A HYPERSPECTRAL-BASED SYSTEM FOR IDENTIFICATION OF COMMON BEAN GENOTYPES RESISTANT TO FOLIAR DISEASES IN TANZANIA

Josiane Irakiza

A Project Report Submitted in Partial Fulfillment of the Requirements of the Award of the Degree of Master of Science in Embedded and Mobile Systems of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

June, 2024

ABSTRACT

Common bean is one of many legumes (family Fabaceae) widely cultivated for their edible seeds, seedpods, and leaves. Despite its benefits and life dependency especially in sub-Saharan Africa, foliar diseases are causing a loss of 20% to 80% of common bean production, and the development of improved common bean seeds resilient to those foliar diseases is still an issue where among the major problem that the bean breeders are facing is manual phenotyping; a slow field process and prone to errors as it depends on the eyes of the viewer. According to the literature, imaging technologies have been introduced to help in different processes for crops and disease management. However, there is a lack of automated mechanisms for phenotyping processes to help breeders in trait data collection, disease scare classification, and analysis of all data collected to identify resilient genotypes digitally. Among existing solutions, there is still also a gap in plant health monitoring during all its growing stages needed by breeders. Therefore, this study developed a unique hyperspectral data-based approach for identifying bean genotypes resistant to foliar diseases and plant health monitoring. Using the Random Forest classifier this study proved the genotype classification in three main breeding categories; Resistant, Medium, and Susceptible. The experiment was conducted in four Regions of Tanzania and three classifiers "Random Forest, XgBoost, and four layers Neuro Network algorithms" were trained and tested with results of 0.96, 0.95, and 0.92 respectively. The model was deployed on the cloud server where it is linked to a web application for easy classification and data analysis. Applying different vegetation indexes including the Chlorophyll Index, Photochemical Reflectance Index, Water Band Index, Modified Chlorophyll Absorption in Reflectance Index, Nitrogen Reflectance Index, Structure Insensitive Pigment Index, and Simple Ratio efficiently proved to be used for plant health insight before disease symptoms are seen. This saves breeders time, reduces errors, and helps them with digital phenotypic data, faster analysis, and easy storage for future references.

DECLARATION

I, Josiane Irakiza, declare to the Senate of the Nelson Mandela African Institution of Science and Technology that this project report is my original work and that it has neither been submitted nor concurrently submitted for a degree award in any other institution.

Josiane Irakiza		12/05/2024
Name of Candidate	**Signature**	**Date**

The above declaration is confirmed by:

Prof. Shubi Kaijage		13/05/2024
Name of Supervisor 1	**Signature**	**Date**

Dr. Judith Leo		13/05/2024
Name of Supervisor 2	**Signature**	**Date**

Miss. Hope Mbelwa		13/05/2024
Name of Supervisor 3	**Signature**	**Date**

COPYRIGHT

CERTIFICATION

The undersigned certify that they have read and hereby recommend for acceptance by the Nelson Mandela African Institution of Science and Technology a project report titled **"*A Hyperspectral-Based System for Identification of Common Bean Genotypes Resistant to Foliar Diseases in Tanzania*"** in partial fulfillment of the requirements for the degree of Master of Science in Embedded and Mobile Systems of the Nelson Mandela African Institution of Science and Technology.

Prof. Shubi Kaijage		13/05/2024
Name of Supervisor 1	**Signature**	**Date**

Dr. Judith Leo		13/05/2024
Name of Supervisor 2	**Signature**	**Date**

Miss. Hope Mbelwa		13/05/2024
Name of Supervisor 3	**Signature**	**Date**

ACKNOWLEDGEMENTS

First and foremost, I would like to praise and thank God, the Almighty, for his grace, strength, sustenance, and most importantly, his faithfulness and love from the beginning of my academic career to this master's level.

Aside from my efforts, the support and direction of many individuals played a significant role in the successful accomplishment of this study, to whom I would like to take this opportunity to express my gratitude.

I express my tremendous appreciation to my amiable, ever-supportive, and humble academic supervisors, Prof. Shubi Kaijage, Dr. Judith Leo, and Miss. Hope Mbelwa, for their invaluable contributions and instructions in developing this project.

I am highly thankful to my industrial supervisors Dr. David Guerena and Dr. Teshale Mamo from the Alliance of Bioversity and CIAT, who took their time to guide me during this research from the beginning to the completion of this work.

Special thanks are extended to the Nelson Mandela African Institution of Science of Technology (NM-AIST), the Centre of Excellence for ICT in East Africa (CENIT@EA), and Deutscher Akademischer Austauschdienst (DAAD) Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH (GIZ)'s scholarship program, and also deep thanks to The Alliance of Bioversity and International Center for Tropical Agriculture (CIAT) for funding the project and for providing me with a facility and a conducive environment that helped me to complete this project successfully.

I would like to acknowledge the help and support from the ARTEMIS team especially project leader Dr. David Guerena, Dr. Stephen Mutuvi, who had been mentoring me during this journey, and my office mates Ellena Girma, Shekinah Henday, Leah Saul, Minaeli Mjema and my sister Gaudensia Katunzi for their daily help and support.

I would like to express my gratitude to my family, Hakizamungu Medard, Mukandepanda Demetrie, Dushimirimana Egide, Izerimana Janviere, Dusabimana Edison, Hakizimana Melane and Mushimiyimana Alice for their unconditional love and support.

Finally, I am deeply thankful to my husband Dieudonne Dusingizimana for his kind support and daily moral encouragement from application to completion of this Master's journey.

DEDICATION

I dedicate this work to my beloved family, my Father Hakizamungu Medard, my Mother Mukandepanda Demetrie, my brothers and sisters, my husband, my friend Rahab Uwamahoro, and I also dedicate it to my teammates at the Alliance of Bioversity International and CIAT. Your love and support is uncomparable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ALS	Angular Leaf Spot
API	Application Programming Interface
AUC	Area Under the Curve
AVG	Average
AWS	Amazon Web Server
CBB	Common Bean Bacterial Blight
CENIT@EA	Centre of Execellence for ICT in East Africa
CI	Chrolophyll Index
CIAT	International Center for Tropical Agriculture
CSS	Cascading Style Sheets
DAAD	Deutscher Akademischer Austauschdienst Deutsche
FAO	Food and Agriculture Organization
FN	False Negative
FP	False Positive
GIZ	Gesellschaft fur Internationale Zusammenarbeit GmmbH
GPUs	Graphics Processing Units
HTML	Hyper Text Markup Language
IDE	Integrated Development Environment
LAI	Leaf Area Index
LED	Light Emitting Diode
MCARI	Modified Chlorophyll Absorption in Reflective Index
NDVI	Normalized Difference Vegetation Index
NIR	Near Infrared
NM-AIST	Nelson Mandela African Institution of Science and Technology
NRI	Nitrogen Reflectance Index

ODK	Open Data Kit
OOB	Out of Bug
PhD	Doctor of Philosophy
PHP	Hypertext Preprocessor
PRI	Photochemichal Reflectance
RAM	Random Access Memory
REP	Replications
ROC	Receiver Operating Characteristics
SIPI	Structure Insentive Pigment Index
SOP	Standard Operating Procedure
SP	Sub Plot
SQL	Structured Query Language
SR	Simple Ratio
SWIR	Short-Wave Infrared
TARI	Tanzania Agriculture Research Institute
TN	True Negative
TP	True Positive
TPUs	Tensor Processing Units
UML	Unified Modeling Language
UNESCO	United Nations Educational, Scientific and Cultural Organization
URL	Uniform Resource Locator
USB	Universal Serial Bus
VI	Vegetation Index
WBI	Water Band Index
WP	Whole Plot
XGBoost	Extreme Gradient Boost
XP	Extreme Programming

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

Common bean (Phaseolus vulgaris), originated in Central and South America at approximately 6000 BC (Nigatie, 2021) Since then, beans have been one of the world's most important crops for nutrition and health, especially in sub-Saharan Africa. Beans are one of the most important crops for nutrition and health in the world, especially in sub-Saharan Africa. In 2021, 27 million tons of beans were produced worldwide, 7 million tons were produced in Africa, 5 million tons in East Africa, and 1 million tons in Tanzania (FAO, 2023). In Tanzania, concentrated production of common bean is found in the Kagera region of Tanzania and southwest Uganda (Farrow & Muthoni-Andriatsitohaina, 2020).

Common beans are high in protein and other essential amino acids and minerals such as thiamin, riboflavin, and iron (Britannica, 2021) with a contribution to human nutrition of 7.3% calories, and 71% protein from legumes in diets (Ndimbo *et al.,* 2022). Given all these benefits to people, climatic change, high population expansion, harsh weather conditions, and considerable loss of arable land, and water supplies pose great challenges to common beans and agriculture in general, including biotic and abiotic pressures (Tavakoli *et al.,* 2022).

To overcome these challenges, bean breeding programs were established prioritizing the development of new seed varieties to ensure high production and yield stability under adverse environmental conditions (Li & Yan, 2020). Additionally, scientists are working to identify and incorporate resistance genes to create more resilient seed varieties in order to reduce plant stress and yield losses and meet the consumption demand around the world (Tavakoli *et al.,* 2021).

Seed is a genetic material that serves as the initial link in the food chain, a source of life, and a future plant (Guardian, 2011). However, beans' aerial and underground parts are affected by a number of foliar diseases that cause significant yield losses throughout the world where globally, diseases severely reduce the quality and yield of common beans by 20–100% (Nguyen *et al.,* 2021) leading to hunger and this can be fought by giving access to and usage

of improved seeds to smallholder farmers to boost agricultural productivity, hence improving lives (Singh & Schwartz, 2010; Belay *et al.,* 2022).

High-yielding and resilient seed varieties are known from their observable traits which are measured through the process of phenotyping. Plant phenotyping is a new discipline that integrates several approaches and procedures to assess plant properties (such as growth, morphology, architecture, function, and composition) and manual phenotyping is still a key impediment to plant breeding progress (Teressa, 2019); as green coloration mostly used for observation varies significantly depending on ambient conditions, and the eyes of the viewer (Carvalho *et al.,* 2021). The signs of unhealthy plants considered are mostly the visible ones on the plant's leaves, such as uneven leaf forms and diminished growth color changes (Guo *et al.,* 2021), and it does not give insight into what is causing those visible symptoms before breeders can see them resulting in a lower yield if not treated early (Sinha & Shekhawat, 2020).

To enable efficient disease control, different technological innovations and methods have been introduced to help agriculture improve. However, in Tanzania, the adaptation of technology in the breeding system is still low as most existing systems are based on visible facts that do not reflect plant diseases at their pre-symptoms stage needed by breeders (Ahmad *et al.,* 2023). Early disease detection in crops and digital phenotyping which are ones of important technological solutions for ease and quick breeding results are still problems. When it comes to accurately analyzing the contents of the plant, eyes' view, red, green, and blue camera-based imaging technology is not consistent compared to near-infrared (NIR) and short-wave infrared (SWIR) reflectance (Nguyen *et al.,* 2021).

To increase and optimize the utilization of better seeds to adapt to climate change and meet the consumption demand (Fuglie *et al.,* 2020). This study introduces the use of hyperspectral data (Zhang *et al.,* 2020), and the use of artificial intelligence to help breeders in the identification of traits that will help them in the early identification of bean varieties resistant to foliar diseases (Terentev *et al.,* 2022), and decision making using automated phenotypic data analysis (Cheshkova, 2022).

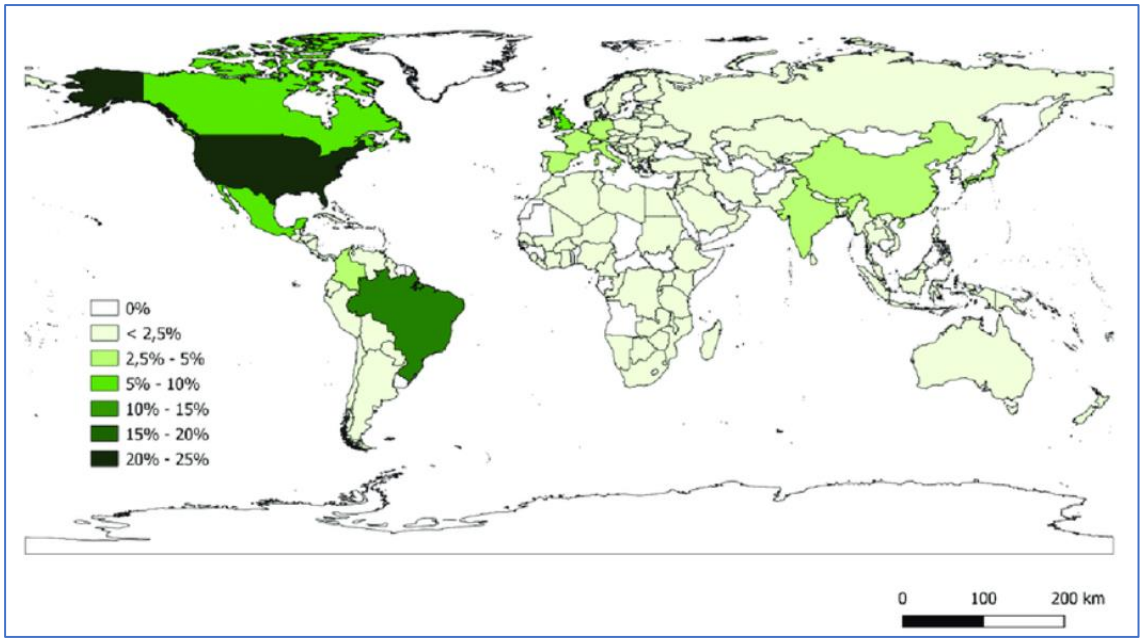

Figure 1: Common bean production around the world (Spatti *et al.*, 2022)

1.2 Statement of the Problem

Despite efforts from the public sector to finance the expansion and development of the agriculture sector, each year farmers still face several difficulties.

The majority of diseases affecting beans in Tanzania and across the world are brought on by bacterial, viral, and fungal agents. Even if they germinate, infected bean seeds usually become discolored, chlorotic, mushy, and decaying. In 1-3 weeks, they become brown or possibly die. On bean genotypes, various Pythium species cause seed deterioration, pre-emergency, and post-emergency conditions (Binagwa, 2019).

It has taken a lot of effort and time to breed resistant cultivars against these foliar diseases known from their observable traits which are measured through the process of phenotyping. Many technology experts have attempted to assess the issue and try to develop different methods to detect plant diseases. In this regard, a variety of supervised and unsupervised learning approaches have been applied to tackle disease detection using leaf images.

However, gaps remain in the manual phenotyping for common bean breeders which is a tiring slow field process and has a lot of data errors, data loss, and data biases as it depends on the eyes of the viewer with manual paper writing as data collection tools. The Alliance is trying to

overcome this in the ARTEMIS project with the goal of Imaging Technology for Food Secure Future.

This project focused on assisting common bean breeders with a faster way for disease identification and common bean varieties classification using spectroscopy which was found to be the best way for assessing plant health with spectral data. This shorten breeding field process times due to faster diseases scoring and easy phenotypic data analysis.

Figure 2: Foliar diseases in beans and manual phenotyping

1.3 Rationale of the Study

Reflectance is the portion of light that is reflected back toward the observer after striking a leaf. Spectra, which are commonly used to shorten the name of the reflectance spectrum, are the amounts of energy reflected at each light frequency. Chlorophyll and other pigments involved in photosynthetic processes are the primary determinant of reflectance. The leaf's structural discontinuities determine the amount of reflectance in the near-infrared (NIR), or between 700 and 1300 nm, where there are no prominent absorption features (Martínez-Martínez *et al.*, 2018).

Chlorophyll is a pigment found in chloroplasts that plays an important role in photosynthesis. Chlorophyll content per leaf area can be used to predict plant health and consequently plant stress (Chung *et al.*, 2018).

Many spectral methods of analysis measure the reflectance in the wavelengths corresponding to greenness. While important spectroscopy within the visible wavelengths of light tends to capture plant diseases only after considerable symptom progression. The NIR has proven to be useful as it can detect disease appearance before visual symptoms appear (Danzi *et al.*, 2022).

Additionally, wavelength can be used to track vegetation stress brought on by a variety of conditions, including disease, nutrient deficiency, and drought (Wemmert *et al.*, 2020).

Many of the developed plant health diagnostic systems rely on smartphone or camera images within the visible spectrum with red, green and blue ranges to aid in disease detection (Digital & Capture, 2018). With the use of spectroscopy, we were able to analyze multiple wavelength parts of light, allowing for a more in-depth examination of the pigments and other chemical components of the leaf which is needed by breeders.

The existing process of identifying the foliar diseases by breeders at the Alliance was a very long and tedious process where the researcher had to go to the screen houses and fields observe the plant, conduct visual scoring write them down, and then go back to the office for analysis. There are known errors in this system as the visual interpretation is highly variable by individual and by fatigue.

With this project, we removed all the processes so that everything can be done automatically in a single platform with higher accuracy of data collection and analysis from screen houses and fields.

The system is more robust with fewer external influencing factors such as lighting conditions, camera angles, or image resolution impacts. The spectrometer was calibrated with white tape to provide more detailed information about the chemical composition of common bean leaves which allowed the whole light range to be detected even the ones invisible to human eyes. The leaves were scanned directly using the black tape to remove other color interferences which led to faster identification of diseases.

1.4 Objectives of the Study

1.4.1 Main Objective

To develop a hyperspectral-based system for the identification and classification of common bean genotypes that are resistant to foliar diseases.

1.4.2 Specific Objectives

The study aimed to achieve the following specific objectives:

(i) To identify system requirements.

(ii) To develop the system with the model that classifies disease damage at different levels.

(iii) To validate the performance of the developed system.

1.5 Research Questions

With the objective of automating the common bean phenotyping process and breeding data analysis, this research was answering the following main scientific questions:

(i) What are the requirements for developing a hyperspectral-based system for the identification of common bean genotypes resistant to foliar diseases?

(ii) How can we develop a system that will identify the resistivity of common bean genotypes?

(iii) How will the performance of the developed system be validated?

1.6 Significance of the Study

Among the major fungal diseases, Angular Leaf Spot (ALS) and Common Bacterial Bligh (CBB) are some of the most destructive due to their propensity for intermittently dry weather, warm cool weather, and high humidity. When a susceptible host is present, the pathogen can colonize various parts of bean plants, including the leaves, pods, and seeds.

The findings of this study, as well as the development of the classification of foliar diseases damage to bean plants and categorizing different varieties of common bean genotypes, will help the breeders and the country in general so that farmers can be given trusted seeds, restoring hope to common bean growers who are depressed due to the expenses and damages brought on by production-related diseases. Therefore, the solution will lessen the workload of the nation's few researchers and skilled breeders because it is automated and will not require human intervention.

1.7 Delineation of the Study

This study aimed to develop a system that can help breeders identify foliar diseases in common beans and classify genotypes that are resistant to those diseases. The system was developed

using a dataset collected using a handheld spectrometer to capture different wavelengths so that the whole plant physiology could be assessed and monitored timely by the breeders. System results show that it is suitable to be used for detecting Angular Leaf Spots and common bacterial blight in beans and classifying different genotypes according to their levels of resistivity in bean plants.

However, it is worth notifying that other foliar diseases like Common Bean Rust and Bean Common Mosaic Virus were not identified in this study as the timeline was limited due to the long process of preparing fields, inoculum, planting, and waiting for the diseases to show up while we were using limited fields and screen houses.

CHAPTER TWO

LITERATURE REVIEW

2.1 Definition of Key Terms

2.1.1 Plant Breeding

Plant breeding, also known as cultivar development, crop improvement, or seed improvement, is the scientifically driven creative process of creating new plant varieties (NAPB, 2019). This is accomplished by envisioning an ideal plant with the greatest number of desirable characteristics (Borra-Serrano *et al.*, 2022) Breeders take into account various attributes such as resistance to pests and diseases, ability to withstand heat, salinity in the soil, or frost, suitable dimensions, maturity period, and numerous other general and specific traits that enhance environmental adaptation, facilitate cultivation and handling, increase yield, and improve quality (Allard, 2023). The same applies to common beans as one of the most consumed legumes around the world. Before being introduced into the market, new bean types have to pass a rigorous assessment. In order to gather farmer and merchant preferences (Bishaw & Gastel, 2009), bean breeders conduct tests on-farm, and in controlled conditions. They do this by using participatory variety selection techniques. The national variety list of the participating countries is then updated with the approved varieties (Bishaw & Gastel, 2009).

2.1.2 Foliar Diseases

Foliar disease is a kind of plant disease that affects the leaves of plants. It is frequently a reaction to an irritant and is often a fungal or fungal-like entity (Ricks, 2022). One of the most prevalent foliar bean diseases is angular leaf spot (ALS), which is caused by Pseudocercospora griseola, common bacterial blight (CBB), which is caused by Xanthomonas axonopodis pv phaseoli and its fuscan variants, bean rust, which is caused by Uromyces append (Onyango, 2023). There are many effects of diseases on plants but most of the effect on the plants is yield reduction which in many cases leads to famine and food insecurity (Chen *et al.*, 2021a). World production loss due to foliar diseases is counted as more than 30% (Singh *et al.*, 2021). Another loss related to foliar diseases is the cost related to disease

control and treatment where a study shows that these other extra costs related to plant disease control are more like 15% of the crop production (Elad & Pertot, 2014).

2.1.3 Angular Leaf Spot

One of the most destructive diseases of the common bean in tropical and subtropical production areas is Angular Leaf Spot (ALS), which is brought on by Pseudocercospora griseola (Nay, 2019). This disease develops quickly in warm temperatures around 25℃, but it can also occur in moderate to warm temperatures around 16-28℃ when accompanied by wet weather or high humidity alternating with dry, and windy conditions (Greenlife, 2023). The disease infects the leaves, petioles, stems, and pods of plants. Lesions appear on leaves as brown spots with a tan or silvery center that are initially confined to the leaf tissue between major veins, giving it an angular appearance (Greenlife, 2023).

2.1.4 Common Bacterial Blight

Common Bacterial Blight of bean (CBB), a serious disease that is still challenging to eradicate, is caused by the two Xanthomonas species, Xanthomonas citri pv. fuscans (Xcf) and Xanthomonas phaseoli pv. phaseoli (Xpp). These pathogens are members of several Xanthomonas species and have experienced a dynamic evolutionary history that included the horizontal transfer of genes-producing elements likely necessary for common bean adaption and pathogenicity (Chen *et al.,* 2021a). The CBB spots develop to cause the entire leaf to die and the plant to defoliate (IPM, 2018). Losses resulting from CBB are likewise predicted to be between 10% and 40% (Girma *et al.,* 2022).

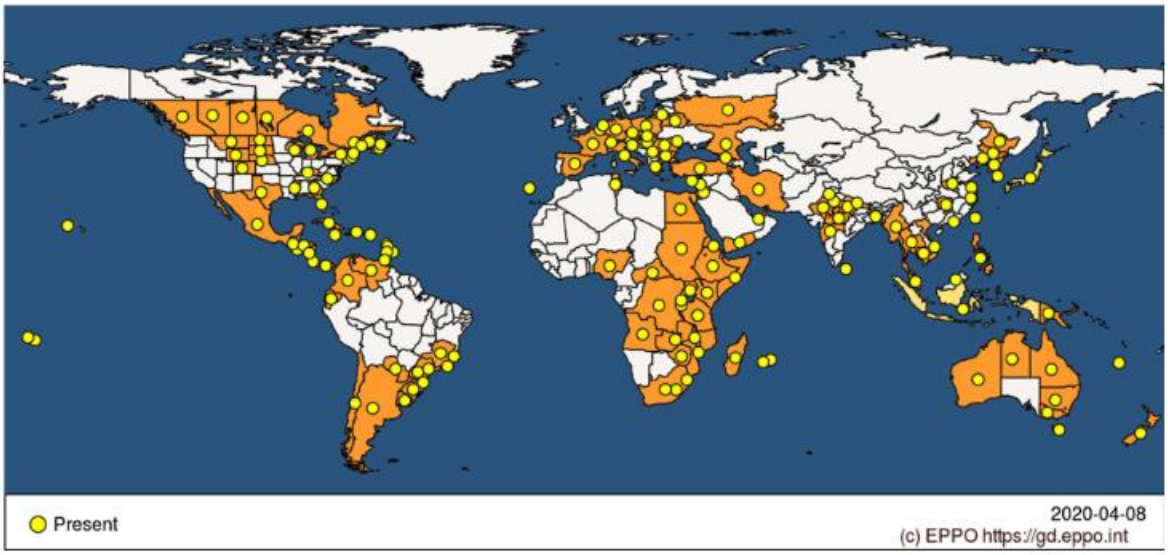

Figure 3: Common bacterial blight of bean distribution (Chen *et al.*, 2021b)

2.1.5 Spectral Reflectance Indices

Vegetation (spectral reflectance) indices are spectral transformations of two or more bands of spectral imagery that are intended to improve vegetation properties and canopy structure (Svanback & Bolnick, 2019). Different indices employ various wavelength combinations and calculating methods to get a numerical number that indicates the health of the plant (Radočaj *et al.*, 2023) and they may be used to calculate the relative abundance of particular features of interest about a plant (Geospatial, 2023). They are generated from vegetation's reflectance qualities and each VI is intended to highlight a specific vegetation trait (Geospatial, 2023). Studies show that in precision agriculture, depending on the vegetation Indices you have used, you can be able to detect various aspects of plant growth and easily monitor chlorophyll content, water, nitrogen, and other plant contents (Kurbanov & Zakharova, 2020).

2.2 Related Works

Different studies, research, reports, and systems were conducted to solve the disease identification in plants and phenotypic data analysis. The following are some of the literatures reviewed from different studies and it was done checking the usability of their methodologies, limitations, and the updates that can be done to improve technology usage in agriculture specifically helping common bean breeders with foliar diseases identification.

Intending to detect diseased and healthy bean crops (Sahu *et al.*, 2021) by using leaves images captured by smartphones, used a dataset of 1296 leaf images from AI Lab-Makerere to train the classifier of bean crops. To automatically extract the features from the images fed to the trained network, two Deep Learning models, Google Net and VGG16, were used. The authors found that Google Net outperforms VGG16 with an accuracy of 95.31% whereas VGG16 was 93%.

In the study made by Elfatimi *et al.* (2022) with the purpose of classifying healthy and diseased beans, 1296 images were a dataset used from GitHub to classify them. The data were separated into three classes such as Healthy class with 428 examples, Angular Leaf Spot with 432 and the last class is Bean Rust with 436. They used MobileNet and MobileNetV2 models for bean disease classification. The results of the study were 97% accuracy while using MobileNet.

To detect leghemoglobin in legumes, Yerokun and Onyesolu (2021) developed a fuzzy expert system. This was a rule-based forward chain technique that was applied and came up with a neuro-fuzzy expert system that can be used to detect leghemoglobin in legumes. The study resulted in the development of an expert system that can be used by both experts and non-experts in legume diseases with an accuracy of 99.5%. However, the system was developed based on visible symptoms which led to late decision-making.

A study made by Feng *et al.* (2021) with the aim to develop a diseases classifier of rice, they used hyperspectral imaging data to rain different CNN models and it was a success of more than 97% accuracy. Their study was limited by a shortage of data where only a sample of 250 data was available for training, testing and finetuning three different types of models.

By using wavelength light to examine the wavelength of the light source that affects the growth of Brassica rapa L (pak choi) (Gao *et al.*, 2019). The lamps used in the study were an 11-watt UV light, an 8-watt white LED light, a red LED light, green and blue LED lights. A biometric measurement was performed by measuring stem height, leaf width, and leaf number. The observations lasted 14 days, beginning on the first day of planting. According to the measurements, the pak choi can grow well under the white LED light and it cannot grow well in other light sources due to a lack of wavelength required by the plant.

In their study with the target to diagnose plant diseases, Almeida *et al.* (2021) and Sahu and Pandey (2023) used spatial fuzzy c-means to improve the accuracy of the classification in

disease diagnosis among different plants leaves. A dataset of 54 303 leaves images were used for training and testing the system. Their study resulted with 90% of accuracy and as others the system was only dependent on visual symptoms and annotations.

In a report released by the Department of Agricultural and Biological Engineering at the University of Florida, Shamshiri (2008) using the Normalized Difference Vegetation Index they discovered different main issues in citrus and wheat plants. The use of NDVI can't be the main point to look at when it comes to early detection of diseases in different plants as the greenness might be higher while the inside is going down.

Another study was conducted by Yong *et al.* (2023) on the research on Automatic Disease Detection of Basal Stem Rot Using Deep Learning and Hyperspectral Imaging, they came up with results showing how these two combined methods the infected wavelengths showed no differences with confidence of 95%. Their main analysis was done on the 938 nm part of the wavelength which is not a representative of the whole wave bands when it comes to assessing all possible patterns showing the diseases in plants.

To detect abnormalities in plants in their greenhouses or natural environment, Maniyath *et al.* (2018) came up with an algorithm to help classify the issue. One hundred and sixty (160) images of papaya leaves were used to train the Random Forest model and the model could classify with roughly 70%, and the Support vector machine with an accuracy of 40%.

Another crop-specific annotated dataset for coffee diseases detection was obtained from the survey of deep learning techniques for plant disease diagnosis by Ahmad *et al.* (2023) with a dataset of 1560 images taken with a 5 MP smartphone camera in an Ecuadorian field with 390 coffee plants, each coffee plant yielded four images, which were annotated with ground truth using an open-source tool. The VGG16 model was found to be 86.51% accurate in predicting the severity.

In his study, Rodríguez *et al.* (2019) tried to look for common beans resistant to Angular Leaf Spot (ALS). Their study was done on 181 bean genotypes in different countries including The United States, Puerto Rico, Honduras, Ecuador, Colombia, Tanzania, Malawi, and Angola among the countries represented. Their study was about isolating genotypes in the controlled condition with the use of improved varieties that combine resistance genes of Andean and Mesoamerican origin to control this disease. Sixteen (16) lines were resistant to

biotic and abiotic stresses, this knowledge can assist common bean breeding programs in pyramiding genes from the Andean and Mesoamerican gene pools to produce varieties with long-lasting resistance to this disease.

By using the deep learning model and the MobileNet model, Elfatimi *et al.* (2022) developed a model on different bean disease classifications using mobile phone leaves image taking and processing. The model was developed using the GitHub public dataset and gave results of 97% model accuracy using MobileNet algorithm.

Dashti *et al.* (2019) in their study about checking if remote sensing in dryland may not lead to bad decision. Their study was a success with the insight that the remote sensing with the models trained with data of different seasons and areas may lead to that misinterpretation of nitrogen content in plants. However, they were considering the remote sensing data from spectrometers statistically calculating data according to remotely captures bands not direct spectrometers to a specific plant.

To detect Healthy Brown Spot, Hispa, and Leaf Blast diseases in rice with imaging technology, a rice dataset of 3355 images were used. The dataset made up of images taken in a controlled lab setting with uniform white backgrounds. As a result, training robust deep learning models capable of identifying rice diseases in the field will be difficult. The study was among others done by Shrivastava and Pradhan (2021).

With the aim of checking agricultural insects and diseases pests that mostly affects plants, Matthew used radar remote sensing to detects insects and pests (Cock *et al.,* 2016). Their study came up with the conclusion that there are many ways to detect them using Radar imagery. It was a great research though it is still an issue when it comes to African breeders to afford radar imaging and control.

Another work carried out by Girma *et al.* (2022) to identify common beans resistant to ALS and CBB was conducted in Ethiopia. Their study was about trying different three replications and doing the analysis later. It was an amazing study where eight genotypes, namely DAB-388, DAB-478, DRKDDRB-70, DRKDDRB-81, NUA-225, NUA-517, NUA-536, and NUA-577 attained relatively low disease severity. However, it was stressful to do the laboratory work and field work without any technology tools to ease their work.

A research made by Vanitha and Padma (2021), they used a Fuzzy Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) for detecting four diseases on leaves of cotton. Their objective was to check which diseases were mostly attacking the cotton and the study gave a success of Verticilium wilt, Grey mildew being the highest disease attacking cotton. Their study was a great success with higher accuracy in detecting those diseases with the expert system usage which increased their decision making but it was only based on visible aspects of the diseases.

In their study, Garriga *et al.* (2017) with a concern to check if Predicted Trait-Values Really More Important Than Directly Identifying the Elite Genotypes. Using spectrometer data on different wheat genotypes, the total number of 348 genotypes were in the experiment and assessed using different vegetation indices. With more that 90% of accuracy they were able to differentiate those wheat genotypes with their ability to react on different diseases. They concluded by advising the adoption of spectrometer measurements for effective data analysis on disease aspects however they used higher atmospheric spectrometers which are not easy for Tanzanian breeders to afford.

2.3 Technical Gap

Various technological advancements and approaches have been implemented to facilitate the effective management of plant diseases. As was discussed in the previous part, numerous researchers developed a variety of superb technological approaches to address the problem of identifying diseases in various plants.

Most existing solutions are image-based technologies that use cameras and primarily rely on capturing green reflection. These technologies have proven to be less reliable for accurate digital plant health analysis, particularly in assessing biochemical contents. As of now, no research has been found to assist common bean breeders with automating the phenotyping process, identifying foliar diseases, classifying genotypes, or helping with data analysis in a single system.

2.4 Developed System

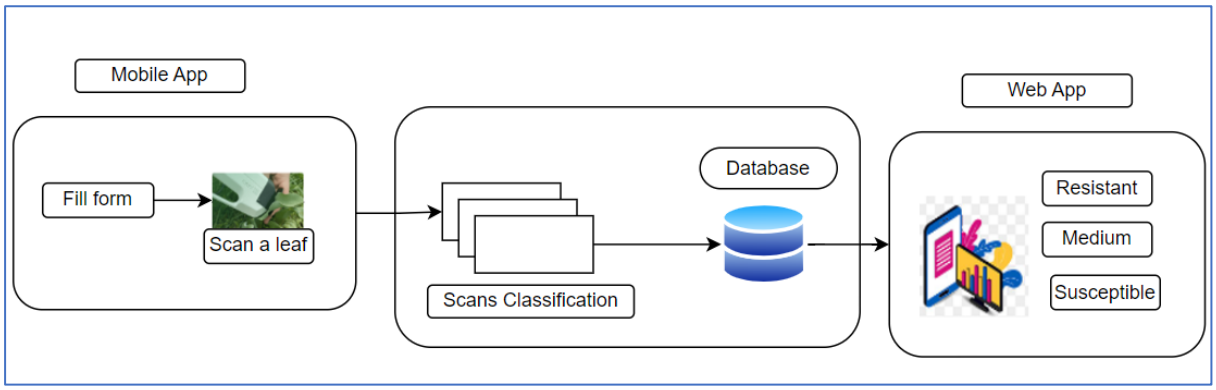

Figure 4: Developed system

The system is divided into three primary components, as shown in Fig. 4; the website, the database, and the mobile application. The end user use the mobile application to get data from the field using a form that has distinct details about the genotypes they are gathering data for. Scanning the plant leaves is a crucial step in filling out the form. The scans are uploaded from the field to the cloud server, where the model is trained to process and classify the images. Following the classification of the scans, the bleeder can review all of the field data that was gathered, categorized as resistant, medium, or vulnerable to foliar diseases on the web application.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Project Case Study

This research was carried out in Tanzania, focusing on various common bean breeding areas in greenhouses and open fields of the Alliance and Tanzania Agriculture Research Institute (TARI). The study took place in the areas of high concentration of common bean production around Tanzania. They are categorized into three main regions considered as the production of common beans in the country where according to the Ministry of Agriculture in Tanzania (Ministry of Agriculture, 2023) those regions are found in the North zone, the Southern Highlands zone, and Lake Zone. Regions under which this study was carried out were the

Arusha and Manyara Regions found in the Northern Zone, the Mbeya Region found in the southern highland zone and Kagera Region found in the Lake Zone.

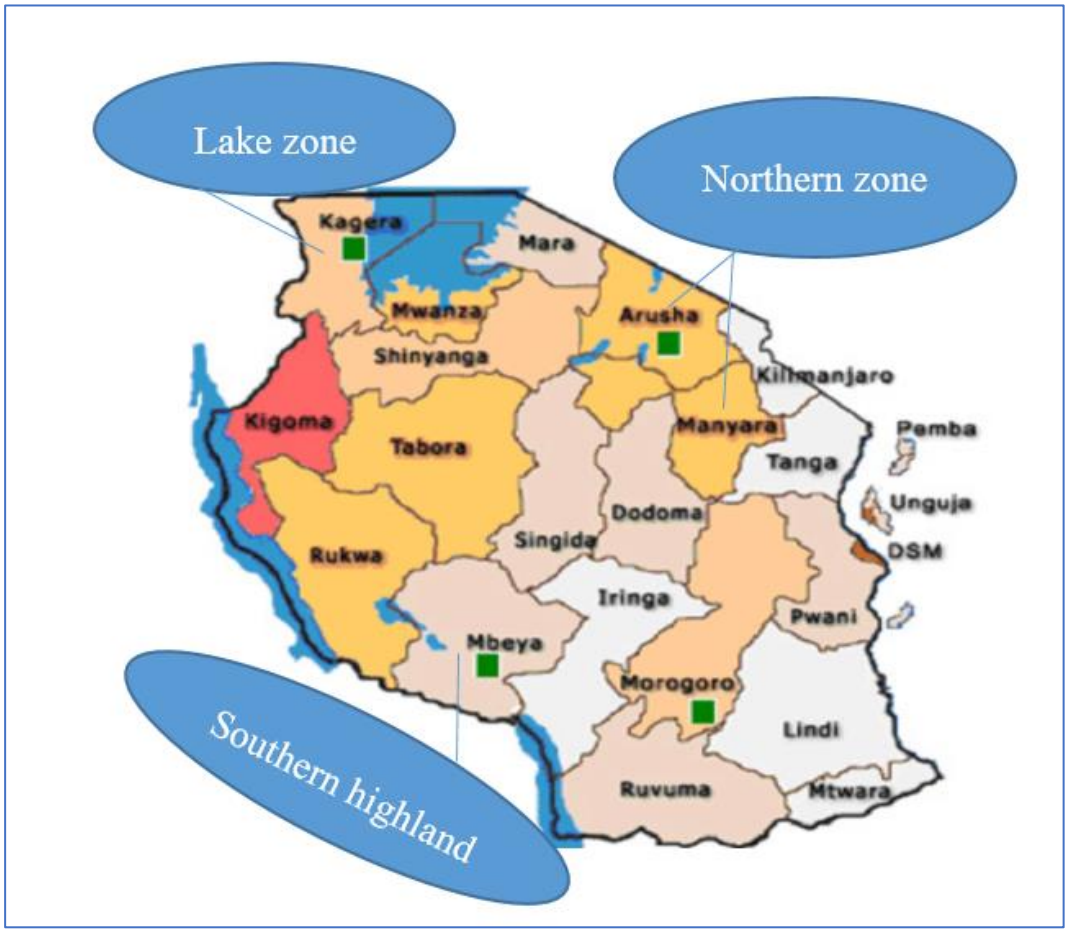

Figure 5: Research study area

3.2 Target Process

This project aimed to change the breeding phenotyping process from how they were used to manually score diseases, data recording and manually analysis of collected data to identify resistant genotypes.

3.3 Sampling Technique and Sample Size

The purposive method was used as a sampling technique because it involves intentionally selecting participants who meet certain criteria or characteristics that are relevant to the research question or objectives (Nikolopoulou, 2022). This method allows researchers to target specific populations or individuals that are more likely to provide valuable insights for the study (Palinkas *et al.,* 2015). In this study, four sites Arusha, Manyara, Mbeya, and Bukoba

were used for requirements. Breeders were purposely the sample of the study as the project was aiming to help them in the phenotyping of ALS and CBB while they are working on producing improved common bean varieties and sixty respondents were involved in providing the gathered requirements. These 60 respondants are the total population of Tanzanian breeders involved in Common Bean breeding system in the stated zones where this research was conducted.

3.4 System Requirements

3.4.1 Data Collection Methods and Tools

To collect the system requirements, qualitative techniques such as focus groups, interviews, and observations were employed.

(i) Primary data source

Interview

Different common bean breeders and field technicians were asked different formulated questions related to their demography, phenotyping process, the need for the system in their daily work, and what the system should be doing that is not possible for them with their manual phenotyping. This helped us in getting to understand what is happening with genotypes identification field processes which led us to get the functional requirements of the system. The interview took approximately an hour per person and answers were recorded manually by writing. Main questions asked are found on Appendix 1.

Focus group discussion

To supplement interviews, closed focus group discussions were also applied to have better results in obtaining the needed requirements. With this group discussion, we applied a list of 35 questions as seen in Appendix 1 where they were divided into groups from digital literacy, breeding processes, phenotyping, and data capacity. The answers were captured by taking notes on each question asked and different views from different respondents.

Observation

After conducting interviews and discussing with the breeders and field technicians, we continued with them in different fields where we had a chance to observe the process of manual phenotyping following Appendix 2 and checking errors that are being made as it depends on the eyes of the viewer for them to score the plant. We also focused on checking time spent in the field doing the scoring of diseases and data records.

(ii) Secondary data source

This involved searching around the previous work done on disease detection, genotype identification using technology, and digital phenotyping among different crops. We focused on checking if there are existing works done related to common bean automated phenotyping, methodologies used, and limitations they are facing that need to be addressed.

3.4.2 Data Analysis

Qualitative data analysis was done by assessing non-numerical information, which was text on our side to uncover patterns, themes, and insights of all collected information. It was done by developing a coding scheme based on in-depth interviews, focus group talks, and observation records. After that, a Python library Matplotlib was used to plot different views from the data gathered with the main patterns of interest for this study. From the results found after analysis, we realized the need for automating the phenotyping process for the common bean breeding system in Tanzania so that it will be easier to identify the resistant varieties of common beans in the country.

3.5 System Development Approach

Through this work, the Agile XP methodology was used. This agile approach was used for the development of the system. This is a software development methodology that enables developers to create a prototype, show it to users, and then modify it based on their feedback before the release of the whole system with its five core values including simplicity, feedback, communication, respect and courage (Agile Alliance, 2023). Throughout the software development lifecycle, it encourages iterative development and testing. This method produces user-friendly, efficient software promptly.

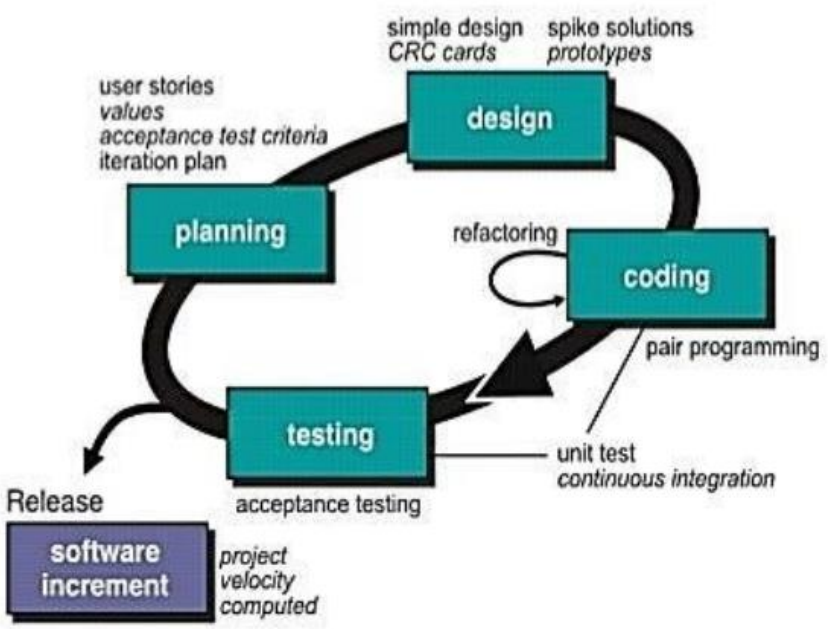

Figure 6: Extreme programming cycle (Noerlina *et al.*, 2020)

Planning: This is the part where we gathered user stories focusing on which ones are providing features to be considered in developing the system so that it will be useful to them.

Design: At this level, we provided a conceptual design of how we are planning to have the system working from backend to frontend with all needed flows of data and how all things will be following each other according to the user stories gathered during the planning phase.

Coding with Refactoring: Throughout this stage, the code for the system was produced, following to secure coding standards and engaging in refactoring on a regular basis to improve the code's design and maintainability. Pair programming was used to assure high-quality code and foster knowledge exchange.

Testing: During this phase, testing was carried out to ensure that the system worked correctly and securely in a variety of circumstances. This entailed filling out new forms in the field, capturing new scans and testing the model performance, as well as validating the veracity of the results on the website. Tests were carried out on a regular basis, and issues were corrected as they arose.

Release: This step entailed delivering and deploying the system. Continuous integration and continuous delivery were used to automate the development, testing, and deployment processes. Feedbacks were received and used to improve the system in subsequent iterations.

3.6 System Design

System design is the process of establishing system characteristics such as modules, architecture, components and their interfaces, and data based on specified requirements. It is the process of discovering, establishing, and designing systems to satisfy the specific aims and criteria of a firm or organization (Collegenote, 2023).

3.6.1 Block Diagram

Figure 7 shows the blog diagram of the system.

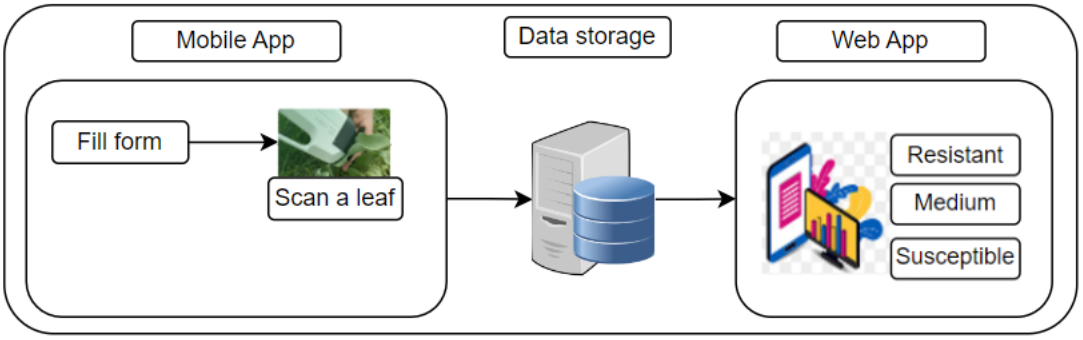

Figure 7: Block diagram

3.6.2 Activity Diagram

Activity diagrams are useful in modelling systems and processes, as they allow us to visualize and understand the flow of activities within the system. For this work, the process of activities are shown in Fig. 8.

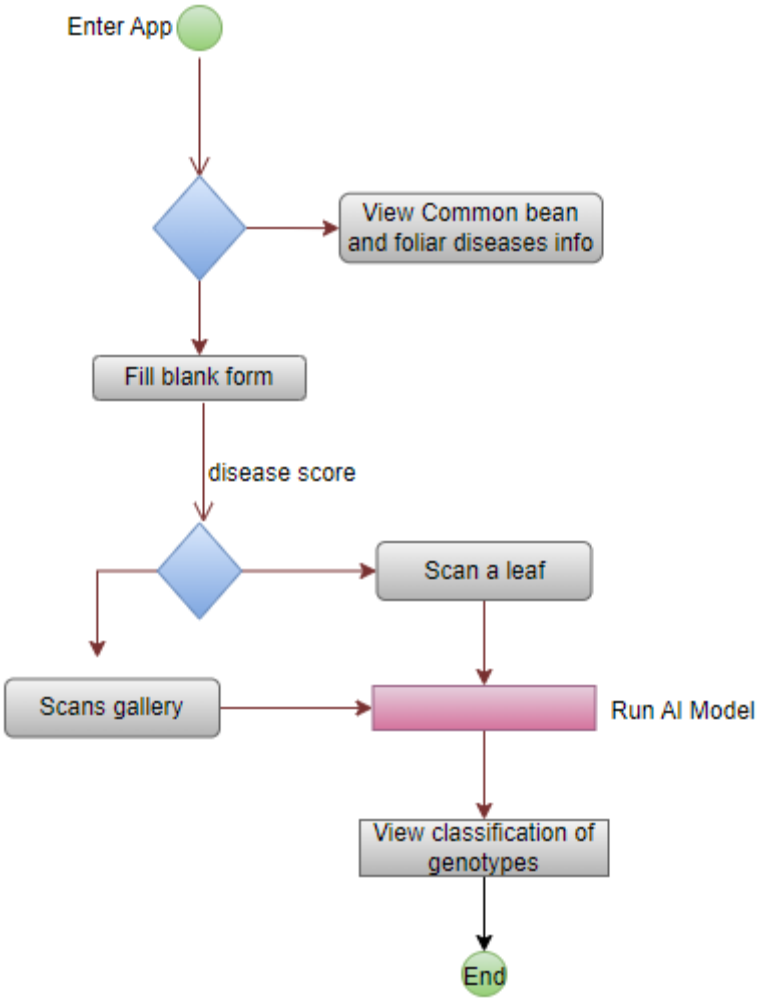

Figure 8: Activity diagram

As shown in Fig. 8, the user will enter the mobile application and then register the project with their username and password. After logging in, the user will have access to the data collection of diseases in question by filling out the genotypes information and then the system will take them to the scanning part using the spectrometer which will be connected to that mobile phone. After scanning the user will be able to save the filled form and data will be sent to the server for further analysis to be shown on the web application where they will be able to check the classification of different genotypes that were involved in data collection.

3.6.3 Sequence Diagram

A sequence diagram (Fig. 9) is a Unified Modeling Language (UML) diagram that depicts interactions between objects in sequential order. It represents the messages exchanged between

objects in a particular scenario or use case and shows the flow of control from one object to another over time.

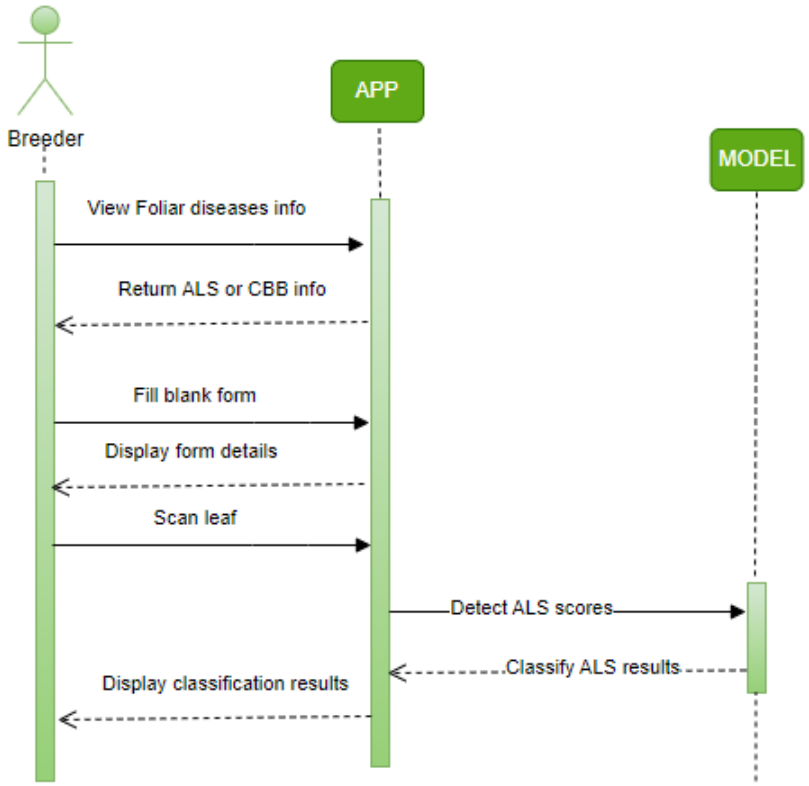

Figure 9: Sequence diagram

The sequence diagram (Fig. 9) shows how the internal flow of the data works and communication between different modules of the system. Once the user requests information from the application, the information is given to them by the mobile and web applications, the same applies when the user wants to fill out the data collection information form. When it reaches the scanning part the application opens the spectrometer to scan the leaves, after scanning the leaves of different plants data are sent on the server where the classifier will classify those scans. Those data are then used by the web application to display different reports needed by the users.

3.6.4 Use Case Diagram

A use case diagram represents the information about your system's users (also known as actors) and how they interact with it (Lucidchart, 2023). The use case for the Hyperspectral-based

system is shown in Fig. 10 and all the use cases from login, to view common bean information, filling the field form, scanning the leaves and displaying the classification are explained in Table 1.

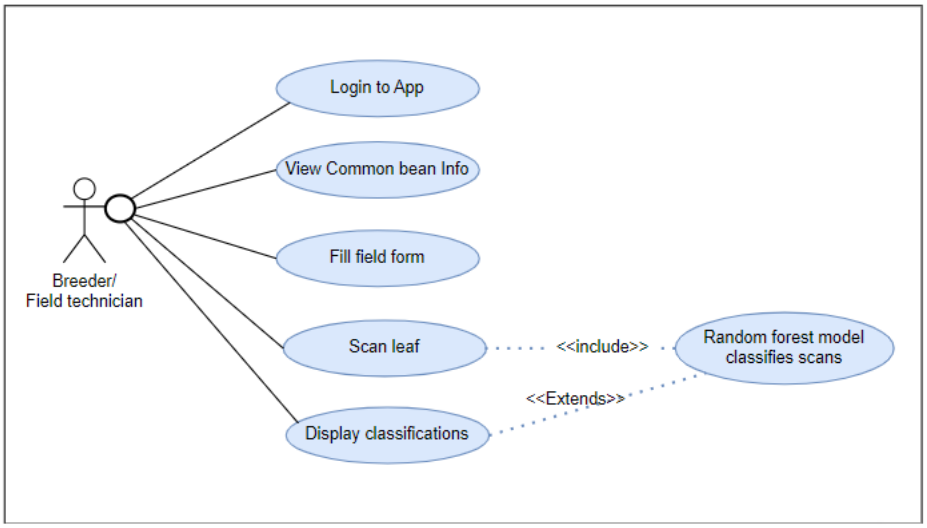

Figure 10: Sequence diagram

Table 1: Cases description

Use cases	Description	Actor
Login to Applications	The user is able to login to the mobile and web applications	Breeders and field technicians
View common beans information	The user can view different information regarding common beans and foliar diseases in beans	Anyone who wish
Fill field form	The user is able to fill the filed form with all planted genotype information	Breeders and field technicians
Scan a leaf	The system user can scan the leaf using the spectrometer	Breeders and Field Technicians
Display classifications	The users view the classifications results of scanned plants.	Breeders and Field Technicians

3.7 System Development

3.7.1 Hardware Requirements and Tools Used

(i) The QED Scan Spectrum

ScanSpectrum is incredibly portable and small enough to fit in your palm. It connects to both a USB power bank and your own Android phone. Everything you require is included in a lightweight (0.8 kg) suitcase that may fit within your backpack.

Figure 11: Spectrometer

- Operation System Compatibility: Android Version 3.2 to highest
- Spectral range: 400-1000 nm
- Measurement time: 4 seconds
- Spectral image size: 1920 px
- Weight (full kit): 860 g
- Operating temperature: 0-70°C
- Power requirements: 5 V, 2 A USB source
- Working modes: transmittance using standard cuvettes, reflectance (qed, 2023).

Figure 12: Spectrometer connected to Android Phone

(ii) Galaxy A14

- Operating System: Android 13
- Storage: 128 GB
- RAM: 4 GB
- Display: 6.6 inches
- Battery: Li-Po 5000 mAh, non-removable

(iii) Power Bank

Powerology ppbcha07

- Capacity: 30000 mAh
- USB-C PD Input: 45 W 5 V/3 A, 9 V/3 A, 12 V/3 A, 15 V/3 A, 20 V/2.25 A
- USB-C PD Output: 45 W 5 V/3 A, 9 V/3 A, 12 V/3 A, 15 V/3 A, 20 V/2.25 A
- USB-A Out 1 & 2: 18 W 5 V/3 A, 9 V/2 A, 12 V/1.5 A
- Cable: 90 cm Charging Cable
- Cable Type: USB-C to USB-C
- Total Output: 45 W Max

3.7.2 Software Requirements

(i) Backend Development

Python

Python is a powerful and easy-to-learn programming language. It features efficient high-level data structures and a straightforward yet powerful method for object-oriented programming. Python is an excellent language for scripting and rapid application development on a variety of platforms because of its fantastic syntax, dynamic typing, and interpreted nature (Python, 2023). That is why it was a suitable programming language to develop different parts of the system.

Google Colaboratory

Colaboratory is a hosted Jupyter Notebook service that offers free access to processing power, including GPUs and TPUs, and doesn't require any setup. Colab works particularly effectively in the fields of education, data science, and machine learning (Colab, 2023). It was used for the training and testing of the models used in this study.

XAMPP

In this project, XAMPP acts as the local server environment for hosting and testing the system. It is made up of the Perl programming language, PHP for server-side scripting, MySQL database management system, and Apache web server. Because XAMPP facilitates a seamless

connection between the server and the other system modules, it offers a dependable testing environment for developers.

Hypertext Preprocessor (PHP)

Hypertext preprocessor (PHP) is a server-side programming language popular in web development. It was utilized in this project to develop the web application that interfaced with the database and performed the business logic.

Codegniter

CodeIgniter, a powerful PHP framework, is used as the server-side scripting language to manage data processing, user authentication, and database connection. It includes built-in support for AES encryption, which reinforces the system's security features. CodeIgniter ensures secure and efficient data handling, which contributes to the overall system reliability.

Draw.io

Draw.io, a strong diagramming tool, was vital for developing thorough flowcharts, use case diagrams, and other schematic representations in this project. This web-based tool simplified the visual representation of system activity, user interactions, and hardware connections.

(ii) Fronten Development

Java script

JavaScript is a client-side programming language used in web development. It was used in this project to provide interaction and functionality to the user interface of the web application. In this project, WebSocket technology, which was implemented in JavaScript, was used to offer real-time communication between the online application and the mobile application. WebSockets enable real-time data transmission between the server and the client by providing a persistent, bidirectional communication channel via a single, long-lived connection.

Bootstrap

Bootstrap is essential for generating a responsive and mobile-friendly design for the system. As a front-end framework, Bootstrap provides a collection of pre-designed components and styles that are used to improve the visual appeal and responsiveness of the user interface.

Hypertext Markup Language (HTML)

Hypertext Markup Language (HTML) and CSS (Cascading Style Sheets) are important components in this project since they create the system's user interface. The HTML defines web page components and arranges text and layout, whereas CSS increases visual appearance and styling. They collaborate to create an easy-to-use interface that allows users and the system to connect easily.

Opem Dta Kit (ODK)

Open Data Kit (ODK) allows you to create powerful forms to collect data from anywhere. This was used to create the form that is being used in android to collect field data and capture the scans for diseases detection.

(iii) Database Design

MySQL

MySQL is a powerful relational database management system that is open source and free. In this project, it was utilized to store and manage client, field data storage, and system settings data.

3.7.3 Development of Standards Operating Procedures (SoPs)

For the development of the proposed system, the first thing was to develop SoPs for the system to comply with the breeding processes and protocols as it was a new methodology introduced in the work environment.

(i) Scans per leaf

Regarding the size of the spectrometer window and the size of the bean leaf, this point was made by breeders and pathologists to determine which number of scans will be used to represent the whole leaf of the selected plant, according to the growing stage of the plant. Directed by phenotyping protocols, at the first trifoliate while the size of leaves is still small, two to three scans will be taken around the leaf focusing on the area of lesions concentration if any. At the maturity stage where the plant has big leaves, six to eight scans will be considered as a representative of the leaf following the same process of lesions and making sure all angles are kept as seen in Fig. 13.

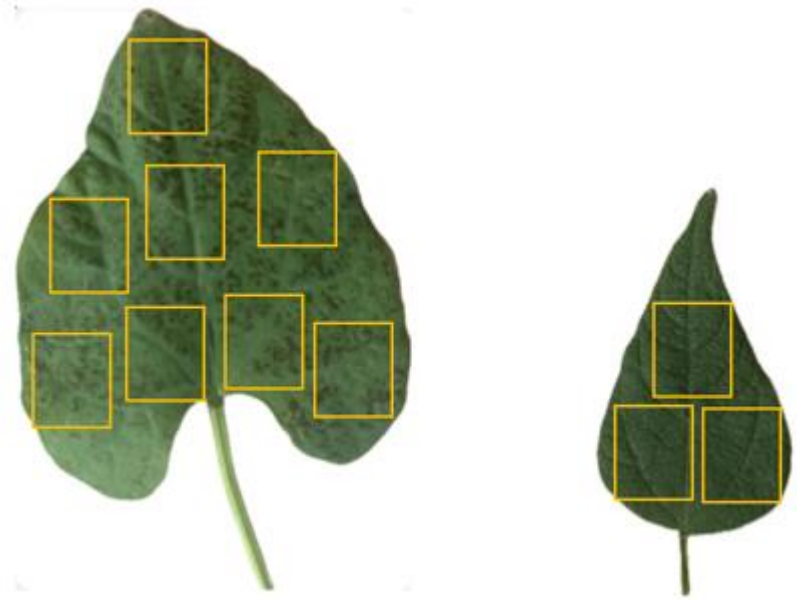

Figure 13: Scans per leaf

(ii) Leaves per Plant

After deciding of scans that are representing the leaf, the following point was figuring out how many leaves will represent the whole plant. Considering that plants have different number of leaves at different growing stages, at this point with breeders following plant pathology assessment protocols, at the early stages of the plant three leaves of the first trifoliate must be scanned and at the maturity stages only four middle leaves will be considered since the old leaves are not considered while scoring plants. New leaves will also not be scanned since they don't get affected at their early stages and the focus has also to be given on leaves with lesions.

(iii) Plants per Genotype

For the planted plants to represent the whole genotype, 80% of the planted plants must be among the selected sample for scanning.

(iv) Scoring Procedure

After checking all the needed parts of the data collection, the followed part was the process of checking how the scores will be considered according to breeding system. As there will be many scans of the plant and it might happen than all the scans are not having the same scores, the protocol of pathology illustrates that the highest score that is found on the plant is the one considered to be the disease score level on that plant and mapping the highest starts from the susceptible class to medium then resistant.

(v) Scoring a Plant

The pathology manual scoring protocol was adopted to classify the scans of the same plant. That means, if different scans of the same plant have different categories of diseases scores, the highest score among them was considered. If one plant has both medium and resistant scores in their classes, the plant is considered medium class, if it has susceptible and medium the plant is considered susceptible, and if it has all the three classes the plant will still be considered susceptible. That was the protocol followed in scoring every plant that happened to have different scores.

(vi) Scoring a Plot/Genotype

This was also done using the same current manual pathology technique of identifying resistant and susceptible genotypes. This is accomplished by scoring the genotype by percentages based on all of the plants scored on that same genotype, where for a genotype to be considered resistant, all plants classified as resistant must be greater than 80%, the same for the medium category, and all remaining genotypes are considered susceptible. The calculations are based on the percentage of each class of plants of that genotype, as given in the following formula.

$$\mathrm{X} = \sum_{k=1}^{n} \mathrm{K} * 100/\mathrm{N}$$

Where:

X= Percentage of the K class

K= Plant one of class

n=Number of K plants

N= Total number of plants in the same genotype

3.7.4 Field Setup

After developing SoPs for the data collection, the next step was to set up fields where the experiments were made. This includes the planting material process of inoculation and all other field-related works as seen in Figs. 14 and 15.

Figure 14: Screen house

Figure 15: Open field

(i) Planting Material

To get data to be used in training and testing of the system, trials were set with ten different common bean genotypes in the screen houses and open fields. Those genotypes used in the trial were: Kigoma, Mont Calm, Mexico 54, Uyole 03, Seliani13, Tari Bean5, Tari Bean 6, Cal 143, Uyole 96, Uyole 94, Jesca and Cal 96. Those genotypes were used as they are the ones under experiment in Tanzania and they are in the process of being improved for foliar disease resilience and higher yields production at the moment.

Figure 16: Common beans varieties

(ii) Screen house experiment design and planting

Five screen houses were set two from Arusha, two in Mbeya, and one in Kagera. A pot experiment arranged in a split-plot design with three replications was set, with disease infestation being the main plots and bean genotypes sub-plots. The experiment had main plot treatments for ALS and CBB, where 10 bean genotypes (sub-plots) were planted in three replicates in the main plots. Three seeds of a bean genotype were planted in each pot. The moisture of soils in the pots was kept at field capacity during the entire experimental period by irrigation. ALS and CBB were artificially inoculated for all genotypes in all experiments.

To supplement those screen houses, four open fields were also established two fields in Arusha, one in Manyara, and another one in Mbeya. The setup of the planting both for screen houses and open filled are summarized in Plots 1 to 6.

Plot 1:

Plot	1101	1102	1103	1104	1105	1106	1107	1108	1109	1110
Rep	1	1	1	1	1	1	1	1	1	1
WP	1	1	1	1	1	1	1	1	1	1
SP	1	2	3	4	5	6	7	8	9	10
Dis	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB
Gen	10	4	2	9	7	8	5	1	6	3

Plot 2:

Plot	1201	1202	1203	1204	1205	1206	1207	1208	1209	1210
Rep	1	1	1	1	1	1	1	1	1	1
WP	2	2	2	2	2	2	2	2	2	2
SP	1	2	3	4	5	6	7	8	9	10

Dis	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS
Gen	6	1	7	3	5	9	2	10	8	4

Plot 3:

Plot	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110
Rep	2	2	2	2	2	2	2	2	2	2
WP	1	1	1	1	1	1	1	1	1	1
SP	1	2	3	4	5	6	7	8	9	10
Dis	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS
Gen	10	7	2	6	1	8	4	9	5	3

Plot 4:

Plot	2201	2202	2203	2204	2205	2206	2207	2208	2209	2210
Rep	2	2	2	2	2	2	2	2	2	2
WP	2	2	2	2	2	2	2	2	2	2
SP	1	2	3	4	5	6	7	8	9	10
Dis	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB
Gen	1	6	10	3	7	9	5	2	4	8

Plot 5:

Plot	3101	3102	3103	3104	3105	3106	3107	3108	3109	3110
Rep	3	3	3	3	3	3	3	3	3	3
WP	1	1	1	1	1	1	1	1	1	1
SP	1	2	3	4	5	6	7	8	9	10
Dis	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS	ALS
Gen	5	2	1	9	4	10	8	6	3	7

Plot 6:

Plot	3201	3202	3203	3204	3205	3206	3207	3208	3209	3210
Rep	3	3	3	3	3	3	3	3	3	3

WP	2	2	2	2	2	2	2	2	2	2
SP	1	2	3	4	5	6	7	8	9	10
Dis	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB	CBB
Gen	8	9	5	10	6	4	7	1	3	2

Where:

WP : Whole Plot

SP : Sub Plot

Rep : Replications

Dis : Disease

Gen : Genotype

ALS : Angular Leaf Spot

CBB : Common Bean Bacterial Blight

Figure 17: Plantations

(iii) Inoculation

After planting those ten genotypes according to breeding SoPs, the inoculation was the next step in trial follow-up. They were also done depending on disease inoculation protocols where for ALS it was done on the 14th day after planting in screen houses, and on the open field it was done between 17 to 21 days depending on when the first trifoliate came up. For the CBB, in the screen houses, it was the same as for ALS where it was done on the 14th day after plantation and 21 to 27 days in the open fields depending on the first trifoliate. Plants were then evaluated 7-10 days after inoculation before and when the disease started showing up on leaves.

Figure 18: Inoculating plants

3.8 Dataset

With the usage of a powered handheld spectrometer connected to an Android phone, data were collected from different mentioned fields and the collected data were about ten genotypes stated. The spectrometer was first calibrated to capture all the light reflectance using the white tape which represents the combination of all colors, and then after calibration, the spectrometer was used to capture leaf reflectance directly with the black tape placed at the bottom of the leaf to avoid any light interference as black is a representation of no color at the object. With those setups, we started scanning leaves following the made-up scanning SoPs, within the stated four sites a set of 1105 scans were collected and divided into different sets to be understood by the

computer program these scans were composed of 1 270 750 bands considered individually as part of the wavelength.

Figure 19: Data collection

3.8.1 Data Processing

Data processing is a stage in the data analysis process that takes raw data and converts it into a format that computers and machine learning can understand for analysis (Mesevage, 2021). With this work, the data were processed by labeling them with names that will make it easier for the algorithms to understand and classify them according to the labels. The newly labeled data were then used as input to the models. We collected 1105 scans as wavelengths and each scan has 1150 bands considered individually. Those data were corrected in four Regions in Arusha we corrected 505 scans, 100 scans in Manyara, 320 in Mbeya, and 180 scans were captured in Bukoba Region. All those data were corrected manually and also manually labeled according to the mentioned three categories.

3.8.2 Data Labeling

The collected bands were manually labeled according to different scoring scales from one to nine then as the classes of the output results genotypes had to be identified as resistant, medium, or susceptible to the selected two foliar diseases, and those wavelengths were labeled according to those main three categories. The summary is found in Table 2. All data were labelled with the help of experts in plant pathology and breeders from data collection to data analysis and validation.

Table 2: Data description

Regions	Data collected	Resistant	Medium	Susceptible
Arusha	580 750 bands	200 000 bands	280 000 bands	100 750 bands
Manyara	115 000 bands	20 000 bands	55 000 bands	40 000 bands
Mbeya	368 000 bands	150 000 bands	90 000 bands	128 000 bands
Bukoba	207 000 bands	80 000 bands	69 000 bands	58 000 bands

3.9 System Frameworks

3.9.1 Framework Design of the System

This framework shows the logical flow of the system from the field to the web view where the user uses a mobile application for field data collection and scan leaves with a spectrometer, after that the random forest model will perform the classification of the scans in the three main categories needed by breeders, and then with the web application breeders and field technicians will be able to view collected data and different field data analysis and at the end, the user will be able to check genotypes identified as resistant, medium or susceptible as shown in Fg. 20.

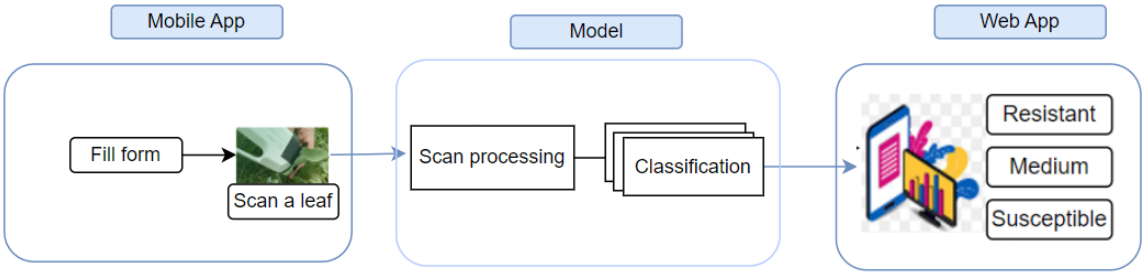

Figure 20: Framework of the system

3.9.2 Framework Design of the Classifier

The Classifier framework shows how the flow of the classifier works where different set of bands were collected and divided into two main groups one which counts 80% was used for training the model and the other part which was 20% of the whole collected bands was used for testing the model performance. After the first training and testing the model's parameters were updated to get the model tuned for us to get the optimized model. The model was deployed

on the server where the web and mobile applications were storing data for easy access and retrieval of data.

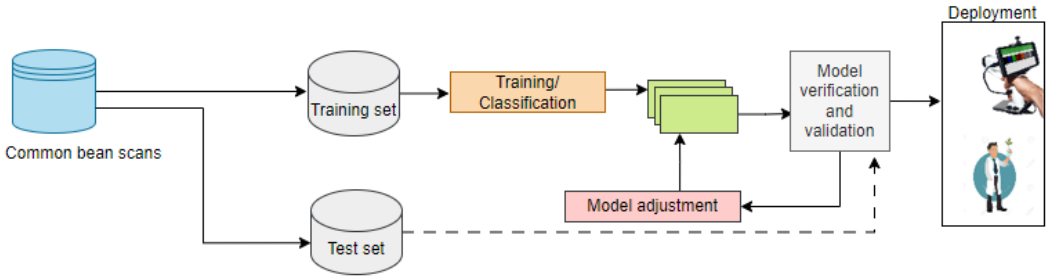

Figure 21: Framework of classifier

3.10 Algorithms Used

We used different existing algorithms including random forest, XgBoost, and Neuro Network. Those models were used before and proved outstanding higher performance on complex datasets, robust to overfitting, and proved to be effective in classification and regression issues. The usage of random forest demonstrated higher performance in different studies conducted by researchers like Saini (2023) in his study of using Random forest and Adaboost (adaptive boost) models using vegetation indexes for features analysis and Random forest outperformed others with higher accuracy. Another random forest model was used by Wójtowicz *et al.* (2021) to check the wheat disease where they find out that with the selected wavelength the accuracy was more than 96%.

3.10.1 Random Forest

Random forests are a tree-based machine-learning systems that harness the power of many decision trees. Tin Kam Ho, who led the Statistics and Learning Research Department at Bell Laboratories at the time, developed the first such method in 1995 and Leo Breiman and Adele Cutler then expanded (Dataiku, 2021). Random Forest is a popular machine learning algorithm from the supervised learning technique. It can be used for both classification and regression problems in machine learning. Its foundation is the idea of ensemble learning, a technique that combines multiple classifiers to solve a challenging issue and enhance the functionality of the model (Javapoint, 2023). According to Reza *et al.* (2016) of the University of California, A random forest is created by combining different tree predictors so that every tree in the forest

is dependent on the values of a random vector that is randomly sampled and has the same distribution for every tree. They continued explaining that a random vector Θk is created for the kth tree, independent of the previous random vectors Θ1, ... , Θk−1 but with the same distribution; and a tree is formed using the training set and Θk, resulting in a classifier h(x,Θk) where x is an input vector (Reza *et al.*, 2016).

From a computational perspective, Random Forests are advantageous because they can easily handle parallel implementation, handle both regression and (multi-class) classification naturally, train and predict relatively quickly, rely only on one or two tuning parameters, and have an integrated estimate of the generalization error (Cutler *et al.*, 2012). The random forest predicts the result based on the majority vote of predictions from each decision tree, as opposed to relying solely on one (Javapoint, 2023).

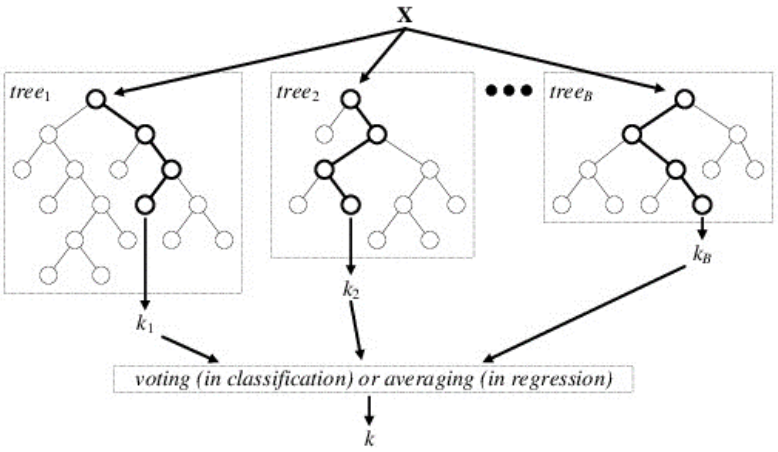

Figure 22: Random forest architecture (Vaiciukynas, 2023)

3.10.2 Extreme Gradient Boost

The XGBoost is a supervised learning gradient-boosting technique released in 2014. It is an open-source library that can train and test models on massive datasets. It has become the machine learning algorithm of choice for data scientists and machine learning developers and can be applied in a variety of fields (Simplilearn, 2023). The XGBoost, like other boosting approaches, works by sequentially adding new models to the ensemble. However, unlike bagging methods such as Random Forest, which grow trees in parallel, boosting approaches train models one after the other, with each new tree assisting in the correction of faults generated by the previously trained tree (Wohlwend, 2023). The system's significance has been

widely recognized in a variety of machine learning and data mining challenges like the challenge hosted by the machine learning competition site Kaggle where 17 of the 29 challenge-winning solutions 3 published on Kaggle's blog in 2015 employed XGBoost (Chen & Guestrin, 2016).

3.11 Training of Models

3.11.1 Training Random Forest Algorithm

A Random Forest Algorithm was trained for the aim of classifying scans in the three categories needed by breeders. All the 1105 scans composed of 1150 variables of the wavelength which made 1 270 750 bands were divided into the training set and testing set. Where, as we have said the training set was 80% of the total set of 1 016 600 bands and the testing set of 20% which is 254 150 bands. All data were labelled according to what the algorithm can understand so that it will be easier for the classification. Hyper parameters used for the random forest model as shown in Table 3.

Table 3: Random forest parameters

S/N	Hyperparameters	Values
1.	N_Estimators	100
2.	Criterion	Gini
3.	Max_depth	None
4.	Min_samples_split	2
5.	Min_samples_leaf	1
6.	Min_weight_fraction_leaf	0.0
7.	Max_features	Auto
8.	Maxleaf_nodes	None
9.	Min_impurity_decrease	0.0
10.	Bootstrap	True
11.	Oob_score	False
12.	N_jobs	None

13.	Random_state	42
14.	Verbose	0
15.	Warm_start	False
16.	Class_weight	None

3.11.2 Training the Extreme Gradient Boost Algorithm

An XgBoost Algorithm was trained with the intention of categorizing scans into the three groups required by breeders. All 1105 scans were separated into training and testing sets, which were made up of 1150 wavelength variables totalling 1 270 750 bands. As previously stated, the training set comprised 80% of the overall set of 1 016 600 bands, whereas the testing set comprised 20% representing 254 150 bands. All data were labeled according to what the system could understand, making classification easier. The Hyperparameters for the XGBoost model are shown in Table 4.

Table 4: The XGBoost parameters

S/N	Hyperparameters	Values
1.	N_Estimators	100
2.	Learning_rate	0.3
3.	Max_Depth	6
4.	Min_Child_Weight	1
5.	Subsample	1
6.	Colsample_bytree	1
7.	Colsample_bylevel	1
8.	Gamma	0
9.	Reg_alpha	0
10.	Reg_lambda	1
11.	Scale_pos_weight	1
12.	Base_score	0.5
13.	Objective	'binary:logistic'

14.	Booster	'gbtree'
15.	Tree_method	'auto'
16.	n_jobs	1
17.	Random_state	0
18.	Use_label_encoder	True
19.	Eval_metric	error

3.11.3 Training the Neuro Network Algorithm

Neuro network models were also found very accurate in classifying those complex dataset as they use neuros and different layers in tackling and checking features in those dataset to get the needed results with reasonable accuracy. All 1105 scans were separated into training and testing sets, which were made up of 1150 wavelength variables totaling 1 270 750 bands. As previously stated, the training set comprised 80% of the overall set of 1 016 600 bands, whereas the testing set comprised 20% representing 254 150 bands. The Hyperparameters used to get higher accuracy of the Neuro Network model were shown in Table 5.

Table 5: Neuro network parameters

S/N	Hyperparameters	Values
1.	Input layer	128 neurons
2.	Hidden layers	One with 64 neurons
3.	Output layer	The number of neurons matches the number of classes in the one-hot encoded labels.
4.	Dropout rate	0.2
5.	Optimizer	Adam
6.	Loss function	Categorical cross-entropy
7.	Metrics	Accuracy
8.	Number of epochs	300
9.	Batch size	32

3.12 Model Performance Evaluation

In machine learning, model assessment is the process of determining a model's performance using metrics-driven analysis (Iguazio, 2023). It is a process of analyzing the performance, strengths and limitations of the trained model using different criteria. In this work, the criteria used to evaluate the random forest model performance were precision, recall, F1-score, confusion matrix, receiver operating characteristics, and out-of-bag (oob) over a number of trees, classes error rates and mean average precision.

3.12.1 Precision

Precision is one of the metrics used to evaluate model's performance, it indicates how frequently a machine learning model predicts the positive class accurately (AI, 2023) it mainly helps in knowing how positive predictions are correct. Precision is defined as the number of true positives divided by the total number of positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Where:

TP (True Positive): Number of all cases classified as positive. For our study, they are those data that are truly detected as resistant, medium, or susceptible.

FP (False Positive): Those detected as Medium, Resistant, or Susceptible to foliar diseases while they were not in either of the classes.

3.12.2 Recall

Recall is a metric that indicates how frequently a machine learning model accurately detects positive examples (true positives) from all of the actual positive samples in the dataset. Recall is computed by dividing the total number of positive cases by the number of true positives. The latter consists of false negative results (missed cases) and true positives (successfully detected cases) (AI, 2023).

$$\text{Recall} = \frac{TP}{TP+FN}$$

Where:

TP (True Positive): Number of all cases classified as positive. For our study, they are those data that are truly detected as resistant, medium, or susceptible.

FN (False Negative): Number of all cases in which the projected value is negative when the real value is positive.

3.12.3 The F1 Score

The F-score, commonly known as the F1-score, is a model's accuracy on a dataset. It is employed in the evaluation of binary classification systems. It is a method of aggregating the model's recall and precision, it can also be defined as the harmonic mean of these two metrics (Wood, 2023) which makes it ideal for unbalanced datasets (Singh, 2023).

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.12.4 Confusion Matrix

A confusion matrix is a machine learning performance evaluation tool that represents the accuracy of a classification model. It shows the quantity of false positives, false negatives, true positives, and true negatives. It provides us with direction to reroute our path and aids in the evaluation of our model's performance and areas of failure (Bhandari, 2023).

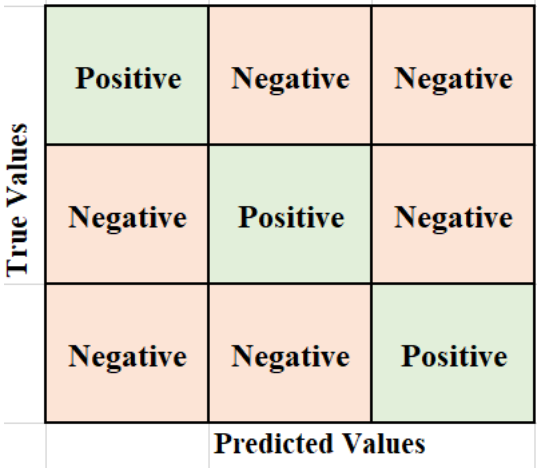

Figure 23: Confusion matrix for multiclass

3.12.5 Class Error Rate

The class error rate (ERR), also known as the per-class error rate or classification error rate per class, is a metric that evaluates the performance of a machine learning model in a multiclass classification issue on a per-class basis. It describes a metric used to quantify how much a model's predictions differ from the actual model (Ting, 2011) . the frequency with which the model misclassifies instances for each specific class. It is evaluated using the following formula.

$$ERR = \frac{FP + FN}{P + N}$$

Where:

ERR: Error Rate

FP+FN: Total number of Missclassifications

P+N: Total number of all Instances

3.12.6 Algorithm Selection

When creating a functional prediction model or a model to comprehend the data-generating mechanism, model selection is an essential stage (Zhang & Yang, 2015). This was done by comparing different evaluation metrics, training time, and classification response time to check which model performs better than others.

3.13 Model Deployment

Model deployment refers to the process of integrating a machine learning model into an already-existing production environment in order to use data to inform actionable business choices (Data Robot, 2023). This allows users, developers, or systems to access the model's predictions so they may utilize the data to inform decisions and communicate with their applications (Iguazio, 2024).

The developed model was deployed on the Amazon Web Server (AWS) due to the complexity of the dataset time spent on processing scans of each data captured from the field and for easy analysis for web display. This was done by getting the model API and give it to the web

application for cloning data and having classes of the genotypes automatically without users' intervention as needed by the end users.

3.14 Vegetation Indexes

Vegetation indexes also known as Spectral indices are used to calculate the relative abundance of particular features of interest about a plant (Geospatial, 2023). During this study following are different indexes were used.

3.14.1 Chlorophyll Index (CI)

The chlorophyll index determines the overall chlorophyll content of the leaves. The CIgreen and CIred-edge readings are sensitive to minor changes in chlorophyll concentration and are consistent across most species (Hiphen, 2023).

CI is calculated using the following formula:

$$CL = \frac{R880}{R590} - 1$$

Where:

CL: Chlorophyll Index

R880: Leaf Reflectance at NIR band

R590: Leaf Reflectance at Green band

3.14.2 Photochemical Reflectance Index (PRI)

Photochemical reflectance index (PRI) as a narrow-band vegetation index as a proxy for photosynthetic activity in vegetation (Sasagawa *et al.*, 2022). This index is a reflectance measurement that is sensitive to changes in carotenoid pigments. Carotenoid pigments indicate the rate of carbon dioxide uptake by plants per unit energy received. It is calculated using the following formula:

$$PRI = \frac{R570 - R530}{R570 + R530}$$

Where:

PRI: Photochemical Reflectance Index

R570: Leaf Reflectance at 570 band

R530: Leaf Reflectance at 570 band

3.14.3 Water Band Index (WBI)

This index measures reflectivity and is sensitive to variations in water state. As the water content of plants increases, so does the degree of absorption at 970 nm in comparison to 900 nm (Geospacial, 2023). It is measured using the formula:

$$WBI = \frac{R900}{R970}$$

Where:

WBI: Water Band Index

R900: Leaf Reflectance at 900 wave

R970: Leaf Reflectance at 970 wave

3.14.4 Modified Chlorophyll Absorption in Reflective Index (MCARI)

The MCARI provides a measurement of the depth of absorption of chlorophyll and is highly responsive to changes in both the Leaf Area Index (LAI) and chlorophyll concentrations. Conditions related to illumination, background reflectance from soil, and other non-photosynthetic materials do not alter MCARI results (Hiphen-plant, 2023). It is calculated in the following way:

$$MCARI = \frac{(R850 - R710) - 0.2(R850 - R570)}{R710}$$

Where:

MCARI: Modified Chlorophyll Absorption in Reflective Index

R850: Leaf Reflectance at 850 wave

R710: Leaf Reflectance at 710 wave

R570: Leaf Reflectance at 570 wave

3.14.5 Nitrogen Reflectance Index (NRI)

The Nitrogen Reflectance Index (NRI) is used to determine the nitrogen level in plants. Similarly to NDVI, green indicates a high quantity of nitrogen while red indicates a low level of nitrogen. The presence of red spots on the map may suggest a nitrogen deficit (Bohl, 2021).

$$NRI = \frac{R570 - R670}{R570 + R670}$$

Where:

NRI: Nitrogen Reflectance Index

R570: Reflectance in green wave

R670: Reflectance in red wave

3.14.6 Structure Insensitive Pigment Index (SIPI)

This index, which measures reflectance, is intended to maximize its sensitivity to the bulk carotenoids ratio. It provides an estimate of the carotenoids to chlorophyll ratio. A crop disease that frequently causes vegetation to lose chlorophyll could be indicated by elevated SIPI results (high carotenoids and low chlorophyll) (Geospacial, 2023). It is defined by the following formula:

$$SIPI = \frac{R800 - R445}{R800 + R680}$$

Where:

SIPI: Structure Insensitive Pigment Index

R800: NIR Reflectance

R445: Blue Reflectance

R680: Red Reflectance

3.14.7 Simple Ratio (SR)

The Ratio Vegetation measure (RVI) or Simple Ratio (SR) is the most basic ratio-based measure. The reflectance in the NIR band divided by the reflectance in the red band yields this index. This value could be very helpful in differentiating between stressed and non-stressed vegetation.

$$SR = \frac{R850}{R675}$$

Where:

SR: Simple Ratio

R850: Reflectance in NIR

R675: Reflectance on red band

3.15 System Testing

3.15.1 Unity Testing

This is the testing of each module and sub-module of the system and it was done individually checking if the module meets the requirements as its own. During the development of each module, we used release testing of each increment where we had three main modules one of mobile form for data collection, classification model and website.

After creating the ODK form, it was tested using different mobile phones with different Android versions and it was working as needed by the end users for them to be able to use normal phones they use in their daily lives without adding additional costs for them to use the developed system. After testing the data collection form and its usability on different phones and maintainability by checking if there is new data to be added, if there is an error while filling out the form and so other requirements from the use, the second part was about checking the classifier. The model was trained and tested using unseen data that were not used during

training or testing and the colouration of the data with the ground truth was a success for the breeders to agree on the developed model to be used as their digital scoring tool. After agreeing on the classifier of genotype according to disease levels, the last part of development and testing was the website where we checked if all parts mentioned in the requirements gathering were there with the parts of data needed.

3.15.2 Integration Testing

Following the development of each module, the next step involved system integration, allowing data to be linked from the field to the website. Following the Agile XP protocol, which involves presenting each release to customers, waiting for their confirmation, and then conducting a testing phase with them to see if there is any part of the system that needs to be revised, the release was made available to users after linking all the desired parts of the system.

3.16 System Validation

Following system testing, the next procedure is system validation. This was accomplished by collaborating with the breeders on how to use the system from the field to the reports they desired from the system.

This was accomplished by filling out several forms from the field, scanning leaves of various genotypes, and uploading data to the server. They then checked the website to see if the same information captured in the field could be found there as it was mentioned in the requirements, and if the classifier had given the classes as needed, which was validated by comparing the system class to the visual scores and the reports with the required analysis.

After their confirmation of the initial requirements and the operational system, the users were provided with confidential questionnaires to complete in order to validate the system's functionality. The primary functional and non-functional needs were acquired through sets of questions on these questionnaires in order to affirm or deny the use of the system for digital phenotyping in order to find resilient common bean genotypes in breeding concepts.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results of Requirements Gathering

4.1.1 Findings from Interviews

(i) Demography of respondents

This part was mostly used to analyze the targeted end users to see if they will be able to use the system or if there is no need to develop a system to avoid usage failure. Among the requirements gathered for this stage were the gender of respondents, age, and education levels to assess the targeted end users for their capacity and ability in technology usage. The main objective of this interview section was to get respondents wishes on what the system will be solving that is not easy for their current processes.

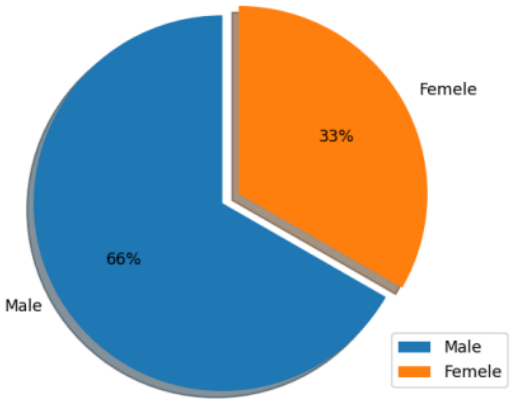

Figure 24: Demography of respondents

Among the interviewed personnel, 66% were men and 33% were women according to the UNESCO report of April 2023 (UNESCO, 2023) this number gives the system a go-ahead to work also considering the GSM's 2019 Mobile Report, 77% of women own a mobile phone

compared to 86% of males in Tanzania (The United Republic Of Tanzania "Realising Competitiveness and Industrialisation for Human Development" Ministry of Finance and Planning, 2021).

This means the targeted group and the targeted country give us the hope to have the system being used. According to Aganecy (2023) first and second specific objectives, Tanzania is aiming to expand on previous accomplishments toward the Tanzania Development Vision 2025 goal of making Tanzania a semi-industrialized, middle-income country, and Strengthening capacity building in science, and technology.

With this requirements gathering we figured out that Tanzanian breeders are of 25 years to 35 years old average age. According to a recent study, 23% of the smartphone users in Tanzania are the younger generation of 18 to 24 years, 73.7% are 25 to 34 years old, 1.7% are 35-44 years old, and 0.8% are the remaining (Start.io, 2023). This means our working environment regarding the age shown in Table 6, the part we are working with are those in the age range that are familiar with smartphone usage in Tanzania which will be easier for the system adaptation and usage.

Regarding the education levels of all respondents, it gives the ease of the system to be used and helpful to them as seen in Table 6, the lowest level of education we have is an advanced diploma in general agriculture and the highest is a PhD. This means the training, testing, and validation of the system will be easier as we are working with intellectual people who have a deep understanding and knowledge of what they are doing in the beans breeding system.

Table 6: Age and education levels of respondents

Age and education levels in the common bean breeding system		
Ages	**Education Levels**	**Percentage**
20-30	Advanced Diploma	8%
20-30	Bachelor	50%
31-40	Bachelor	8%
31-40	Masters	27%
41-50	PhD	7%

(ii) Identified Requirements

Some of the main questions asked during interviews were regarding the need for the system in breeders' daily work and what main problems they are facing that the system should be solving. Answers were mainly about the time spent in the field, the bias of manual phenotyping, changing of data results according to the scorer in the field, and also paper-based data collection which mostly lead to the loss of data, errors in typing, and prone to unreliable decision-making.

Functional requirements

After a long list of interview questions with different individuals involved in the breeding system, the following are the main functional requirements for the system developed to ease their work and help in the production of improved common bean seeds around Tanzania.

Table 7: Functional requirements

S/N	Requirements
1.	The system must be able to help breeders in data collection using smart phones
2.	The system must have a website view for easy data access and storage
3.	The system has to provide collected data with selected traits and their scales
4.	The system must differentiate genotype data at different growing stages
5.	The system has to identify resistant, medium and susceptible varieties according to targeted traits
6.	The system must specify the person, Region and dates of data collection
7.	The system must provide biochemical contents even if there are no visible symptoms

Non-Functional requirements

These requirements describes the system capacity to be used, constraints and operational capacities.

Table 8: Non-Functional requirements

S/N	Requirements
1.	The system shall be reliable for common bean phenotyping
2.	The system’s results need to be efficient

3.	The system must have tolerable performance
4.	Data collection form has to be compatible with different Android phones
5.	The web application has to be accessible on different browsers
6.	The system must allow new data entry traits when needed
7.	The system has to be user friendly

4.1.2 Results of the Focus Group Discussions

Different questions were given to breeders and field technicians to determine the need for automating the phenotyping process in bean breeding. Digital literacy, breeding processes, phenotyping, tools available for data collection, and data capacity were among the discussed topics. Following are some of the key points discussed with breeders around Tanzania.

Data collection tools

One of the key point about this question was to figure out how the data collection was usually being done to check if there are existing methodologies that can be upgraded or if there is any technology in place that is being used for phenotyping and data analysis.

From the analysis of the discussion results, 70% of the data collection is done by using papers where a breeder or a field technician goes to the field with a paper and writes the findings down for further analysis when they get back to their offices.

Twenty five percent (25%) of the answers was the usage of the field books which are the books designed for data collection, the breeder or field technician also goes to the fields with those field books and fills out the observable phenotypic data continues with the analysis when they get back too.

Within the discussion made 1% of respondents use digital tools where they have software that helps in doing data collection only for specific reports not for phenotyping purpose.

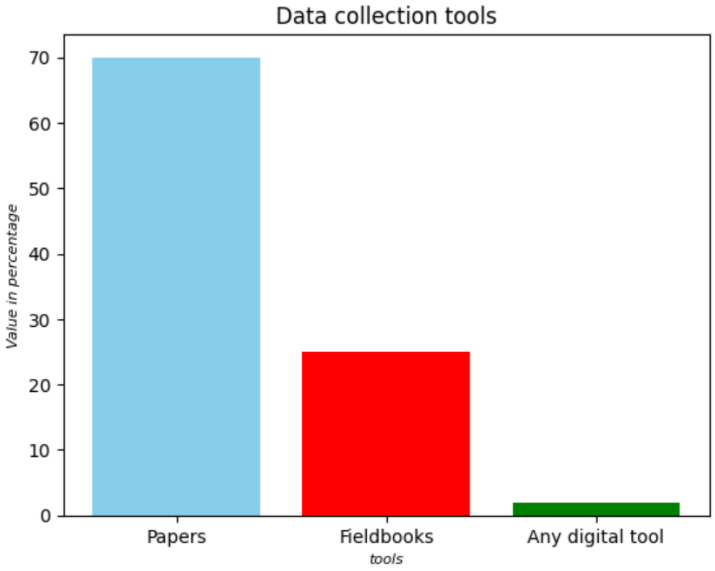

Figure 25: Data collection tools

Digital tools in use

This question was brought in the discussion to check how breeders are familiar with digital tools, with what they use them for and how they benefit from them.

More than 90% are using mobile phones for their daily lives communicating with other breeders on certain decisions and discussing with them in different platforms available on mobiles (Fig.26). The 80% of breeders use computers mostly excel in their daily work especially for data analysis where they come from field with their papers or field books and type down in excel the phenotyping data then continue with analysis using excel. The other mostly used digital tool is Microsoft word and the internet and we didn't find any breeder who doesn't use any digital tool in their daily work.

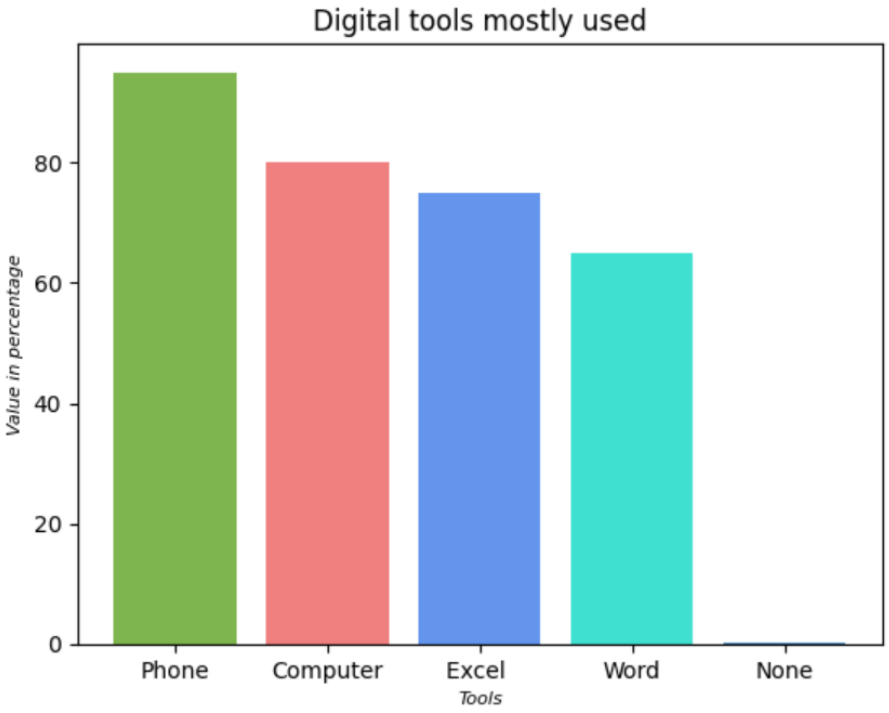

Figure 26: Digital tools commonly used

Willing to adopt new system

Another important parameter that was asked in the discussion was to check if breeders need any system for their work and why 99.3% agreed on adopting the new system that can help them ease the work and processes so that the identification of resistant genotypes can be faster and smoother (Fig. 27). The 0.7% disagreed on that point saying that many have been tempted to give them systems but they only end in plans they didn't receive any.

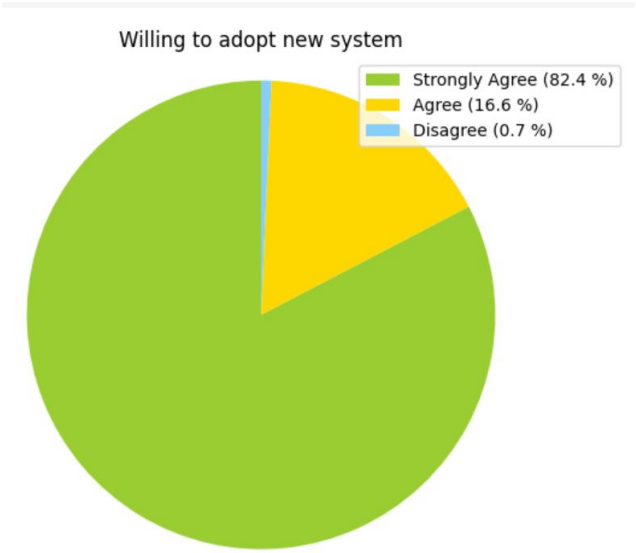

Figure 27: Willing to adopt new systems

4.1.3 Results of Observations

Observation of field processes were also used where different breeders were observed at different time for field works results, time used in the fields and data gathered by different persons.

We have observed that many breeders and field technicians have different data and time used in the field is too much which explains why manual phenotyping is slow, tiring and prone to error (Fig. 28).

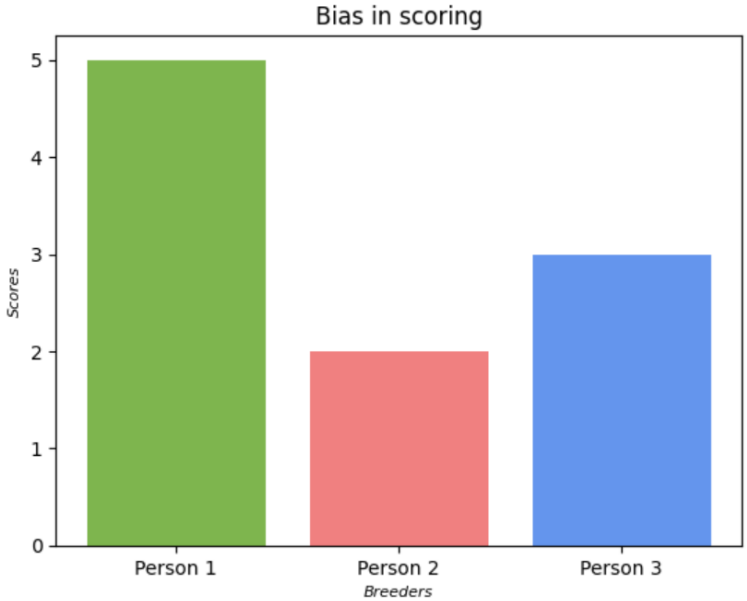

Figure 28: Bias in scoring

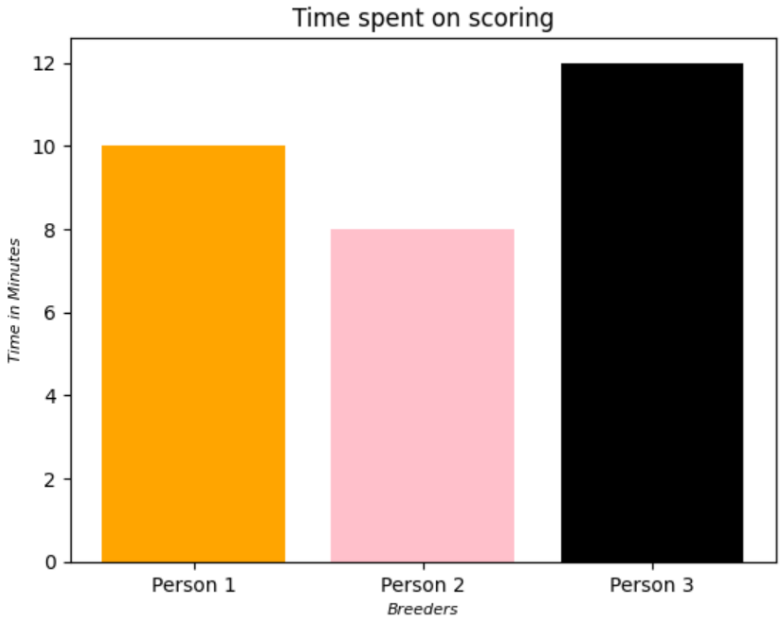

Figure 29: Time spent on scoring

4.2 Results of System Development

4.2.1 Website Developed

As one of the main requirements gathered was to provide a website where data collected can be viewed with different reports and analysis for easy access and storage for future references, following are the main parts of the developed website as required by the breeders.

(i) Dashboard

This part is showing the project objectives and what the breeders and field technicians are using the system for (Fig. 30).

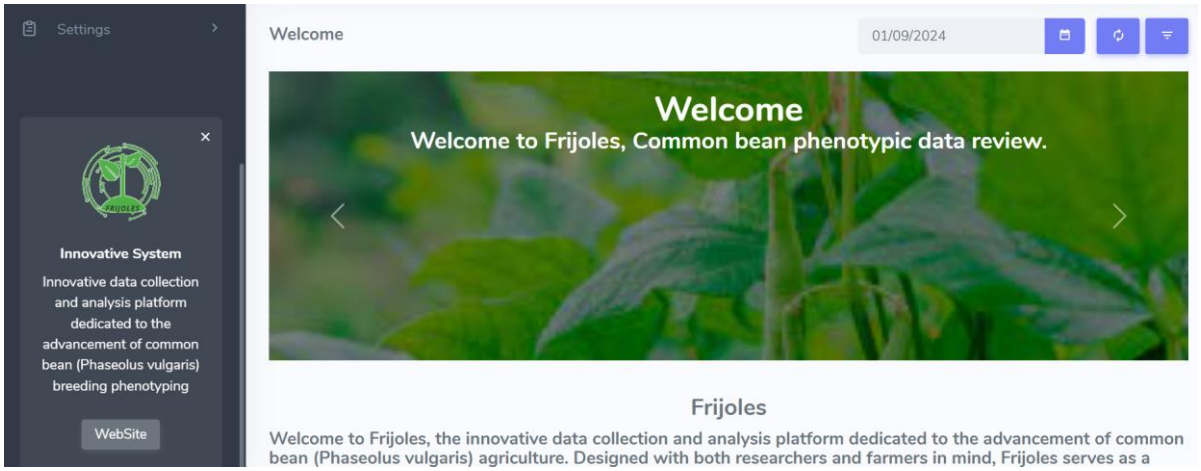

Figure 30: Project overview

This overview part also helps users to easy navigate to where the data are and settings where there is user management, logs control and system settings (Fig. 31).

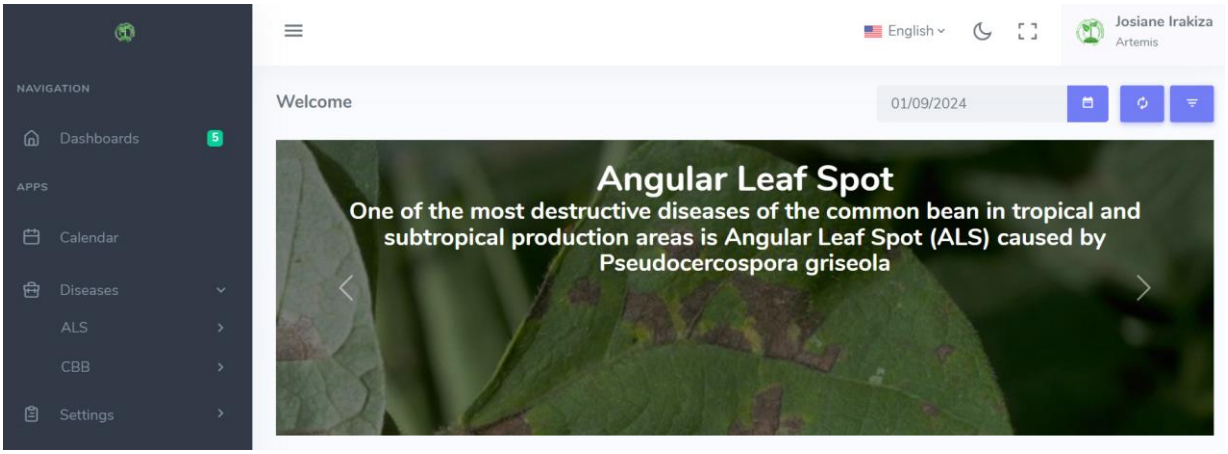

Figure 31: System navigations

(ii) Diseases

The diseases section is where all the needed data analysis are found. The user has the option to choose which disease they want to check data for, after that he/she has four main analyses that provide him/her with a deep understanding of what the data collected means without going through the manual process they used to pass through. Those main categories are the view of collected data, genotypes classification, genotypes per region, and plots summary of the analyzed data.

This part helps the users with different reports as indicated in the requirements where we have the data collected from the fields.

Collected Data

This section allows the user to search different data collected by date, where the user will get access to different forms on different dates as their wishes all forms filled and classification of all scans done on that particular plants they have view and printing option as in their requirements. This part elaborates data collected from the field which provides the user with the view of how many genotypes were identified to be resistant to the targeted foliar diseases either ALS or CBB. The user first select which dates they want to check data as shown in Fig. 32.

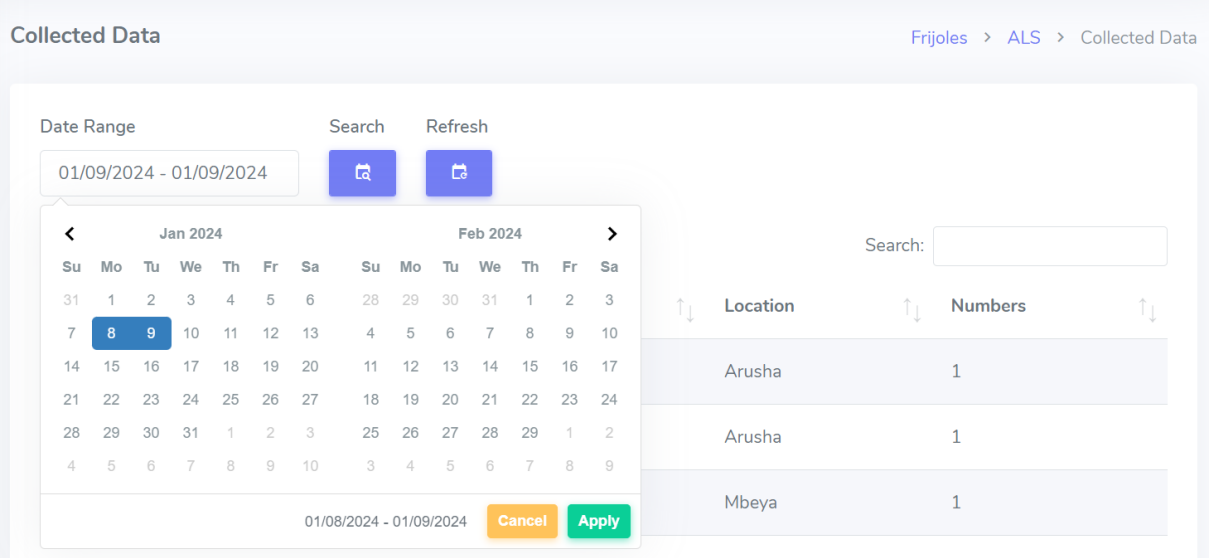

Figure 32: Dates selection

After selecting the dates, the system provides the breeders with the date of data collection region where data was collected, growing stages of the plants when the data were collected and number of forms that were filled on the same growing stage (Fig. 33).

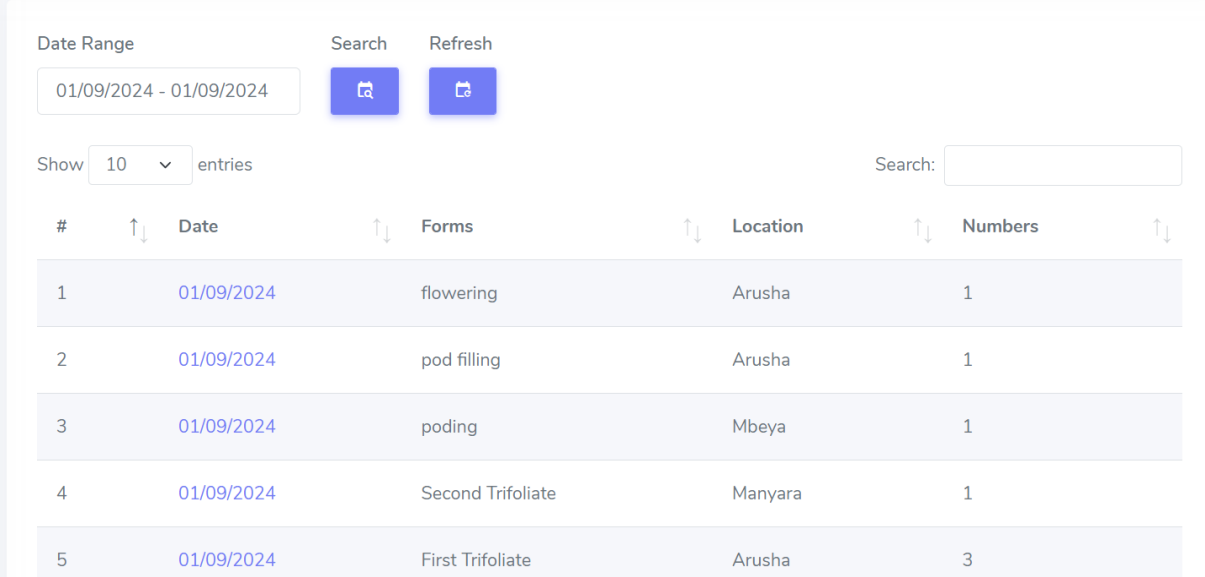

Figure 33: The vvcollected data 1

After the display of all collected data, the user can now check details of all the collected forms on a particular date where they click on that date depending on a growing stage they want to check for. After checking the user is given more information about the data collection including the person who did data collection in the field, location of the data collection as they are on the same growing stages but captured from different rigions and then actions where the user can choose either to view data, print that form or delete the form if it was mistakenly uploaded and it is only allowed for admin not everyone is allowed to delete data from the system (Fig. 34).

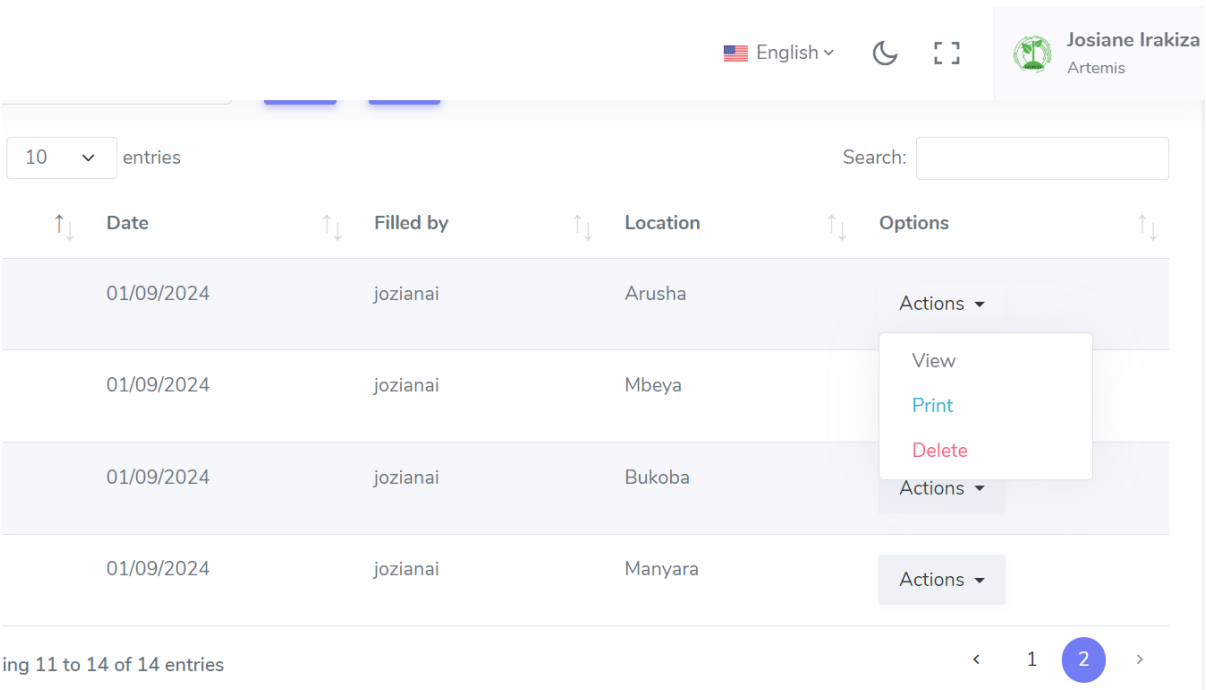

Figure 34: Growing stage data review

Once a view is selected from the actions dropdown list, the user can then view the name of the Genotype that was chosen in data collection, date and time of data collection, plant number of that genotype, number of scans captured on that particular plant, the overall class of that plant and the visual score of the breeder (Fig. 35).

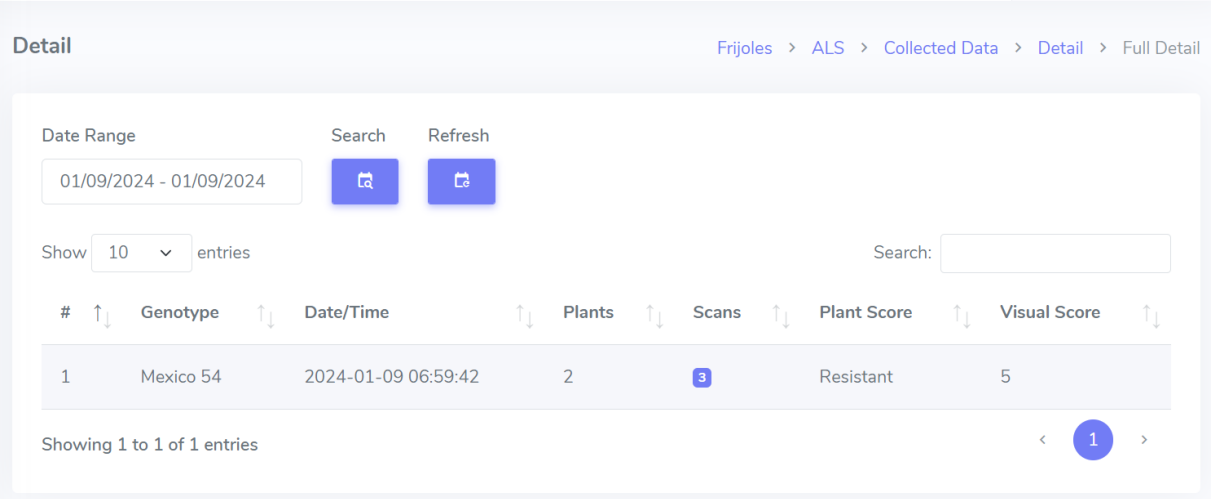

Figure 35: Details of the form

After checking all details of the form with the overall class from the model, the user can also check details of each scan's score that provided the overall score of the plant as the plant always has many scans (Fig. 36).

Plant Number	Scans	Scores	Plant Score
2	1	Resistant	
	2	Resistant	Resistant
	3	Resistant	

Figure 36: Scans classes

Genotypes Classification

This part helps the users with the deep analysis of all collected data of different genotypes and gives the users the lists of all genotypes classified in different levels according to the classifier where they are found as resistant, medium or susceptible. The system analyses them according to which genotype has more plants with resustant class compared to others after classifying the scans at each growing stages (Fig. 37).

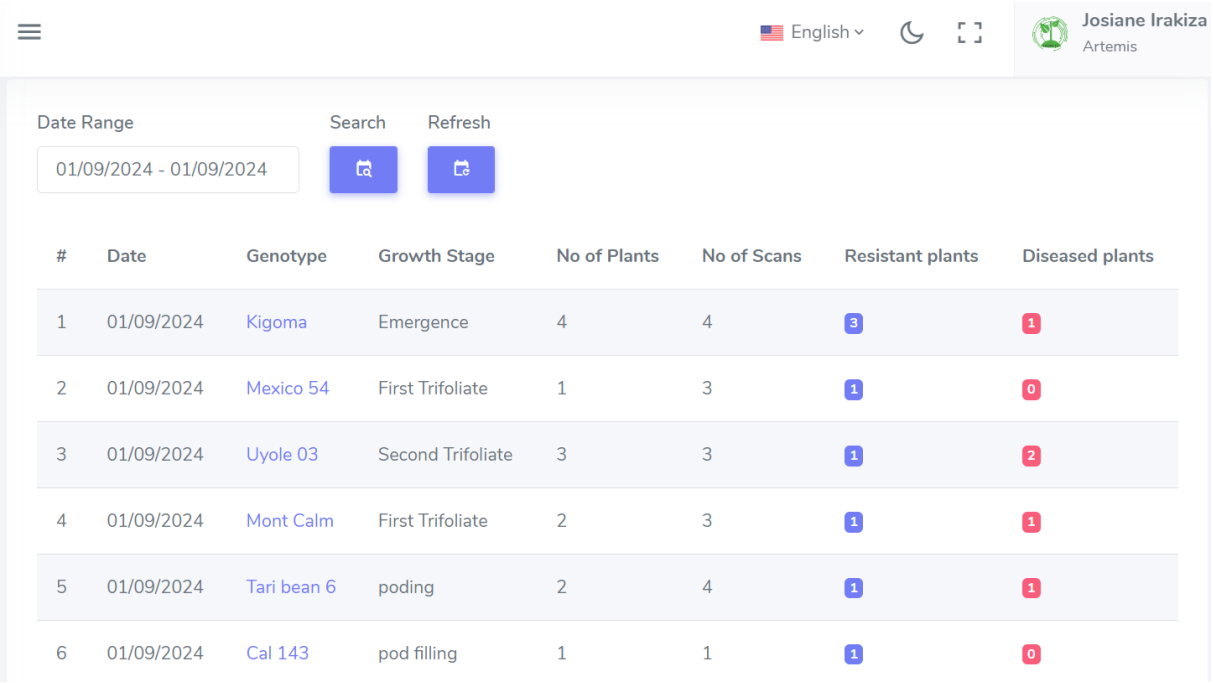

Figure 37: Genotype classification

Genotypes per Region

This section is helping users in knowing which region is more favoring the resistivity or susceptibility of certain genotypes so that the breeders will know which part of the country to plant which genotypes to get higher yields resistant to diseases and favorable to climate of that particular region (Fig. 38).

2024-01-09	Mont Calm	Arusha	0%	Susceptible
2024-01-09	Mont Calm	Bukoba	100%	Resistant
2024-01-09	Mexico 54	Arusha	100%	Resistant
2024-01-09	Uyole 03	Manyara	100%	Resistant
2024-01-16	Uyole 03	Arusha	0%	Susceptible
2024-01-09	Tari bean 6	Arusha	0%	Susceptible
2024-01-09	Tari bean 6	Mbeya	100%	Resistant
2024-01-09	Cal 143	Arusha	100%	Resistant

Figure 38: Genotypes per region

Plants Health Monitoring

This was accomplished by tracking the genotypes' health from the first trifoliate to maturity level using several vegetative indices. The plant health monitoring was done by checking biochemical contents, especially chlorophyll, water absorption, photochemical reflectance, nitrogen content and others. This report help the breeders understand which growing stage the plant reaches and loose more of its biochemical content and which genotype is more resistant to a certain biochemical that is one of the needed part in breeding system as they are always searching for those strong traits to produce resilient seeds.

Summary

In the summary sub-part, the users have different plots showing them the performance of the genotypes in the experiment. The plots demonstrate the plants of the genotypes per growing stages, regions and users have option to download the data regarding all the collected data including genotypes and plants number used on that particular genotype. It has part of the comparision of genotype contribution to overall data collected, and number plants at different growing stages (Figs. 39, 40 and 41).

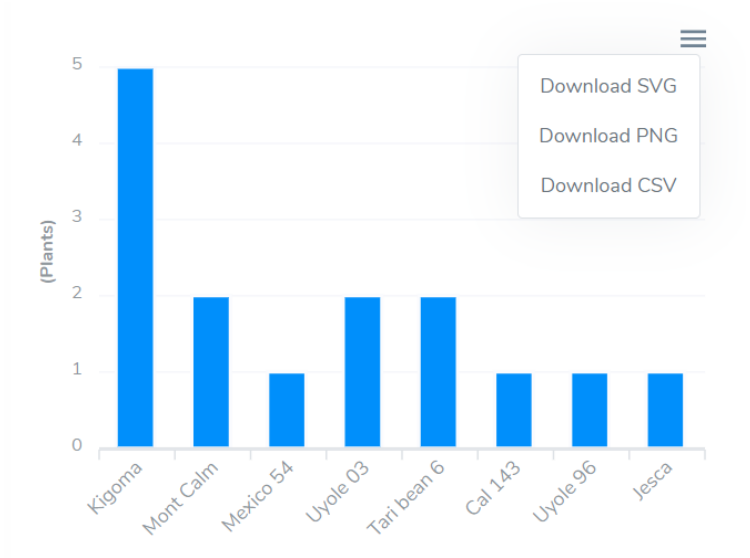

Figure 39: Genotype performance

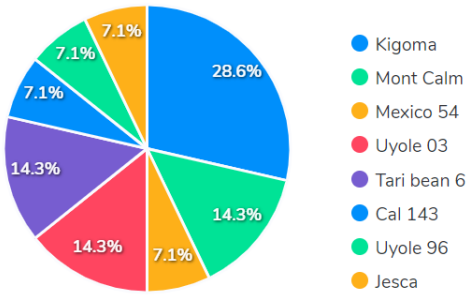

Figure 40: Data collected summary

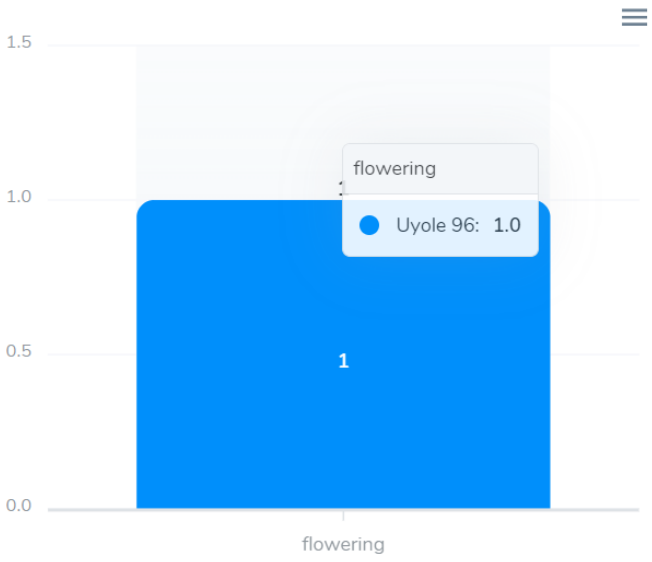

Figure 41: Genotype per growing stage

4.2.2 Classifier Results

Upon inspecting the performance plots of the model trained on the full spectrum of wavelengths, it has evidently borne fruit. The comprehensive set of features available from the entire wavelength range likely endowed the model with more discriminative power, allowing for superior differentiation between classes.

In conclusion, leveraging the full spectrum of wavelengths for training has significantly uplifted the model's performance. The following plots testify to the heightened accuracy, precision, and recall achieved. The results underscore the importance of employing a rich feature set, especially in spectral data, to harness the full potential of the classification model. The first part was comparing different trained models performance for us to know which one to use and the details are found in Table 9.

Table 9: Model selection

Algorithm	Training Time	Classification time	Accuracy	Precision	Error rate	Recall
Random Forest	36 hrs	3 Sec	0.96	0.97	0.21	0.96
XgBoost	72 hrs	3 Sec	0.95	0.96	0.28	0.96
Neuro Network	45 hrs	5 Sec	0.94		Loss: 0.1122	Val_Accuracy: 0.94

After checking those evaluation metrics, the Random Forest showed the better performance with lower training time, and low response time in classifying scans.

Table 10: Model evaluation scores

Class	Precision	Recall	F1-Scoe	Error rate	Support
Resistant	0.98	0.96	0.97	0.21	641
Medium	0.96	0.96	0.96	0.28	640
Susceptible	0.96	0.97	0.97	0.22	641
Accuracy	-	-	0.96	-	1922
mAP	-	-	0.99	-	-
Macro avg	0.96	0.96	0.96	-	1922

Weighted avg	0.96	0.96	0.96	-	1922

(i) Precision-recall curve for multiclass

The precision-recall curve is closely hugging the upper right corner for all classes. This indicates excellent performance, with high precision across varying levels of recall. Such a trend underscores the model's capability to achieve high true positive rates while maintaining low false positives for each class (Fig. 42).

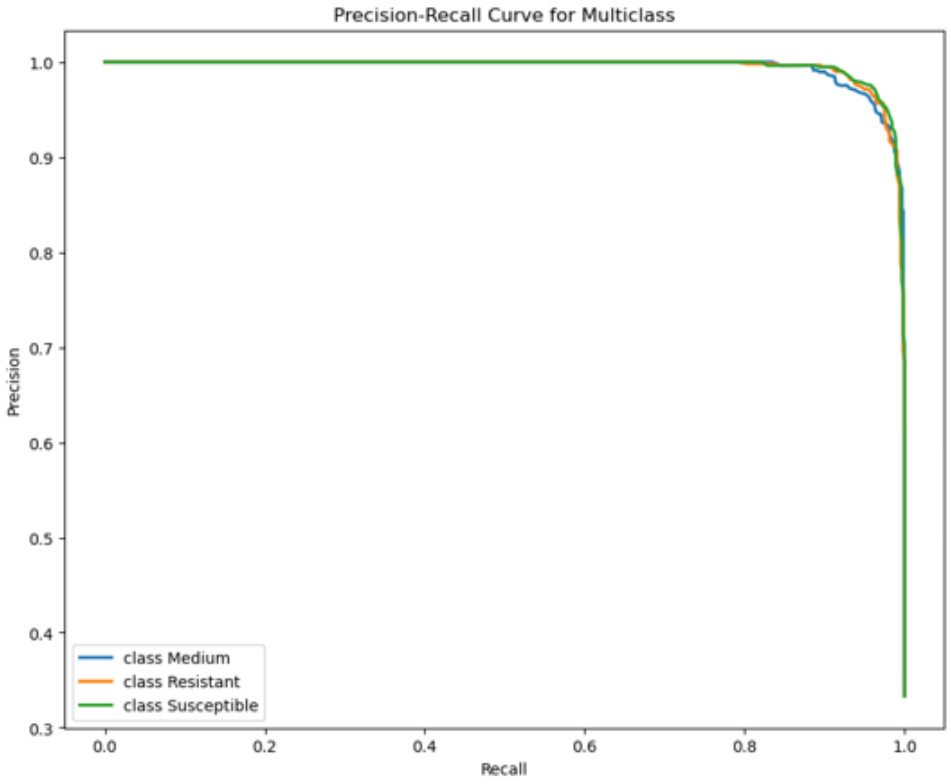

Figure 42: Precision-recall for multiclass

(ii) Receiver operating characteristic (ROC) for multiclass

The ROC curves for all classes converge towards the top left corner with an area under the curve (AUC) of 1.00 for each. An AUC of 1.00 is indicative of perfect classification. The results, therefore, highlight the model's exceptional discriminative power for the different classes (Fig. 43).

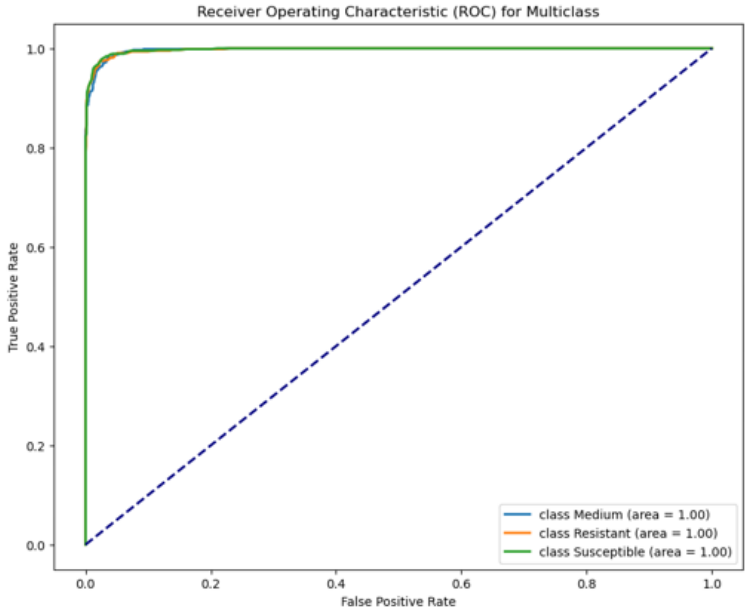

Figure 43: The ROC for multiclass

(iii) Confusion matrix for multiclass

A glance at the matrix reveals a predominantly diagonal pattern, signifying that a large number of observations have been correctly classified. While there are some off-diagonal elements, representing misclassifications, their numbers are relatively low. The matrix, thus, emphasizes the model's strong classification performance across all classes (Fig. 44).

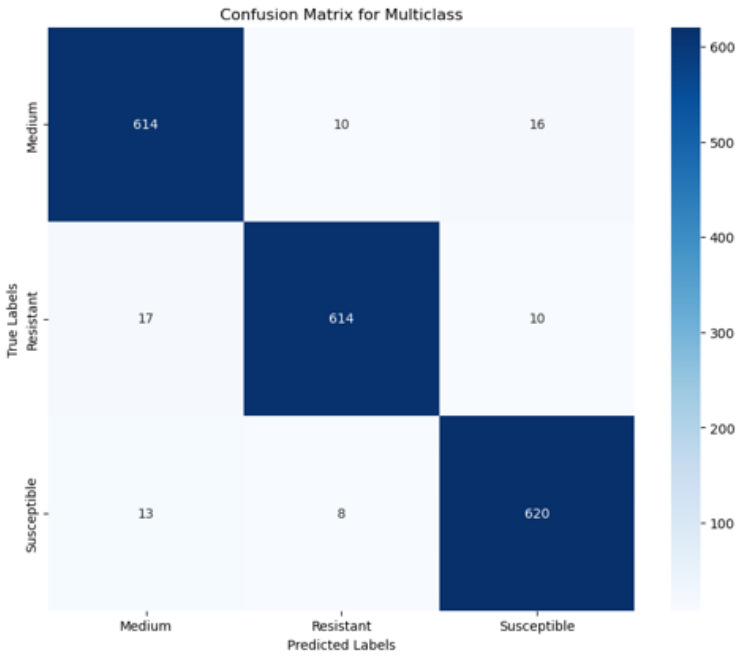

Figure 44: Confusion matrix results

(iv) Out of bug (OOB) accuracy versus number of trees

The accuracy curve demonstrates a rapid ascent before plateauing as the number of trees increases. This indicates that the ensemble model benefited from adding more trees initially, with diminishing returns after a certain point. The plateau at high accuracy levels suggests robust performance and stable predictions (Fig. 45).

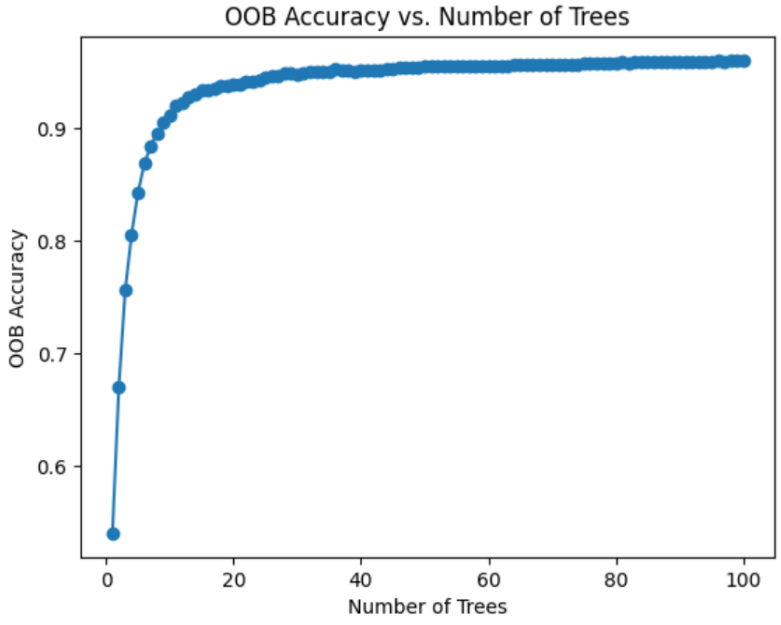

Figure 45: Out of bug versus trees

4.2.3 Developed mobile application data collection form

The mobile application form was developed based on ODK concepts and following are main features of the form from the beginning to the sending of the data to the web.

Upon accessing the mobile application built on ODK, users are prompted to configure the project on the initial screen by providing the URL of the web application and the user's credentials. After having all needed details of the project, the user will be able to download new form that is deployed on the website.

The user will then fill out the form with all needed information about the genotypes they are gathering data for. All information are provided for them to choose in order to reduce time spent on the field and easy the phenotyping process. Once filled out, they can examine draft forms, send completed forms, and review the forms they have already sent. Figure 46 shows the screens of the mobile application form that are displayed.

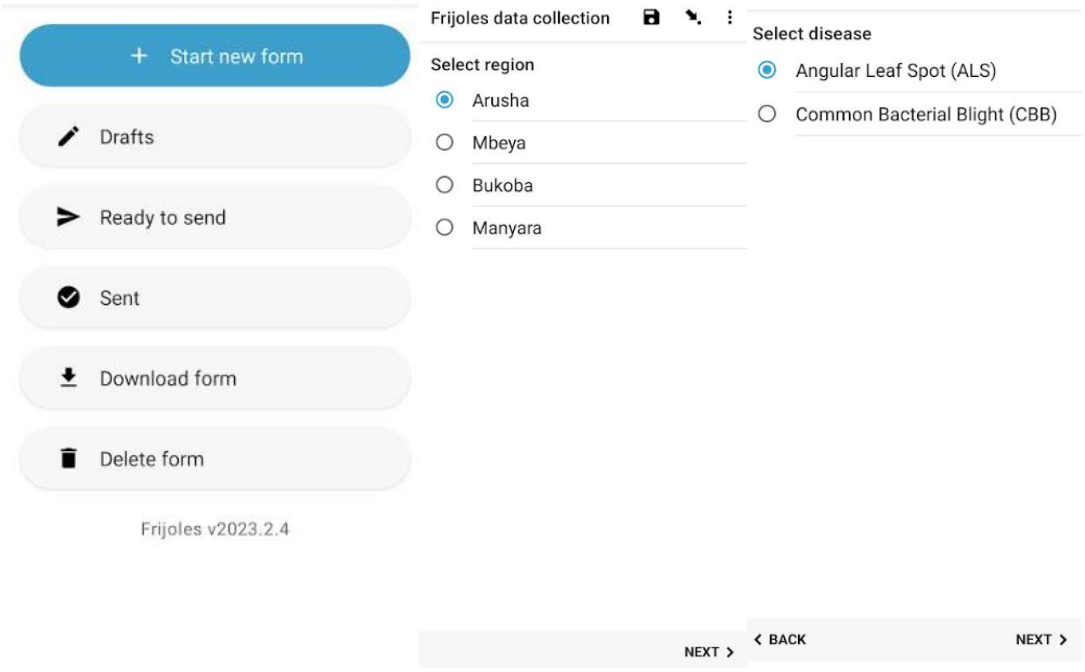

Figure 46: Form filling 1

Once the user begins filling out the form, he or she must first select the region from which they are collecting data, followed by the ailment for which they are collecting data. After selecting

the disease and region, the user will choose the genotype he/she wants to phenotype and number of plant they are on that form, as illustrated in Fig. 47.

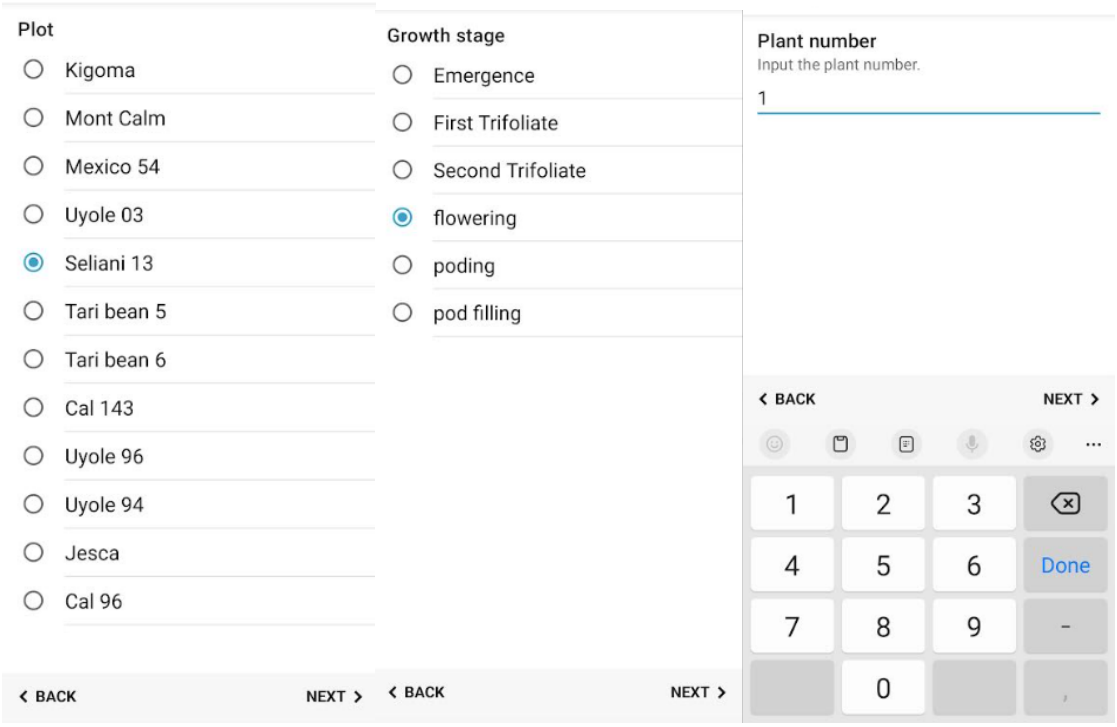

Figure 47: Form filling 2

When the genotype to assess, the growing stage of the plant, and the number of a particular plant they are phenotyping, they are asked to provide the visual score for the future data validation and checking the correlation between the model prediction and the breeder scoring as the system is still a new integration in the breeding processes. After entering the score of the plant, the users are then on the scanning part of leaves of that plant where they launch the spectrometer, then they have to choose which mode they want to assess the plant at the diseases identification it is always Reflectance mode to capture all biochemical contents of the plant.

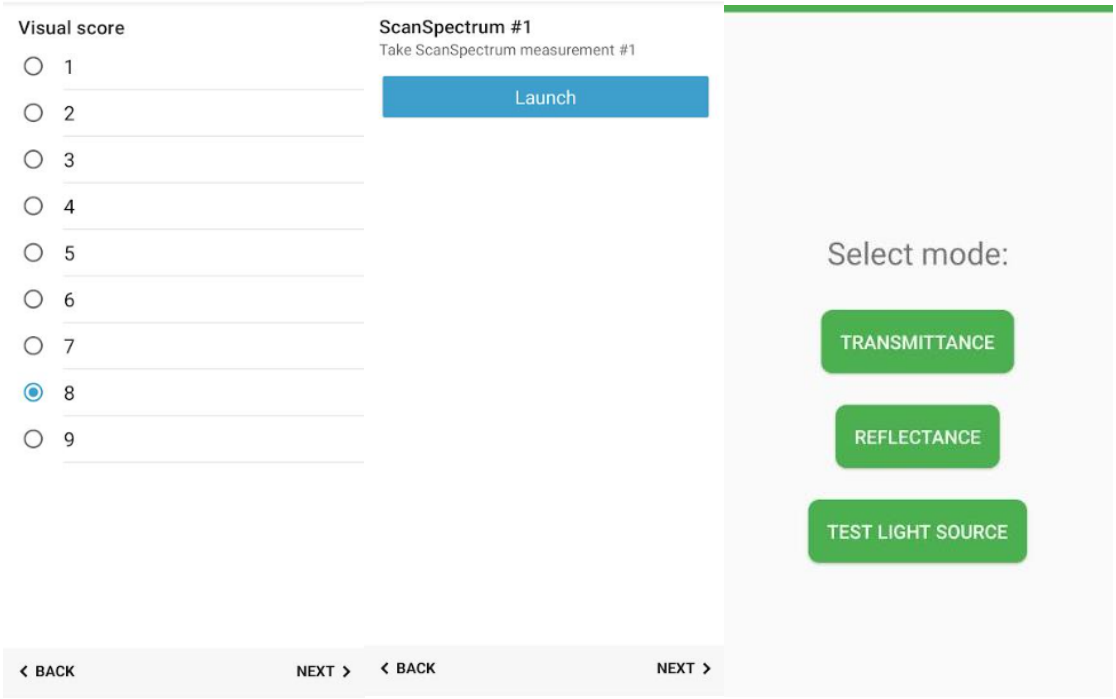

Figure 48: Form filling 3

After choosing the reflectance mode of the spectrometer, the spectrometer is calibrated to capture all the lights transmitted within the leaf that is done by placing the white tape on the light of the spectrometer and wait for it to be calibrated. After the spectrometer is calibrated to capture all reflectance of the leaves, it asks the user to press the scanning button of the spectrometer then wait for it to finish to scan. The scanning is a process that takes three to five seconds at each scan.

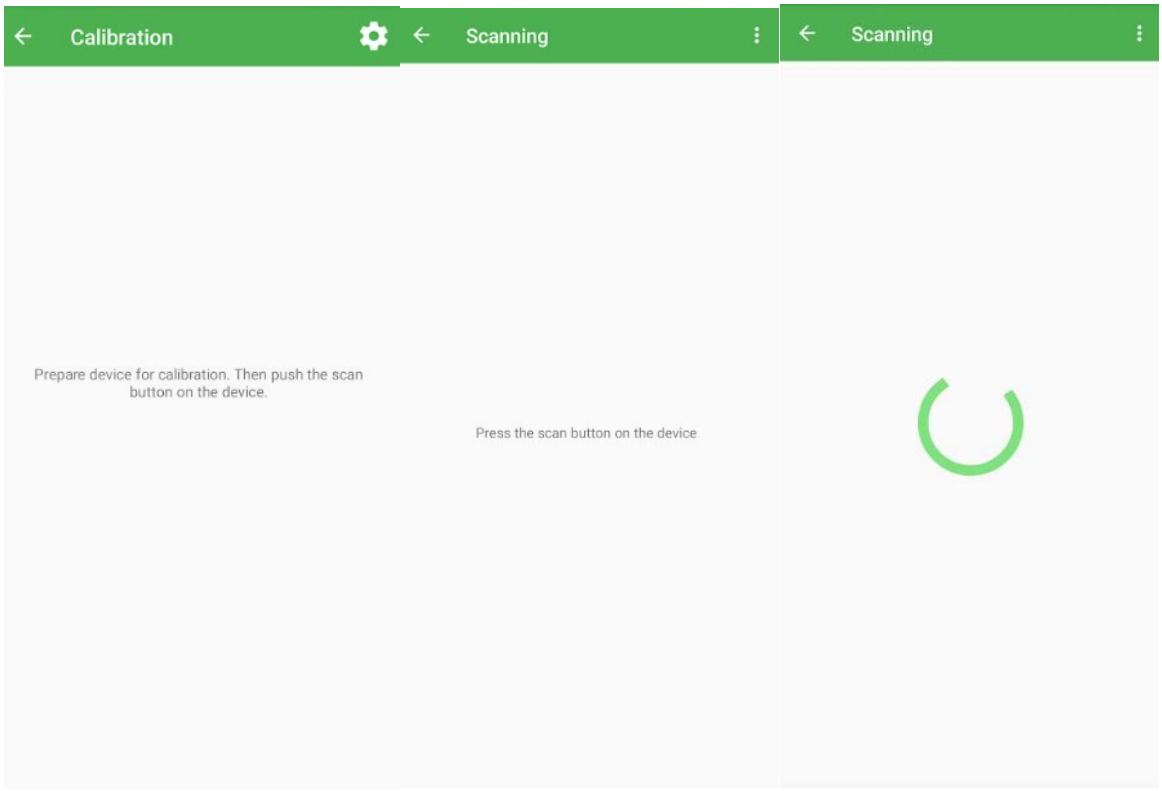

Figure 49: Scanning a leaf

Once the spectrometer finishes scanning the leaf it provides the user with the wavelength view together with the NDVI of the scan. The user agrees on the captured scan and also has the option of rescan in case there was an error. The captured scan will be deleted and the user scan a new wavelength. After accepting the captured wavelength the Json file is saved. The user finalize the form and send the data to the web where the json file is captured and sent to the model for classification, then all data including the classes from the model can be seen on the web with needed further analysis as shown in Fig. 50.

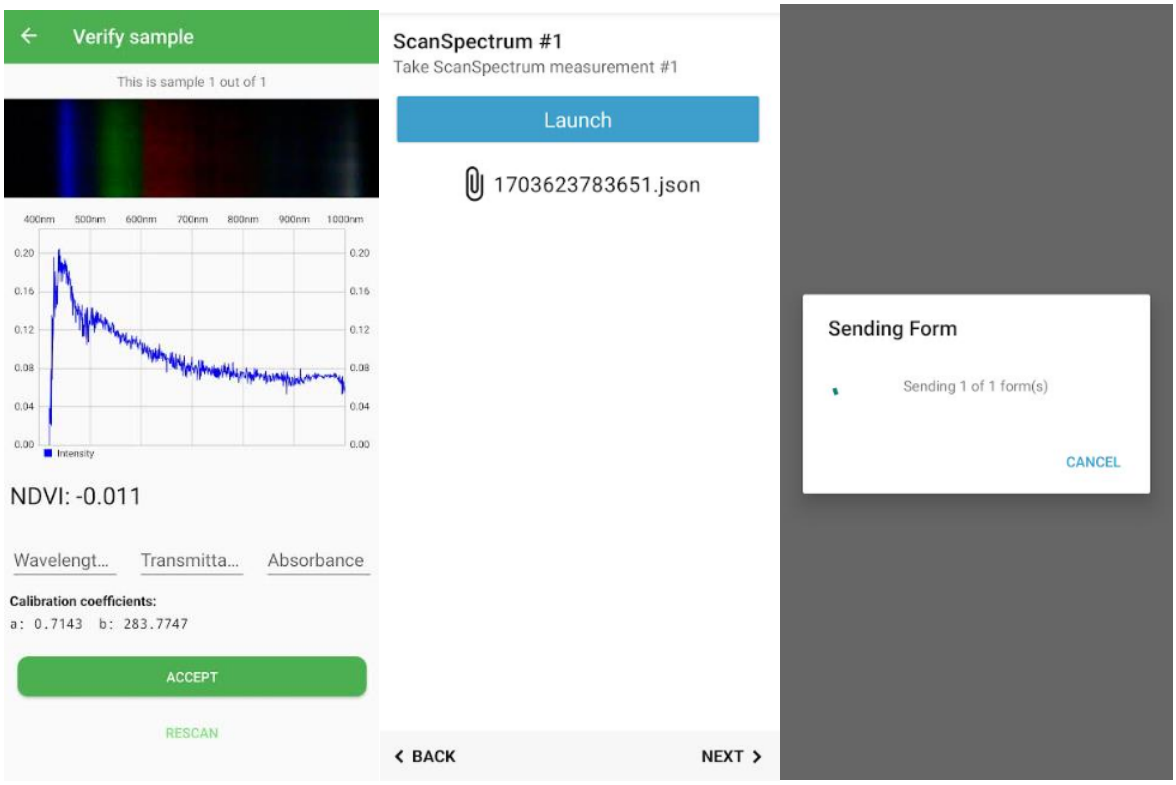

Figure 50: Saving scans and sending data

4.3 System Validation and Users Acceptance Results

User acceptance is a process of the user checking the developed system and checking if it meets their provided wishes to the system developers. It was done by filling out questionnaires on how they have seen the system compared to what they want their system to be providing to them questionnaire sample can be found in the appendix of this report. After collecting all the questionnaires from the users, the followed part was analysis of all the provided answers to check if the system really met users' expectations or if it was a failure. This was done using python language with pandas and Matplotlib libraries to interpret those results in terms of data and plots that are easier to understand. A total number of 40 respondents provided their views about the system usage as shown in the following reports.

Table 11: User acceptance results

S/N	Requirement	Strongly Agree	Agree	Disagree
1.	The System meets overall requirements provided	90%	10%	0%
2.	The system can be used for data collection in the field	100%	0%	0%
3.	The system can provide overview of all data collected on a website	100%	0%	0%
4.	The system can provide biochemical content of the plant	80%	20%	0%
5.	The system provide reports with search and print options	100%	0%	0%
6.	The system allow the scanning of leaves	100%	0%	0%
7.	The system provides the genotype classes according to diseases levels	70%	30%	0%
8.	The system is easy to use	75%	20%	5%

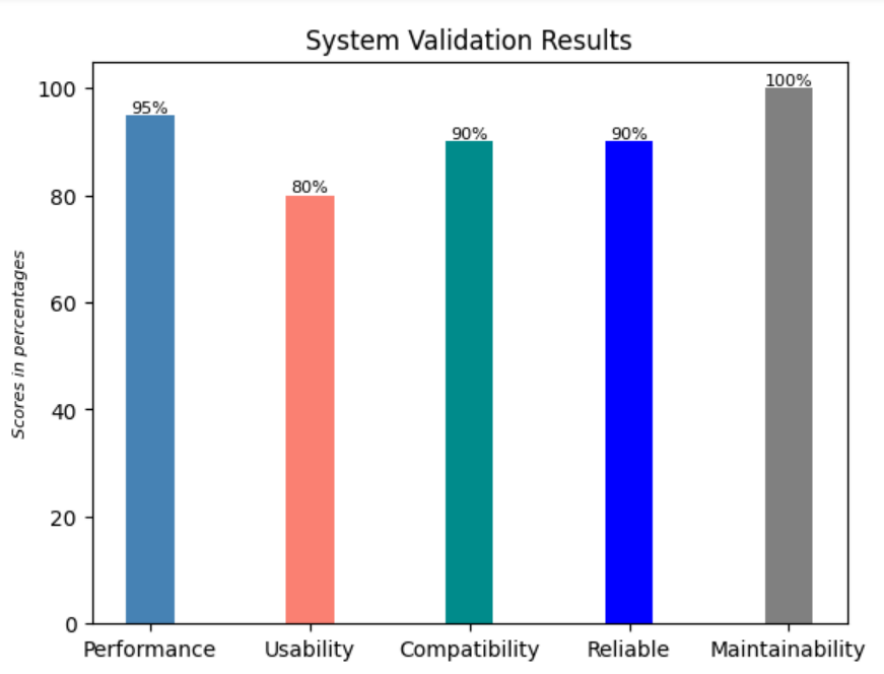

Figure 51: System validation results

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The utilization of hyperspectral data, spanning a wide spectral range from 350 nm to 1600 nm, has proven to be instrumental in capturing intricate biochemical variations associated with foliar diseases. The Random Forest algorithm, known for its robustness and ability to handle complex datasets, has effectively leveraged this wealth of information to make accurate disease identifications.

In conclusion, the hyperspectral-based system utilizing the Random Forest algorithm has demonstrated remarkable success in the identification of common bean genotypes resistant to foliar diseases, achieving an impressive accuracy rate of 96%. This noteworthy level of accuracy signifies the system's efficacy in discriminating between different spectral signatures associated with resistant, medium and susceptible genotypes.

This high level of accuracy not only enhances our capability to swiftly and precisely detect foliar diseases in beans but also holds promise for the implementation of proactive and targeted intervention strategies in agricultural practices. By enabling early and accurate disease identification, this hyperspectral-based system has the potential to contribute significantly to the optimization of crop management, minimizing the impact of diseases on bean yield and overall crop health.

As we move forward, the integration of hyperspectral technology with advanced machine learning algorithms, exemplified by Random Forest, continues to pave the way for innovative solutions in precision agriculture. The success of this system underscores the importance of harnessing cutting-edge technologies for sustainable and efficient crop production, ushering in a new era of disease management and agricultural productivity.

5.2 Recommendations

We recommend all Agricultural Research-based Institutions, Higher learning Institutions, and Ministry of Agriculture for adoption of a hyperspectral-based system for the identification of genotypes resistant to foliar diseases in beans. This innovative technology has demonstrated tremendous potential in revolutionizing the field of agricultural research and crop management by capturing data across a broad spectral range (350 nm to 1150 nm), hyperspectral imaging provides a comprehensive view of the biochemical and physiological characteristics of bean plants. This depth of information is invaluable for understanding the nuanced responses of different genotypes to foliar diseases, enabling researchers to make informed decisions regarding disease resistance. This forward-looking approach not only enhances research capabilities but also positions agricultural practices at the forefront of technological innovation, ensuring resilience and adaptability in the face of evolving challenges.

The integration of hyperspectral data with advanced analytics and machine learning facilitates data-driven decision-making processes. Researchers can leverage this information to identify patterns, correlations, and key indicators of disease resistance, enhancing the efficiency of breeding programs and accelerating the development of disease-resistant bean genotypes.

5.3 Future Work

To keep on improving the agriculture sector in producing resilient with higher yields production varieties to meet consumption demand and adapt to climatic changes, we recommend that this research be greatly expanded for the use of hyperspectral systems to include more remaining foliar diseases, field trials across the whole country, and and more crops to examine the viability and effectiveness of utilizing hyperspectral data in the development of reliable and useful systems for disease identification in a variety of leguminous crops like soybeans, peas, lentils, chickpeas and others by pursuing these pathways.

REFERENCES

Aganecy, P. (2023). *Online Gender-Based Violence: A silent flesh-eating cancer in Tanzania*. https://www.peaceagency.org/online-gender-based-violence-in-tanzania.

Agile Alliance. (2023). *What is Extreme Programming (XP)?Agile Alliance*. https://www.agilealliance.org/glossary/xp.

Ahmad, A., Saraswat, D., & El-Gamal, A. (2023). A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. *Smart Agricultural Technology*, *3*(2022), 100083.

AI, E. (2023). *Accuracy vs. precision vs. recall in machine learning: what's the difference?* https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall.

Allard, R. . (2023). *Plant breeding | History, Applications, & Methods | Britannica*. https://www.britannica.com/science/plant-breeding.

Almeida, C. P. de, de Carvalho Paulino, J. F., Bonfante, G. F. J., Perseguini, J. M. K. C., Santos, I. L., Gonçalves, J. G. R., Patrício, F. R. A., Taniguti, C. H., Gesteira, G. S., Garcia, A. A. F., Song, Q., Carbonell, S. A. M., Chiorato, A. F., & Benchimol-Reis, L. L. (2021). Angular Leaf Spot Resistance Loci Associated With Different Plant Growth Stages in Common Bean. *Frontiers in Plant Science*, *12*, 1–18.

Belay, A. J., Salau, A. O., Ashagrie, M., & Haile, M. B. (2022). Development of a chickpea disease detection and classification model using deep learning. *Informatics in Medicine Unlocked*, *31*, 100970. https://doi.org/10.1016/j.imu.2022.100970.

Bhandari, A. (2023). *Understanding Confusion Matrix in Machine Learning (2024)*. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning.

Binagwa, P. H., Bonsi, C. K., & Msolla, S. N. (2016). Evaluation of common bean (Phaseolus

vulgaris) genotypes for resistance to root rot disease caused by Pythium aphanidermatum and Pythium splendens under screen house conditions. *Evaluation*, *6*(6), 36-43.

Bishaw, Z., & Van-Gastel, A. J. (2009). Variety release and policy options. *Plant Breeding and Farmer Participation*, *21*, 565-587.

Bohl, S. (2021). *What is NDVI and how to use crop imaging in remote sensing?* https://blog.eagronom.com/ndvi-remote-sensing.

Borra-Serrano, I., Van-Laere, K., Lootens, P., & Leus, L. (2022). Breeding and selection of nursery plants assisted by high-throughput field phenotyping using UAV imagery: Case studies with sweet box (Sarcococca) and garden rose (Rosa). *Horticulturae*, *8*(12), 1186. https://doi.org/10.3390/horticulturae8121186.

Britannica, T. E. (2021). *Common bean legumes*. Common Bean Vegetable.

Chandra, A., & Mursitama, T. N. (2020). Sales Revenue Sharing Model using Dynamics NAV Modification in Health Industries. *IOP Conference Series: Earth and Environmental Science*, *426*(1), 012162. https://doi.org/10.1088/1755-1315/426/1/012162.

Chen, N. W. G., Ruh, M., Darrasse, A., Foucher, J., Briand, M., Costa, J., Studholme, D. J., & Jacques, M. A. (2021a). Common bacterial blight of bean: a model of seed transmission and pathological convergence. *Molecular Plant Pathology*, *22*(12), 1464–1480. https://doi.org/10.1111/mpp.13067.

Chen, N. W. G., Ruh, M., Darrasse, A., Foucher, J., Briand, M., Costa, J., Studholme, D. J., & Jacques, M. A. (2021b). Common bacterial blight of bean: a model of seed transmission and pathological convergence. *Molecular Plant Pathology*, *22*(12), 1464–1480. https://doi.org/10.1111/mpp.13067.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785.

Chung, S., Breshears, L. E., & Yoon, J. Y. (2018). Smartphone near infrared monitoring of

plant stress. *Computers and Electronics in Agriculture*, *154*, 93-98.

Cock, M. J. W., Rui, T., Liu Zhi, L. Z., Huan, W. H., McGillivray, L. A., Thomas, S. E., & Feng, Z. (2016). The main agricultural insect and disease pests of China and implications for the use of remote sensing for their management. *CABI Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, *11*(2016), 1-23.

Colab. (2023). *colab.google*. https://colab.google.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Ensemble Machine Learning. *Ensemble Machine Learning, January*. https://doi.org/10.1007/978-1-4419-9326-7.

Danzi, D., De Paola, D., Petrozza, A., Summerer, S., Cellini, F., Pignone, D., & Janni, M. (2022). The use of near-infrared imaging (NIR) as a fast non-destructive screening tool to identify drought-tolerant wheat genotypes. *Agriculture*, *12*(4), 537.

Dashti, H., Glenn, N. F., Ustin, S., Mitchell, J. J., Qi, Y., Ilangakoon, N. T., Flores, A. N., Silvan-Cardenas, J. L., Zhao, K., Spaete, L. P., & De-Graaff, M. A. (2019). Empirical Methods for Remote Sensing of Nitrogen in Drylands May Lead to Unreliable Interpretation of Ecosystem Function. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(6), 3993–4004. https://doi.org/10.1109/TGRS.2018.2889318.

Data Robot. (2023). *Machine Learning Model Deployment | DataRobot AI Wiki*. https://www.datarobot.com/wiki/machine-learning-model-deployment.

Dataiku. (2021). *Decision Tree and Random Forest Algorithms: Decision Drivers - History of Data Science*. https://www.historyofdatascience.com/decision-tree-and-random-forest-algorithms-decision-drivers.

Digital, H., & Capture, C. (2018). *Open Your Shutters*. *June*, 1–11.

Donges, Niklas, J. P. (2022). *What Is Transfer Learning? A Guide for Deep Learning | Built In*. https://builtin.com/data-science/transfer-learning.

Elad, Y., & Pertot, I. (2014). Climate Change Impacts on Plant Pathogens and Plant Diseases. *Journal of Crop Improvement*, *28*(1), 99–139.

Elfatimi, E., Eryigit, R., & Elfatimi, L. (2022). Beans Leaf Diseases Classification Using MobileNet Models. *IEEE Access*, *10*, 9471–9482.

Evaldas Vaiciukynas. (2023). *Architecture of the random forest model. | Download Scientific Diagram*. https://www.researchgate.net/figure/Architecture-of-the-random-forest-model_fig1_301638643.

FAO. (2023). *Faostat*.

Farrow, A., & Muthoni-Andriatsitohaina, R. (2020). ATLAS Common bean production in Africa second edition. In *World*, *2*(2020).

Feng, L., Wu, B., He, Y., & Zhang, C. (2021). Hyperspectral imaging combined with deep transfer learning for rice disease detection. *Frontiers in Plant Science*, *12*, 693521. https://doi.org/10.3389/fpls.2021.693521.

Garriga, M., Romero-Bravo, S., Estrada, F., Escobar, A., Matus, I. A., del Pozo, A., Astudillo, C. A., & Lobos, G. A. (2017). Assessing wheat traits by spectral reflectance: Do we really need to focus on predicted trait-values or directly identify the elite genotypes group? *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.00280.

Geospatial, N. (2023). *Vegetation Indices*.

Girma, F., Fininsa, C., Terefe, H., & Amsalu, B. (2022). Evaluation of common bean (Phaseolus vulgaris) genotypes for resistance to common bacterial blight and angular leaf spot diseases, and agronomic performances. *Heliyon*, *8*(8). https://doi.org/10.1016/j.heliyon.2022.e10425.

Greenlife. (2023). *Bean Angular Leaf Spot*. https://www.greenlife.co.ke/bean-angular-leaf-spot.

Hiphen-plant. (2023). *Plant Phenotyping Vegetation Indices for Chlorophyll - Blog Hiphen*. https://www.hiphen-plant.com/vegetation-indices-chlorophyll/3612.

Hiphen. (2023). *Plant Phenotyping Vegetation Indices for Chlorophyll - Blog Hiphen*. https://www.hiphen-plant.com/vegetation-indices-chlorophyll/3612.

Iguazio. (2023). *What is Model Evaluation*. https://www.iguazio.com/glossary/model-evaluation.

Iguazio. (2024). *What is Model Deployment*. https://www.iguazio.com/glossary/model-deployment.

IPM. (2018). *Common Bacterial Blight / Dry Beans / Agriculture: Pest Management Guidelines / UC Statewide IPM Program*. https://ipm.ucanr.edu/agriculture/dry-beans/common-bacterial-blight.

Javapoint. (2023). *Machine Learning Random Forest Algorithm - Javatpoint*. https://www.javatpoint.com/machine-learning-random-forest-algorithm.

Kumar Sahu, S., & Pandey, M. (2023). An optimal hybrid multiclass SVM for plant leaf disease detection using spatial Fuzzy C-Means model. *Expert Systems with Applications*, *214*, 118989. https://doi.org/10.1016/J.ESWA.2022.118989.

Kurbanov, R. K., & Zakharova, N. I. (2020). Application of Vegetation Indexes to Assess the Condition of Crops. *Agricultural Machinery and Technologies*, *14*(4), 4–11. https://doi.org/10.22314/2073-7599-2020-14-4-4-11.

Li, Q., & Yan, J. (2020). Sustainable agriculture in the era of omics: Knowledge-driven crop breeding. *Genome Biology*, *21*(1), 5–9. https://doi.org/10.1186/s13059-020-02073-5.

Lucidchart. (2023). *UML Use Case Diagram Tutorial*. https://www.lucidchart.com/pages/uml-use-case-diagram.

Maiza, R., & Kurnia, D. (2019). The Influence of Light Wavelengths Toward the Growth of Brassica rapa L. *Journal of Physics: Conference Series*, *1245*(1), 012089. https://doi.org/10.1088/1742-6596/1245/1/012089.

Maniyath, S. R., Vinod, P. V., Niveditha, M., Pooja, R., Prasad Bhat, N., Shashank, N., & Hebbar, R. (2018). Plant disease detection using machine learning. *Proceedings - 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control*, 41–45. https://doi.org/10.1109/ICDI3C.2018.00017.

Martínez-Martínez, V., Gomez-Gil, J., Machado, M. L., & Pinto, F. A. C. (2018). Leaf and

canopy reflectance spectrometry applied to the estimation of angular leaf spot disease severity of common bean crops. *PLoS ONE*, *13*(4), 1–18.

Mesevage, T. G. (2021). *What Is Data Preprocessing & What Are The Steps Involved?* https://monkeylearn.com/blog/data-preprocessing.

Ministry of Agriculture. (2023). *Crops Suitability. Ministry of Agriculture*.

NAPB. (2019). *What is Plant Breeding? National Association of Plant Breeders*. https://www.plantbreeding.org/about-us/what-is-plant-breeding.

Nay, M. M. (2019). *Angular Leaf Spot and Ascochyta Disease Resistance in Common Bean-Characterization and Application for Breeding* (Doctoral dissertation, ETH Zurich).

Ndimbo, M., Shida, N., Mbiu, J., Kilango, M., Madata, C., Binagwa, P., & Kasuga, R. (2022). Variety Catalogue of Common Beans (Phaseolus Vulgaris) in Tanzania.

Nguyen, C., Sagan, V., Maimaitiyiming, M., Maimaitijiang, M., Bhadra, S., & Kwasniewski, M. T. (2021). Early detection of plant viral disease using hyperspectral imaging and deep learning. *Sensors (Switzerland)*, *21*(3), 1–23. https://doi.org/10.3390/s21030742.

Nigatie, T. Z. (2021). Review on effect of N and P fertilizer rates on yield and yield components of common bean [Phaseolus vulgaris (L.)] varieties. *International Journal of Research in Agronomy*, *4*(1), 32–40. https://doi.org/10.33545/2618060x.2021.v4.i1a.46.

Nikolopoulou, K. (2022). *What Is Purposive Sampling? Definition & Examples*. https://www.scribbr.com/methodology/purposive-sampling.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., Hoagwood, K., Angeles, L., & Northwest, K. P. (2015). "Dentists face added drug regulation. *Dental Survey*, *44*(12), 73. https://doi.org/10.1007/s10488-013-0528-y.Purposeful.

Patricia Onyango. (2023). *Affordable phenotyping for evaluating foliar diseases in common bean | PABRA*. https://www.pabra-africa.org/affordable-phenotyping-for-evaluating-foliar-diseases-in-common-bean.

Python. (2023). *The Python Tutorial*. https://docs.python.org/3/tutorial/index.html.

Qed. (2023). *ScanSpectrum | Home page*.

Radočaj, D., Šiljeg, A., Marinović, R., & Jurišić, M. (2023). State of major vegetation indices in precision agriculture studies indexed in web of science: A review. *Agriculture*, *13*(3), 707. https://doi.org/10.3390/agriculture13030707.

Reza, M., Miri, S., & Javidan, R. (2016). A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset. *International Journal of Advanced Computer Science and Applications*, *7*(6), 1–33.

Ricks. (2022). *What Are Foliar Diseases? Ricks Plant Health Care*. https://ricksplanthealthcare.com/2022/04/what-are-foliar-diseases-and-how-to-prevent-them.

Rodríguez, D., Beaver, J., De Jensen, C. E., & Porch, T. (2019). Identification of resistance sources of common bean (Phaseolus vulgaris L.) to angular leaf spot (pseudocercospora griseola). *Revista Facultad Nacional de Agronomia Medellin*, *72*(2), 8785–8791. https://doi.org/10.15446/rfnam.v72n2.70238.

Sahu, P., Chug, A., Singh, A. P., Singh, D., & Singh, R. P. (2021). Deep Learning Models for Beans Crop Diseases: Classification and Visualization Techniques. *International Journal of Modern Agriculture*, *10*(1), 796–812.

Saini, R. (2023). Integrating Vegetation Indices and Spectral Features for Vegetation Mapping from Multispectral Satellite Imagery Using AdaBoost and Random Forest Machine Learning Classifiers. *Geomatics and Environmental Engineering*, *17*(1), 57–74. https://doi.org/10.7494/geom.2023.17.1.57.

Sasagawa, T., Akitsu, T. K., Ide, R., Takagi, K., Takanashi, S., Nakaji, T., & Nasahara, K. N. (2022). Accuracy Assessment of Photochemical Reflectance Index (PRI) and Chlorophyll Carotenoid Index (CCI) Derived from GCOM-C/SGLI with In Situ Data. *Remote Sensing*, *14*(21), 1–26. https://doi.org/10.3390/rs14215352.

Shamshiri, R. (2008). *Plant Disease Detection Based On Spectral Band Selection Study Report*.

Shrivastava, V. K., & Pradhan, M. K. (2021). Rice plant disease classification using color

features: a machine learning paradigm. *Journal of Plant Pathology*, *103*(1), 17–26. https://doi.org/10.1007/S42161-020-00683-3.

Simplilearn. (2023). *What is XGBoost? An Introduction to XGBoost Algorithm in Machine Learning | Simplilearn*. https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article.

Singh, P. P., Kumar, A., Gupta, V., & Prakash, B. (2021). Recent advancement in plant disease management. *Food Security and Plant Disease Management*, 1–18.

Singh, S. P., & Schwartz, H. F. (2010). Breeding common bean for resistance to diseases: A review. *Crop Science*, *50*(6), 2199–2223. https://doi.org/10.2135/cropsci2009.03.0163.

Singh, V. (2023). *How to Calculate the F1 Score in Machine Learning*. *v*.

Spatti, A. C., Bezerra, L. M. C., Fredo, C. E., Bin, A., Paulino, J. F. de C., Chiorato, A. F., Carbonell, S. A. M., & Correia, G. G. (2022). A century of common bean: bibliometrics and scientific production. *Cadernos de Ciência & Tecnologia*, *39*(1), 26949. https://doi.org/10.35977/0104-1096.cct2022.v39.26949.

Start.io. (2023). *Start.io. Smartphone Users in Tanzania Audience*. https://www.start.io/audience/smartphone-users-in-tanzania.

Svanback, R., & Bolnick, D. I. (2019). Food Speciation. *Encyclopedia of Ecology*, 1636–1642.

Tavakoli, H., Alirezazadeh, P., Hedayatipour, A., Nasib, B. A. H., & Landwehr, N. (2021). Leaf image-based classification of some common bean cultivars using discriminative convolutional neural networks. *Computers and Electronics in Agriculture*, *181*(960115364), 105935. https://doi.org/10.1016/j.compag.2020.105935.

Ting, K. M. (2011). Error Rate. *Encyclopedia of Machine Learning*, 331–331. https://doi.org/10.1007/978-0-387-30164-8_262/COVER.

UNESCO. (2023). *Adolescent girls and young women in Tanzania expand digital literacy and skills.* http://www.unesco.org/en/articles/adolescent-girls-and-young-women-tanzania-expand-digital-literacy-and-skills.

Vanitha, S. (2021). Decision support model for prioritization of cotton plant diseases using integrated fahp-topsis approach. *Turkish Journal of Computer and Mathematics Education*, *12*(10), 7587-7596.

Wemmert, S., Ketter, R., Rahnenführer, J., Beerenwinkel, N., Strowitzki, M., Feiden, W., Hartmann, C., Lengauer, T., Stockhammer, F., Zang, K. D., Meese, E., Steudel, W. I., Von-Deimling, A., & Urbschat, S. (2020). Patients with high-grade gliomas harboring deletions of chromosomes 9p and 10q benefit from temozolomide treatment. *Neoplasia*, *7*(10), 883–893. https://doi.org/10.1593/neo.05307.

Wohlwend, B. (2023). *Decision Tree, Random Forest, and XGBoost: An Exploration into the Heart of Machine Learning*. https://medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948.

Wójtowicz, A., Piekarczyk, J., Czernecki, B., & Ratajkiewicz, H. (2021). A random forest model for the classification of wheat and rye leaf rust symptoms based on pure spectra at leaf scale. *Journal of Photochemistry and Photobiology B: Biology*, *223*, 112278.

Wood, T. (2023). *F-Score Definition | DeepAI*. https://deepai.org/machine-learning-glossary-and-terms/f-score.

Yerokun, O. M., & Onyesolu, M. O. (2021). On the Development of Neuro-Fuzzy Expert System for Detection of Leghemoglobin (NFESDL) in Legumes. *Advances in Multidisciplinary & Scientific Research Journal Publication*, *9*(1), 1–14.

Yong, L. Z., Khairunniza-Bejo, S., Jahari, M., & Muharam, F. M. (2023). Automatic Disease Detection of Basal Stem Rot Using Deep Learning and Hyperspectral Imaging. *Agriculture (Switzerland)*, *13*(1). https://doi.org/10.3390/agriculture13010069.

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112.

APPENDICES

Appendix 1: Interview and Focus Group Discussion Guide

BREEDERS/ FIELD TECHNICIANS QUESTIONS – 2023

1. **Tell me about your background and your job**
    1. Names:
    2. Region:
    3. Sex:
    4. Age (range):
    5. Educational level:
    6. How long have you been in this position?
    7. To whom do you report? Who reports to you:

2. **Digital literacy and digital tools at work:**
    1. Computer literacy (Do you have any knowledge of using computers?)
        Yes ☐
        No ☐
    2. Which type of phone do you have?
        Android ☐
        iOS ☐
    3. How often do you use digital tools? E.g. WhatsApp, Facebook, excel, etc.

Daily ☐
Weekly ☐
Monthly ☐

4. Why do you like or dislike using digital tools?
   ……..
   ……...
   …….
5. Do you use any digital tools at work? How often? What do you enjoy and what do you find frustrating?
   ……
   ……
   ……
6. Who helps you when you have any struggles with the tools?
   Friends
   Colleagues
   Online search
   Family
   Others(Specify)
7. What are your most important tasks while using those tools?
   ……..
   ……..
8. Is there anything you wish you could do with these products that is currently impossible?
   ……..
   ……..
   ……..
9. Are there any ways these products do not support your current needs?
   Yes(Explain):……
   ……..
   ……..

3. Breeding ecosystem and resources

1. How many people there are in your team?
   ……

2. How many trials do you manage? How many are offsite??
   …… / ……..

3. For the remote stations: who is in charge of managing it? How do they report back to you and how often?
   ……

4. How do you use technology to collaborate with other breeders or research institutions in your data analysis process?
   ……
5. Do you have any favorite collaboration platforms or tools?

   …….

4. Phenotyping data ("<u>data journey</u>") – data capacities

1. Breeding objectives (overall): what guides your scoring decision and which decisions you might take?
   …….

2. How often do you (or the technicians) go to the field?
   ……..

3. Data collection: how often is data collected during the season? who is typically in charge of collecting it?
   ……

4. How is this data collected?
   ……
5. How do you validate the data? (e.g. if different people collecting it).
   ……

6. Data processing and analysis: how is the data analyzed?
   ……....

7. By whom?

……

Storytelling (narrative questions)

Trust in digital technologies and new methodologies

- Your past experience in starting the usage of a new digital tool
  ……..
- Tell us about a specific moment in which you were very happy about using a digital tool/methodology
  ……..
  ……..
- Tell us about a specific moment in which you felt insecure about a digital tool/methodology
  ……
  ……
  ……..
- Have you ever felt supported by a digital tool/methodology?
  ……
  ……
- Have you felt that a digital tool was making your experience harder?
  ……..

New tool (need for the system questions)

- Willing to adopt new digital tools in your work?
  ……
  ……
- What are the main issues the system should be solving?
  ……
  ……
- What are the main features to be available for you in the system?
  ……
  ……
- What do you wish to be added that you were not able to do now?

……
……

END!

Appendix 2: Observation Guidelines

BREEDERS/ FIELD TECHNICIANS OBASERVATION GUIDELINE – 2023

1. Field setups

1. Is it field/ screen house
   - field ☐
   - screen house ☐

2. Observe differences between diseases

2. Equipment and Technology

1. Is there any equipment used in data collection
   - Yes ☐
   - No ☐

3. Disease scoring capacity

1. Observe three to five different persons scoring same genotype at different times and compare their answers

……/……/……

4. **Time spent on scoring a plant**
    1. Observe three to five different persons scoring same genotype at different times and compare time used on the activity
    ……/……/……
5. **Data management**

    1. Check how data are recorded from the field
    ……
    2. Check how data are stored
    ……

    3. Check data analysis methodologies in use
    ……

END!

Appendix 3: Validation Questionnaire

USER ACCEPTANCE QUESTIONNAIRE – 2024

1. **Tell me about your background and your job**
    1. Names:
    2. Region:
    3. Sex:
    4. Age (range):
    5. Educational level:
    6. How long have you been in this position?
    7. To whom do you report? Who reports to you:

(iv) The system meets the overall provided requirements

Strongly agree ☐
Agree ☐
Disagree ☐

(v) The system helps in field data collection

Strongly agree ☐
Agree ☐

Disagree ☐

(vi) Data collection can be done using your own smart phone

Strongly agree ☐
Agree ☐
Disagree ☐

(vii) The system has all data about the genotypes

Strongly agree ☐
Agree ☐
Disagree ☐

(viii) The system allows to scan leaves using spectrometer

Strongly agree ☐
Agree ☐
Disagree ☐

(ix) The system has a website to view data collected in the field

Strongly agree ☐
Agree ☐
Disagree ☐

(x) The system show all collected data with genotype classes

Strongly agree ☐
Agree ☐
Disagree ☐

(xi) The system provides search and print options for reports

Strongly agree ☐
Agree ☐
Disagree ☐

(xii) The system is easy to use

Strongly agree ☐
Agree ☐
Disagree ☐

VALIDATION QUESTIONNAIRE – 2024

1. **Tell me about your background and your job**
    1. Names:
    2. Region:
    3. Sex:
    4. Age (range):
    5. Educational level:
    6. How long have you been in this position?
    7. To whom do you report? Who reports to you:
2. **The system's results are reliable**

Strongly agree ☐
Agree ☐
Disagree ☐

3. **The system's performance is tolerable**

Strongly agree ☐
Agree ☐
Disagree ☐

4. **The system can work on different mobile phones**

Strongly agree ☐
Agree ☐
Disagree ☐

5. **The system is available all times**

Strongly agree ☐
Agree ☐
Disagree ☐

6. **The system track users**

Strongly agree ☐
Agree ☐
Disagree ☐

END!

Appendix 4: Used Codes

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="utf-8" />
    <title>Log In | Frijoles</title>
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <meta content="A fully featured admin theme which can be used to build CRM, CMS, etc." name="description" />
    <meta content="Coderthemes" name="author" />

    <!-- App favicon -->
    <link rel="shortcut icon" href="https://atrustz.000webhostapp.com/Jo//assets/images/logo.png">

    <!-- Theme Config Js -->
    <script src="https://atrustz.000webhostapp.com/Jo//assets/js/hyper-config.js"></script>

    <!-- App css -->
    <link href="https://atrustz.000webhostapp.com/Jo//assets/css/app-saas.min.css" rel="stylesheet" type="text/css" id="app-style" />

    <!-- Icons css -->
    <link href="https://atrustz.000webhostapp.com/Jo//assets/css/icons.min.css" rel="stylesheet" type="text/css" />
</head>

<body class="authentication-bg pb-0">

    <div class="auth-fluid">
        <!--Auth fluid left content -->
        <div class="auth-fluid-form-box">
            <div class="card-body d-flex flex-column h-100 gap-3">

                <!-- Logo -->
                <div class="auth-brand text-center text-lg-start">
                    <a href="javascript:void(0)" class="logo-dark text-center">
                        <span><img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo.png" alt="dark logo" height="50"></span>
                    </a>
```

```
    <body>
        <!-- Begin page -->
        <div class="wrapper"><!-- ========== Topbar Start ========== -->
<div class="navbar-custom">
    <div class="topbar container-fluid">
        <div class="d-flex align-items-center gap-lg-2 gap-1">

            <!-- Topbar Brand Logo -->
            <div class="logo-topbar">
                <!-- Logo light -->
                <a href="javascript:void(0)" class="logo-light">
                    <span class="logo-lg">
                        <img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo.png" alt="logo">
                    </span>
                    <span class="logo-sm">
                        <img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo-sm.png" alt="small logo">
                    </span>
                </a>

                <!-- Logo Dark -->
                <a href="javascript:void(0)" class="logo-dark">
                    <span class="logo-lg">
                        <img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo-dark.png" alt="dark logo">
                    </span>
                    <span class="logo-sm">
                        <img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo-dark-sm.png" alt="small logo">
                    </span>
                </a>
            </div>

            <!-- Sidebar Menu Toggle Button -->
            <button class="button-toggle-menu">
                <i class="mdi mdi-menu"></i>
            </button>


            <!-- Horizontal Menu Toggle Button -->
            <button class="navbar-toggle" data-bs-toggle="collapse" data-bs-target="#topnav-menu-content">
                <div class="lines">
                    <span></span>
                    <span></span>
                    <span></span>
                </div>
            </button>

            <!-- Topbar Search Form -->

        </div>

        <ul class="topbar-menu d-flex align-items-center gap-3">
            <li class="dropdown d-lg-none">
                <a class="nav-link dropdown-toggle arrow-none" data-bs-toggle="dropdown" href="javascript:void(0)" role="button" aria-haspopup="false" aria-expanded="false"
                    <i class="ri-search-line font-22"></i>
                </a>
                <div class="dropdown-menu dropdown-menu-animated dropdown-lg p-0">
                    <form class="p-3">
                        <input type="search" class="form-control" placeholder="Search ..." aria-label="Recipient's username">
                    </form>
                </div>
            </li>

            <li class="dropdown">
                <a class="nav-link dropdown-toggle arrow-none" data-bs-toggle="dropdown" href="javascript:void(0)" role="button" aria-haspopup="false" aria-expanded="false"
                    <img src="https://atrustz.000webhostapp.com/Jo//assets/images/flags/us.jpg" alt="user-image" class="me-0 me-sm-1" height="12">
                    <span class="align-middle d-none d-lg-inline-block">English</span> <i class="mdi mdi-chevron-down d-none d-sm-inline-block align-middle"></i>
                </a>
                <div class="dropdown-menu dropdown-menu-end dropdown-menu-animated">
```

```
    <body>
        <!-- Begin page -->
        <div class="wrapper"><!-- ========== Topbar Start ========== -->
<div class="navbar-custom">
    <div class="topbar container-fluid">
        <div class="d-flex align-items-center gap-lg-2 gap-1">

            <!-- Topbar Brand Logo -->
            <div class="logo-topbar">
                <!-- Logo light -->
                <a href="javascript:void(0)" class="logo-light">
                    <span class="logo-lg">
                        <img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo.png" alt="logo">
                    </span>
                    <span class="logo-sm">
                        <img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo-sm.png" alt="small logo">
                    </span>
                </a>

                <!-- Logo Dark -->
                <a href="javascript:void(0)" class="logo-dark">
                    <span class="logo-lg">
                        <img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo-dark.png" alt="dark logo">
                    </span>
                    <span class="logo-sm">
                        <img src="https://atrustz.000webhostapp.com/Jo//assets/images/logo-dark-sm.png" alt="small logo">
                    </span>
                </a>
            </div>

            <!-- Sidebar Menu Toggle Button -->
            <button class="button-toggle-menu">
                <i class="mdi mdi-menu"></i>


     <div class="collapse" id="sidebarProjects">
         <ul class="side-nav-second-level">
             <li class="side-nav-item">
                 <a data-bs-toggle="collapse" href="#ALS" aria-expanded="false" aria-controls="ALS">
                     <span> ALS </span>
                     <span class="menu-arrow"></span>
                 </a>
                 <div class="collapse" id="ALS">
                     <ul class="side-nav-third-level">
                         <li>
                             <a href="https://atrustz.000webhostapp.com/Jo/ALS">Collected data </a>
                         </li>
                         <li>
                             <a href="https://atrustz.000webhostapp.com/Jo/GenotypeClassfication">Forms</a>
                         </li>
                         <li>
                             <a href="https://atrustz.000webhostapp.com/Jo/ClassificationPercentage">Gen. classification</a>
                         </li>
                         <li>
                             <a href="https://atrustz.000webhostapp.com/Jo/GenotypePerRegion">Gen. per region</a>
                         </li>
                         <li>
                             <a href="https://atrustz.000webhostapp.com/Jo/ALSSummary">Summary</a>
                         </li>
                     </ul>
                 </div>


     <li class="side-nav-item">
         <a data-bs-toggle="collapse" href="#CBB" aria-expanded="false" aria-controls="CBB">
             <span> CBB </span>
             <span class="menu-arrow"></span>
         </a>
         <div class="collapse" id="CBB">
             <ul class="side-nav-third-level">
                 <li>
                     <a href="https://atrustz.000webhostapp.com/Jo/CBB">Collected data </a>
                 </li>
                 <li>
                     <a href="https://atrustz.000webhostapp.com/Jo/GenotypeClassficationCBB">Forms</a>
                 </li>
                 <li>
                     <a href="https://atrustz.000webhostapp.com/Jo/ClassificationPercentageCBB">Gen. classification</a>
                 </li>
                 <li>
                     <a href="https://atrustz.000webhostapp.com/Jo/GenotypePerRegionCBB">Gen. per region</a>
                 </li>
                 <li>
                     <a href="https://atrustz.000webhostapp.com/Jo/CBBSummary">Summary</a>
                 </li>
             </ul>
```

```
<div class="collapse" id="sidebarTasks">
    <ul class="side-nav-second-level">
        <li>
            <a href="https://atrustz.000webhostapp.com/Jo/CI">CI</a>
        </li>
        <li>
            <a href="https://atrustz.000webhostapp.com/Jo/MCARI">MCARI</a>
        </li>
        <li>
            <a href="https://atrustz.000webhostapp.com/Jo/NRI">NRI</a>
        </li>
        <li>
            <a href="https://atrustz.000webhostapp.com/Jo/PRI">PRI</a>
        </li>
        <li>
            <a href="https://atrustz.000webhostapp.com/Jo/SIPI">SIPI</a>
        </li>
        <li>
            <a href="https://atrustz.000webhostapp.com/Jo/SR">SR</a>
        </li>
        <li>
            <a href="https://atrustz.000webhostapp.com/Jo/WBI">WBI</a>
        </li>
    </ul>
</div>
```

```
class ALS extends MY_Controller {

    public function index()
    {

        if (!empty($this->input->post('dateRange'))) {
            $dateRange=explode('-', $this->input->post('dateRange'));
            $startDate = date('Y-m-d', strtotime($dateRange[0])); // Replace 'your_start_date' with the actual start date
            $endDate = date('Y-m-d', strtotime($dateRange[1]));     // Replace 'your_end_date' with the actual end date

            $critere=" AND cd.Date BETWEEN '".$startDate."' AND '".$endDate."' ";
        }else{
            $critere='';
        }

        $data['collecteddata']=$this->Model->readQuery("SELECT DATE_FORMAT(cd.Date, '%m/%d/%Y') AS Date, gs.GrowthStage_Name, r.Region_Name, COUN
        $this->load->view('ALS_View',$data);

    }

    function getDetailPerRegionAndGrowthStage($date){
        $dateOnly = str_replace('_', '', $date);

        // print_r($dateOnly);
        $data['details']=$this->Model->readQuery("SELECT DATE_FORMAT(cd.Date, '%m/%d/%Y') AS Date, gs.GrowthStage_Name, r.Region_Name, cd.Region_
        $this->load->view('Detail_View',$data);
    }
```

```
function fullDetail($CollectedData_ID){
  $data['ViewDetails']=$this->Model->readQuery("SELECT * FROM collecteddata cd join spectrumscan sp ON cd.CollectedData_ID=sp.CollectedData
  $plantScore=$this->determineResult($data['ViewDetails']);
  $data['plantScore']=$plantScore;
    $this->load->view('FullDetail_View',$data);
}

function getPlantDetail(){
    $CollectedData_ID=$this->input->post('CollectedData_ID');
    $plants=$this->Model->readQuery("SELECT sp.SpectrumScan_ID,sp.CollectedData_ID,sp.ModelResult,cd.PlantNumber FROM spectrumscan sp JOIN
    $plantScore=$this->determineResult($plants);
    $totalScan=count($plants);
    $j=1;
    $i=1;

    foreach ($plants as $row) {

        if ($j== 1) {
            echo "<tr>";
            echo "<td rowspan=" . count($plants) . ">" . $row['PlantNumber'] . "</td>";
        } else {
            echo "<tr>";
        }
```

```
function determineResult($plants) {
    $modelResults = array_column($plants, 'ModelResult');

    if (in_array('Susceptible', $modelResults)) {
        return 'Susceptible';
    } elseif (in_array('Medium', $modelResults)) {
        return 'Medium';
    } elseif (in_array('Resistant', $modelResults)) {
        return 'Resistant';
    }

    return ''; // Return a default value if none of the conditions are met
}


function printCollectedData(){
  $CollectedData_ID=$this->input->post('CollectedData_ID');
  $data['ViewDetails']=$this->Model->readQuery("SELECT * FROM collecteddata cd join spectrumscan sp ON cd.CollectedData_ID=sp.CollectedData
  $plantScore=$this->determineResult($data['ViewDetails']);
  $data['plantScore']=$plantScore;
  $this->load->view('PrintData_View',$data);
}
```

```php
$chartData = json_encode([
    'series' => [
        ['name' => 'Plant Count', 'data' => array_column($genotypeData, 'PlantCount')]
    ],
    'chart' => [
        'type' => 'bar',
        'height' => 350
    ],
    'plotOptions' => [
        'bar' => [
            'horizontal' => false,
            'columnWidth' => '55%',
            'endingShape' => 'rounded'
        ]
    ],
    'dataLabels' => [
        'enabled' => false
    ],
    'stroke' => [
        'show' => true,
        'width' => 2,
        'colors' => ['transparent']
    ],
    'xaxis' => [
        'categories' => array_column($genotypeData, 'Genotype_Name'),
    ],
    'yaxis' => [
        'title' => [
            'text' => ' (Plants)'
```

```python
import tensorflow as tf
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
label_encoder = LabelEncoder()
y_train_encoded = label_encoder.fit_transform(y_train)
y_test_encoded = label_encoder.transform(y_test)
y_train_onehot = tf.keras.utils.to_categorical(y_train_encoded)
y_test_onehot = tf.keras.utils.to_categorical(y_test_encoded)
model = tf.keras.Sequential([
    tf.keras.layers.Dense(128, activation='relu', input_shape=(X_train_scaled.shape[1],)),
    tf.keras.layers.Dropout(0.2),
    tf.keras.layers.Dense(64, activation='relu'),
    tf.keras.layers.Dropout(0.2),
    tf.keras.layers.Dense(y_train_onehot.shape[1], activation='softmax')
])

model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])


model.fit(X_train_scaled, y_train_onehot, epochs=300, batch_size=32, validation_data=(X_test_scaled, y_test_onehot))
```

```
Epoch 296/300
241/241 [==============================] - 1s 3ms/step - loss: 0.1015 - accuracy: 0.9646 - val_loss: 1.9947 - val_accuracy: 0.9480
Epoch 297/300
241/241 [==============================] - 1s 3ms/step - loss: 0.0893 - accuracy: 0.9656 - val_loss: 1.9837 - val_accuracy: 0.9506
Epoch 298/300
241/241 [==============================] - 1s 3ms/step - loss: 0.0990 - accuracy: 0.9612 - val_loss: 1.9195 - val_accuracy: 0.9521
Epoch 299/300
241/241 [==============================] - 1s 3ms/step - loss: 0.0912 - accuracy: 0.9640 - val_loss: 2.0115 - val_accuracy: 0.9490
Epoch 300/300
241/241 [==============================] - 1s 3ms/step - loss: 0.1122 - accuracy: 0.9564 - val_loss: 2.1328 - val_accuracy: 0.9454
<keras.src.callbacks.History at 0x7917fc478e20>
```

```
[ ] # Save the model architecture and weights
    model.save('neural_network_model.h5')
    joblib.dump(label_encoder, 'nn_label_encoder.joblib')
```

```
import xgboost as xgb
from sklearn.metrics import accuracy_score, classification_report
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()

y_train_encoded = label_encoder.fit_transform(y_train)
y_test_encoded = label_encoder.transform(y_test)
xgb_classifier = xgb.XGBClassifier(random_state=42, use_label_encoder=False, eval_metric='logloss')

xgb_classifier.fit(X_train_scaled, y_train_encoded)
y_pred_encoded = xgb_classifier.predict(X_test_scaled)
y_pred = label_encoder.inverse_transform(y_pred_encoded)

accuracy = accuracy_score(y_test, y_pred)


class_report = classification_report(y_test, y_pred)

accuracy, class_report
```

```
print("Accuracy:", accuracy)
print("Class Report:\n", class_report)
print("Error Rates:", error_rates)
print("Mean Average Precision:", mean_average_precision)
```

```
Accuracy: 0.963579604578564
Class Report:
               precision    recall  f1-score   support

      Medium       0.96      0.96      0.96       640
   Resistant       0.98      0.96      0.97       641
 Susceptible       0.96      0.97      0.97       641

    accuracy                           0.96      1922
   macro avg       0.96      0.96      0.96      1922
weighted avg       0.96      0.96      0.96      1922

Error Rates: {'Medium': 0.02861602497398541, 'Resistant': 0.021331945889698223, 'Susceptible': 0.02289281997918835}
Mean Average Precision: 0.994871092511696
```

```
import pandas as pd

file_paths = {
    'Medium': ['/content/drive/MyDrive/notebooksmodels/Medium1.csv', '/content/drive/MyDrive/notebooksmodels/Medium2.csv'],
    'Resistant': ['/content/drive/MyDrive/notebooksmodels/Resistant1.csv', '/content/drive/MyDrive/notebooksmodels/Resistant2.csv'],
    'Susceptible': ['/content/drive/MyDrive/notebooksmodels/Suceptible1.csv', '/content/drive/MyDrive/notebooksmodels/Suceptible2.csv']
}

dfs = []

for label, paths in file_paths.items():
    for path in paths:
        df = pd.read_csv(path)
        df['label'] = label
        dfs.append(df)
```

RESEARCH OUTPUTS

(i) Research Paper

Irakiza, J., Kaijage, S., Leo, J., Mamo, T., Mbelwa, H., & Guerena, D. (2023). Identification of Resistant Common Bean Genotypes to Foliar Diseases using Hyperspectral Data. In *2023 First International Conference on the Advancements of Artificial Intelligence in African Context*, 1-9.

(ii) Poster Presentation

Appendix 5: Poster Presentation