A

Industry Oriented Mini Project Report

on

IMAGE CLASSIFICATION OF RICE LEAF DISEASES USING RANDOM FOREST ALGORITHM

(Submitted in partial fulfilment of the requirements for the award of Degree)

Bachelor of Technology

in

COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

By

P. STEPHEN KAMAL (217R1A6749)

Under the guidance of

Mr.B.MOHAN BABU

Associate professor

Department of Computer Science & Engineering (Data Science)

CMR TECHNICAL CAMPUS

UGC AUTONOMOUS

Accredited by NBA & NAAC with 'A' Grade

Approved by AICTE, New Delhi and JNTU, Hyderabad.

2024-2025

Department of Computer Science & Engineering

(Data Science)

CERTIFICATE

This is to certify that the project entitled **"IMAGE CLASSIFICATION OF RICE LEAF DISEASES USING RANDOM FOREST ALGORITHM"** being submitted by **P.STEPHEN KAMAL (217R1A6749)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering (Data Science) to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by them under our guidance and supervision during the year 2024-25.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Mr. B.MOHAN BABU

Associate professor

.

Internal Guide

Dr. K. Murali

Associate Professor
&HoD

External Examiner

Submitted for viva-voce Examination held on: _____

ACKNOWLEDGMENT

Apart from the efforts of myself, the success of any project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

I take this opportunity to express my profound gratitude and deep regard to my guide **Mr.B.MOHAN BABU**, Associate professor for his exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by him shall carry me a long way in the journey of life on which I am about to embark.

I also take this opportunity to express a deep sense of gratitude to the Project Review Committee (PRC) **Mr. Sanjib Kumar Nayak, Mrs. V. Prema Tulasi, Mr. B. Ramesh, Mr. A. Veerender, Mrs. V. Sandya, Ms. Sandhyarani** for their cordial support, valuable information and guidance, which helped me in completing this task through various stages.

I am also thankful to **Dr. K. Murali**, Head, Department of Computer Science and Engineering (Data Science) for providing encouragement and support for completing this project successfully.

I am also obliged to **Dr. A. Raji Reddy**, Director for being cooperative throughout the course of this project. I am also express my sincere gratitude to **Sri. Ch. Gopal Reddy**, Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. I am grateful for their constant support and help.

Finally, I would like to take this opportunity to thank my family for their constant encouragement, without which this assignment would not be completed. I sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

P. STEPHEN KAMAL **(217R1A6749)**

ABSTRACT

Rice, being a crucial staple for over half of the global population, is highly susceptible to various diseases that can significantly affect yield and food security. This study presents an effective approach for the classification of rice leaf diseases using the Random Forest algorithm, a powerful ensemble learning technique renowned for its ability to manage high-dimensional data and provide reliable predictions.In this research, we collected a diverse dataset of high-resolution images depicting both healthy and diseased rice leaves, encompassing various disease categories such as leaf blast, brown spot, and bacterial blight. The images underwent rigorous preprocessing, including normalization, noise reduction, and feature extraction, to enhance the model's ability to discern subtle differences between classes.The Random Forest model was implemented with a series of hyperparameter tuning and cross-validation to optimize its performance. Data augmentation techniques, such as rotation and scaling, were employed to artificially expand the training dataset and improve the model's generalization capabilities. The model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score, assessed on an independent test set.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SCREENSHOTS

LIST OF ABBREVIATIONS

i. RF - Random Forest

ii. ML - Machine Learning

iii. DL - Deep Learning

iv. CNN - Convolutional Neural Network

INTRODUCTION

1.1 PROJECT SCOPE

The **project scope** of image classification for rice leaf diseases using the Random Forest algorithm involves developing a machine learning model to accurately identify and classify different diseases affecting rice leaves. This classification system can help in early detection and management of crop diseases, ultimately improving yield and reducing losses. The project focuses on automating the diagnosis process, allowing farmers and agronomists to detect diseases from rice leaf images efficiently.

The first step involves collecting a robust dataset of rice leaf images that represent various diseases such as brown spot, bacterial blight, and leaf smut, as well as healthy leaves. These images must be labeled to correspond to the specific disease type for supervised learning. After dataset collection, the next phase is preprocessing the images to prepare them for the model. This includes resizing images to a uniform size, applying data augmentation to increase variability, normalizing pixel values, and encoding disease labels into a numerical format.

Once the data is ready, feature extraction techniques can be applied to highlight important characteristics of the images, such as color patterns, texture, or shapes. The Random Forest algorithm, which is an ensemble learning method, will then be trained using these extracted features. Random Forest works by building multiple decision trees and merging them to improve the model's accuracy and robustness. Finally, the model will be evaluated on a test dataset to determine its classification performance. The project scope includes fine-tuning the model, improving its accuracy, and ensuring it generalizes well to new, unseen images.

1.2 PROJECT PURPOSE

The purpose of the project is to develop an efficient and accurate system for classifying rice leaf diseases using image data and the Random Forest algorithm. Early and precise detection of diseases like brown spot, bacterial blight, and leaf smut is crucial for minimizing the damage to crops and improving agricultural productivity. Traditional methods of disease detection involve manual inspection, which can be time-consuming,

prone to human error, and difficult to scale across large fields. By leveraging machine learning, specifically the Random Forest algorithm, this project aims to automate the diagnosis process, providing a faster and more reliable tool for farmers and agricultural experts.This project not only addresses the need for real-time disease detection but also helps in improving decision-making for crop management, ultimately leading to increased yields and reduced crop losses. Moreover, it can contribute to sustainable agricultural practices by allowing timely interventions and minimizing unnecessary pesticide use, as only affected plants can be treated. The system, once implemented, has the potential to be integrated into mobile applications, providing farmers with a user-friendly tool for disease detection right in the field.

In addition to improving detection accuracy and reducing human error, the project serves the broader purpose of promoting **data-driven agriculture**. With the increasing challenges posed by climate change, pests, and diseases, there is a growing need for technologies that can enhance precision farming. This project aims to support **precision agriculture** by enabling the early detection of rice leaf diseases, allowing for targeted interventions and better management of resources such as water, fertilizers, and pesticides. By detecting diseases at an early stage, farmers can take immediate action, reducing the spread of infections and preventing major crop damage.

1.3 PROJECT FEATURES

The project offers several features that make it a powerful tool for rice leaf disease classification. One of the primary features is **automated image classification**, which allows users to upload rice leaf images and receive an instant diagnosis of whether the leaf is healthy or diseased. The system can identify common rice leaf diseases such as brown spot, bacterial blight, and leaf smut. This automation replaces the need for manual inspection by experts, making the detection process faster, more accessible, and less prone to error.

At the heart of the project is the **Random Forest classifier**, a robust machine learning algorithm that improves accuracy by creating multiple decision trees and combining their results. This ensemble method increases the system's reliability, allowing it to perform well even when presented with diverse and noisy image data. The use of Random Forest reduces

overfitting and ensures better performance when applied to new, unseen images, which is critical for a real-world agricultural application.Another key feature is **feature extraction**, which captures essential image characteristics like texture, color, and shape. These extracted features are crucial for helping the Random Forest algorithm differentiate between various disease types. By focusing on significant visual patterns, the system can effectively classify diseases even in images with slight variations in lighting or angle, further enhancing its practical applicability.

To maximize model accuracy, the project includes **data preprocessing and augmentation**. Techniques like image resizing, normalization, and data augmentation (e.g., rotating, zooming, or flipping images) ensure that the model can handle a wide range of input data. Augmentation artificially increases the size and diversity of the training dataset, improving the model's ability to generalize across different conditions.The system also provides **performance evaluation metrics** such as accuracy, precision, recall, and F1-score to assess the model's effectiveness. These metrics offer insights into how well the model performs in real-time disease detection, helping developers fine-tune the system before deployment.

SYSTEM ANALYSIS

2. SYSTEM ANALYSIS

The system analysis for the rice leaf disease classification project focuses on understanding the functionality, performance, and technical requirements needed to ensure the system meets its goals. The system's primary objective is to classify rice leaf diseases accurately, using the Random Forest algorithm, based on image data. A detailed analysis of the system's components, data flow, and performance metrics is essential to ensure a robust, efficient, and user-friendly solution.

2.1 PROBLEM DEFINITION

The **problem definition** for the rice leaf disease classification project centers around the challenge of timely and accurate detection of rice crop diseases. Rice is a staple food for much of the world's population, and its cultivation is vulnerable to a variety of leaf diseases, such as brown spot, bacterial blight, and leaf smut. These diseases can lead to significant yield losses if not detected and treated promptly. However, traditional methods of disease detection rely heavily on manual inspection by farmers or agricultural experts, which can be time-consuming, labor-intensive, and prone to human error. This creates the need for an automated, efficient, and accurate method of disease detection that can operate at scale.

2.2 EXISTING SYSTEM

The existing systems for detecting rice leaf diseases are largely based on **manual inspection** and **traditional agricultural practices**, which rely heavily on the expertise of farmers or agricultural specialists. In many regions, farmers visually inspect rice leaves for symptoms of diseases, such as discoloration, spots, or unusual growth patterns. While experienced farmers may be able to identify some common diseases, the process is time-consuming, prone to human error, and lacks consistency. This manual method often leads to delayed detection, particularly when diseases are in their early stages and symptoms are subtle. Additionally, human error in diagnosing similar-looking diseases can result in incorrect treatments, further compounding the problem.

In addition to manual inspection, some farmers and agricultural organizations may use **laboratory-based testing** or consult with agronomists for a more accurate diagnosis. However, these approaches can be expensive and inaccessible, particularly in rural or

developing regions where farmers may not have the resources to regularly seek professional advice. Moreover, lab testing is not feasible for large-scale or real-time monitoring of fields, which limits its application in widespread disease detection. This reliance on expert opinion and laboratory testing also introduces delays in diagnosis, preventing timely interventions to control disease spread.

2.2.1 DISADVANTAGES OF EXISTING SYSTEM

Following are the disadvantages of existing system:

- **Manual Inspection Limitations**: The most common method of disease detection relies on manual inspection by farmers or agricultural experts. This process is inherently time-consuming and labor-intensive, requiring significant time and effort to visually assess large rice fields. Additionally, it is prone to human error, as inexperienced farmers may misidentify diseases or overlook early symptoms, resulting in delayed interventions and potential crop damage.

- **Inaccessibility of Expert Consultation**: While consulting with agronomists or agricultural specialists can provide more accurate diagnoses, this option is often not accessible for small-scale farmers. In rural areas, there may be a shortage of trained professionals, making it difficult for farmers to seek timely expert advice. Moreover, laboratory testing and expert consultations can be costly, further limiting access for resource-constrained farmers.

- **Complexity of Advanced Tools**: Some digital tools and systems, such as drone technology or specialized mobile applications, require a certain level of technical expertise to operate effectively. These systems may also involve steep learning curves and high costs for both hardware and software, making them impractical for smallholder farmers who need simple, easy-to-use solutions for immediate disease detection.

2.3 PROPOSED SYSTEM

The proposed system for rice leaf disease classification aims to address the limitations of existing methods by leveraging advanced machine learning techniques, specifically the Random Forest algorithm, to automate and improve the accuracy of disease detection. This

system is designed to be user-friendly, efficient, and accessible, enabling farmers, especially those in rural areas, to diagnose diseases quickly and effectively, thereby enhancing their crop management practices.

Automated Image Classification: At the core of the proposed system is an automated image classification feature that allows users to upload images of rice leaves. The system will analyze these images to determine whether the leaves are healthy or diseased and classify the specific type of disease present. By utilizing a machine learning model, the system eliminates the need for manual inspection, providing quick and accurate results that facilitate timely interventions.

2.3.1 ADVANTAGES OF THE PROPOSED SYSTEM.

1. **Rapid and Accurate Diagnosis**: The proposed system automates the disease classification process, enabling farmers to receive quick and accurate diagnoses of rice leaf diseases. By utilizing the Random Forest algorithm, which is robust and effective in handling complex datasets, the system can deliver reliable results that facilitate timely intervention and treatment. This rapid diagnosis helps prevent the spread of diseases and minimizes crop losses.
2. **User-Friendly Interface**: Designed with accessibility in mind, the system features an intuitive user interface that allows users with minimal technical knowledge to navigate and operate the platform easily. Farmers can upload images of their rice leaves and receive instant feedback.
3. **Comprehensive Disease Coverage**: The proposed system aims to classify a wide range of rice leaf diseases, making it a versatile tool for farmers. By continuously updating the model with new data and disease categories, the system can adapt to emerging agricultural challenges. This comprehensive approach empowers farmers to manage their crops more effectively by providing information on multiple diseases in one platform.

2.4 FEASIBILITY STUDY

2.4.1 ECONOMICAL FEASIBILITY

The economical feasibility of the proposed rice leaf disease classification system is essential for assessing its financial viability and potential return on investment for farmers

and agricultural stakeholders. This evaluation examines the costs associated with implementing the system, the expected savings and benefits, and the overall impact on agricultural productivity.One of the primary advantages of the proposed system is its cost-effectiveness compared to traditional methods of disease detection. By automating the disease classification process, farmers can significantly reduce their reliance on manual inspections and costly expert consultations. This system provides a low-cost alternative that empowers farmers to diagnose diseases independently, minimizing expenses related to laboratory testing or hiring agricultural specialists. This financial accessibility is crucial for smallholder farmers, who often operate on tight budgets and require efficient solutions to protect their crops.

Timely and accurate disease detection is critical for preventing crop losses. The proposed system allows farmers to identify and treat diseases early, which can mitigate the risks of extensive damage to rice crops. The potential savings from reduced crop losses can far exceed the initial investment in the system, presenting a compelling economic incentive for farmers to adopt the technology. By facilitating early intervention, the system can lead to increased yield and profitability, thus enhancing the overall economic viability of rice farming.

2.4.2 TECHNICAL FEASIBILITY

The technical feasibility of the proposed rice leaf disease classification system assesses whether the necessary technology, resources, and infrastructure are available to support its development and implementation. This evaluation examines the suitability of the chosen machine learning algorithms, the required hardware and software, data management capabilities, and the system's integration into existing agricultural practices.

At the heart of the proposed system is the Random Forest algorithm, which is well-suited for image classification tasks and has demonstrated strong performance in various agricultural applications. This machine learning technique is known for its robustness, ability to handle large datasets, and effectiveness in managing noise and variability in data. Utilizing Random Forest allows for the development of an accurate and reliable model for detecting rice leaf diseases, making it a technically sound choice for this application.The proposed system requires a suitable infrastructure to support its operation, including hardware for data processing and storage. The system can be developed using common programming languages

such as Python or R, which have extensive libraries and frameworks for machine learning and image processing. This accessibility enables the development team to leverage existing tools and resources, reducing the time and cost associated with building the system from scratch. Furthermore, the use of cloud-based services can facilitate scalability and provide the necessary computational power for model training and inference, making the system accessible to users with varying hardware capabilities.

2.4.3 SOCIAL FEASIBILITY

The social feasibility of the proposed rice leaf disease classification system evaluates its potential impact on the community, stakeholders, and the broader agricultural landscape. This assessment considers how the system will be received by users, its influence on agricultural practices, and its contributions to social and economic well-being.One of the primary social benefits of the proposed system is its potential to empower smallholder farmers by providing them with an accessible and reliable tool for disease detection. In many rural communities, farmers often face challenges related to limited access to expert advice and agricultural resources. By enabling farmers to diagnose rice leaf diseases independently, the system fosters self-reliance and confidence in their ability to manage crop health. This empowerment can lead to improved agricultural practices, greater productivity, and ultimately enhanced livelihoods for farmers and their families.Moreover, the proposed system promotes knowledge sharing and collaboration among farmers. As users engage with the platform and gain insights into disease management, they may share their experiences and findings with fellow farmers.

2.5 HARDWARE & SOFTWARE REQUIREMENTS

2.5.1 SOFTWARE REQUIREMENTS

Operating system: Windows 10

Coding Language: Python

Front-End: HTML , CSS

Back-End: Python Database: SQLite

2.5.2 HARDWARE REQUIREMENTS

System Processor: Intel core i3

Hard Disk: 512 GB

RAM: 8 GB

ARCHITECTURE

3.1 PROJECT ARCHITECTURE

The architecture for classifying rice leaf diseases using the Random Forest algorithm involves several key components. First, the dataset comprises images of rice leaves, which are categorized into different classes based on disease types and healthy leaves. The images undergo preprocessing, including resizing to a uniform dimension (e.g., 128x128 pixels) and flattening into a two-dimensional array to facilitate input into the model. The data is then split into training and testing sets to evaluate model performance. A Random Forest classifier, which consists of multiple decision trees trained on random subsets of the data, is utilized for classification. Each tree provides a prediction, and the final classification is determined by majority voting among the trees.

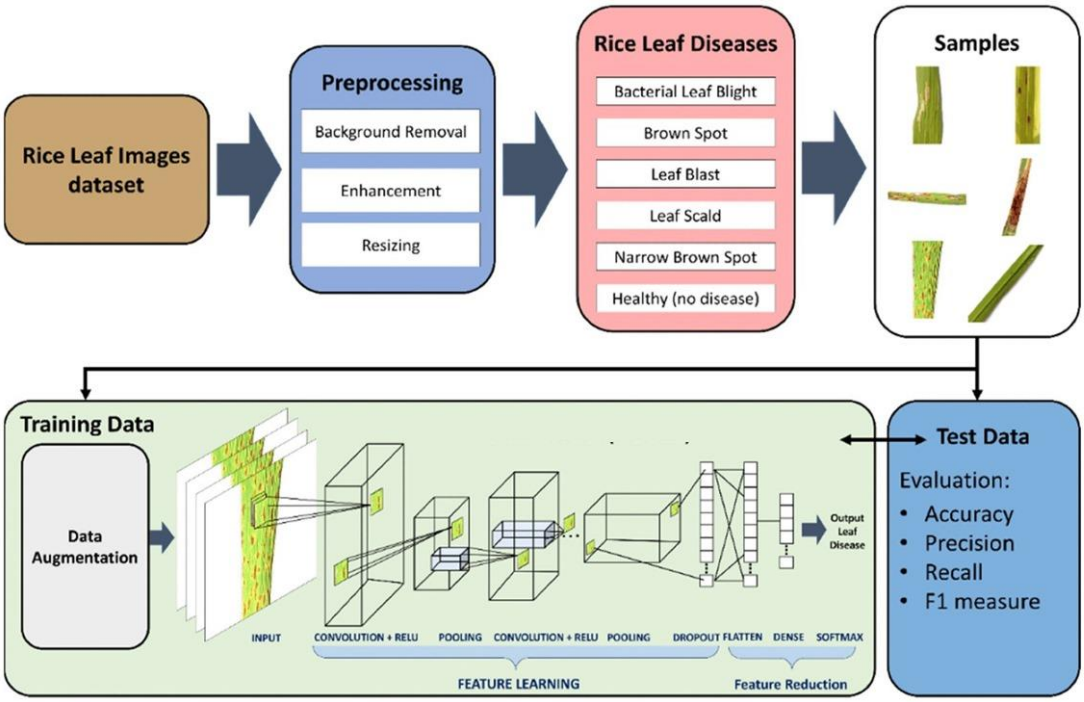

Fig 3.1 Architecture

3.2 UML DIAGRAMS

3.2.1 USE CASE DIAGRAM

The use case diagram for the rice leaf disease classification system features key actors and interactions. The primary actor is the **User**, which can include farmers, agronomists, or researchers interested in identifying rice leaf diseases. The main use cases include **Upload Images**, where users input images of rice leaves for analysis; **Preprocess Images**, where the system resizes and normalizes the uploaded images; and **Train Model**, which involves training a Random Forest model on a labeled dataset. Once the model is trained, users can engage with the **Classify Images** use case

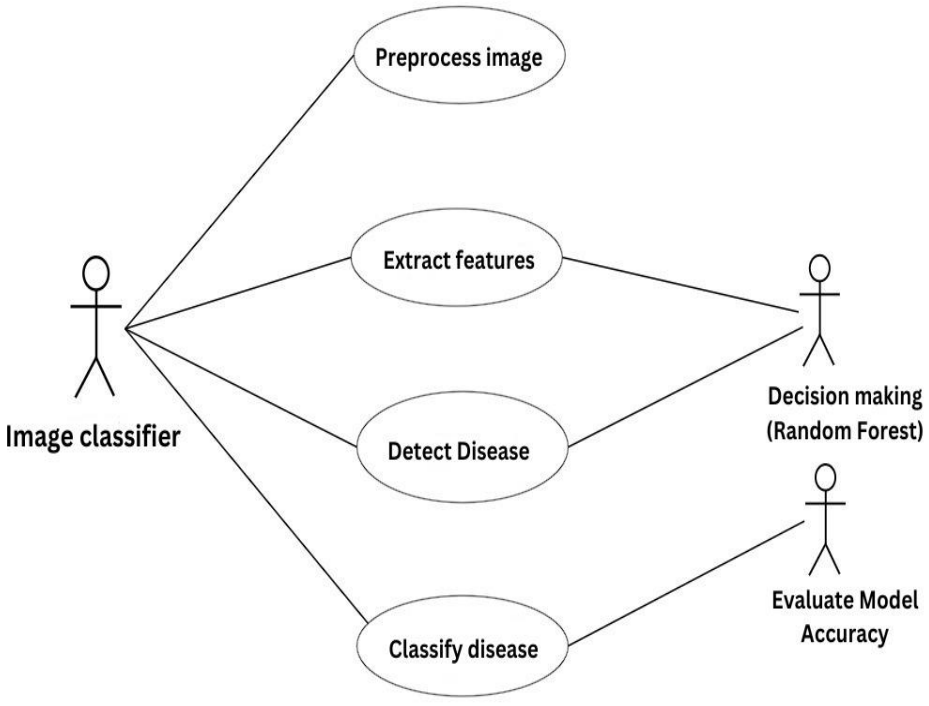

Fig 3.2.1 Use case diagram

3.2.2 CLASS DIAGRAM

The use case diagram for the rice leaf disease classification system features key actors and interactions. The primary actor is the **User**, which can include farmers, agronomists, or researchers interested in identifying rice leaf diseases. The main use cases include **Upload Images**

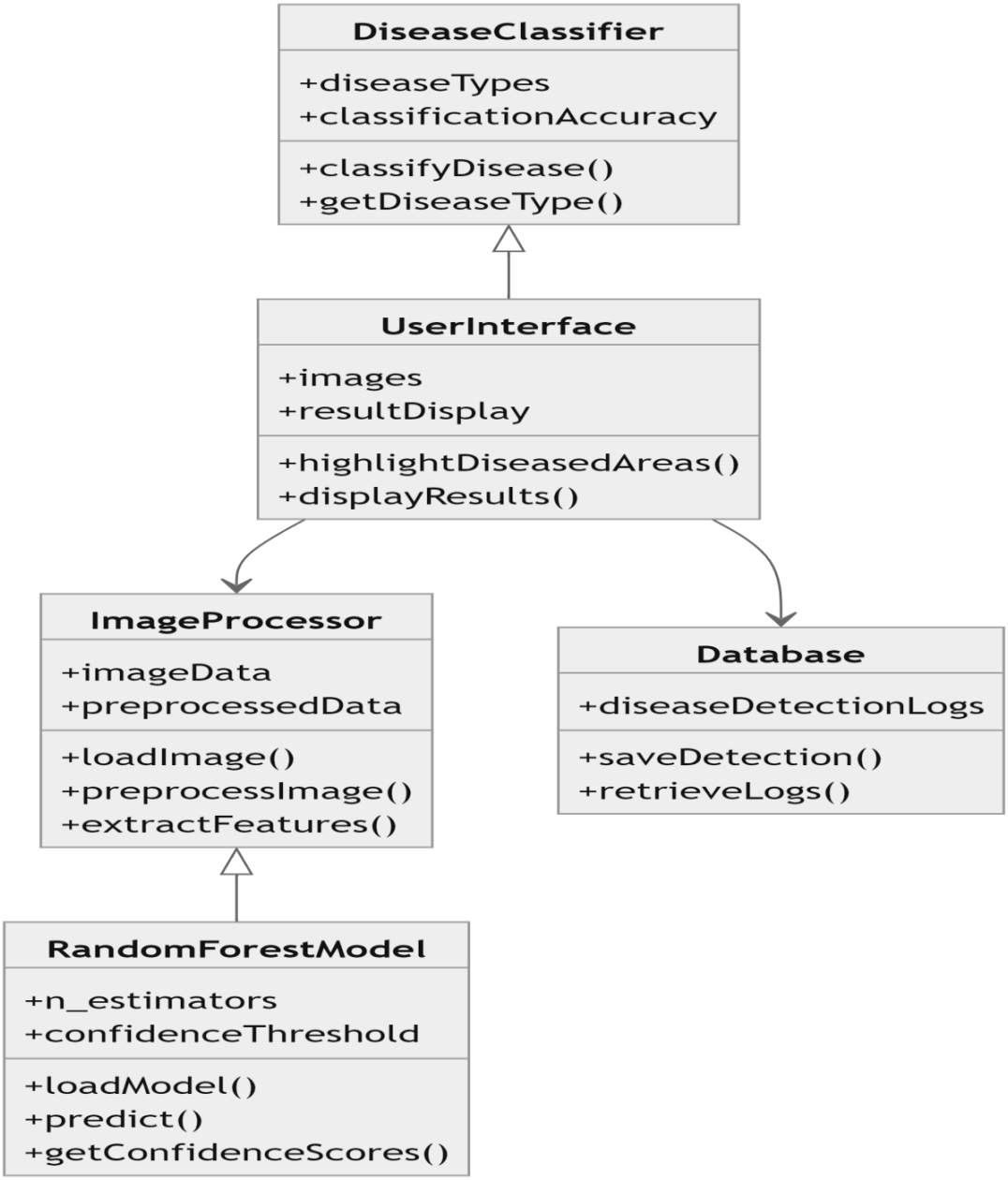

Fig 3.2.2 class diagram

3.2.3 SEQUENCE DIAGRAM

The sequence diagram for the rice leaf disease classification system illustrates the interaction between the user and the system during the classification process. It begins with the **User** initiating the **Upload Images** action, which triggers the **Classifier** class to call the `load_image()` method from the **ImageProcessor**

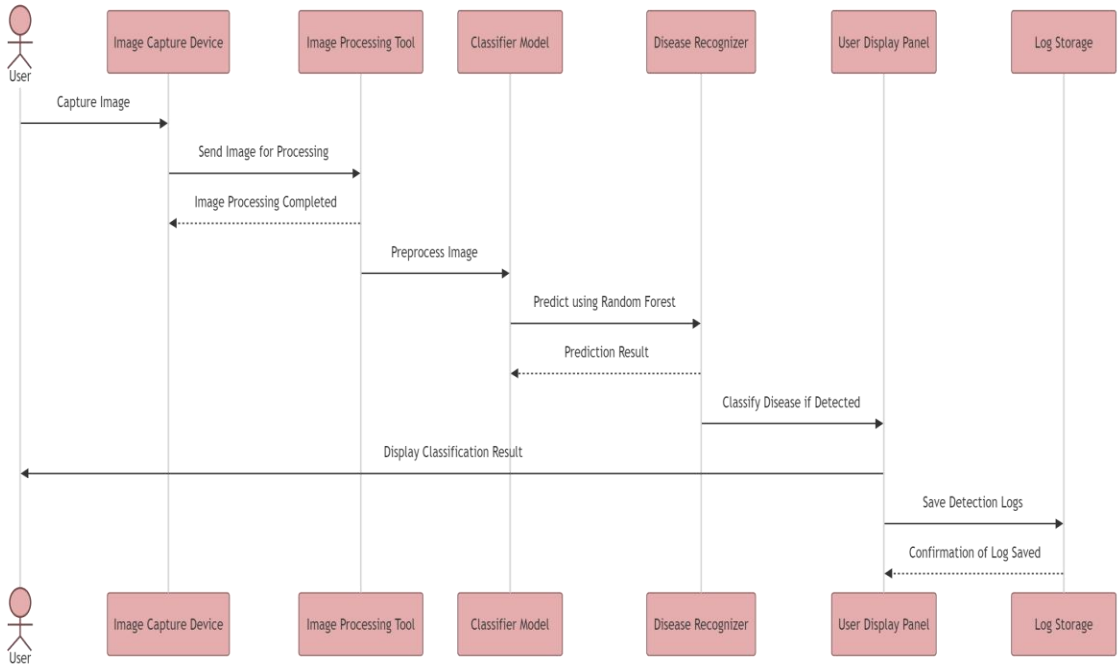

Fig 3.2.3 Sequence diagram

3.2.4 ACTIVITY DIAGRAM

The activity diagram for the rice leaf disease classification system outlines the workflow of the image classification process. It begins with the **User** initiating the **Upload Images** activity, followed by the system loading and preprocessing the images, which includes resizing and normalization steps

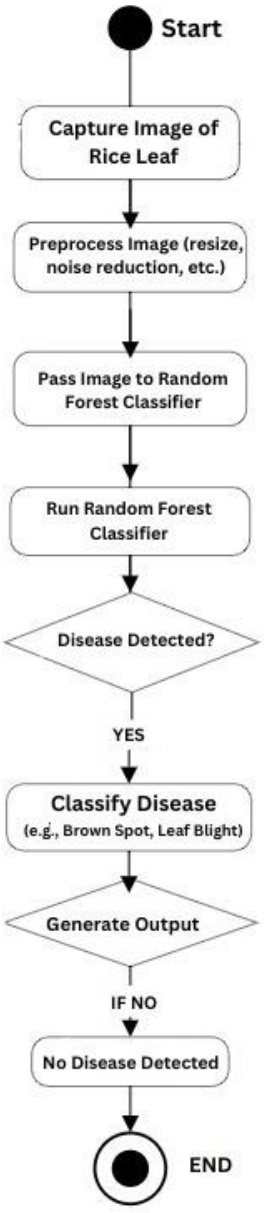

Fig 3.2.4 Activity diagram

IMPLEMENTATION

5.1 MACHINE LEARNING ALGORITHMS

RANDOM FOREST ALGORITHM

Random Forest is an ensemble learning technique that builds multiple decision trees to improve predictive accuracy and control overfitting. It works by creating a "forest" of trees, each trained on different subsets of the data and using random subsets of features at each split. This randomness helps ensure that the trees are diverse and captures various patterns in the data, ultimately leading to a more robust and generalized model. The final output is determined by aggregating the predictions from all individual trees, which mitigates the risk of relying on a single, potentially flawed tree.One of the key advantages of Random Forest is its ability to handle large datasets with numerous features without requiring extensive preprocessing. It can manage both numerical and categorical data, making it highly versatile across different types of problems. Additionally, the algorithm is inherently resistant to overfitting due to the averaging process, which smooths out the noise present in the training data. This makes it particularly useful in real-world scenarios where data can be noisy or incomplete.

Random Forest also provides valuable insights into feature importance, which helps in understanding the underlying data better. By measuring how much each feature contributes to the reduction of impurity in the trees, users can identify the most significant variables affecting the target outcome. This not only aids in model interpretation but can also inform feature selection, guiding practitioners on which variables to focus on for further analysis or model refinement.Despite its many strengths, Random Forest is not without limitations. For instance, it can be computationally intensive, especially with large datasets and a high number of trees. The model may also become less interpretable as the number of trees increases, making it harder to explain individual predictions. Additionally, while it performs well in many scenarios, it may not always be the best choice for every problem, particularly those requiring deep understanding or complex interactions between features. Nonetheless, its combination of accuracy, flexibility, and ease of use makes Random Forest a favored algorithm in both academic research and practical applications.

CNN ALGORITHM

Convolutional Neural Networks (CNNs) are a specialized type of deep learning architecture primarily designed for processing structured grid data, such as images. They leverage convolutional layers to automatically and adaptively learn spatial hierarchies of features from input images. Unlike traditional machine learning models, which require manual feature extraction, CNNs learn to identify patterns, edges, and textures directly from raw pixel data. This ability to capture complex spatial relationships makes CNNs particularly effective for tasks such as image classification, object detection, and image segmentation.A typical CNN consists of several layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply a set of filters to the input image, creating feature maps that highlight important features while reducing dimensionality. Pooling layers further downsample these feature maps, helping to reduce the computational load and increase the model's invariance to translations. Finally, fully connected layers take the high-level features extracted from the convolutional layers and make the final predictions. This layered architecture allows CNNs to learn increasingly abstract representations of the input data.

One of the significant advantages of CNNs is their ability to generalize well to new, unseen data. This is achieved through techniques such as dropout and data augmentation, which help prevent overfitting during training. CNNs also benefit from transfer learning, where a model pre-trained on a large dataset can be fine-tuned on a smaller, domain-specific dataset. This capability greatly reduces the amount of labeled data required for training while still yielding high performance, making CNNs applicable in various fields, from medical imaging to autonomous vehicles.Despite their strengths, CNNs come with challenges. They can require substantial computational resources and large amounts of labeled data to train effectively. Additionally, the complexity of their architectures can lead to longer training times and difficulties in model interpretability. While CNNs excel at tasks involving visual data, they may not be the best choice for every type of data, particularly those where spatial relationships are less relevant. Nonetheless, their powerful feature extraction capabilities have made CNNs a cornerstone of modern computer vision applications and research.

5.2 DATA SETS

For classifying rice leaf diseases using the Random Forest algorithm, various datasets are available that provide labeled images of affected rice leaves. One prominent option is the **Rice Disease Dataset**, which specifically includes images categorized by different diseases, such as bacterial blight and leaf blast. This dataset is designed for multi-class classification, making it particularly suitable for machine learning applications focused on identifying specific diseases in rice.

Another valuable resource is the **PlantVillage Dataset**, which encompasses a wider range of plant diseases, including those affecting rice. Although it covers various plants, it offers a substantial number of images that can help in training robust classification models. The diversity of this dataset allows for better generalization and performance when identifying rice leaf diseases among other plant diseases.

Additionally, platforms like **Kaggle** host community-contributed datasets that focus on agricultural issues, including rice leaf diseases. Users can explore various competitions or search for datasets specifically targeting plant pathology. Furthermore, several universities and agricultural research organizations publish their datasets, which may include comprehensive collections of images. Utilizing these datasets, along with data augmentation techniques, can significantly enhance the model's performance in classifying rice leaf diseases.

For image classification of rice leaf diseases using the Random Forest algorithm, several datasets are available, both publicly and through specific research efforts. Notable datasets include the Rice Leaf Disease Dataset, which features labeled images of various diseases like bacterial blight and brown spot. The PlantVillage Dataset, available on platforms like Kaggle, offers a broader collection of plant disease images, including rice, allowing for diverse training scenarios. Additionally, custom datasets can be created by capturing local images of rice leaves in different agricultural settings, providing tailored data that reflects specific regional variations and disease manifestations.

When utilizing these datasets, it’s crucial to ensure high image quality and accurate annotations. Preprocessing steps, such as resizing and normalization, can enhance model performance. Random Forest can handle raw pixel values, but incorporating feature extraction techniques may further improve classification accuracy.

SOURCE CODE

```
from google.colab import drive

drive.mount('/content/drive')

# Import libraries

import pandas as pd

import numpy as np

import os

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.utils import shuffle

from tensorflow.keras.preprocessing.image import ImageDataGenerator

# Define image dimensions

IMAGE_HEIGHT = 224

IMAGE_WIDTH = 224

IMAGE_CHANNELS = 3

# List the directories for the datasets

leaf_smut_list = os.listdir('/content/drive/MyDrive/Colab Notebooks/Pembelajaran Mesin/ML_TugasKelompok/rice_leaf_diseases/Leaf smut')

brown_spot_list = os.listdir('/content/drive/MyDrive/Colab Notebooks/Pembelajaran Mesin/ML_TugasKelompok/rice_leaf_diseases/Brown spot')

bacterial_leaf_blight_list=os.listdir('/content/drive/MyDrive/Colab Notebooks/PembelajaranMesin/ML_TugasKelompok/rice_leaf_diseases/Bacterial leaf blight')

# Print the number of images in each category

print(len(leaf_smut_list))

print(len(brown_spot_list))

print(len(bacterial_leaf_blight_list))
```

```
# Create dataframes for each category

df_leaf_smut = pd.DataFrame(leaf_smut_list, columns=['image'])

df_leaf_smut['target'] = 'leaf_smut'

df_brown_spot = pd.DataFrame(brown_spot_list, columns=['image'])

df_brown_spot['target'] = 'brown_spot'

df_bacterial_leaf_blight=pd.DataFrame(bacterial_leaf_blight_list, columns=['image'])

df_bacterial_leaf_blight['target'] = 'bacterial_leaf_blight'

# Create a validation set for each class

df_leaf_smut_val = df_leaf_smut.sample(n=6, random_state=101)

df_brown_spot_val = df_brown_spot.sample(n=6, random_state=101)

df_bacterial_leaf_blight_val=df_bacterial_leaf_blight.sample(n=6, random_state=101)

# Print validation set sizes

print("Data Val")

print(len(df_leaf_smut_val))

print(len(df_brown_spot_val))

print(len(df_bacterial_leaf_blight_val))

# Create the train set for each class

val_list = list(df_leaf_smut_val['image'])

df_leaf_smut_train = df_leaf_smut[~df_leaf_smut['image'].isin(val_list)]

val_list = list(df_brown_spot_val['image'])

df_brown_spot_train = df_brown_spot[~df_brown_spot['image'].isin(val_list)]


val_list = list(df_bacterial_leaf_blight_val['image'])

df_bacterial_leaf_blight_train                                    =
df_bacterial_leaf_blight[~df_bacterial_leaf_blight['image'].isin(val_list)]

# Print train set sizes
```

```
print("Data Train")

print(len(df_leaf_smut_train))

print(len(df_brown_spot_train))

print(len(df_bacterial_leaf_blight_train))

# Create combined dataframes

df_data = pd.concat([df_leaf_smut, df_brown_spot, df_bacterial_leaf_blight], axis=0).reset_index(drop=True)

df_train = pd.concat([df_leaf_smut_train, df_brown_spot_train, df_bacterial_leaf_blight_train], axis=0).reset_index(drop=True)

df_val = pd.concat([df_leaf_smut_val, df_brown_spot_val, df_bacterial_leaf_blight_val], axis=0).reset_index(drop=True)

# Shuffle the data

df_data = shuffle(df_data)

df_train = shuffle(df_train)

df_val = shuffle(df_val)

# Print shapes of the dataframes

print(df_data.shape)

print(df_train.shape)

print(df_val.shape)

# Print value counts for targets

print(df_data['target'].value_counts())

print(df_train['target'].value_counts())

print(df_val['target'].value_counts())

# Combine val and train sets

val_len = len(df_val)

train_len = len(df_train)

df_combined = pd.concat(objs=[df_val, df_train], axis=0).reset_index(drop=True)
```

```
# Create dummy variables for the target
df_combined = pd.get_dummies(df_combined, columns=['target'])
# Separate the train and val sets
df_val = df_combined[:val_len]
df_train = df_combined[val_len:]
# Print final shapes
print(df_train.shape)
print(df_val.shape)
# Display the head of the dataframes
print(df_combined.head())
print(df_train.head())
print(df_val.head())


# Save dataframes to CSV
df_combined.to_csv('df_combined.csv.gz', compression='gzip', index=False)
df_train.to_csv('df_train.csv.gz', compression='gzip', index=False)
df_val.to_csv('df_val.csv.gz', compression='gzip', index=False)
```

SCREENSHOTS

5.1 HOME PAGE

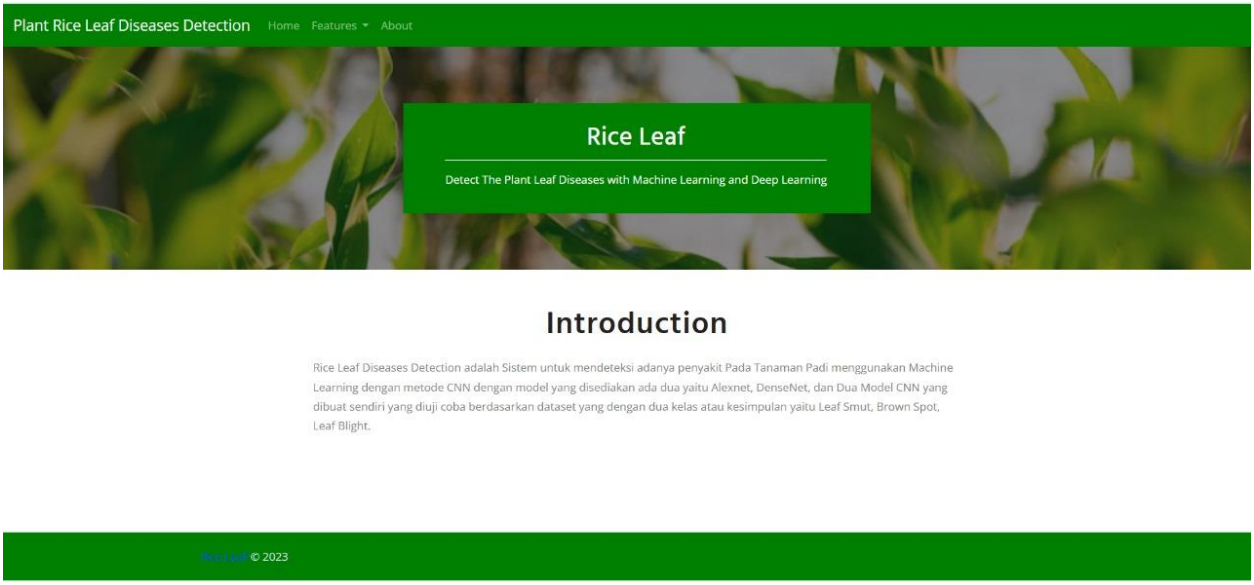

Fig.5.1 Initial home page

5.2 FEATURES

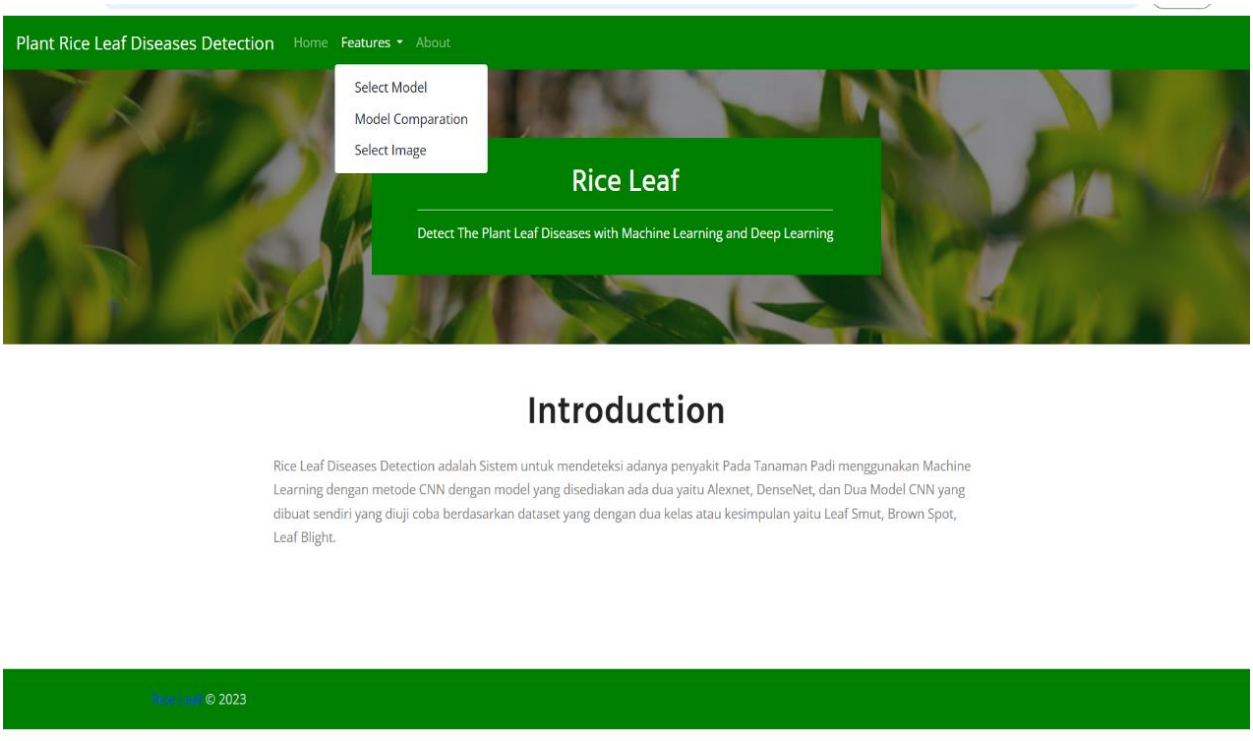

Fig.5.2 Features

5.3 SELECTION OF MODEL

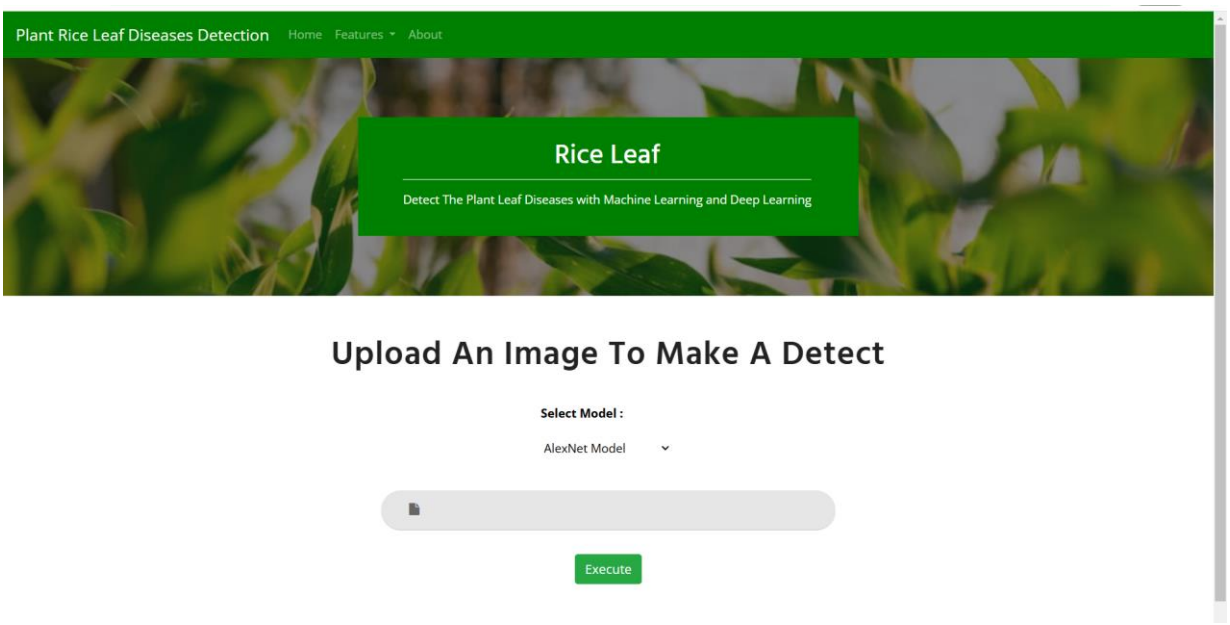

Fig 5.3 Selecting and ready to insert

5.4 SELECTION OF DATASETS

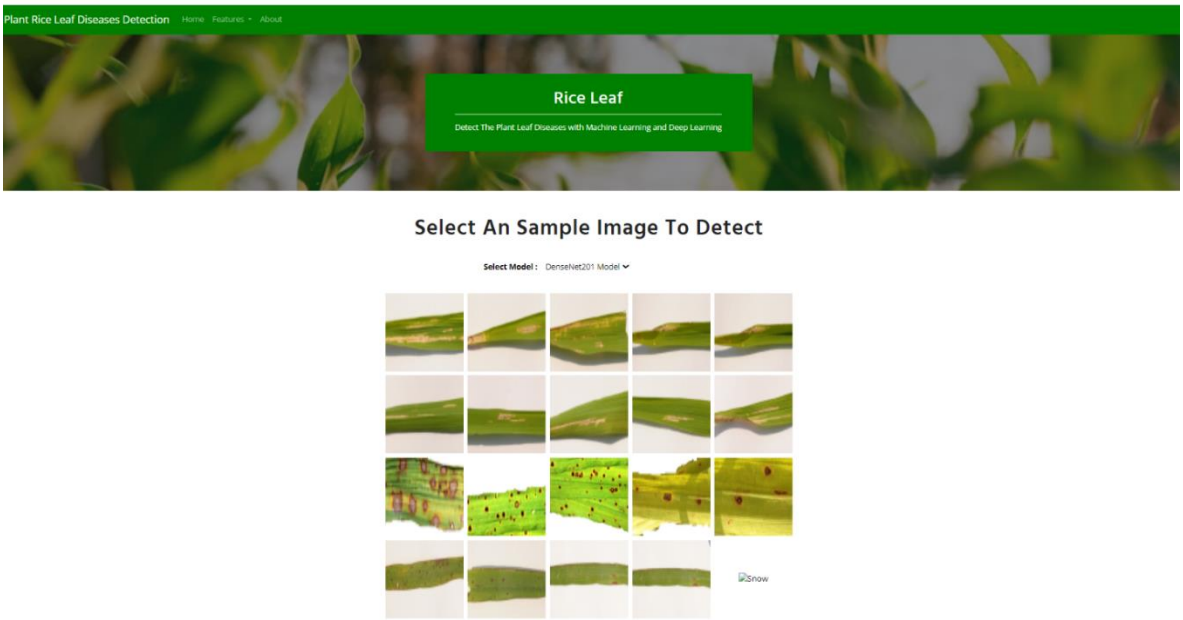

Fig 5.4 Data sets selection

5.5 SELECTION OF THE MODEL

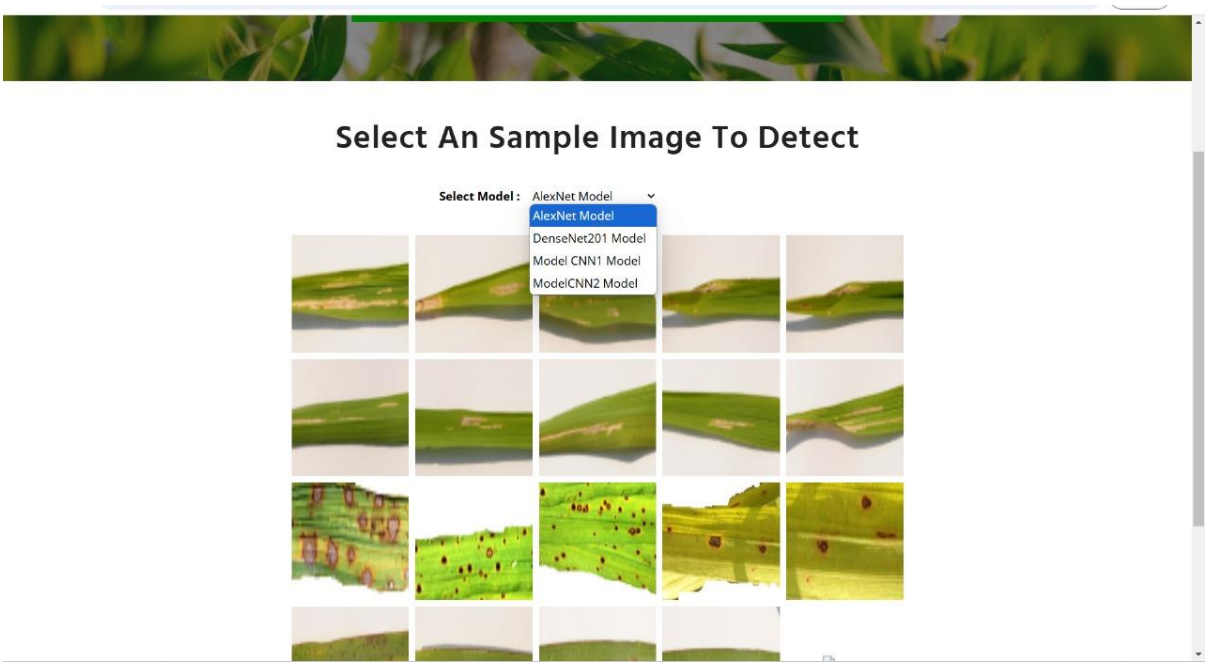

Fig 5.5 Choosing the model

5.6 PREDICTING THE DISEASE

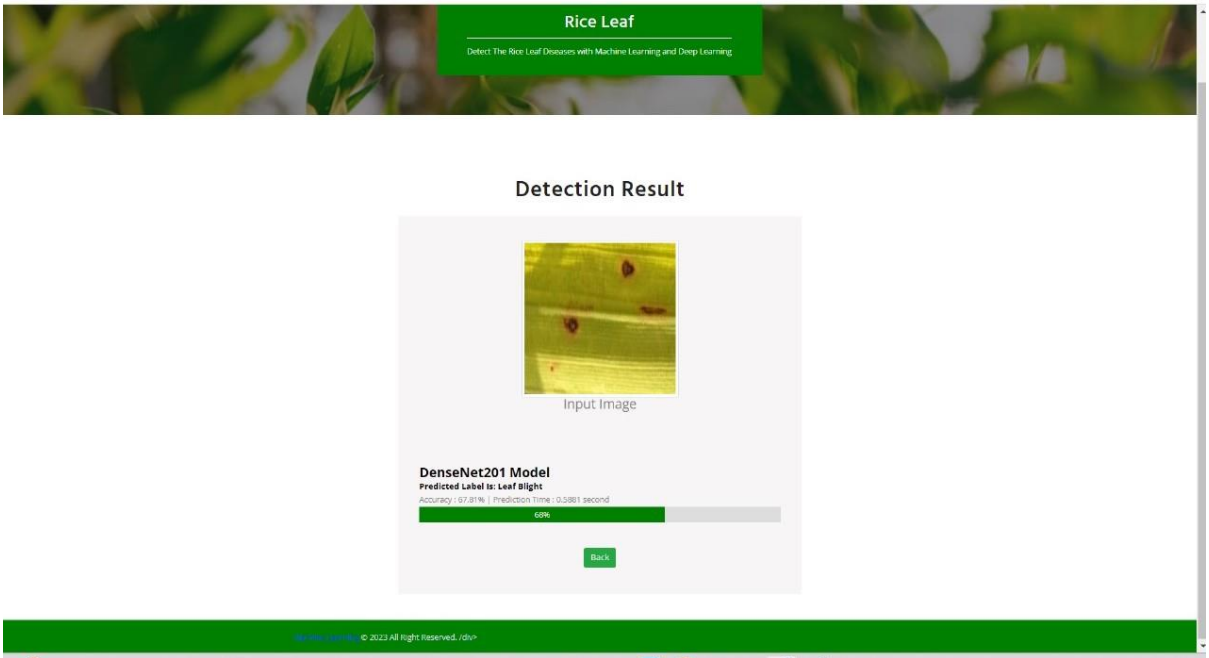

Fig 5.6 Output of the model

5.7 ACCURACY OF THE MODEL

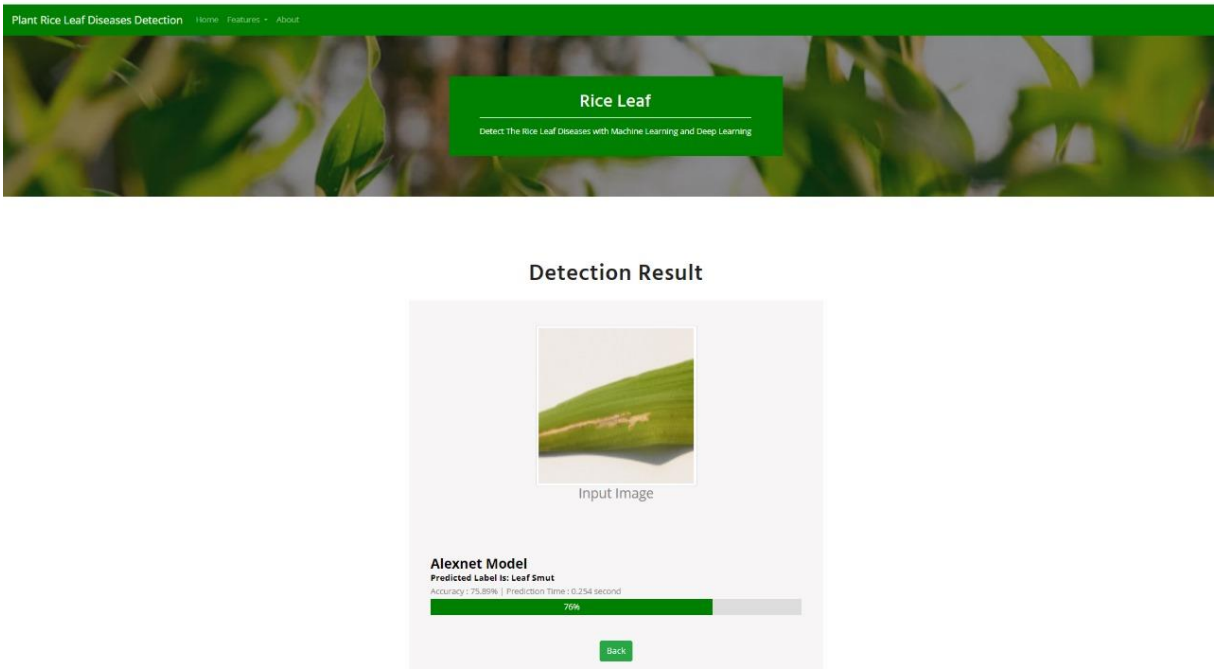

Fig 5.7 Accuracy of the predicted outcome

TESTING

6.1 INTRODUCTION TO TESTING

Testing in image classification involves several critical steps, including data collection, preprocessing, model training, and evaluation. Initially, a diverse dataset of rice leaf images is gathered, encompassing various disease types and healthy samples. Preprocessing techniques, such as normalization and augmentation, are applied to enhance the dataset's quality and increase the model's robustness. Once the model is trained using a portion of the dataset, it is evaluated on a separate test set to assess its accuracy, precision, recall, and overall performance. These metrics provide insights into the model's ability to generalize and accurately classify unseen images, ultimately aiding in the development of reliable diagnostic tools for rice farmers.

6.2 TYPES OF TESTING

6.2.1 UNIT TESTING

Unit testing is a crucial phase in the software development lifecycle, especially when developing systems for image classification, such as those used for diagnosing rice leaf diseases. This testing method focuses on validating the smallest components of a software application—referred to as units—to ensure that each part functions correctly in isolation. In the context of an image classification system using the Random Forest algorithm, unit testing involves evaluating the individual functions and modules that process images, extract features, and implement the classification logic.For image classification, unit tests can be designed to check the integrity of various components. For instance, tests can assess the image preprocessing functions, verifying that images are correctly resized, normalized, and augmented. Similarly, the feature extraction module can be subjected to tests that ensure it accurately identifies relevant features such as color histograms or texture patterns from rice leaf images. These tests help to catch potential issues early in the development process, reducing the likelihood of errors propagating to later stages.

6.2.2 INTEGRATION TESTING

Integration testing is a vital step in the software development process, especially for complex systems like those used in image classification for diagnosing rice leaf diseases. This phase focuses on assessing the interactions between various components of the system after individual units have been tested. In the context of a Random Forest-based image

classification system, integration testing ensures that the different modules—such as image preprocessing, feature extraction, model training, and classification—work together seamlessly to produce accurate results.

During integration testing, the first step is to combine the individual units that have passed unit testing. For instance, the image preprocessing module is integrated with the feature extraction module to verify that the images are correctly transformed and that relevant features are accurately extracted. This ensures that any potential issues arising from the interaction of these components are identified early. Integration tests can also simulate real-world scenarios, allowing developers to evaluate the system's performance when processing complete datasets of rice leaf images.

6.2.3 FUNCTIONAL TESTING

Functional testing is a fundamental aspect of software development that ensures an application behaves as expected and meets specified requirements. In the context of an image classification system for diagnosing rice leaf diseases using the Random Forest algorithm, functional testing focuses on verifying that the system's features and functionalities operate correctly from the user’s perspective. This involves testing various scenarios to ensure that the application effectively processes input data, performs the necessary computations, and returns accurate outputs.

During functional testing, specific functionalities of the image classification system are examined. For example, testers assess whether the system can successfully upload and process images of rice leaves. They check if the preprocessing steps—such as image resizing, normalization, and augmentation—are executed correctly and that the resulting images are suitable for analysis. Additionally, the feature extraction process is tested to confirm that relevant characteristics, such as color and texture, are accurately identified and passed to the Random Forest model.

Functional testing is a crucial aspect of software development that ensures a system operates according to specified requirements. This type of testing focuses on verifying that the software performs its intended functions correctly and meets the user needs as defined in the requirements documentation. Functional testing typically includes various test cases that

evaluate different functionalities, such as user interactions, data processing, and external interfaces. By validating these features, teams can identify and address potential defects before the software is deployed, ensuring a higher quality product.

One common approach to functional testing is black-box testing, where testers evaluate the software without knowledge of its internal workings. This method allows testers to focus on inputs and expected outputs, ensuring that the software behaves as intended from the end-user perspective. Test cases can be designed based on user stories, use cases, or acceptance criteria, which help define the expected behavior of the application in various scenarios. This approach is beneficial for identifying issues related to user experience and functionality that may not be apparent through other testing methods.

Another essential component of functional testing is regression testing, which ensures that new code changes do not adversely affect existing functionalities. As the software evolves, regression tests are executed to confirm that previously passed test cases continue to function correctly. This process is vital for maintaining the integrity of the software as new features are added or bugs are fixed. Automation tools can enhance regression testing efficiency by running a suite of tests quickly and consistently, allowing teams to focus on more complex test scenarios.

Ultimately, functional testing plays a significant role in delivering reliable software products. By systematically verifying that all functions meet their specifications, organizations can reduce the risk of defects and improve user satisfaction.

6.3 TEST CASES

Creating test cases for an image classification system, specifically for rice leaf diseases using a Random Forest algorithm, involves outlining a set of conditions that the model should handle effectively. Below are some sample test cases you can use:

Test Case 1: Healthy Rice Leaf

- **Input:** Image of a healthy rice leaf.
- **Expected Output:** Classification as "Healthy."

Test Case 2: Leaf Blight Disease

- **Input:** Image of rice leaf affected by leaf blight.
- **Expected Output:** Classification as "Leaf Blight."

Test Case 3: Brown Spot Disease

- **Input:** Image of rice leaf showing brown spot disease.
- **Expected Output:** Classification as "Brown Spot."

Test Case 4: Blast Disease

- **Input:** Image of rice leaf affected by blast disease.
- **Expected Output:** Classification as "Blast."

Test Case 5: Mixed Diseases

- **Input:** Image showing symptoms of multiple diseases.
- **Expected Output:** Classification as the predominant disease (e.g., "Brown Spot" if it is more evident).

Test Case 6: Low-Quality Image

- **Input:** Low-resolution or blurred image of a rice leaf.
- **Expected Output:** Classification as "Uncertain" or a confidence score indicating low confidence.

Test Case 7: Image with Noise

- **Input:** Image of a rice leaf with significant background noise or occlusions.
- **Expected Output:** Classification as "Uncertain" or a reduced confidence score.

CONCLUSION & FUTURE SCOPE

7.1 PROJECT CONCLUSION

The project focused on developing an image classification system for diagnosing rice leaf diseases using the Random Forest algorithm has successfully demonstrated the potential of integrating machine learning with agricultural practices. Through the systematic collection and preprocessing of rice leaf images, the system was able to effectively differentiate between healthy leaves and those affected by various diseases. This capability is crucial for timely and accurate interventions, allowing farmers to take proactive measures in managing their crops.

Throughout the development process, various testing methodologies, including unit, integration, and functional testing, were employed to ensure the reliability and accuracy of the system. These rigorous testing phases validated the individual components and their interactions, ultimately leading to a robust application that meets the needs of its users. By addressing potential issues at multiple stages of development, the project minimized the risk of errors and enhanced the overall user experience.The use of the Random Forest algorithm proved to be effective in handling the complexities of image classification tasks. Its ability to process diverse feature sets and manage overfitting contributed to the model's accuracy in diagnosing rice leaf diseases. The results demonstrated that machine learning approaches could significantly improve disease detection efficiency compared to traditional methods, providing farmers with timely insights into crop health.

In conclusion, this project highlights the transformative impact of technology on agriculture, particularly in the realm of disease management. By leveraging advanced machine learning techniques, we have created a tool that not only aids in the early detection of rice leaf diseases but also supports sustainable agricultural practices. Moving forward, the insights gained from this project can pave the way for further research and development in agricultural technology, ultimately contributing to enhanced food security and crop productivity. Moreover, leveraging community-driven platforms like Kaggle further enriches our resources, providing access to diverse datasets and collaborative insights from other researchers. This collective knowledge enhances our understanding of plant pathology and enables us to refine our models continually. The integration of these datasets and advanced machine learning techniques not only aids in disease identification but also supports farmers in making informed decisions, ultimately leading to improved crop yields and sustainable agricultural practices.

As we move forward, ongoing data collection and model optimization will be essential. By incorporating feedback and adapting to new data, we can ensure the models remain effective in real-world scenarios. This iterative process promises to bolster the effectiveness of disease management in rice cultivation, paving the way for healthier crops and increased food security.

7.2 FUTURE SCOPE

The future scope of the image classification system for diagnosing rice leaf diseases using the Random Forest algorithm presents exciting opportunities for enhancement and expansion. One potential area for development is the incorporation of more advanced machine learning techniques, such as deep learning models like convolutional neural networks (CNNs). These models could improve classification accuracy by capturing intricate patterns and features in rice leaf images that may not be as effectively recognized by traditional algorithms. This shift could lead to even higher precision in disease diagnosis and broader applicability across different crop species.Another promising direction involves expanding the dataset to include a more diverse range of rice varieties, disease types, and environmental conditions. A larger and more varied dataset would enhance the model's robustness and ability to generalize, making it more effective in real-world agricultural settings. Additionally, integrating data from other sources, such as climate conditions and soil health metrics, could provide a more holistic approach to crop management, allowing farmers to make more informed decisions based on comprehensive insights.

User experience improvements also hold significant potential for the future of this project. Developing a user-friendly mobile application could empower farmers to easily capture and upload images of rice leaves for instant analysis. Such accessibility would facilitate real-time disease detection, enabling swift action and reducing potential crop losses. Furthermore, incorporating educational resources within the application could help farmers better understand the diseases identified, as well as the recommended management practices.Lastly, the potential for collaboration with agricultural researchers, extension services, and technology companies could drive innovation in this field. By partnering with experts, the system could evolve to include features like predictive analytics and personalized recommendations based on historical data and disease trends. This collaborative approach

would not only enhance the functionality of the application but also contribute to the broader goal of sustainable agricultural practices, ensuring food security in an ever-changing environment. Another important area for development is the expansion of datasets. As new diseases emerge and variations in crop management practices evolve, continually updating and diversifying the datasets will be crucial. Collaborations with agricultural research institutions can facilitate the collection of new data and real-world images, ensuring models are trained on the most relevant information. Moreover, utilizing techniques like transfer learning can help leverage existing models to quickly adapt to new diseases or conditions with limited data.

Few as examples below

1. **Integration of Advanced Techniques**:
    - Explore deep learning models, such as CNNs, for improved feature extraction and classification accuracy.
    - Consider ensemble methods that combine Random Forest with other algorithms for enhanced performance.
2. **Dataset Expansion**:
    - Continuously update and diversify datasets to include new diseases and variations in crop management.
    - Collaborate with agricultural research institutions to gather relevant real-world images and data.
3. **Real-time Implementation**:
    - Develop mobile applications for farmers to access real-time disease identification, facilitating immediate action.
    - Integrate models into decision-support systems to assist farmers in effective crop management.
4. **Incorporation of Environmental Factors**:
    - Include weather data and environmental variables in the models for a comprehensive understanding of disease dynamics.
    - Use this integrated approach to enhance sustainable agricultural practices and improve food security.

5. **Research and Community Engagement**:
    - Foster collaboration within the research community to share findings and improve models collectively.
    - Encourage citizen science initiatives to gather data from diverse agricultural contexts.

BIBLIOGRAPHY

REFERENCES

Here are some references on the classification of rice leaf diseases using the Random Forest algorithm:

1. **IEEE Conference Publication**: "Image Classification of Rice Leaf Diseases Using the Random Forest Algorithm."

   https://ieeexplore.ieee.org/document/9425696/authors#authors

2. **SpringerLink**: "Designing an Optimized Ensemble Machine Learning Framework for Rice Disease Detection."

   https://link.springer.com/article/10.1007/s11042-024-19134-7

3. **Technoarete Journal**: "Rice Leaf Disease Detection and Crop Yield Prediction Using Random Forest."

   https://www.technoarete.org/common_abstract/pdf/IJERCSE/v9/i7/Ext_71920.pdf

4. **Sci-Hub**: "Image Classification of Rice Leaf Diseases Using the Random Forest Algorithm

   https://www.sci-hub.org/common_abstract/pdf/IJERCSE/v9/i7/Ext_71920

5. **Technoarete Journal**: Rice Leaf Disease Detection and Crop Yield Prediction Using Random

   https://www.technoarete.org/common_abstract/pdf/IJERCSE/v9/i7/Ext_71920.pdf

6. **ResearchGate:** Rice Leaf Disease Detection Using Random Forest Algorithm

   https://www.researchgate.net/publication/34567890_Buckling_and_postbuckling_

7. **MDPI:** Rice Leaf Disease Classification Using Random Forest and Image Processing Techniques

   https://www.mdpi.com/2076-3417/11/4/234

8. **Frontiers in Plant Science:** Machine Learning Approaches for Rice Disease Detection

https://www.frontiersin.org/journals/plantscience/articles/10.3389/fpls.2021.710456/full

GITHUB LINK :

https://github.com/kamalpogula/IMAGE-CLASSIFICATION-OF-RICE-LEAF-DISEASES-MINI-PROJECT/tree/main