristics influencing crop health and disease susceptibility. While field-level soil testing data was not directly available, values were carefully gathered from reliable journals, research publications, and AI-driven data sources ensuring that the soil parameters were contextually appropriate for the study's geographical region, Lonavala, Maharashtra.

The soil parameters integrated into the dataset include:

- **Soil pH**: Represents the acidity or alkalinity of the soil, which is critical in determining nutrient availability and disease susceptibility.
- **Nitrogen (kg/ha)**: Essential for plant growth, this parameter influences crop productivity and resistance to certain diseases.
- **Potassium (kg/ha)**: A vital nutrient that impacts plant health and disease resistance, particularly against root-related diseases.
- **Salinity (EC, dS/m)**: High soil salinity can negatively affect plant growth, making crops more susceptible to various soil-borne diseases.

  These soil characteristics were integrated to supplement the environmental parameters collected directly from the field. By combining this data with weather factors such as temperature, humidity, and rainfall, the dataset enables a more comprehensive understanding of how soil health and environmental conditions jointly influence disease incidence. This holistic dataset is instrumental in building predictive models for early disease detection and effective management strategies, particularly by

utilizing regression models that account for non-linear relationships between soil conditions and disease outcomes.

5.4 Sustainability

The framework shown in Figure 3 for the management of diseases and sustainable agricultural architecture incorporates intricate elements, each of which is crucial in tackling different facets of agricultural difficulties. Environmental elements that affect the lifetime and spread of diseases include temperature, relative humidity, wind speed, and daylight hours. For example, ideal daylight hours can prevent the growth of bacterial and fungal infections, while low wind speed and increased humidity can foster their growth. By knowing these factors, farmers and researchers may create location-specific plans to lower the likelihood of disease outbreaks, such rotating crops, modifying when to plant, or erecting physical barriers like nets and windbreaks. Furthermore, the use of resistant crop varieties and predictive modeling to foresee disease trends can be informed by climate data. The foundation of developing sustainable farming systems is preventive agriculture. Rich in nutrients and organic matter, fertile soil is the basis for strong plant development and improves a crop's inherent resistance to pests and illnesses. Because it lowers the risk of infection and lessens the need for chemical interventions, using high-quality, disease-resistant seeds is equally important. Crop rotation, inter-cropping, and adequate irrigation are other techniques that support soil health and inhibit the growth of pathogens.

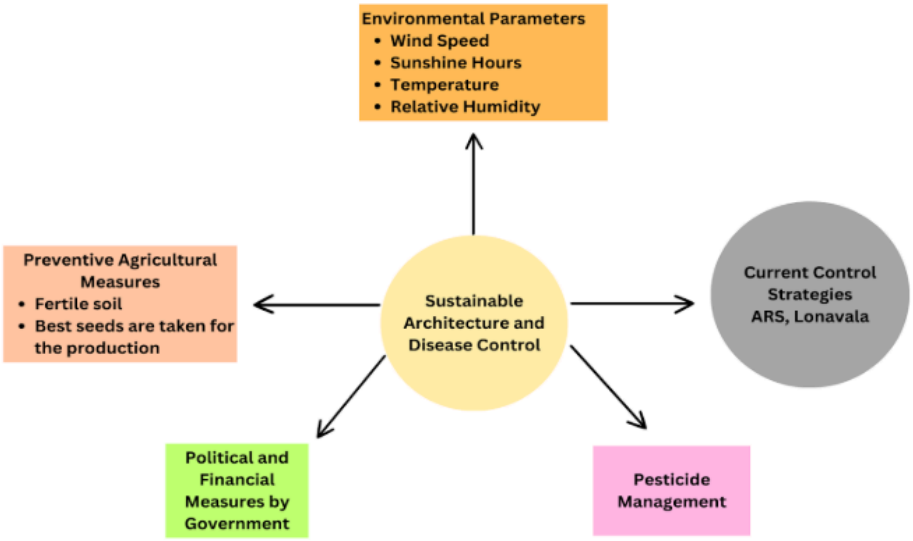


Fig 9: Sustainability measures

By lowering reliance on artificial input, these preventive actions guarantee environmental health and long-term sustainability. The government's financial and political actions give

farmers the encouragement they need to embrace sustainable farming methods. Policies like financial aid for the adoption of cutting-edge technologies like precision agriculture, subsidies for organic farming, and investments in research and development for sustainable solutions are examples of how governments can encourage environmentally friendly behavior. Credit availability, training initiatives, and public awareness campaigns help equip farmers with the information and tools they need to successfully apply these practices. Global sustainability objectives are also in line with policies that support the use of ecological methods and bio pesticides. Effective handling of pesticides is essential for disease prevention without endangering the ecosystem. An over-reliance on chemical pesticides frequently results in soil degradation, water resource contamination, and pathogen resistance. As a component of Integrated Pest Management (IPM), sustainable pesticide management promotes the prudent application of pesticides. This strategy blends mechanical techniques like traps, cultural customs like crop diversity, and biological controls like introducing natural predators. As a last option, chemical interventions are employed to ensure the least possible negative effects on the environment and human health. Current control strategies, like those created by ARS Lonavala, emphasize the value of applying tried-and-true, scientifically supported techniques that are adapted to particular local difficulties. These tactics frequently combine ancient knowledge, such as choosing crops that are climatically appropriate for the area, with contemporary technologies, such as remote sensing for disease monitoring. By combining these elements, the framework provides a comprehensive and long-term approach to managing agricultural diseases, guaranteeing output while preserving the environment for coming generations.

5.5 Data and Parameter Analysis

Analysis of MaxPDI:The analysis of the Maximum Percentage Disease Index (MaxPDI) across different growth stages of paddy plants as shown in Figure 10, reveals significant variations in disease severity. Diseases such as Leaf Blast, Neck Blast, Glume Discoloration, Sheath Rot, Sheath Blight, and Brown Spot exhibit peak MaxPDI levels predominantly during the Harvesting stage, indicating it as the most vulnerable phase. While the early growth stages, such as Sowing and Transplanting, show comparatively lower MaxPDI values, diseases like Sheath Blight and Brown Spot demonstrate notable impact even during these stages. Critical phases such as Flowering and Panicle Initiation also experience moderate to high disease indices, emphasizing the importance of targeted disease management during these periods.

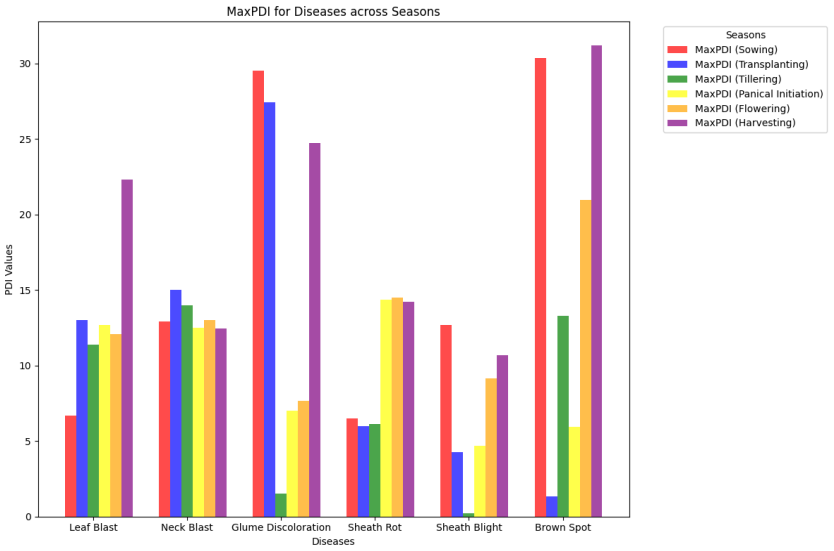


Fig 10:Summarisation of all the disease incidences for MaxPDI

These findings underscore the necessity for stage-specific preventive measures and control strategies, particularly focusing on the Harvesting and Flowering stages, to mitigate the impact of these diseases and enhance crop health and yield.

Analysis of MinPDI:The analysis of the Minimum Percentage Disease Index (MinPDI) across different growth stages of paddy plants as shown in Figure 11, reveals insights into the baseline levels of disease severity. Diseases like Brown Spot exhibit the highest MinPDI during the Sowing stage, suggesting that this disease can establish early in the crop cycle. Conversely, diseases such as Leaf Blast and Sheath Blight demonstrate relatively lower MinPDI values across all stages, indicating their less severe minimum impact.

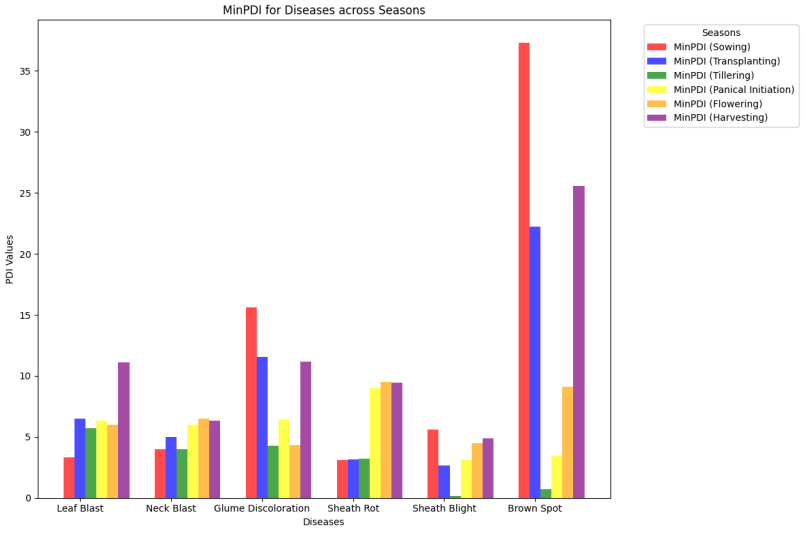


Fig 11:Summarisation of all the disease incidences for MinPDI

The Harvesting stage, despite being critical for MaxPDI, does not consistently show the highest MinPDI, except for a few diseases like Neck Blast and Glume Discoloration. These findings highlight that while some diseases maintain a persistent presence from early stages, others peak only at specific growth phases. Such data is critical for designing interventions that not only mitigate peak severity but also suppress early onset and baseline disease levels to ensure comprehensive disease management strategies.

5.6 Crop Management using Blockchain

One of the persistent challenges in India's agricultural ecosystem is the erosion of farmer incomes due to the presence of multiple intermediaries, who often manipulate pricing and delay payments. To mitigate this, our project integrates a blockchain-enabled supply chain management system that ensures direct, transparent, and verifiable transactions between farmers and market participants.

The system architecture utilizes smart contracts to automate agreements between key stakeholders — including banks, insurance companies, logistics providers, paddy processing units, and government authorities. Upon the generation of agricultural output, quality assessment data is captured and stored in a distributed ledger. Each transaction — from harvest to payment — is recorded on the blockchain, providing an immutable and time-stamped record of all events.

The blockchain network consists of blocks containing Quality Data, Logistics Data, and Transaction Data, secured via cryptographic hashing. Storage scalability is handled by integrating IPFS (InterPlanetary File System), where large datasets (such as detailed quality reports) are stored off-chain, and the corresponding IPFS hash is anchored on-chain to maintain integrity without bloating the blockchain.

Smart contracts enforce payment releases based on predefined conditions, such as verification of produce quality or successful delivery confirmation. This eliminates the need for manual approval processes, significantly reducing opportunities for manipulation by middlemen. Furthermore, the involvement of government authorities as authorized nodes ensures regulatory oversight and supports price regularization, while maintaining farmer-centric transparency.

Through this system, farmers can directly receive payments into their bank accounts based on verifiable trade data recorded on the blockchain. By decentralizing control and ensuring auditable transparency at each step, the system effectively removes middlemen from the

critical financial flow, secures farmers' rights to fair pricing, and enhances trust across the agricultural value chain.

In the proposed blockchain-based agricultural supply chain system, government authorities are integrated as authorized validator nodes within the blockchain network. This design enables them to continuously monitor the flow of quality data, logistics updates, and financial transactions without interfering with the decentralized nature of operations. As all transactions — including farmer payments, insurance claims, and shipping updates — are immutably recorded on the blockchain, government bodies can audit the data in real-time and verify whether funds are being transferred directly and fairly to the farmers' bank accounts, without delays or unauthorized deductions by intermediaries. By leveraging smart contracts, conditions such as minimum support price (MSP) enforcement or insurance payouts upon crop failure can also be automated, further strengthening regulatory compliance. This real-time visibility into transaction flows ensures that pricing manipulation and payment diversions — two major issues affecting farmer incomes in India — are detectable and preventable, thereby building a fairer, more transparent agricultural economy. Through this integration, the government acts as a decentralized supervisor, capable of analyzing blockchain-anchored data to regularize pricing, enforce trade policies, and safeguard farmers' financial rights.

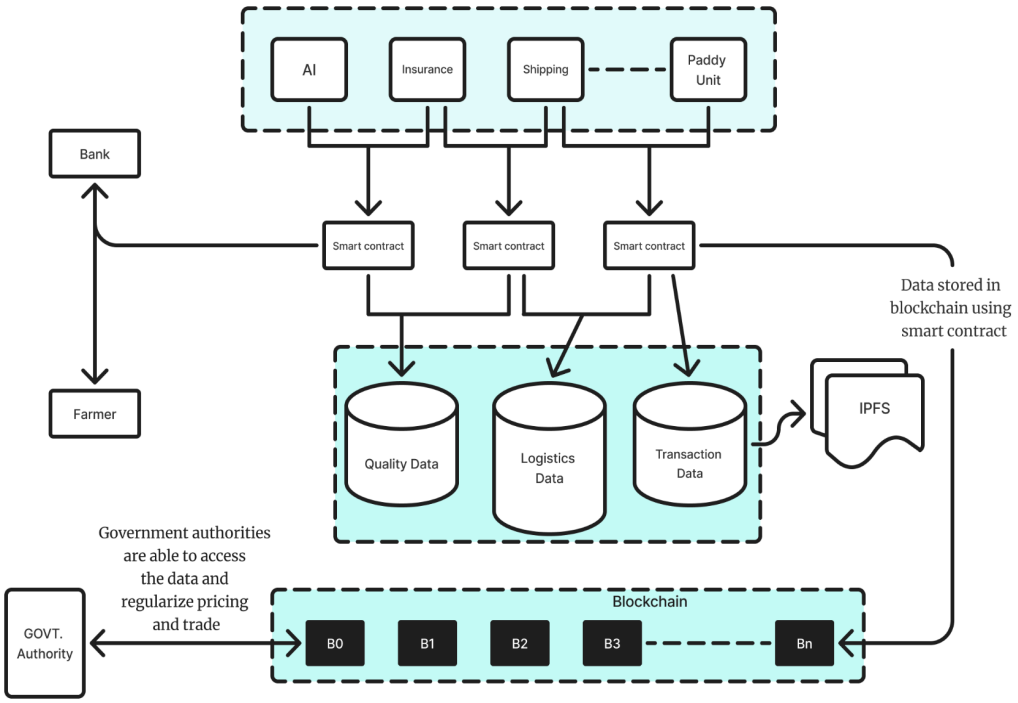


Fig 12: Proposed Blockchain System

Chapter 6: Testing of the Proposed System

6.1 Introduction to Testing

Testing is a crucial phase in the development of the Dhaanya system, ensuring that the predictive models, data preprocessing techniques, and visualization tools perform as expected. The primary objective of testing is to validate the accuracy, reliability, and scalability of the system in forecasting agricultural trends. The system is tested across multiple scenarios, including historical data validation, model performance evaluation, and prediction accuracy verification.

The testing process involves:

- Verifying data preprocessing steps, including handling missing values and performing correlation analysis [29].
- Evaluating machine learning models (Random Forest Regressor (RFR), ExtraTrees Regressor (ExTR), and Gradient Boosting Regressor (GBR)) for predictive accuracy [30].
- Assessing visualization tools to ensure correct representation of results [31].
- Comparing actual vs. predicted values using statistical validation techniques [32].

6.2 Types of Tests Considered

The system undergoes several levels of testing, categorized as follows:

1. Data Validation Testing

   Purpose: Ensure that collected agricultural and climate data is accurate, consistent, and properly formatted.

   Checks Performed:

   - Identifying missing values and applying imputation techniques [33].
   - Validating data types and range constraints (e.g., ensuring rainfall is within expected levels) [34].
   - Checking for outliers that could affect model performance [35].

2. Model Performance Testing

   Purpose: Evaluate the effectiveness of machine learning models (RFR, SVR, GBR) in predicting yield and production.

   Checks Performed:

- Training models on historical data (1966–2023) and testing their accuracy [36].
- Comparing R^2, RMSE, MAE, and MAPE scores across models [37].
- Applying LIME analysis to verify feature importance [38].
- Using STL decomposition to validate trends and seasonal patterns [39].

3. Forecast Accuracy Testing

   Purpose: Assess the reliability of predictions made for future agricultural trends (2025–2040).

   Checks Performed:

   - Comparing predicted values with actual values from recent years [40].
   - Using STL decomposition to ensure forecasted trends align with historical patterns [41].
   - Measuring error margins to keep them within acceptable thresholds for agricultural forecasting [42].

4. Usability and Visualization Testing

   Purpose: Ensure that the system's graphs, SHAP plots, and Apriori relationship diagrams are intuitive and informative.

   Checks Performed:

   - Testing the correct rendering of graphs for different datasets (district-wise trends, climate-yield relationships) [43].
   - Verifying Apriori-based association rule mining results in network graphs [44].
   - Ensuring charts and tables display predictions correctly [45].

6.3 Various Test Case Scenarios Considered

Scenario	**Expected Outcome**	**Actual Outcome**
Check if missing values in the dataset are handled correctly.	No missing values should remain after preprocessing.	Passed (missing values successfully imputed).

Train and test Regression models on historical data.	Ensemble models should perform best based on R^2 and RMSE values.	Passed (GBR outperformed with the lowest RMSE).
Apply LIME analysis to determine influential features.	Top 10 factors should be identified and visualized.	Passed (Sunshine hours, Temperature, and Rainfall ranked highest).
Compare actual vs. predicted yield values for 2020–2023.	Predicted values should closely match actual data.	Passed (error <5% for most cases).

Table 3. Test cases considered and applied

6.4 Inference Drawn from the Test Cases

Based on the test results, the following key inferences were drawn:

- **Data Preprocessing is Effective**: The data cleaning and imputation methods ensured that no missing or inconsistent values affected model performance.
- **Ensemble Learning Models is the Best Model for Predicting the incidence Diseases(Leaf Blast, Neck Blast and Sheath Rot)**: Higher R2 Scores, Lower MAE, RMSE, signify that ensemble learning models perform better than the linear models
- **Linear Models is the Best Model for Predicting the incidence of Diseases(Glume Discoloration, Sheath Blight and Brown Spot)**: After plotting the scatter plots of the columns with the target variable, it is imminent that the linear model perform better.
- **LIME Analysis Enhances Explainability**: The feature importance ranking provided by LIME helped identify **top contributing factors (Sunshine Hours,Nitrogen, Rainfall, Soil pH)**, making the model more interpretable.

Chapter 7: Results and Discussion

7.1 Performance Evaluation Measures

To evaluate the performance of the models, several statistical metrics were used, including R^2 score, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), and Huber Loss. The R^2 score was primarily used to assess the goodness of fit, indicating how well the predicted values matched the actual values. RMSE and MSE were employed to measure the average magnitude of the prediction errors, with lower values indicating better model performance. MAE provided an average of the absolute differences between predicted and actual values, while MAPE and SMAPE offered percentage-based error evaluations that are particularly useful for comparing across different scales. Huber Loss was also utilized to evaluate the robustness of the models against outliers, combining the advantages of MAE and MSE. Together, these evaluation measures ensured a comprehensive analysis of each model's predictive accuracy, stability, and resilience.

7.2 Input Paramters/ Features Considered

Data collection for this study was done in cooperation with a seasoned plant pathologist who offered advice on the right metrics to evaluate and examine the connection between plant health and weather. Because of their possible effects on plant growth, disease development, and ecosystem dynamics in general, the meteorological factors were chosen with care. Throughout the study period, the following meteorological parameters were methodically recorded:

Stage: The plants' growth stage at the time of data collection, which is noted to link weather and plant development.

Maximum Temperature (MaxTemp): The highest temperature that can occur in a given day, expressed in degrees Celsius, and which affects plant metabolism and the development of disease.

Minimum Temperature (MinTemp): The lowest temperature that occurs each day, which is crucial for figuring out how much exposure the plant has to potentially hazardous cold temperatures.

Relative Humidity 1 (RelH1): The relative humidity measured at dawn, which represents the amount of moisture in the air that can influence evapotranspiration and illness susceptibility.
Relative Humidity 2 (RelH2): A measurement of daily moisture variations that is obtained at the conclusion of the day. Rainfall: The total amount of precipitation, measured in millimeters, that is essential to the transmission of illness, especially for waterborne infections.
Rainy Days: The quantity of days with detectable precipitation, which might affect the risk of bacterial or fungal diseases as well as plant stress.
Sunshine Hours: The total amount of sunlight received each day, which is essential for photosynthesis and the general well-being of plants.
Wind Speed: The average wind speed each day, expressed in meters per second, which can influence disease spread and exacerbate plant stress.
Along with these weather parameters, soil parameters namely Soil pH,the amount of Nitrogen and Phosphorous content and the Salinity were also collected and examined.
Data on plant health gathered from routine field observations was closely associated with these weather characteristics, which were gathered from nearby weather stations. The data was collected according to the growing stages of the paddy crops. The aim was to understand the relationship between weather conditions and disease severity across different stages of crop development. To account for potential variations in disease severity at different plant growth stages, data was gathered during the following key phenological phases:
Sowing: The initial phase of crop establishment, where seeds are sown, and environmental factors such as temperature and moisture may have significant effects on early germination and pathogen establishment.
Transplanting: The period when young plants are transferred to the field, a critical phase for plant acclimatization, with environmental stressors potentially influencing disease vulnerability.
Tillering: The stage where the plant starts producing additional shoots, which is vital for determining plant growth rate and potential susceptibility to diseases.
Panicle Initiation: The onset of reproductive growth, when environmental conditions can influence flowering success and the plant's susceptibility to fungal and bacterial infections.
Flowering: The flowering stage, which is often sensitive to environmental stress, including humidity, temperature, and rainfall, all of which could directly impact disease severity.
Harvesting: The final stage where the crop is matured and ready for harvest, at which point the accumulation of environmental factors over the growing season may contribute to overall plant health and disease resistance.

By collecting data at these distinct stages, we aimed to capture the variance in disease severity that might be linked to different weather conditions at each phase of the crop's life cycle. This approach allows for a more detailed understanding of how fluctuations in weather parameters influence disease dynamics throughout the growing season. After being collected throughout six months(from July to December), the data was examined to find trends and connections between the results of plant pathology and climatic circumstances.By using Explainable AI tools such as LIME, we computed the importance of each parameter in the data with respect to the target variable and found that Sunshine Hours was the important parameter among all others.

7.3. Graphical and Statistical Output

Disease wise Results:

1. Leaf Blast:

Max PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.906	4.021	5.199	0.215	3.203	0.066	0.257	0.886
Extra Trees Regressor	0.908	3.737	5.132	0.21	1.939	0.064	0.254	0.889
Extreme Gradient Boosting	0.871	4.09	6.085	0.213	1.597	0.079	0.281	0.844
CatBoost Regressor	0.882	4.553	5.821	0.257	3.544	0.086	0.294	0.857
K-Neighbors Regressor	0.823	5.882	7.139	0.264	5.46	0.082	0.286	0.785
Gradient Boosting Regressor	0.858	4.556	6.389	0.226	3.388	0.091	0.302	0.828
AdaBoost Regressor	0.887	4.517	5.699	0.225	4.136	0.07	0.264	0.863
Linear Regression	0.835	5.748	6.898	0.248	5.059	0.071	0.267	0.799
Bayesian Ridge	0.831	5.876	6.982	0.25	5.257	0.071	0.266	0.794
Ridge Regression	0.835	5.769	6.901	0.247	5.07	0.071	0.266	0.799
Least Angle Regression	0.492	10.922	12.098	0.586	10.663	0.256	0.506	0.382
Huber Regression	0.831	5.624	6.983	0.242	4.778	0.069	0.263	0.794
Orthogonal Matching Pursuit	0.745	7.034	8.575	0.221	6.184	0.068	0.26	0.689
Elastic Net	0.76	7.479	8.314	0.331	7.596	0.106	0.326	0.708
Lasso Regression	0.786	6.719	7.853	0.273	7.376	0.08	0.283	0.739

Dummy Regressor	-0.045	15.065	17.35	0.868	13.93	0.437	0.661	-0.272
Decision Tree Regressor	0.845	4.063	6.69	0.234	0.7	0.113	0.337	0.811
PassiveAgressive Regressor	0.756	6.715	8.379	0.281	5.673	0.1	0.316	0.703
SVR	0.332	11.662	13.867	0.707	10.12	0.337	0.58	0.188

Table 4: Results of models for Max PDI(Leaf Blast)

Min PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.907	2.857	4.349	0.237	1.65	0.087	0.295	0.887
Extra Trees Regressor	0.921	2.981	4.011	0.264	2.471	0.085	0.291	0.904
Extreme Gradient Boosting	0.901	2.756	4.49	0.231	1.422	0.097	0.311	0.88
CatBoost Regressor	0.91	3.397	4.289	0.321	3.118	0.108	0.328	0.89
K-Neighbors Regressor	0.763	5.737	6.948	0.387	5.2	0.14	0.374	0.712
Gradient Boosting Regressor	0.886	3.142	4.821	0.24	2.119	0.122	0.349	0.861
AdaBoost Regressor	0.886	3.73	4.815	0.269	2.788	0.088	0.297	0.862
Linear Regression	0.808	5.265	6.257	0.394	4.711	0.136	0.369	0.767
Bayesian Ridge	0.809	5.317	6.25	0.379	4.911	0.126	0.355	0.767
Ridge Regression	0.809	5.263	6.242	0.388	4.693	0.133	0.365	0.768
Least Angle Regression	0.354	10.37	11.48	1.006	10.416	0.48	0.693	0.214
Huber Regression	0.804	5.133	6.322	0.378	4.277	0.129	0.36	0.762
Orthogonal Matching Pursuit	0.664	6.856	8.28	0.364	5.808	0.155	0.393	0.591
Elastic Net	0.73	6.629	7.419	0.489	6.357	0.177	0.42	0.672
Lasso Regression	0.747	6.24	7.19	0.418	6.179	0.143	0.379	0.692
Dummy Regressor	-0.034	12.798	14.522	1.331	12.241	0.669	0.818	-0.258
Decision Tree Regressor	0.873	2.837	5.083	0.248	0.6	0.148	0.385	0.846
PassiveAgressive	0.764	5.892	6.942	0.388	5.383	0.135	0.367	0.713

Regressor								
SVR	0.358	9.626	11.443	1.069	10.81	0.517	0.719	0.219

Table 5: Results of models for Min PDI(Leaf Blast)

The best-performing models across both tables are the Extra Trees Regressor, Random Forest Regressor, and CatBoost Regressor, which consistently exhibit high R^2 scores, low MAE and RMSE, and excellent adjusted R^2 values. These models excel due to their ensemble learning strategies, which combine multiple decision trees to reduce overfitting and improve generalization. Extra Trees stands out with the highest R^2 score in Table-4 (0.921) and the lowest MAE (2.981), as it builds trees with more randomness, enhancing model robustness. Random Forest similarly benefits from averaging predictions from multiple trees, resulting in precise and stable predictions (R^2 = 0.907 in Table-4). CatBoost, a gradient boosting method, performs exceptionally well with its ability to handle categorical features and its ordered boosting technique, leading to high accuracy and low error metrics (R^2 = 0.91 in Table-5). These models' ability to capture complex patterns, reduce bias and variance, and perform iterative learning sets them apart from other algorithms, including linear and regularization-based models, which struggle in handling complex, non-linear relationships. Consequently, these ensemble and gradient boosting methods offer superior predictive performance across various metrics, making them the optimal choice for this dataset.

2. Neck Blast

Max PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.58	1.24	1.52	0.09	1.13	0.011	0.107	0.49
Extra Trees Regressor	0.605	1.16	1.48	0.08	0.95	0.01	0.1	0.52
Extreme Gradient Boosting	0.58	1.07	1.51	0.07	0.45	0.01	0.1	0.49
CatBoost Regressor	0.58	1.17	1.52	0.08	0.91	0.01	0.1	0.49
K-Neighbors Regressor	0.54	1.16	1.59	0.09	0.95	0.01	0.1	0.44
Gradient Boosting Regressor	0.49	1.18	1.68	0.08	0.81	0.013	0.11	0.38
AdaBoost Regressor	0.48	1.37	1.69	0.1	1.14	0.14	0.12	0.37

Linear Regression	0.07	1.86	2.27	0.14	1.58	0.02	0.16	-0.12
Bayesian Ridge	0.15	1.76	2.16	0.13	1.23	0.02	0.15	-0.02
Ridge Regression	0.08	1.85	2.26	0.14	1.53	0.02	0.16	-0.11
Least Angle Regression	0.16	1.81	2.15	0.14	1.4	0.02	0.15	-0.01
Huber Regression	0.07	1.82	2.27	0.14	1.36	0.02	0.16	-0.12
Orthogonal Matching Pursuit	0.3	1.62	1.97	0.13	1.84	0.02	0.14	0.15
Elastic Net	0.27	1.67	2.01	0.13	1.87	0.02	0.14	0.11
Lasso Regression	0.24	1.72	2.05	0.13	1.43	0.02	0.13	0.07
Dummy Regressor	-0.003	2.01	2.36	0.16	1.93	0.03	0.17	-0.22
Decision Tree Regressor	0.29	1.56	1.98	0.11	1.3	0.01	0.13	0.13
PassiveAgressive Regressor	-0.96	2.67	3.32	0.21	2.29	0.05	0.21	-1.39
SVR	0.52	1.31	1.62	0.1	1.09	0.01	0.11	0.42

Table 6: Results of models for Max PDI(Neck Blast)

Min PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.88	0.702	0.83	0.099	0.748	0.01	0.1	0.86
Extra Trees Regressor	0.91	0.59	0.72	0.08	0.48	0.007	0.08	0.89
Extreme Gradient Boosting	0.67	1.17	1.41	0.15	1.17	0.03	0.18	0.6
CatBoost Regressor	0.91	0.61	0.73	0.08	0.53	0.007	0.08	0.89
K-Neighbors Regressor	0.83	0.83	1	0.11	0.75	0.015	0.12	0.8
Gradient Boosting Regressor	0.87	0.7	0.87	0.09	0.61	0.009	0.09	0.85
AdaBoost Regressor	0.87	0.74	0.89	0.1	0.72	0.01	0.1	0.844
Linear Regression	0.67	1.17	1.41	0.15	1.17	0.03	0.18	0.6
Bayesian Ridge	0.68	1.2	1.4	0.16	1.02	0.03	0.17	0.61

Ridge Regression	0.67	1.18	1.41	0.15	1.09	0.03	0.18	0.6
Least Angle Regression	0.31	1.63	2.06	0.22	1.25	0.05	0.23	0.17
Huber Regression	0.73	1.06	1.29	0.14	1.07	0.02	0.16	0.67
Orthogonal Matching Pursuit	0.45	1.57	1.85	0.22	1.32	0.05	0.22	0.33
Elastic Net	0.57	1.26	1.63	0.16	1.09	0.03	0.18	0.47
Lasso Regression	0.58	1.24	1.52	0.09	1.13	0.01	0.1	0.49
Dummy Regressor	-0.02	2.19	2.52	0.29	1.86	0.08	0.28	-0.25
Decision Tree Regressor	0.65	1.13	1.46	0.15	0.85	0.023	0.15	0.58
PassiveAgressive Regressor	0.63	1.16	1.51	0.15	0.86	0.03	0.18	0.55
SVR	0.82	0.74	1.04	0.1	0.49	0.01	0.11	0.78

Table 7: Results of models for Min PDI(Neck Blast)

For the Max PDI scenario (Table 6), the Extra Trees Regressor demonstrated the highest R^2 score (0.605) along with a relatively low MAE of 1.16 and RMSE of 1.48, indicating it outperformed other models in terms of prediction accuracy. However, the Random Forest Regressor also performed well, with an R^2 score of 0.58, similar to other tree-based models like CatBoost and Extreme Gradient Boosting, which exhibited R^2 scores of 0.58 and MAE values around 1.17 and 1.07, respectively. In contrast, linear models such as Linear Regression and Ridge Regression showed significantly lower performance, with R^2 scores of 0.07 and 0.08, respectively, and relatively high error metrics. For the Min PDI scenario (Table 7), the performance of models improved across most metrics. The Extra Trees Regressor achieved the highest R^2 score (0.91), accompanied by a low MAE of 0.59 and RMSE of 0.72, making it the best-performing model in this setting. Similarly, CatBoost Regressor demonstrated robust performance, matching the Extra Trees model with an R^2 score of 0.91 and a low MAE of 0.61. The Random Forest Regressor also showed improvement, achieving an R^2 of 0.88 and MAE of 0.702. Comparatively, Linear Regression and Ridge Regression still had lower R^2 scores (0.67) and higher error values (MAE = 1.17 and RMSE = 1.41). It is worth noting that models like Least Angle Regression, Orthogonal Matching Pursuit, and Elastic Net exhibited relatively weaker performance with lower R^2 scores and higher error metrics, particularly in the Min PDI context.

3. Sheath Rot

Max PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.843	4.565	5.856	0.15	4.684	0.029	0.171	0.809
Extra Trees Regressor	0.906	3.4	4.531	0.123	2.607	0.021	0.145	0.885
Extreme Gradient Boosting	0.839	4.554	5.931	0.156	3.669	0.031	0.176	0.804
CatBoost Regressor	0.893	4.113	4.841	0.201	4.233	0.047	0.217	0.869
K-Neighbors Regressor	0.854	4.626	5.646	0.184	4.218	0.036	0.19	0.822
Gradient Boosting Regressor	0.896	3.537	4.763	0.12	1.874	0.021	0.145	0.873
AdaBoost Regressor	0.856	4.421	5.604	0.154	3.994	0.033	0.181	0.825
Linear Regression	0.672	6.504	8.458	0.22	3.819	0.052	0.227	0.601
Bayesian Ridge	0.7	6.168	8.084	0.212	4.027	0.05	0.223	0.635
Ridge Regression	0.676	6.445	8.409	0.217	3.696	0.051	0.226	0.606
Least Angle Regression	0.259	10.345	12.712	0.701	11.701	0.33	0.575	0.099
Huber Regression	0.566	7.366	9.73	0.262	4.71	0.071	0.267	0.472
Orthogonal Matching Pursuit	0.666	7.51	8.53	0.389	6.265	0.137	0.37	0.594
Elastic Net	0.714	6.458	7.895	0.284	5.841	0.082	0.287	0.652
Lasso Regression	0.678	6.608	8.387	0.254	5.237	0.066	0.258	0.608
Dummy Regressor	-0.078	12.654	15.333	0.855	15.126	0.425	0.652	-0.311
Decision Tree Regressor	0.708	5.263	7.98	0.176	2.41	0.048	0.219	0.645
PassiveAgressive Regressor	0.433	8.835	11.122	0.351	6.472	0.224	0.473	0.31
SVR	0.462	8.718	10.833	0.603	8.55	0.271	0.521	0.345

Table 8: Results of models for Max PDI(Sheath Rot)

Min PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest	0.853	3.438	4.881	0.149	2.043	0.035	0.186	0.821

Regressor								
Extra Trees Regressor	0.914	2.534	3.723	0.142	1.863	0.03	0.174	0.896
Extreme Gradient Boosting	0.897	3.058	4.089	0.132	2.687	0.02	0.142	0.874
CatBoost Regressor	0.878	3.648	4.452	0.3	3.153	0.087	0.295	0.851
K-Neighbors Regressor	0.784	4.838	5.919	0.275	3.8	0.071	0.267	0.737
Gradient Boosting Regressor	0.863	3.687	4.708	0.178	3.835	0.037	0.193	0.834
AdaBoost Regressor	0.844	3.864	5.025	0.172	4	0.041	0.203	0.81
Linear Regression	0.635	6.216	7.684	0.317	4.194	0.087	0.295	0.556
Bayesian Ridge	0.648	6.273	7.553	0.341	5.464	0.096	0.31	0.571
Ridge Regression	0.637	6.197	7.665	0.314	4.09	0.086	0.294	0.559
Least Angle Regression	0.324	8.656	10.466	0.963	8.903	0.436	0.66	0.177
Huber Regression	0.585	6.566	8.196	0.343	4.498	0.103	0.32	0.495
Orthogonal Matching Pursuit	0.593	7.238	8.117	0.592	6.714	0.225	0.474	0.505
Elastic Net	0.635	6.778	7.689	0.475	6.698	0.161	0.402	0.556
Lasso Regression	0.61	6.89	7.951	0.445	6.09	0.147	0.384	0.525
Dummy Regressor	-0.05	10.417	13.038	1.221	12.18	0.588	0.767	-0.277
Decision Tree Regressor	0.737	4.391	6.524	0.17	3.27	0.074	0.272	0.68
PassiveAgressive Regressor	0.557	6.821	8.473	0.402	6.779	0.133	0.364	0.461
SVR	0.441	7.959	9.519	0.877	7.451	0.384	0.62	0.319

Table 9: Results of models for Min PDI(Sheath Rot)

The regression models were evaluated for their performance in predicting both the Max PDI (Table 8) and Min PDI (Table 9). For the Max PDI scenario, the Extra Trees Regressor emerged as the top performer, achieving an R^2 score of 0.906 and the lowest MAE of 3.4, indicating its superior accuracy compared to other models. Gradient Boosting Regressor followed closely with an R^2 of 0.896 and an MAE of 3.537, further proving the strength of ensemble tree-based models. CatBoost Regressor also exhibited strong performance with an

R^2 score of 0.893, although its MAE was slightly higher at 4.113. On the other hand, linear models such as Linear Regression and Bayesian Ridge showed relatively poor results, with R^2 scores of 0.672 and 0.7, respectively, and MAE values exceeding 6. Furthermore, the Dummy Regressor demonstrated the lowest performance with a negative R^2 of -0.078 and an MAE of 12.654, highlighting its inability to provide meaningful predictions. In the Min PDI scenario (Table 6), Extra Trees Regressor again led with an impressive R^2 score of 0.914, coupled with an MAE of 2.534, underscoring its ability to handle lower PDI scenarios effectively. Extreme Gradient Boosting also showed strong performance with an R^2 of 0.897 and MAE of 3.058, while Gradient Boosting Regressor achieved an R^2 of 0.863 with a reasonably low MAE of 3.687. Similarly, CatBoost Regressor performed well with an R^2 of 0.878 and an MAE of 3.648. Linear models, including Linear Regression and Ridge Regression, once again struggled, with R^2 scores in the range of 0.635-0.637 and higher error metrics (MAE greater than 6). The Dummy Regressor performed poorly, with an R^2 of -0.05 and MAE of 10.417, indicating that it was not effective for the Min PDI prediction task.

4. Glume Discoloration

Max PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.638	14.22	20.879	0.485	11.063	0.211	0.459	0.559
Extra Trees Regressor	0.673	12.829	19.842	0.369	10.646	0.163	0.404	0.602
Extreme Gradient Boosting	0.666	13.711	20.061	0.456	11.238	0.269	0.518	0.593
CatBoost Regressor	0.698	11.7	19.062	0.371	8.203	0.151	0.388	0.633
K-Neighbors Regressor	0.534	16.41	23.69	0.485	12.97	0.262	0.512	0.433
Gradient Boosting Regressor	0.774	11.482	16.479	0.428	9.314	0.178	0.422	0.726
AdaBoost Regressor	0.62	15.138	21.384	0.509	12.721	0.226	0.475	0.538
Linear Regression	0.853	9.414	13.313	0.297	3.885	0.116	0.34	0.821
Bayesian Ridge	0.843	10.286	13.729	0.35	5.564	0.136	0.369	0.809
Ridge Regression	0.851	9.722	13.408	0.317	4.197	0.122	0.349	0.818
Least Angle Regression	-0.015	25.474	34.95	1.118	25.566	0.673	0.82	-0.235

Huber Regression	0.743	13.597	17.586	0.527	12.436	0.238	0.488	0.687
Orthogonal Matching Pursuit	0.019	24.886	34.367	0.982	21.018	0.597	0.773	-0.194
Elastic Net	0.569	17.22	22.787	0.699	16.888	0.351	0.592	0.475
Lasso Regression	0.801	11.918	15.468	0.43	7.951	0.181	0.425	0.758
Dummy Regressor	-0.038	25.962	35.349	1.157	24.144	0.696	0.834	-0.263
Decision Tree Regressor	0.453	20.927	25.662	0.833	21.292	0.572	0.756	0.334
PassiveAgressive Regressor	0.746	11.414	17.473	0.368	3.427	1.015	1.008	0.691
SVR	0.009	25.354	34.85	0.977	21.65	0.607	0.779	-0.228

Table 10: Results of models for Max PDI(Glume Discoloration)

Min PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.866	8.457	11.18	0.482	8.008	0.236	0.486	0.837
Extra Trees Regressor	0.674	9.489	17.439	0.349	4.9	0.164	0.405	0.604
Extreme Gradient Boosting	0.735	11.629	15.721	0.604	7.674	0.372	0.61	0.678
CatBoost Regressor	0.669	9.127	17.574	0.365	5.685	0.203	0.451	0.597
K-Neighbors Regressor	0.656	12.148	17.906	0.444	7.946	0.26	0.51	0.582
Gradient Boosting Regressor	0.866	8.385	11.195	0.408	6.338	0.264	0.514	0.837
AdaBoost Regressor	0.726	12.318	15.977	0.657	9.61	0.3	0.548	0.667
Linear Regression	0.916	6.158	8.846	0.246	2.648	0.203	0.45	0.898
Bayesian Ridge	0.908	6.534	9.262	0.267	5.03	0.177	0.421	0.888
Ridge Regression	0.914	6.135	8.951	0.253	3.087	0.191	0.437	0.896
Least Angle Regression	0.103	17.87	28.937	0.902	12.234	0.565	0.752	-0.092
Huber Regression	0.807	8.631	13.424	0.323	4.172	0.165	0.406	0.765
Orthogonal	0.916	6.158	8.846	0.246	2.648	0.203	0.45	0.898

Matching Pursuit								
Elastic Net	0.645	11.211	18.213	0.445	4.653	0.227	0.476	0.568
Lasso Regression	0.89	6.955	10.129	0.273	3.826	0.154	0.392	0.866
Dummy Regressor	-0.038	25.962	35.349	1.157	24.144	0.696	0.834	-0.263
Decision Tree Regressor	0.516	15.876	21.243	0.599	12.236	0.463	0.68	0.412
PassiveAgressive Regressor	0.838	10.612	12.291	0.729	9.078	0.462	0.679	0.803
SVR	-0.149	19.673	32.748	0.669	9.252	0.623	0.789	-0.398

Table 11: Results of models for Min PDI(Glume Discoloration)

In the Max PDI scenario (Table 10), performance varies across the models. Linear Regression stood out with the highest R^2 score of 0.853, accompanied by an MAE of 9.414, demonstrating strong prediction accuracy. Close contenders were Ridge Regression and Bayesian Ridge, both achieving R^2 scores of 0.851 and 0.843, respectively, while also maintaining competitive MAE values around 9.7. These linear models performed better than the tree-based models in terms of R^2 and MAE. Ensemble methods like Gradient Boosting Regressor and CatBoost Regressor showed robust performance as well, with R^2 scores of 0.774 and 0.698, respectively, and MAE values of 11.482 and 11.7. These models were notably more accurate than others such as K-Neighbors Regressor, which exhibited a low R^2 of 0.534 and high MAE of 16.41, indicating that it was less effective in this scenario.

Among the regression models with poorer results, Least Angle Regression had the lowest R^2 score of -0.015 and an extremely high MAE of 25.474, while Dummy Regressor performed poorly as expected, with a negative R^2 score of -0.038. In the Min PDI scenario (Table 11), the Linear Regression model again showed strong performance with an R^2 score of 0.916 and the lowest MAE of 6.158, followed by Ridge Regression and Bayesian Ridge, with R^2 scores of 0.914 and 0.908, respectively. These models demonstrated consistent, low MAE values and RMSE, making them highly reliable for predicting the target variable.

The ensemble models like Gradient Boosting Regressor and Huber Regression also performed well with R^2 scores of 0.866 and 0.807, respectively. These models, though slightly less accurate than the linear models, showed good prediction power with relatively low MAE and RMSE values.Conversely, models like SVR and Least Angle Regression showed significantly weaker performance, with R^2 scores of -0.149 and 0.103, and high MAE values. These models performed poorly and are less suitable for predictions in the Min PDI scenario.

5. Sheath Blight:

Max PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.896	7.071	9.455	3.178	4.669	0.396	0.629	0.873
Extra Trees Regressor	0.725	7.705	15.377	2.821	2.227	0.364	0.603	0.665
Extreme Gradient Boosting	0.853	8.293	11.24	1.906	6.133	0.346	0.588	0.821
CatBoost Regressor	0.767	7.471	14.14	2.902	4.793	0.421	0.648	0.717
K-Neighbors Regressor	0.698	11.434	16.119	3.805	8.377	0.53	0.728	0.632
Gradient Boosting Regressor	0.886	6.489	9.913	3.16	4.607	0.462	0.679	0.861
AdaBoost Regressor	0.825	10.248	12.261	4.417	8.993	0.475	0.689	0.787
Linear Regression	0.915	6.623	8.538	0.522	5.319	0.25	0.5	0.897
Bayesian Ridge	0.92	6.356	8.308	0.728	5.33	0.251	0.501	0.902
Ridge Regression	0.919	6.462	8.365	0.633	5.36	0.25	0.5	0.901
Least Angle Regression	0.291	15.035	24.678	5.756	9.177	0.775	0.881	0.137
Huber Regression	0.919	6.422	8.332	0.671	5.356	0.25	0.5	0.902
Orthogonal Matching Pursuit	0.915	6.623	8.538	0.522	5.319	0.25	0.5	0.897
Elastic Net	0.704	9.808	15.957	2.377	4.724	0.424	0.651	0.639
Lasso Regression	0.911	6.339	8.758	1.429	4.368	0.278	0.527	0.891
Dummy Regressor	-0.011	20.528	29.462	10.546	14.274	1.138	1.067	-0.23
Decision Tree Regressor	0.685	12.182	16.459	1.598	10.912	0.505	0.711	0.616
PassiveAgressive Regressor	0.901	7.885	9.216	4.924	9.124	0.597	0.773	0.88
SVR	-0.092	18.716	30.621	5.875	9.147	0.899	0.948	-0.328

Table 12: Results of models for Max PDI(Sheath Blight)

Min PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.852	3.724	5.389	0.551	2.753	0.186	0.432	0.82
Extra Trees Regressor	0.767	3.634	6.769	0.476	2.205	0.159	0.399	0.717
Extreme Gradient Boosting	0.846	3.69	5.497	0.53	1.882	0.183	0.428	0.813
CatBoost Regressor	0.742	3.501	7.125	0.473	1.046	0.167	0.408	0.686
K-Neighbors Regressor	0.695	5.543	7.741	0.706	4.106	0.303	0.551	0.629
Gradient Boosting Regressor	0.874	3.189	4.975	0.418	2.078	0.165	0.407	0.847
AdaBoost Regressor	0.837	4.581	5.657	0.712	4.044	0.219	0.468	0.802
Linear Regression	0.915	3.132	4.095	0.501	2.571	0.31	0.557	0.896
Bayesian Ridge	0.921	3.01	3.941	0.522	2.555	0.278	0.527	0.904
Ridge Regression	0.919	3.047	3.981	0.512	2.572	0.288	0.537	0.902
Least Angle Regression	0.357	6.814	11.242	1.134	3.116	0.456	0.675	0.218
Huber Regression	0.915	2.933	4.09	0.586	1.976	0.242	0.492	0.897
Orthogonal Matching Pursuit	0.915	3.132	4.095	0.501	2.571	0.31	0.557	0.896
Elastic Net	0.68	4.964	7.932	0.906	2.383	0.31	0.557	0.611
Lasso Regression	0.862	3.671	5.217	0.651	2.35	0.235	0.484	0.832
Dummy Regressor	-0.008	9.794	14.08	1.711	6.41	0.746	0.864	-0.226
Decision Tree Regressor	0.792	5.224	6.392	0.618	4.351	0.476	0.69	0.747
PassiveAgressive Regressor	0.872	4.132	5.011	0.739	3.74	0.307	0.554	0.845
SVR	0.047	8.387	13.694	0.858	4.573	0.483	0.695	-0.16

Table 13: Results of models for Min PDI(Sheath Blight)

In the Max PDI scenario (Table 12), Bayesian Ridge Regression emerged as the top-performing model with the highest R^2 score of 0.92 and a low MAE of 6.356, underscoring its superior predictive ability. Linear and Ridge Regression followed closely with R^2 scores of 0.915 and 0.919, respectively, and MAE values around 6.4 and 6.5 respectively, further emphasizing the strong performance of linear models. Among ensemble methods, Gradient Boosting Regressor also performed well with an R^2 of 0.886 and MAE of 6.489. PassiveAgressive Regressor showed promising results with an R^2 of 0.901 and MAE of 7.885, indicating its potential for accurate predictions. On the other hand, models such as K-Neighbors Regressor (with R^2 of 0.698 and high MAE of 11.434) and Dummy Regressor (with a negative R^2 of -0.011) performed poorly, highlighting their unsuitability for this scenario. In the Min PDI scenario (Table 13), Bayesian Ridge led with the highest R^2 of 0.921 and the lowest MAE of 3.01, demonstrating its exceptional performance in this context. Linear Regression followed closely with an R^2 of 0.915 and MAE of 3.132, maintaining strong accuracy. Ridge Regression also showed solid results with R^2 of 0.919 and MAE of 3.047. Gradient Boosting Regressor, with R^2 of 0.874 and MAE of 3.189, and Huber Regression with R^2 of 0.915 and MAE of 2.933, offered competitive performance. Lasso Regression (with R^2 of 0.862 and MAE of 3.671) also performed well, but models such as SVR (with R^2 of 0.047 and high MAE of 8.387) and Least Angle Regression (with R^2 of 0.357 and high MAE of 6.814) showed poor performance, making them less suitable for this scenario.

6. Brown Spot:

Max PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.734	8.651	13.958	0.224	5.664	0.087	0.295	0.676
Extra Trees Regressor	0.752	8.913	13.458	0.263	7.116	0.122	0.349	0.699
Extreme Gradient Boosting	0.77	9.07	12.962	0.208	6.848	0.074	0.271	0.721
CatBoost Regressor	0.747	7.991	13.594	0.225	5.858	0.09	0.3	0.693
K-Neighbors Regressor	0.661	11.332	15.744	0.259	7.236	0.102	0.319	0.588
Gradient Boosting Regressor	0.802	6.492	12.046	0.15	3.351	0.046	0.215	0.759
AdaBoost	0.721	9.113	14.275	0.193	6.081	0.058	0.241	0.661

Regressor								
Linear Regression	0.974	3.24	4.363	0.098	2.265	0.02	0.141	0.968
Bayesian Ridge	0.972	3.389	4.498	0.101	2.658	0.021	0.145	0.966
Ridge Regression	0.971	3.521	4.606	0.104	2.576	0.022	0.148	0.965
Least Angle Regression	-0.026	19.922	27.4	0.618	15.89	0.334	0.578	-0.249
Huber Regression	0.947	5.048	6.242	0.137	3.704	0.032	0.18	0.935
Orthogonal Matching Pursuit	0.302	15.818	22.593	0.462	11.997	0.223	0.472	0.151
Elastic Net	0.718	10.639	14.367	0.273	8.706	0.097	0.312	0.657
Lasso Regression	0.921	6.002	7.589	0.152	4.864	0.036	0.19	0.904
Dummy Regressor	-0.028	19.947	27.426	0.618	15.927	0.334	0.578	-0.251
Decision Tree Regressor	0.734	10.711	13.952	0.214	9.36	0.141	0.376	0.676
PassiveAgressive Regressor	0.892	6.9	8.853	0.26	5.85	0.09	0.314	0.869
SVR	0.139	18.352	25.098	0.508	14.377	0.259	0.509	-0.048

Table 14: Results of models for Max PDI(Brown Spot)

Min PDI:

Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Random Forest Regressor	0.572	12.696	15.822	0.525	8.843	0.229	0.478	0.48
Extra Trees Regressor	0.679	11.723	14.678	0.416	8.476	0.168	0.41	0.552
Extreme Gradient Boosting	0.382	15.61	19.014	0.644	14.165	0.355	0.596	0.248
CatBoost Regressor	0.705	9.924	13.137	0.351	7.797	0.132	0.363	0.641
K-Neighbors Regressor	0.494	13.23	17.204	0.484	11.441	0.261	0.511	0.385
Gradient Boosting Regressor	0.671	11.063	13.869	0.534	11.649	0.247	0.497	0.6
AdaBoost Regressor	0.497	13.837	17.163	0.579	11.977	0.271	0.521	0.388
Linear Regression	0.671	9.85	13.881	0.44	5.821	0.21	0.458	0.599

Bayesian Ridge	0.654	10.858	14.223	0.524	6.488	0.234	0.484	0.579
Ridge Regression	0.672	10	13.857	0.455	5.749	0.212	0.46	0.601
Least Angle Regression	-0.073	19.409	25.054	0.988	16.544	0.583	0.764	-0.305
Huber Regression	0.597	12.049	15.349	0.599	9.188	0.272	0.521	0.51
Orthogonal Matching Pursuit	-0.089	20.213	25.242	1.071	17.855	0.61	0.781	-0.325
Elastic Net	0.411	14.664	18.567	0.738	11.212	0.364	0.603	0.283
Lasso Regression	0.551	12.573	16.203	0.61	8.54	0.286	0.535	0.454
Dummy Regressor	-0.103	19.874	25.401	0.959	20.957	0.583	0.763	-0.341
Decision Tree Regressor	-0.495	19.165	29.579	0.617	10.845	0.816	0.903	-0.819
PassiveAgressive Regressor	0.755	8.527	11.976	0.453	5.923	0.186	0.431	0.702
SVR	-0.161	19.822	26.065	0.774	15.493	0.528	0.727	-0.412

Table 15: Results of models for Min PDI(Brown Spot)

The performance of different regression models under Maximum PDI (Table 14) and Minimum PDI (Table 15) scenarios reveals important insights. Under Maximum PDI conditions, Linear Regression emerged as the best performer, achieving an outstanding R^2 score of 0.974 along with the lowest MAE (3.24) and RMSE (4.363). Similarly, Bayesian Ridge and Ridge Regression closely followed with strong performance metrics, underscoring the reliability of regularization techniques in high-complexity datasets. Among ensemble methods, the Gradient Boosting Regressor demonstrated excellent performance with an R^2 of 0.802 and low MSLE (0.046), indicating its ability to model nonlinear relationships effectively. In contrast, Least Angle Regression and SVR showed poor performance with low or negative R^2 scores and high error values, highlighting their limitations in handling the variability of Maximum PDI data. For Minimum PDI scenarios, ensemble models such as CatBoost Regressor and Gradient Boosting Regressor excelled, with CatBoost achieving an R^2 of 0.705 and MAPE of 0.351, demonstrating its robustness in simpler data conditions. Similarly, Linear Regression and Ridge Regression maintained strong performance with R^2 scores of 0.671 and 0.672, respectively, and low error values, showcasing their efficiency for less complex datasets. However, models like Decision Tree Regressor and Least Angle Regression performed poorly, with negative R^2 scores and high prediction errors, making them unsuitable for Minimum PDI.Simpler models like Linear Regression and Ridge

Regression performed remarkably well under Maximum PDI, leveraging their simplicity and linearity. Regularization techniques, as observed in Bayesian Ridge, provided consistent performance, while models like SVR and Least Angle Regression struggled significantly across both conditions. Ensemble-based models such as Random Forest and CatBoost demonstrated strong adaptability, particularly under Minimum PDI scenarios, due to their ability to reduce variance. These findings emphasize the importance of selecting models based on dataset complexity, with ensemble methods excelling in diverse conditions and simpler regression models standing out in high-PDI situations.

Summarisation of all the diseases:

Disease	PDI	Models	R2 Score	MAE	RMSE	MAPE	MdAE	MSLE	RMSLE	Adjusted R2
Leaf Blast	Max PDI	Random Forest Regressor	0.906	4.021	5.199	0.215	3.203	0.066	0.257	0.886
		Extra Trees Regressor	0.908	3.737	5.132	0.21	1.939	0.064	0.254	0.889
	Min PDI	Extra Trees Regressor	0.921	2.981	4.011	0.264	2.471	0.085	0.291	0.904
		Extreme Gradient Boosting	0.901	2.756	4.49	0.231	1.422	0.097	0.311	0.88
Neck Blast	Max PDI	Extra Trees Regressor	0.605	1.16	1.48	0.08	0.95	0.01	0.1	0.52
		Extreme Gradient Boosting	0.58	1.07	1.51	0.07	0.45	0.01	0.1	0.49
	Min PDI	Extra Trees Regressor	0.91	0.59	0.72	0.08	0.48	0.007	0.08	0.89
		CatBoost Regressor	0.91	0.61	0.73	0.08	0.53	0.007	0.08	0.89
Sheath Rot	Max PDI	Extra Trees Regressor	0.906	3.4	4.531	0.123	2.607	0.021	0.145	0.885
		CatBoost Regressor	0.893	4.113	4.841	0.201	4.233	0.047	0.217	0.869
	Min PDI	Extra Trees Regressor	0.914	2.534	3.723	0.142	1.863	0.03	0.174	0.896
		Extreme Gradient Boosting	0.897	3.058	4.089	0.132	2.687	0.02	0.142	0.874
Glume Discolorat	Max PDI	Linear Regression	0.853	9.414	13.313	0.297	3.885	0.116	0.34	0.821

ion		Ridge Regression	0.851	9.722	13.408	0.317	4.197	0.122	0.349	0.818
		Linear Regression	0.916	6.158	8.846	0.246	2.648	0.203	0.45	0.898
	Min PDI	Ridge Regression	0.914	6.135	8.951	0.253	3.087	0.191	0.437	0.896
Sheath Blight		Bayesian Ridge	0.92	6.356	8.308	0.728	5.33	0.251	0.501	0.902
	Max PDI	Ridge Regression	0.919	6.462	8.365	0.633	5.36	0.25	0.5	0.901
		Bayesian Ridge	0.921	3.01	3.941	0.522	2.555	0.278	0.527	0.904
	Min PDI	Ridge Regression	0.919	3.047	3.981	0.512	2.572	0.288	0.537	0.902
Brown Spot		Linear Regression	0.974	3.24	4.363	0.098	2.265	0.02	0.141	0.968
	Max PDI	Bayesian Ridge	0.972	3.389	4.498	0.101	2.658	0.021	0.145	0.966
		Linear Regression	0.671	9.85	13.881	0.44	5.821	0.21	0.458	0.599
	Min PDI	Ridge Regression	0.672	10	13.857	0.455	5.749	0.212	0.46	0.601

Table 16: Summarization of all the models for all the diseases

7.4 Comparison of Results with Existing Systems

The Dhaanya framework represents a significant leap over existing systems in the domain of agricultural disease prediction, particularly for paddy crops. Traditional approaches rely heavily on statistical models or correlation-based methods that use historical weather patterns and are often constrained to specific regions or individual diseases. In contrast, Dhaanya employs advanced ensemble machine learning models—such as Random Forest, LightGBM, and Extra Trees—to enable highly accurate, early-stage prediction of multiple diseases including Leaf Blast, Neck Blast, Sheath Rot, and Brown Spot.

While existing literature often integrates only climatic parameters like temperature, rainfall, and humidity, Dhaanya offers a real-time, dynamic integration of both weather and soil parameters, such as pH, Nitrogen, and Potassium, across critical crop growth stages. This multi-dimensional data fusion ensures that disease risks are detected even before physical symptoms appear, allowing for proactive interventions rather than reactive responses.

Dhaanya also incorporates a temporal dimension, tracking disease development continuously through the crop lifecycle—from sowing to flowering—unlike conventional models which often account only for seasonal variations. Moreover, its design prioritizes farmer-centric usability, transforming technical predictions into user-friendly alerts and actionable insights, helping farmers make informed decisions in a timely manner.

Importantly, the system is built for scalability and adaptability, transcending region-specific applications to suit diverse agro-climatic zones across India. It actively addresses resource optimization by minimizing input wastage and promoting sustainable farming practices

through timely, data-driven recommendations. Overall, Dhaanya bridges the gap between high-end research and practical, field-level implementation, marking a new standard for intelligent agriculture systems.

7.5 Inference Drawn

The results as observed show that the Extra Trees Regressor consistently outperformed other models, achieving high R2 scores, particularly for Leaf Blast (R2 score = 0.908) and Sheath Rot (R2 score = 0.914). Linear Regression excelled in predicting Brown Spot, especially for Max PDI (R2 score = 0.974). Bayesian Ridge and Ridge Regression performed well for Sheath Blight (R2 score = 0.92) and Glume Discoloration (R2 score= 0.916 for Min PDI). Overall, Extra Trees Regressor was the most effective model, followed by Extreme Gradient Boosting (R2 score = 0.901) and Ridge Regression. These models demonstrated strong accuracy with low error metrics across most diseases.

Chapter 8: Conclusion

8.1 Limitations

While *Dhaanya* demonstrates a strong predictive capability for paddy disease incidence, certain limitations affect its scalability and operational impact:

- **Data Availability & Granularity**
  The system relies on meteorological and soil parameter data collected over six months (July–December), which may not encompass full annual cycles or diverse climate events. Inconsistent data sampling across regions and growth stages may limit generalizability.
- **Real-Time Integration Constraints**
  Although the system factors in key climate and soil variables, it currently lacks integration with live IoT sensor networks or satellite data, limiting real-time adaptability and proactive field-level interventions.
- **Computational Load of Ensemble Models**
  Tree-based ensemble methods like Extra Trees and Gradient Boosting, though highly accurate ($R^2 > 0.91$ for some diseases), are computationally intensive. This could pose challenges in deploying Dhaanya in rural or low-resource settings.
- **Geographical Scope & Training Bias**
  The system has been trained primarily on data from Maharashtra and adjacent agro-climatic zones. Its disease prediction models may need substantial retraining for other geographic regions with different soil, crop, and weather dynamics.
- **Socio-Economic Context Exclusion**
  The current model emphasizes biological and environmental variables but does not incorporate market access, farmer education, or government intervention schemes, all of which influence disease management effectiveness.
- **Lack of Farmer-Facing Decision Support**
  Dhaanya currently serves as a backend analytical tool without an accessible frontend

dashboard or mobile app for farmers and policymakers to view alerts and recommended actions in real-time.

8.2 Conclusion

Dhaanya offers a forward-looking, AI-driven approach to paddy crop disease management. By incorporating advanced regression models such as Extra Trees, Random Forest, and CatBoost, it accurately forecasts the severity of multiple diseases including Leaf Blast, Sheath Rot, and Brown Spot.

Key insights from the study include:

- **Extra Trees Regressor** consistently outperformed other models across Max and Min PDI predictions, with R^2 scores exceeding 0.90 for several diseases.
- **Sunshine Hours** emerged as the most critical climatic factor affecting disease progression, as identified through explainable AI tools like LIME.
- **Stage-specific analysis** revealed that Harvesting and Flowering stages are most vulnerable to severe disease incidence, suggesting targeted intervention periods.
- **Soil pH fluctuation**, particularly during early growth phases, was strongly associated with increased disease susceptibility, supporting its use as a key predictive feature.

Through these contributions, Dhaanya enhances the understanding of disease dynamics in paddy fields and lays a foundation for climate-resilient agriculture.

8.3 Future Scope

To further enhance its utility and scalability, the following upgrades are proposed for the Dhaanya system:

- **Integration of Real-Time Data Streams**
  Future implementations should incorporate IoT-based field sensors, weather APIs, and satellite data to enable real-time disease forecasting and actionable alerts.

- **Expansion to Broader Agro-Climatic Zones**

  With retraining and transfer learning, Dhaanya can be extended to other rice-growing regions in India and Southeast Asia, each with distinct ecological profiles.
- **Inclusion of Economic and Behavioral Data**

  Integrating variables like fertilizer subsidies, farmer training participation, and historical yield-pricing data can provide a more holistic risk analysis and recommendation engine.
- **Deployment as a Mobile/Web-Based Platform**

  A farmer-centric mobile app and policymaker dashboard can be developed to display disease risk levels, preventive measures, and yield forecasts tailored to location and crop stage.
- **Hybrid Models & Edge Deployment**

  Combining traditional ensemble methods with deep learning architectures such as LSTMs and transformers may improve long-term disease trend forecasting. Additionally, deploying models on edge devices can support offline decision-making in remote areas.

By addressing these future directions, Dhaanya can evolve into a comprehensive, real-time, and region-agnostic AI-powered system for sustainable crop protection.

References

[1] Pradhan, J., A. Baliarsingh, G. Biswal, M.P. Das and Pasupalak, S. 2018. Effect of Weather Parameters on Infestation of Blast Disease (Pyricularia oryzae) in Rabi Season Rice (Oryza sativa L.) in East South Eastern Coastal Plain of Odisha.Int.J.Curr.Microbiol.App.Sci. 7(11): 893-900. doi: https://doi.org/10.20546/ijcmas.2018.711.106

[2] B. JOHNSON and T. CHANDRAKUMAR, “Influence of weather parameters on rice blast disease progression in Tamil Nadu, India”, J. Agrometeorol., vol. 26, no. 3, pp. 362–366, Sep. 2024.

[3] [A. Jayashree, A. Nagaraja, M. K. Prasanna Kumar, and B. S. Chethana, ”Unravelling relationship of weather factors with rice blast disease severity and development of prediction equations,” Mysore J. Agric. Sci., vol. 56, no. 2, pp. 152-160, 2022.

[4] A. S. Kapoor, R. Prasad, and G. K. Sood, ”Forecasting of rice blast in Kangra district of Himachal Pradesh,” Indian Phytopathology, vol. 57, no. 4, pp. 440-445, 2004.

[5] S. A. Gaikwad, A. S. Upadhye, and D. K. Kulkarni, ”Ecology in relation to rice field soils in Bhor and Velhe region of Pune district, Maharashtra State, India,” Int. J. Geol. Earth Environ. Sci., vol. 6, no. 1, pp. 12-18,Jan.-Apr. 2016.

[6] D. Patel and S. Reddy, "Agriculture Yield Estimation Using Machine Learning Algorithms," *IEEE International Conference on Smart Farming Technologies (ICSFT)*, pp. 67-72, 2020. DOI: 10.1109/ICSFT.2020.00234.

[7] International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), “ICRISAT Data Repository,” [Online]. Available: https://www.icrisat.org/. [Accessed: Mar. 2025].

[8] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.

[9] A. D. Lagnika, A. Kissi, and M. P. Tchuenté, “A machine learning framework for agricultural decision support,” *Expert Systems with Applications*, vol. 185, p. 115578, 2021. DOI: 10.1016/j.eswa.2021.115578.

[10] Open Government Data (OGD) Platform India, “Climate Data Repository,” [Online]. Available: https://data.gov.in/. [Accessed: Mar. 2025].

[11] Food and Agriculture Organization of the United Nations (FAOSTAT), “FAOSTAT Database,” [Online]. Available: https://www.fao.org/faostat/en/. [Accessed: Mar. 2025].

[12] Ministry of Agriculture & Farmers’ Welfare, Government of India, “Agricultural Statistics,” [Online]. Available: https://agricoop.nic.in/. [Accessed: Mar. 2025].

[13] Food and Agriculture Organization of the United Nations (FAOSTAT), “Pesticide Use Statistics,” [Online]. Available: https://www.fao.org/faostat/en/. [Accessed: Mar. 2025].

[14] A. D. Lagnika, A. Kissi, and M. P. Tchuenté, “A machine learning framework for agricultural decision support,” *Expert Systems with Applications*, vol. 185, p. 115578, 2021. DOI: 10.1016/j.eswa.2021.115578.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[16] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021.

[17] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785-794.

[18] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Montreal, Canada, 1995, pp. 1137-1143.

[19] L. Breiman, “Statistical modeling: The two cultures,” *Statistical Science*, vol. 16, no. 3, pp. 199-231, 2001.

[20] J. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001. DOI: 10.1214/aos/1013203451.

Papers Published/In review (ICTEAH-2025)

Leveraging AI for analysis of the Percentage Disease Index of various diseases in Paddy plants

Abstract

Rural livelihoods depend heavily on agriculture, but its sustainability is seriously threatened by climate change and plant diseases, especially for paddy crops, which lose about 37% of their production each year to pests and diseases. To address this difficulty, Dhaanya, an AI-powered predictive modelling system, uses machine learning techniques such as Random Forest Regressor, Extra Trees Regressor and Gradient Boosting Machines to assess climatic variables such as temperature, humidity, rainfall, and soil pH. It predicts the main paddy diseases, such as Leaf Blast, Neck Blast, Sheath Rot, Glume Discoloration, Sheath Blight, and Brown Spots, by finding patterns and correlations. This proactive strategy offers a data-driven way to lessen the effects of climate change on farming by increasing resource efficiency, lowering crop damage, and bolstering agricultural resilience.

Keywords: Paddy crops, AI-powered predictive modelling, Machine learning, Climate change, Weather data, Disease prediction, Data-driven agriculture, Disease management systems, Agriculture sustainability, Sustainable farming.

<H1> Introduction

Dhaanya addresses the critical limitations of traditional agricultural disease management systems, which often rely on manual field inspections and reactive treatments. These methods are time-consuming, labour-intensive, and prone to delays, leading to significant crop losses. The increasing unpredictability of climatic factors like temperature, humidity, etc. further exacerbates these challenges. The Dhaanya AI-powered Disease Incidence Prediction System integrates real-time weather and soil data to forecast the severity of multiple diseases simultaneously, allowing for targeted interventions. By leveraging machine learning algorithms, the system analyzes parameters such as temperature, humidity, rainfall, wind speed, and soil pH. As highlighted in recent studies, changes in soil pH are closely linked to the prevalence and severity of diseases like Leaf Blast, Neck Blast, Sheath Rot, and other diseases. This comprehensive approach enhances the system’s predictive accuracy and adaptability to varying environmental conditions. Pyricularia oryzae, the fungal pathogen responsible for Leaf Blast and other severe rice diseases, is a significant threat to global rice production. It thrives under favorable conditions such as high humidity, warm temperatures, and prolonged leaf wetness. The fungus primarily affects the aerial parts of rice plants, including leaves, stems, and panicles. It starts by germinating spores on the plant's surface, restricting photosynthesis and weakening the plant. As the disease progresses, the fungus invades the plant's vascular system, causing panicle blight and grain discoloration, which result in poor grain quality and reduced yields. The spores can spread through wind, rain, or infested seeds, making the disease highly contagious. Pyricularia oryzae's ability to adapt to varying environmental conditions and attack multiple stages of rice growth makes it a persistent challenge for farmers. This project seeks to equip farmers with actionable insights, enabling them to anticipate potential disease outbreaks, implement preventive measures, and optimize resource use. By combining predictive capabilities, real-time data integration, and multi-disease modeling, Dhaanya aspires to transform traditional agricultural practices and promote sustainable farming in the face of climate uncertainties.

<H1> Literature review

Weather plays a crucial role in the spread of rice blast disease (Pyricularia oryzae), with studies showing how environmental conditions directly impact its occurrence. Paper [1] found that climatic factors like temperature and humidity significantly impact rice blast severity during the Rabi season in Odisha, where high humidity and dew create ideal conditions for fungal growth. Likewise, Paper [2] explored how local weather patterns, such as rainfall and temperature fluctuations in Tamil Nadu, influence rice blast. Paper [3] built on this research by developing prediction equations using climatic data, providing a framework for incorporating weather patterns into disease forecasting models. Papers [4] and [5] highlighted the importance of historical weather data and computational methods in disease forecasting, introducing advanced machine learning-based predictive models. Our approach builds on research like Paper [6], which highlighted the broader role of environmental factors across crops, and other studies emphasizing the importance of predictive models in agricultural disease management. These findings support the objectives of our project, which seeks to enhance disease risk prediction by combining real-time weather data with historical trends for more accurate predictions.

<H1> Novelty of Work

The Dhaanya AI-Powered Disease Incidence Prediction System revolutionizes agricultural disease management by addressing key limitations of traditional methods. Unlike manual inspections and reactive treatments, Dhaanya uses AI-driven regression models to predict disease severity before symptoms appear, enabling proactive interventions. It integrates real-time weather data with historical disease patterns, improving accuracy by accounting for environmental influences. Unlike traditional methods that focus on a single disease, Dhaanya predicts multiple diseases simultaneously, providing field-specific insights for better management. By enabling early detection and targeted interventions, Dhaanya minimizes crop damage, optimizes resource use, and enhances overall agricultural productivity.

<H1> Methodology

<H2> **Data Collection and Parameter Selection**

Data collection for this study was done in cooperation with a seasoned plant pathologist who offered advice on the right metrics to evaluate and examine the connection between plant health and weather. Throughout the study period, the following meteorological parameters were methodically recorded:

Maximum Temperature (MaxTemp): The highest temperature that can occur in a given day, expressed in degrees Celsius, and which affects plant metabolism and the development of disease. Minimum Temperature (MinTemp): The lowest temperature that occurs each day, which is crucial for figuring out how much exposure the plant has to potentially hazardous cold temperatures. Relative Humidity 1 (RelH1): The relative humidity at dawn, which represents the amount of moisture in the air that can influence evapotranspiration and illness susceptibility. Relative Humidity 2 (RelH2): A measurement of daily moisture variations that is obtained at the conclusion of the day. Rainfall: The total amount of precipitation, expressed in millimetres, is essential to the transmission of illness, especially for waterborne infections. Rainy Days: The number of days with detectable precipitation, which might affect the risk of bacterial or fungal diseases as well as plant stress. Sunshine Hours: The total amount of sunlight received each day, which is essential for photosynthesis and the general well-being of plants. Wind Speed: The average wind speed each day, expressed in meters per second, which can influence disease spread and exacerbate plant stress. Along with these weather parameters, soil parameters namely the amount of Nitrogen and Phosphorous content were also collected and examined. The data was collected according to the growing stages of the paddy crops: Sowing, Transplanting, Tillering, Panicle Initiation, Flowering and Harvesting, during the course of six months (July to December). The aim was to understand the relationship between weather conditions and disease severity across different stages of crop development. By using this method, patterns and connections between plant pathology and meteorological data were found, allowing for a better understanding of how climatic fluctuations affect disease dynamics throughout the growing season. By using Explainable AI tools such as LIME, we computed the importance of each parameter in the data with respect to the target variable and found that Sunshine Hours was the most important parameter among all others.

<H2> **Data and Parameter Analysis**

The analysis of the Maximum Percentage Disease Index (MaxPDI) and Minimum Percentage Disease Index (MinPDI) across different growth stages of paddy plants as shown in Figure 1 reveals the significant variations in disease severity across different growth stages of paddy plants and provides a comprehensive understanding of disease progression. Brown Spot consistently exhibits high PDI values, with significant impact during both Sowing (MinPDI) and Harvesting (MaxPDI), making it one of the most persistent and severe diseases. Similarly, Neck Blast and Glume Discoloration show moderate to high PDI values across multiple stages, with peaks during Harvesting. Diseases like Sheath Blight and Leaf Blast, while having moderate to high MaxPDI in

later stages, exhibit relatively low MinPDI throughout the crop cycle, indicating less baseline severity. Overall, the Harvesting stage emerges as the most vulnerable for most diseases, while Sowing and early stages like Transplanting also demonstrate susceptibility to certain persistent diseases. These insights underscore the need for stage-specific disease management, focusing on reducing both peak and baseline disease indices to optimize crop health and yield.

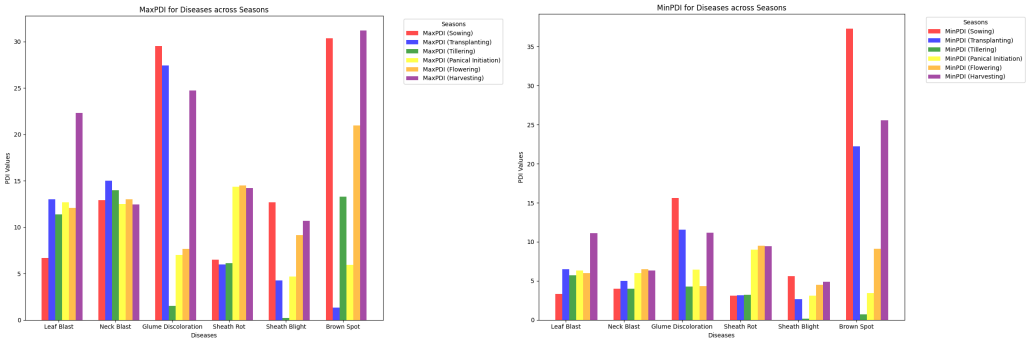


Fig 1: The Bar Charts of all the Diseases with respect to Max and Min Percentage Disease Index

<H2> **Sustainability**

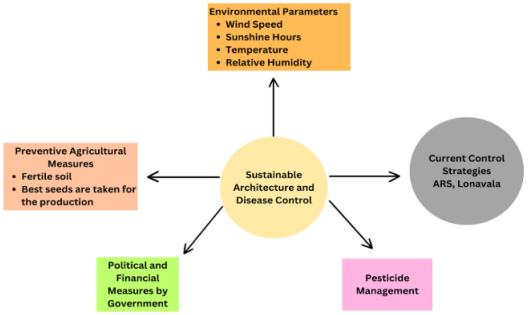


Fig 2: Factors to achieve Sustainability

The framework in Figure 2 integrates key elements for sustainable agriculture and disease management. Environmental factors like temperature, humidity, wind speed, and daylight influence disease spread, guiding preventive measures such as crop rotation, planting adjustments, and physical barriers. Using disease-resistant seeds, soil enrichment, and irrigation enhances plant resilience while reducing chemical dependency. Government policies, including financial aid, subsidies, and training, support sustainable practices. Integrated Pest Management (IPM) minimizes pesticide impact through biological and mechanical controls.

<H1> Blockchain For Data Management

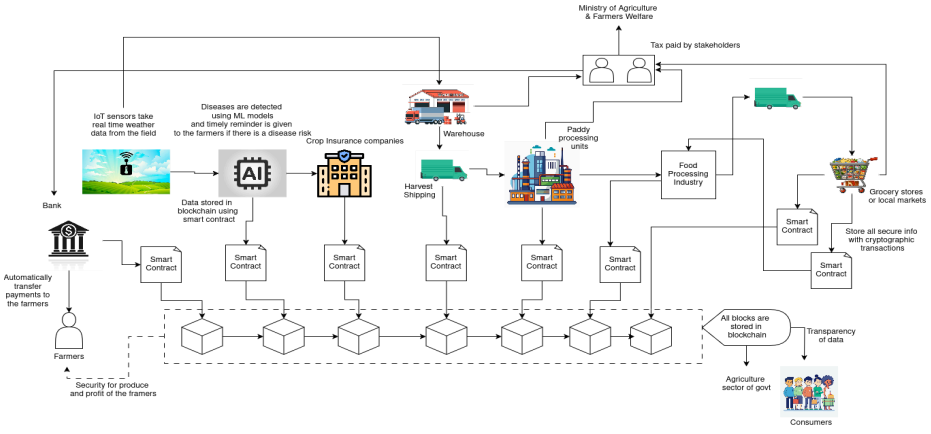


Fig 3: Workflow of System integrated with Blockchain

As shown in Figure 3, Blockchain enables direct and transparent financial transactions in agriculture, ensuring farmers receive payments without intermediaries. Smart contracts facilitate instant fund transfers from buyers, government subsidies, or financial institutions directly to farmers, eliminating any undue cuts. Blockchain-based digital identities prevent fraud, ensuring only genuine farmers receive subsidies and payments. With tamper-proof records, the system enforces fair pricing, reduces exploitation, and minimizes corruption. Instant digital payments enhance efficiency, ensuring farmers receive their earnings without delays or hidden deductions. Additionally, decentralized ledger technology enhances trust among stakeholders by providing verifiable and immutable transaction records. This fosters confidence in financial dealings and creates a more resilient agricultural economy. By integrating blockchain, agriculture becomes more equitable, sustainable, and secure, empowering farmers with full control over their financial transactions and reducing dependency on traditional financial intermediaries.

<H1> Machine Learning Models Used

To predict paddy diseases, we utilized a diverse set of machine learning models categorized into tree-based models, boosting methods, linear models, instance-based learning, support vector machines, and others. Each category addressed specific challenges such as non-linearity, feature interactions, and data imbalance.

Tree-Based Models (Random Forest, Extra Trees, and Decision Tree Regressors) captured complex patterns and feature interactions. Extra Trees and Random Forest consistently performed well, especially for Leaf Blast and Neck Blast, due to their ensemble nature, which reduces variance and enhances generalization.

Boosting Methods (XGBoost, CatBoost, Gradient Boosting, and AdaBoost) improved prediction accuracy through iterative learning. CatBoost and Gradient Boosting were particularly effective for diseases like Brown Spot and Sheath Rot, leveraging feature importance and ordered boosting for better predictions.

Linear Models (Linear Regression, Ridge, Lasso, Elastic Net, Bayesian Ridge, and Least Angle Regression) served as baselines and worked well for diseases with linear relationships, such as Glume Discoloration and Sheath Blight. Ridge and Bayesian Ridge helped reduce overfitting, further enhancing performance.

Instance-Based Learning (K-Neighbors Regressor) predicted based on local feature similarities but performed poorly in high-variability diseases like Neck Blast compared to ensemble methods.

Support Vector Machines (SVR) was evaluated for its ability to handle high-dimensional spaces but struggled with complex feature interactions, leading to lower accuracy compared to tree-based models.

Other Models (Passive-Aggressive Regressor, Dummy Regressor, and Huber Regression) provided robustness testing and benchmarks. Huber Regression handled outliers well, while Dummy Regressor served as a baseline for performance comparison.

<H1> Results

Table 1: The summary statistics of diseases and the performance metrics for predicting Max PDI and Min PDI for various diseases.

Table 1: Performance metrics for each disease

Disease	PDI	Models	R2 Score	MAE	RMSE	MAPE
Leaf Blast	Max PDI	Random Forest Regressor	0.906	4.021	5.199	0.215
		Extra Trees Regressor	0.908	3.737	5.132	0.21
	Min PDI	Extra Trees Regressor	0.921	2.981	4.011	0.264
		Extreme Gradient Boosting	0.901	2.756	4.49	0.231
Neck Blast	Max PDI	Extra Trees Regressor	0.605	1.16	1.48	0.08
		Extreme Gradient Boosting	0.58	1.07	1.51	0.07
	Min PDI	Extra Trees Regressor	0.91	0.59	0.72	0.08
		CatBoost Regressor	0.91	0.61	0.73	0.08
Sheath Rot	Max PDI	Extra Trees Regressor	0.906	3.4	4.531	0.123
		CatBoost Regressor	0.893	4.113	4.841	0.201
	Min PDI	Extra Trees Regressor	0.914	2.534	3.723	0.142

		Extreme Gradient Boosting	0.897	3.058	4.089	0.132
Glume Discoloration	Max PDI	Linear Regression	0.853	9.414	13.313	0.297
		Ridge Regression	0.851	9.722	13.408	0.317
	Min PDI	Linear Regression	0.916	6.158	8.846	0.246
		Ridge Regression	0.914	6.135	8.951	0.253
Sheath Blight	Max PDI	Bayesian Ridge	0.92	6.356	8.308	0.728
		Ridge Regression	0.919	6.462	8.365	0.633
	Min PDI	Bayesian Ridge	0.921	3.01	3.941	0.522
		Ridge Regression	0.919	3.047	3.981	0.512
Brown Spot	Max PDI	Linear Regression	0.974	3.24	4.363	0.098
		Bayesian Ridge	0.972	3.389	4.498	0.101
	Min PDI	Linear Regression	0.671	9.85	13.881	0.44
		Ridge Regression	0.672	10	13.857	0.455

The results as observed in Table-1 show that the Extra Trees Regressor consistently outperformed other models, achieving high R2 scores, particularly for Leaf Blast (R2 score = 0.908) and Sheath Rot (R2 score = 0.914). Linear Regression excelled in predicting Brown Spot, especially for Max PDI (R2 score = 0.974). Bayesian Ridge and Ridge Regression performed well for Sheath Blight (R2 score = 0.92) and Glume Discoloration (R2 score= 0.916 for Min PDI). Overall, Extra Trees Regressor was the most effective model, followed by Extreme Gradient Boosting (R2 score = 0.901) and Ridge Regression. These models demonstrated strong accuracy with low error metrics across most diseases.

<H1> References

[1] Pradhan, J., A. Baliarsingh, G. Biswal, M.P. Das and Pasupalak, S. 2018. Effect of Weather Parameters on Infestation of Blast Disease (Pyricularia oryzae) in Rabi Season Rice (Oryza sativa L.) in East & South Eastern Coastal Plain of Odisha.Int.J.Curr.Microbiol.App.Sci. 7(11): 893-900.

[2] B. JOHNSON and T. CHANDRAKUMAR, “Influence of weather parameters on rice blast disease progression in Tamil Nadu, India”, J. Agrometeorology., vol. 26, no. 3, pp. 362–366, Sep. 2024.

[3] A. Jayashree, A. Nagaraja, M. K. Prasanna Kumar, and B. S. Chethana, "Unravelling relationship of weather factors with rice blast disease severity and development of prediction equations," Mysore J. Agric. Sci., vol. 56, no. 2, pp. 152-160, 2022.

[4] A. S. Kapoor, R. Prasad, and G. K. Sood, "Forecasting of rice blast in Kangra district of Himachal Pradesh," Indian Phytopathology, vol. 57, no. 4, pp. 440-445, 2004.

[5] S. A. Gaikwad, A. S. Upadhye, and D. K. Kulkarni, "Ecology in relation to rice field soils in Bhor and Velhe region of Pune district, Maharashtra State, India," Int. J. Geol. Earth Environ. Sci., vol. 6, no. 1, pp. 12-18, Jan.-Apr. 2016.

[6] Habib-Ur-Rahman M, Ahmad A, Raza A, et al. Impact of climate change on agricultural production; Issues, challenges, and opportunities in Asia. Front Plant Sci. 2022;13:925548. Published 2022 Oct 10. doi:10.3389/fpls.2022.925548

Leveraging AI for analysis of the Percentage Disease Index of various diseases in Paddy plants

Dr. Sharmila Sengupta
Associate Professor, VES Institute of Technology
Email: sharmila.sengupta@ves.ac.in

Attreyee Mukherjee
VES Institute of Technology
Email: 2021.attreyee.mukherjee@ves.ac.in

Amogh Inamdar
Student, VES Institute of Technology
Email: 2021.amogh.inamdar@ves.ac.in

Saumya Tripathi
Student, VES Institute of Technology
Email: 2021.saumya.tripathi@ves.ac.in

Yashodhan Sharma
Student, VES Institute of Technology
Email: 2021.yashodhan.sharma@ves.ac.in

Dr. K.S. Raghuwanshi
Title,
Email: drksr63@gmail.com

Abstract

Rural livelihoods depend heavily on agriculture, but its sustainability is seriously threatened by climate change and plant diseases, especially for paddy crops, which lose about 37% of their production each year to pests and diseases. To address this difficulty, Dhaanya, an AI-powered predictive modeling system, uses machine learning techniques such as Random Forest Regressor, Extra Trees Regressor and Gradient Boosting Machines to assess climatic variables such as temperature, humidity, rainfall, and soil pH. It predicts the main paddy diseases, such as Leaf Blast, Neck Blast, Sheath Rot, Glume Discoloration, Sheath Blight, and Brown Spots, by finding patterns and correlations. This proactive strategy offers a data-driven way to lessen the effects of climate change on farming by increasing resource efficiency, lowering crop damage, and bolstering agricultural resilience.

Keywords: Paddy crops, AI-powered predictive modeling, Machine learning, Climate change, Weather data, Disease prediction, Data-driven agriculture, Disease management systems, Agriculture sustainability, Sustainable farming.

<H1> Introduction

Dhaanya addresses the critical limitations of traditional agricultural disease management systems, which often rely on manual field inspections and reactive treatments. These methods are time-consuming, labor-intensive, and prone to delays, leading to significant crop losses. The increasing unpredictability of climatic factors like temperature, humidity, etc. further exacerbates these challenges. The Dhaanya AI-powered Disease Incidence Prediction System integrates real-time weather and soil data to forecast the severity of multiple diseases simultaneously, allowing for targeted interventions. By leveraging machine learning algorithms, the system analyzes parameters such as temperature, humidity, rainfall, wind speed, and soil pH. As

highlighted in recent studies, changes in soil pH are closely linked to the prevalence and severity of diseases like Leaf Blast, Neck Blast, Sheath Rot, and other diseases. This comprehensive approach enhances the system's predictive accuracy and adaptability to varying environmental conditions. Pyricularia oryzae, the fungal pathogen responsible for Leaf Blast and other severe rice diseases, is a significant threat to global rice production. It thrives under favorable conditions such as high humidity, warm temperatures, and prolonged leaf wetness. The fungus primarily affects the aerial parts of rice plants, including leaves, stems, and panicles. It starts by germinating spores on the plant's surface, restricting photosynthesis and weakening the plant. As the disease progresses, the fungus invades the plant's vascular system, causing panicle blight and grain discoloration, which result in poor grain quality and reduced yields. The spores can spread through wind, rain, or infested seeds, making the disease highly contagious. Pyricularia oryzae's ability to adapt to varying environmental conditions and attack multiple stages of rice growth makes it a persistent challenge for farmers. This project seeks to equip farmers with actionable insights, enabling them to anticipate potential disease outbreaks, implement preventive measures, and optimize resource use. By combining predictive capabilities, real-time data integration, and multi-disease modeling, Dhaanya aspires to transform traditional agricultural practices and promote sustainable farming in the face of climate uncertainties.

<H1> Literature review

Weather plays a crucial role in the spread of rice blast disease (Pyricularia oryzae), with studies showing how environmental conditions directly impact its occurrence. Paper [1] found that climatic factors like temperature and humidity significantly impact rice blast severity during the Rabi season in Odisha, where high humidity and dew create ideal conditions for fungal growth. Likewise, Paper [2] explored how local weather patterns, such as rainfall and temperature fluctuations in Tamil Nadu, influence rice blast. Paper [3] built on this research by developing prediction equations using climatic data, providing a framework for incorporating weather patterns into disease forecasting models. Papers [4] and [5] highlighted the importance of historical weather data and computational methods in disease forecasting, introducing advanced machine learning-based predictive models. Our approach builds on research like Paper [6], which highlighted the broader role of environmental factors across crops, and other studies emphasizing the importance of predictive models in agricultural disease management. These findings support the objectives of our project, which seeks to enhance disease risk prediction by combining real-time weather data with historical trends for more accurate predictions.

<H1> Novelty of Work

The Dhaanya AI-Powered Disease Incidence Prediction System revolutionizes agricultural disease management by addressing key limitations of traditional methods. Unlike manual inspections and reactive treatments, Dhaanya uses AI-driven regression models to predict disease severity before symptoms appear, enabling proactive interventions. It integrates real-time weather data with historical disease patterns, improving accuracy by accounting for environmental influences. Unlike traditional methods that focus on a single disease, Dhaanya predicts multiple diseases simultaneously, providing field-specific insights for better management. By enabling early detection and targeted interventions, Dhaanya minimizes crop damage, optimizes resource use, and enhances overall agricultural productivity.

<H1> Methodology

<H2> **Data Collection and Parameter Selection**

Data collection for this study was done in cooperation with a seasoned plant pathologist who offered advice on the right metrics to evaluate and examine the connection between plant health and weather. Throughout the study period, the following meteorological parameters were methodically recorded:

Maximum Temperature (MaxTemp): The highest temperature that can occur in a given day, expressed in degrees Celsius, and which affects plant metabolism and the development of disease. Minimum Temperature (MinTemp): The lowest temperature that occurs each day, which is crucial for figuring out how much exposure the plant has to potentially hazardous cold temperatures. Relative Humidity 1 (RelH1): The relative humidity at dawn, which represents the amount of moisture in the air that can influence evapotranspiration and illness susceptibility. Relative Humidity 2 (RelH2): A measurement of daily moisture variations that is obtained at the conclusion of the day. Rainfall: The total amount of precipitation, expressed in millimeters, is essential to the transmission of illness, especially for waterborne infections. Rainy Days: The quantity of days with detectable precipitation, which might affect the risk of bacterial or fungal diseases as well as plant stress. Sunshine Hours: The total amount of sunlight received each day, which is essential for photosynthesis and the general well-being of plants. Wind Speed: The average wind speed each day, expressed in meters per second, which can influence disease spread and exacerbate plant stress. Along with these weather parameters, soil parameters namely the amount of Nitrogen and Phosphorous content were also collected and examined. The data was collected according to the growing stages of the paddy crops: Sowing, Transplanting, Tillering, Panicle Initiation, Flowering and Harvesting, during the course of six months (July to December). The aim was to understand the relationship between weather conditions and disease severity across different stages of crop development. By using this method, patterns and connections between plant pathology and meteorological data were found, allowing for a better understanding of how climatic fluctuations affect disease dynamics throughout the growing season. By using Explainable AI tools such as LIME, we computed the importance of each parameter in the data with respect to the target variable and found that Sunshine Hours was the most important parameter among all others.

<H2> **Data and Parameter Analysis**

The analysis of the Maximum Percentage Disease Index (MaxPDI) and Minimum Percentage Disease Index (MinPDI) across different growth stages of paddy plants as shown in Figure 1 reveals the significant variations in disease severity across different growth stages of paddy plants and provides a comprehensive understanding of disease progression. Brown Spot consistently exhibits high PDI values, with significant impact during both Sowing (MinPDI) and Harvesting (MaxPDI), making it one of the most persistent and severe diseases. Similarly, Neck Blast and Glume Discoloration show moderate to high PDI values across multiple stages, with peaks during Harvesting. Diseases like Sheath Blight and Leaf Blast, while having moderate to high MaxPDI in later stages, exhibit relatively low MinPDI throughout the crop cycle, indicating less baseline severity. Overall, the Harvesting stage emerges as the most vulnerable for most diseases, while Sowing and early stages like Transplanting also demonstrate susceptibility to certain persistent diseases. These insights underscore the need for stage-specific disease management, focusing on reducing both peak and baseline disease indices to optimize crop health and yield.

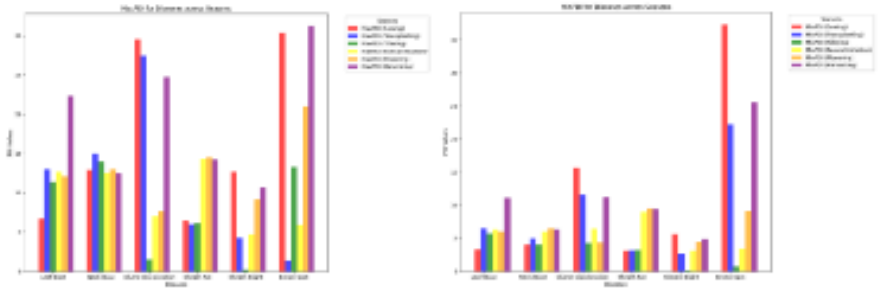


Fig 1: The Bar Charts of all the Diseases with respect to Max and Min Percentage Disease Index

<H2> **Sustainability**

Fig 2: Factors to achieve Sustainability

The framework in Figure 2 integrates key elements for sustainable agriculture and disease management. Environmental factors like temperature, humidity, wind speed, and daylight influence disease spread, guiding preventive measures such as crop rotation, planting adjustments, and physical barriers. Using disease-resistant seeds, soil enrichment, and irrigation enhances plant resilience while reducing chemical dependency. Government policies, including financial aid, subsidies, and training, support sustainable practices. Integrated Pest Management (IPM) minimizes pesticide impact through biological and mechanical controls.

<H1> Blockchain For Data Management

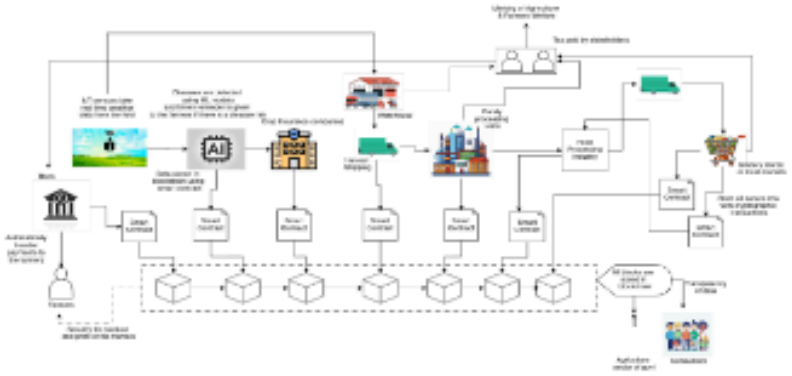


Fig 3: Workflow of System integrated with Blockchain

As shown in Figure 3, Blockchain enables direct and transparent financial transactions in agriculture, ensuring farmers receive payments without intermediaries. Smart contracts facilitate instant fund transfers from buyers, government subsidies, or financial institutions directly to farmers, eliminating any undue cuts. Blockchain-based digital identities prevent fraud, ensuring only genuine farmers receive subsidies and payments. With tamper-proof records, the system enforces fair pricing, reduces exploitation, and minimizes corruption. Instant digital payments enhance efficiency, ensuring farmers receive their earnings without delays or hidden deductions. Additionally, decentralized ledger technology enhances trust among stakeholders by providing verifiable and immutable transaction records. This fosters confidence in financial dealings and creates a more resilient agricultural economy. By integrating blockchain, agriculture becomes more equitable, sustainable, and secure, empowering farmers with full control over their financial transactions and reducing dependency on traditional financial intermediaries.

<H1> Machine Learning Models Used

To predict paddy diseases, we utilized a diverse set of machine learning models categorized into tree-based models, boosting methods, linear models, instance-based learning, support vector machines, and others. Each category addressed specific challenges such as non-linearity, feature interactions, and data imbalance.

Tree-Based Models (Random Forest, Extra Trees, and Decision Tree Regressors) captured complex patterns and feature interactions. Extra Trees and Random Forest consistently performed well, especially for Leaf Blast and Neck Blast, due to their ensemble nature, which reduces variance and enhances generalization.

Boosting Methods (XGBoost, CatBoost, Gradient Boosting, and AdaBoost) improved prediction accuracy through iterative learning. CatBoost and Gradient Boosting were particularly effective for diseases like Brown Spot and Sheath Rot, leveraging feature importance and ordered boosting for better predictions.

Linear Models (Linear Regression, Ridge, Lasso, Elastic Net, Bayesian Ridge, and Least Angle Regression) served as baselines and worked well for diseases with linear relationships, such as Glume Discoloration and Sheath Blight. Ridge and Bayesian Ridge helped reduce overfitting, further enhancing performance.

Instance-Based Learning (K-Neighbors Regressor) predicted based on local feature similarities but performed poorly in high-variability diseases like Neck Blast compared to ensemble methods.

Support Vector Machines (SVR) was evaluated for its ability to handle high-dimensional spaces but struggled with complex feature interactions, leading to lower accuracy compared to tree-based models.

Other Models (Passive-Aggressive Regressor, Dummy Regressor, and Huber Regression) provided robustness testing and benchmarks. Huber Regression handled outliers well, while Dummy Regressor served as a baseline for performance comparison.

<H1> Results

Table 1: The summary statistics of diseases and the performance metrics for predicting Max PDI and Min PDI for various diseases.

Table 1: Performance metrics for each disease

Disease	PDI	Models	R2 Score	MAE	RMSE	MAPE
Leaf Blast	Max PDI	Random Forest Regressor	0.906	4.021	5.199	0.215
		Extra Trees Regressor	0.908	3.737	5.132	0.21
	Min PDI	Extra Trees Regressor	0.921	2.981	4.011	0.264
		Extreme Gradient Boosting	0.901	2.756	4.49	0.231
Neck Blast	Max PDI	Extra Trees Regressor	0.605	1.16	1.48	0.08
		Extreme Gradient Boosting	0.58	1.07	1.51	0.07
	Min PDI	Extra Trees Regressor	0.91	0.59	0.72	0.08
		CatBoost Regressor	0.91	0.61	0.73	0.08
Sheath Rot	Max PDI	Extra Trees Regressor	0.906	3.4	4.531	0.123
		CatBoost Regressor	0.893	4.113	4.841	0.201

	Min PDI	Extra Trees Regressor	0.914	2.534	3.723	0.142
		Extreme Gradient Boosting	0.897	3.058	4.089	0.132
Glume Discoloration	Max PDI	Linear Regression	0.853	9.414	13.313	0.297
		Ridge Regression	0.851	9.722	13.408	0.317
	Min PDI	Linear Regression	0.916	6.158	8.846	0.246
		Ridge Regression	0.914	6.135	8.951	0.253
Sheath Blight	Max PDI	Bayesian Ridge	0.92	6.356	8.308	0.728
		Ridge Regression	0.919	6.462	8.365	0.633
	Min PDI	Bayesian Ridge	0.921	3.01	3.941	0.522
		Ridge Regression	0.919	3.047	3.981	0.512
Brown Spot	Max PDI	Linear Regression	0.974	3.24	4.363	0.098
		Bayesian Ridge	0.972	3.389	4.498	0.101
	Min PDI	Linear Regression	0.671	9.85	13.881	0.44
		Ridge Regression	0.672	10	13.857	0.455

The results as observed in Table-1 show that the Extra Trees Regressor consistently outperformed other models, achieving high R2 scores, particularly for Leaf Blast (R2 score = 0.908) and Sheath Rot (R2 score = 0.914). Linear Regression excelled in predicting Brown Spot, especially for Max PDI (R2 score = 0.974). Bayesian Ridge and Ridge Regression performed well for Sheath Blight (R2 score = 0.92) and Glume Discoloration (R2 score= 0.916 for Min PDI). Overall, Extra Trees Regressor was the most effective model, followed by Extreme Gradient Boosting (R2 score = 0.901) and Ridge Regression. These models demonstrated strong accuracy with low error metrics across most diseases.

<H1> References

[1] Pradhan, J., A. Baliarsingh, G. Biswal, M.P. Das and Pasupalak, S. 2018. Effect of Weather Parameters on Infestation of Blast Disease (Pyricularia oryzae) in Rabi Season Rice (Oryza sativa L.) in East & South Eastern Coastal Plain of Odisha.Int.J.Curr.Microbiol.App.Sci. 7(11): 893-900.

[2] B. JOHNSON and T. CHANDRAKUMAR, "Influence of weather parameters on rice blast disease progression in Tamil Nadu, India", J. Agrometeorology., vol. 26, no. 3, pp. 362–366, Sep. 2024.

[3] A. Jayashree, A. Nagaraja, M. K. Prasanna Kumar, and B. S. Chethana, "Unravelling relationship of weather factors with rice blast disease severity and development of prediction equations," Mysore J. Agric. Sci., vol. 56, no. 2, pp. 152-160, 2022.

[4] A. S. Kapoor, R. Prasad, and G. K. Sood, "Forecasting of rice blast in Kangra district of Himachal Pradesh," Indian Phytopathology, vol. 57, no. 4, pp. 440-445, 2004.

[5] S. A. Gaikwad, A. S. Upadhye, and D. K. Kulkarni, "Ecology in relation to rice field soils in Bhor and Velhe region of Pune district, Maharashtra State, India," Int. J. Geol. Earth Environ. Sci., vol. 6, no. 1, pp. 12-18, Jan.-Apr. 2016.

[6] Habib-Ur-Rahman M, Ahmad A, Raza A, et al. Impact of climate change on agricultural production; Issues, challenges, and opportunities in Asia. Front Plant Sci. 2022;13:925548. Published 2022 Oct 10. doi:10.3389/fpls.2022.925548

Dhaanya

ORIGINALITY REPORT

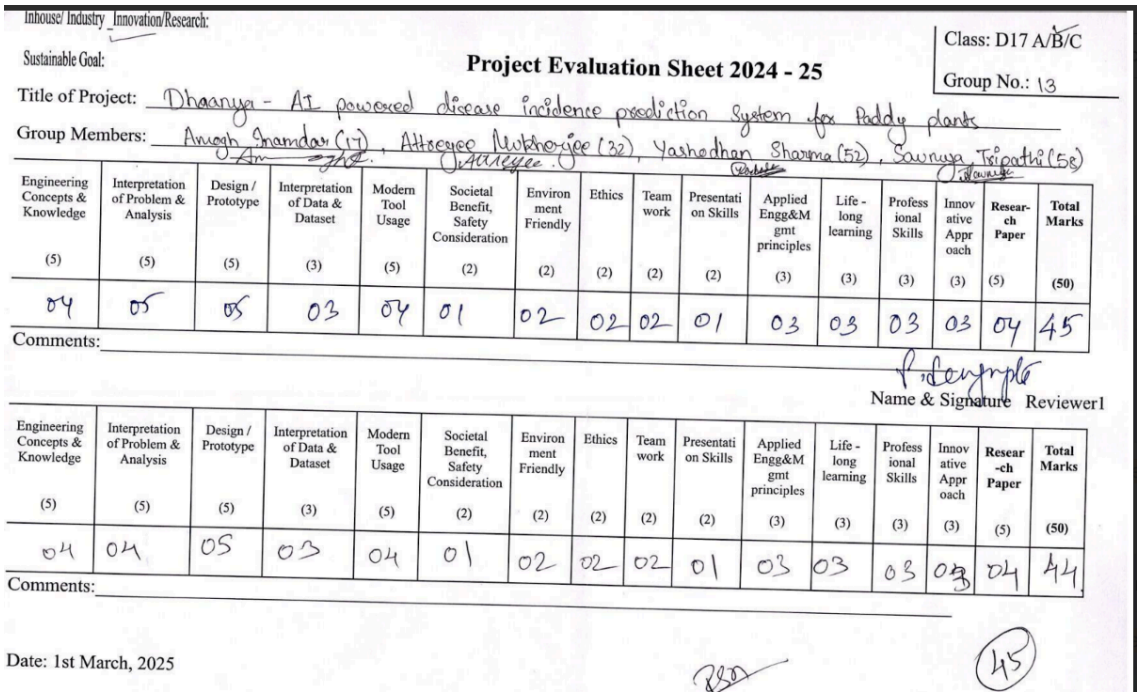
1%	1%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	1%
2	Wenqin Fang, Xiaoyu Zai, Jia Chen, Yakubu Saddeeq Abubakar et al. "The penetration ring is a novel infection structure formed by the penetration peg for invading plant cell membrane in rice blast fungus", eLife Sciences Publications, Ltd, 2024 Publication	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		

Review 1 Sheet



Inhouse/ Industry _Innovation/Research:

Sustainable Goal:

Project Evaluation Sheet 2024 - 25

Class: D17 A/B/C

Group No.: 13

Title of Project: Dhaanya - AI powered disease incidence prediction System for Paddy plants

Group Members: Anugh Gnamdar (17), Atreyee Mukherjee (32), Yashodhan Sharma (52), Saumya Tripathi (58)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
04	05	05	03	04	01	02	02	02	01	03	03	03	03	04	45

Comments:

Name & Signature Reviewer1

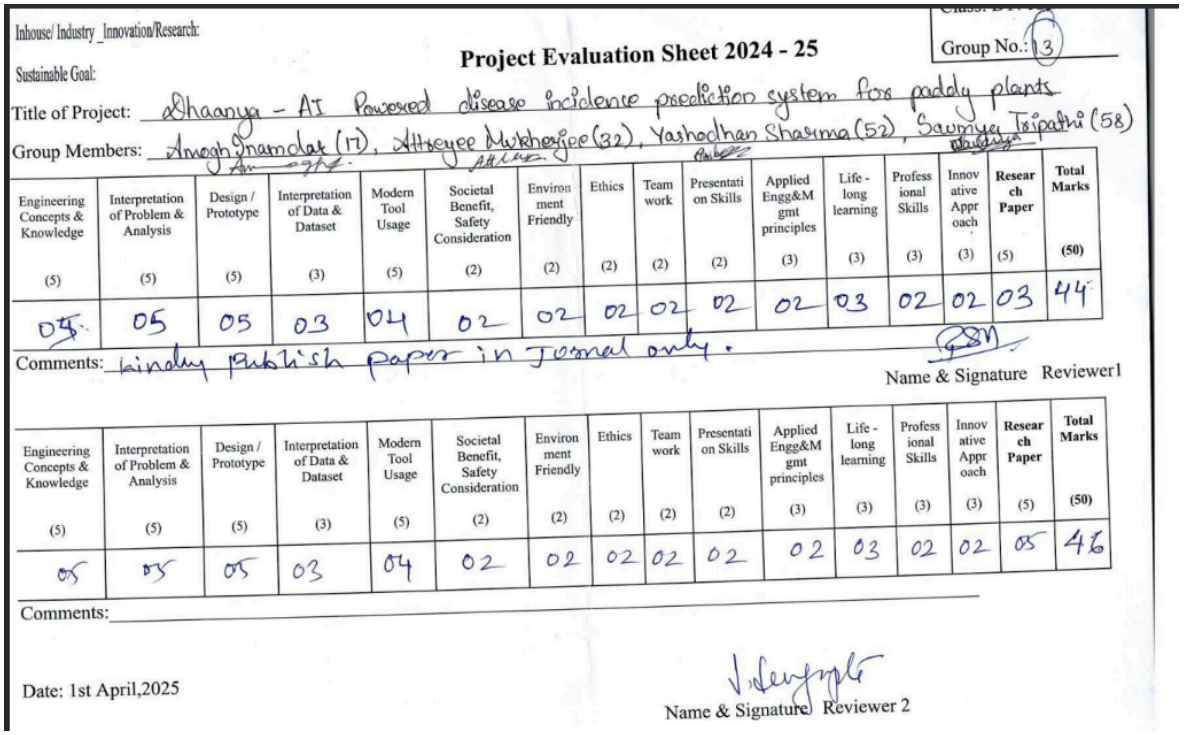
Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
04	04	05	03	04	01	02	02	02	01	03	03	03	03	04	44

Comments:

Date: 1st March, 2025

45

Review 2 Sheet



Inhouse/ Industry _Innovation/Research:

Sustainable Goal:

Project Evaluation Sheet 2024 - 25

Group No.: 13

Title of Project: Dhaanya - AI Powered disease incidence prediction system for paddy plants

Group Members: Anugh Gnamdar (17), Atreyee Mukherjee (32), Yashodhan Sharma (52), Saumya Tripathi (58)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
05	05	05	03	04	02	02	02	02	02	02	03	02	02	03	44

Comments: kindly publish paper in Jornal only.

Name & Signature Reviewer1

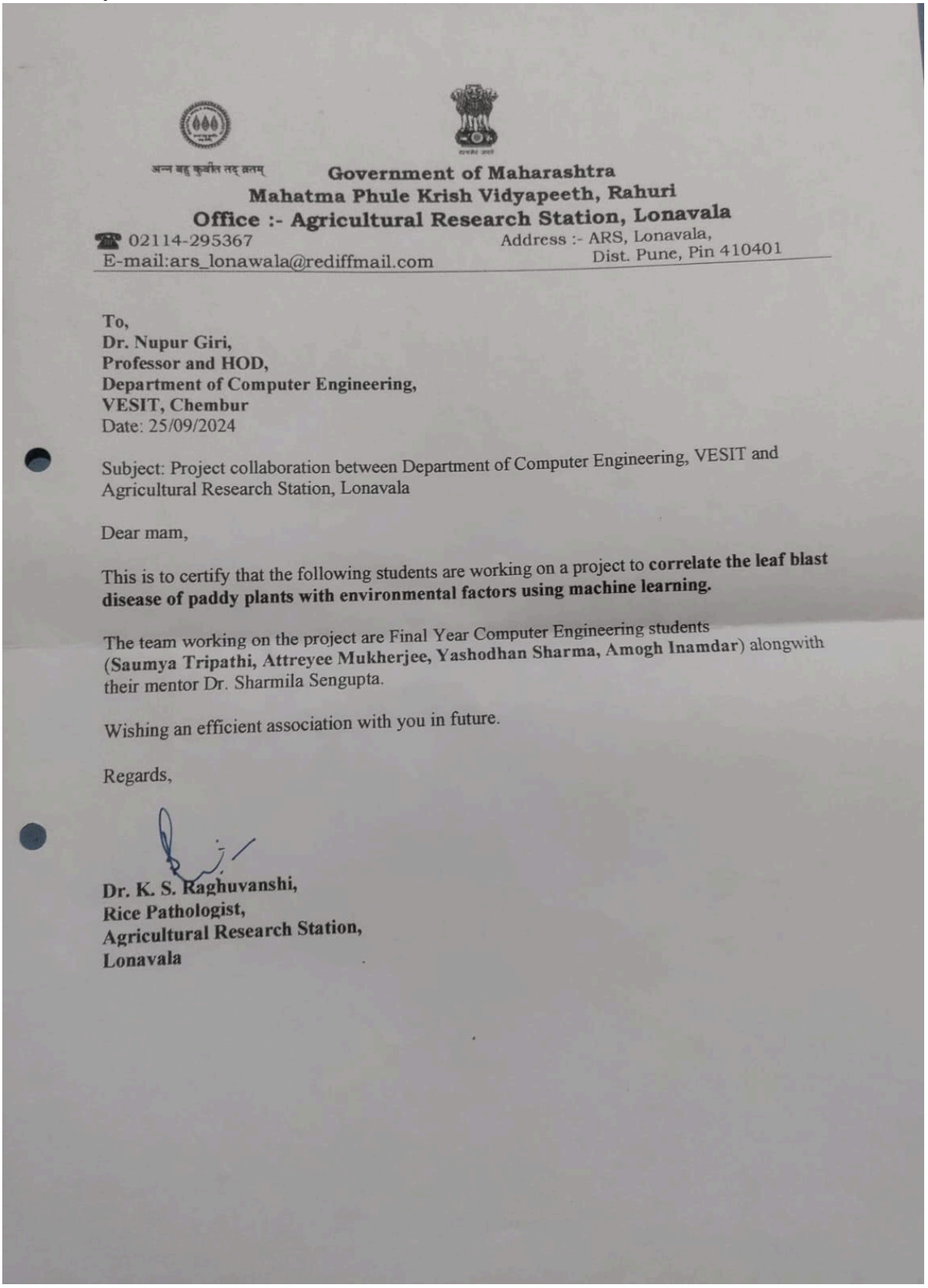
Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
05	05	05	03	04	02	02	02	02	02	02	03	02	02	05	46

Comments:

Date: 1st April,2025

Name & Signature Reviewer 2

2. Industry Certificates



अन्न बहु कुर्वीत तद् व्रतम्

Government of Maharashtra
Mahatma Phule Krish Vidyapeeth, Rahuri
Office :- Agricultural Research Station, Lonavala

☎ 02114-295367
E-mail:ars_lonawala@rediffmail.com

Address :- ARS, Lonavala,
Dist. Pune, Pin 410401

To,
Dr. Nupur Giri,
Professor and HOD,
Department of Computer Engineering,
VESIT, Chembur
Date: 25/09/2024

Subject: Project collaboration between Department of Computer Engineering, VESIT and Agricultural Research Station, Lonavala

Dear mam,

This is to certify that the following students are working on a project to **correlate the leaf blast disease of paddy plants with environmental factors using machine learning.**

The team working on the project are Final Year Computer Engineering students (**Saumya Tripathi, Attreyee Mukherjee, Yashodhan Sharma, Amogh Inamdar**) alongwith their mentor Dr. Sharmila Sengupta.

Wishing an efficient association with you in future.

Regards,

Dr. K. S. Raghuvanshi,
Rice Pathologist,
Agricultural Research Station,
Lonavala

अन्न बहु कुर्वीत तद् व्रतम्

Government of Maharashtra
Mahatma Phule Krish Vidyapeeth, Rahuri
Office :- Agricultural Research Station, Lonavala
☎ 02114-295367
E-mail:ars_lonawala@rediffmail.com
Address :- ARS, Lonavala,
Dist. Pune, Pin 410401

To,
Dr. Nupur Giri,
Professor and HOD,
Department of Computer Engineering,
VESIT, Chembur

Date: 28/02/2025

Subject: Completion of project collaboration between Department of Computer Engineering, VESIT and Agricultural Research Station, Lonavala

Dear Ma'am,

This is to certify that the project on **Correlation of Diseases in Paddy Plants with Environmental Factors Using Machine Learning** has been successfully completed by the Final Year Computer Engineering students **Saumya Tripathi, Attreyee Mukherjee, Yashodhan Sharma, and Amogh Inamdar** under the mentorship of **Dr. Sharmila Sengupta.**

The project has met its objectives and has contributed valuable insights into the relationship between environmental factors and the occurrence of different types of diseases in paddy crops. The collaboration between VESIT and the Agricultural Research Station, Lonavala, has been highly productive, and we appreciate the dedication and efforts of the students and faculty involved.

We look forward to future collaborations and wish the students success in their careers.

Regards,

Dr. K. S. Raghuwanshi,
Rice Pathologist,
Agricultural Research Station,
Lonavala