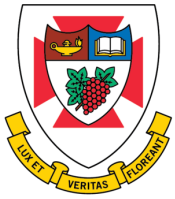
The University of Winnipeg

Department of Applied Computer Science

MSc in Applied Computer Science and Society



THE UNIVERSITY OF
WINNIPEG

Project Title: SIFRA - A Cassava Diseases Detection Model

Instructor: Professor Sheela Ramanna

Report Written by:
Razia Zaman Ela
Student ID. 3181095

Academic year 2023/2024

Project: SIFRA *

Razia Zaman Ela[1]

The University of Winnipeg, 515 Portage Ave, Winnipeg, MB R3B 2E9

Abstract. Human civilization is greatly affected both positively and negatively by the rapid advancement of Artificial Intelligence (AI). Authorized and domain-specific skilled entities should supervise the usage of Artificial Intelligence and Automated tools. Ethical and noble implementation of Artificial Intelligence may lead to discoveries of unresolved mysteries. The objective of the project is to build a responsible AI that will serve the purpose of the greater good. This project is about building an image-based prediction Model that will automatically predict the class of the image if it is properly labeled with domain experts. To build the model a dataset from a Kaggle competition was selected that contains leaves images of cassava plants. Cassava plant diseases are one of the major threats to the world's food security. In West Africa, it was reported as being one of the major causes of famine. The dataset was highly imbalanced. However, in real-life scenarios, data will be mostly imbalanced or unknown. Primarily CNN was selected to check how the model performs on the data or whether is there a scope for improvement on the data. It showed accuracy<= 54%. The techniques applied to this project showed a value of 100% for each category.

Keywords: Machine Learning · Artificial Intelligence · Digital Image Processing· Prediction · Classification

1 Introduction

Agriculture is a sector that is directly connected with human civilization, culture as well as traditions. Climate change drastically interferes with Natural or human-made calamities which leads scientists to research and invent new technology. Over the past few decades, there has been a continuous revolution and applications in the field of the Internet of Things (IoT). IoT is a field where scientists, engineers, and, end-users build techniques, technology, and solutions to tackle problems in real-life scenarios. Specifically, African Agriculture production is a matter of increasing concern to the Major International organization[2]. Populated regions for example Africa are facing food insecurity due to the rising demand for food and climate change. In this study [2] researchers build a prediction system based on machine learning models that will predict the yield of six crops called: maize, seed cotton, bananas, yams, maize, rice, and cassava. Cassava is a root vegetable that holds the position of the third-largest food

* Supported by the University of Winnipeg.

in tropical regions across Africa, South America, and, Asia. Various types of diseases impact cassava yield. Cassava diseases have social and economic consequences for example threats to livelihood, and food security, and may result in famine. Studies support that advanced farming practices and disease monitoring can decrease the influence of cassava diseases in the field.[4]

So, This project focuses on building a system to classify various states of cassava leaves from supervised image data. [5]. As per the paper [1], IoT (Internet of Things) exploits the "Smart farm" model. The study focuses on designing and deployment of practical tasks that ranges from crop harvest forecasting to missing or wrong sensors data reconstruction, exploiting and comparing various machine learning techniques. The results of the study show the direction to employ better efforts and investments.

2 Related Works

Another Study [2] suggested that World and African agricultural production in particular are of increasing concern to major international organizations. The study claimed farmers and agricultural decision-makers need advanced tools to help them make quick decisions that will impact the quality of agricultural yields. The impact of climate change has been observed on agricultural production. They proposed a prediction system based on machine learning to predict the yield of six crops, namely: rice, maize, cassava, seed cotton, yams, and bananas, at the country level in the area of West African countries throughout the year. They combined climatic data, weather data, agricultural yields, and chemical data to help decision-makers and farmers predict the annual crop yields in their country. Promising results are found using three machine learning models. They applied a hyper-parameter tuning technique throughout cross-validation to get a better model that does not face overfitting.

Inspired by the Kaggle competition which was part of the Fine-Grained Visual Categorization workshop at CVPR 2019 (Conference on Computer Vision and Pattern Recognition) a study [5] aimed at detecting cassava diseases using 5 fine-grained cassava leaf disease categories with 10,000 labeled images collected during a regular survey in Uganda. Traditionally, this detection is done mostly through physical inspection and supervision of cassava plants in the garden by farmers or agricultural extension workers. To automate formal methodology this study detects cassava infection earlier to help farmers apply preventive techniques to the non-infected cassava plants in order to improve on yields which subsequently increases African food basket leading to food security which fights famine. This paper, focused on techniques to achieve better results using deep convolutional neural networks from scratch.

3 Theoretical Framework

Overall, the theoretical framework encompasses key concepts and techniques involved in image preprocessing, data partitioning, classifier modeling, and model evaluation. It provides a structured approach for understanding and interpreting the experimental process and results, guiding future research and application of machine learning techniques for image classification tasks. Figure 1 shows the theoretical framework of the project

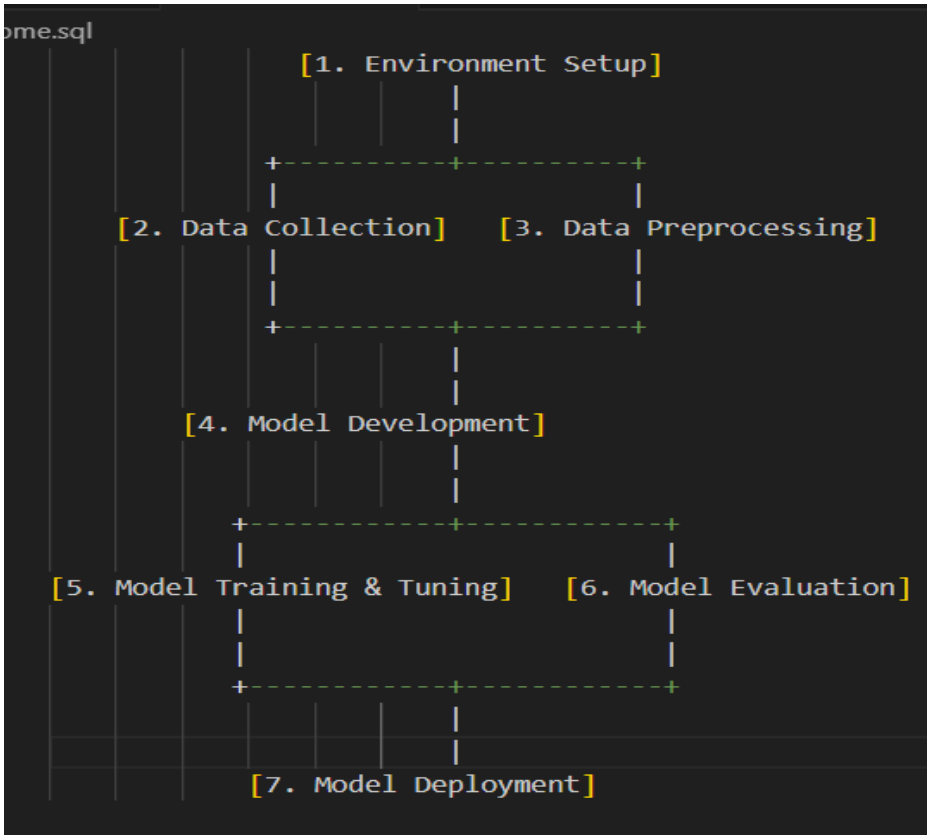


Fig. 1: Theoretical framework

4 Environment Setup

The implementation strategy of a paper was followed to check whether the computing environment was compatible with producing promising accuracy value [5] The implementation strategy of the study was partially followed and the model

classified the data with an accuracy value less than 100%. So, the computing environment was efficient in building the project. Fig 2 shows the implementation workflow of the project.

System Specifications The computing environment of the project has the following configuration:

- Python version: 3.10.12
- CPU: x86_64 architecture
- Physical cores: 1
- Total cores: 2
- Total RAM: 12 GB
- Available RAM: 11 GB
- GPU: No GPU detected.

5 Data Collection

Images used in this project are gathered from a competition on Kaggle, a popular platform for data scientists. The competition was a part of FGVC6 workshop) known as the fine-grained visual categorization workshop at CVPR 2019. The dataset [3] comprises 5 fine-grained cassava leaf disease was classified with 9436 labeled images collected during a survey in Uganda. As per instructions, farmers took images of their gardens and the experts of the National Crops Resources Research Institute (NaCRRI) along with the AI lab at Makarere University in Kampala annotated the images. So, only the leaf portion of the cassava plant was captured during the data collection process. So, there are 9436 labeled images and 12595 unlabeled images of cassava leaves.

The dataset includes 3 files as follows:

- train.zip : size 776
- test.zip : size 514 MB
- extra_images.zip : size 1.04 GB

train.zip has a root folder named train and 5 sub-folders that contain corresponding images for the 5 different classes and the names of those sub-folders are cbb,cbsd,cgm, cmd, and healthy respectively. **test.zip** and **extra.zip** contain images of cassava plants. Figure 3 shows the folder structure of the **train.zip**

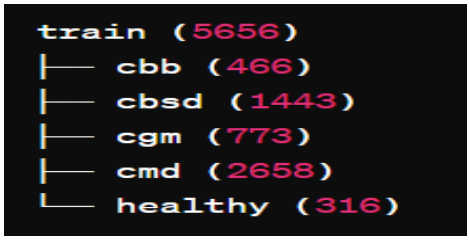


Fig. 2: Data Directory folder structure

Here, cbb, cbsd, cgm, cmd folders contain the images of cassava leaves that are affected by Cassava Bacterial Blight, Cassava Brown Streak Disease, Cassava Green Mite, Cassava Mosaic Disease respectively. And, healthy folder contains images of healthy cassava leaves.

(a) Cassava Bacterial Blight

(b) Cassava Brown Streak Disease

(c) Cassava Green Mite

(d) Cassava Mosaic Disease

(e) Cassava healthy leaf

Fig. 3: Cassava Plant Data

6 Implementation

The collected dataset contains 4 sets of disease data and 1 set of healthy data. The objective of the project is to predict a target class from the cassava plant leaf images. So, it is decided to select one dataset of a particular class, train the model, and examine the value of the classification report. Figure 4 shows the model implementation strategy of the project.

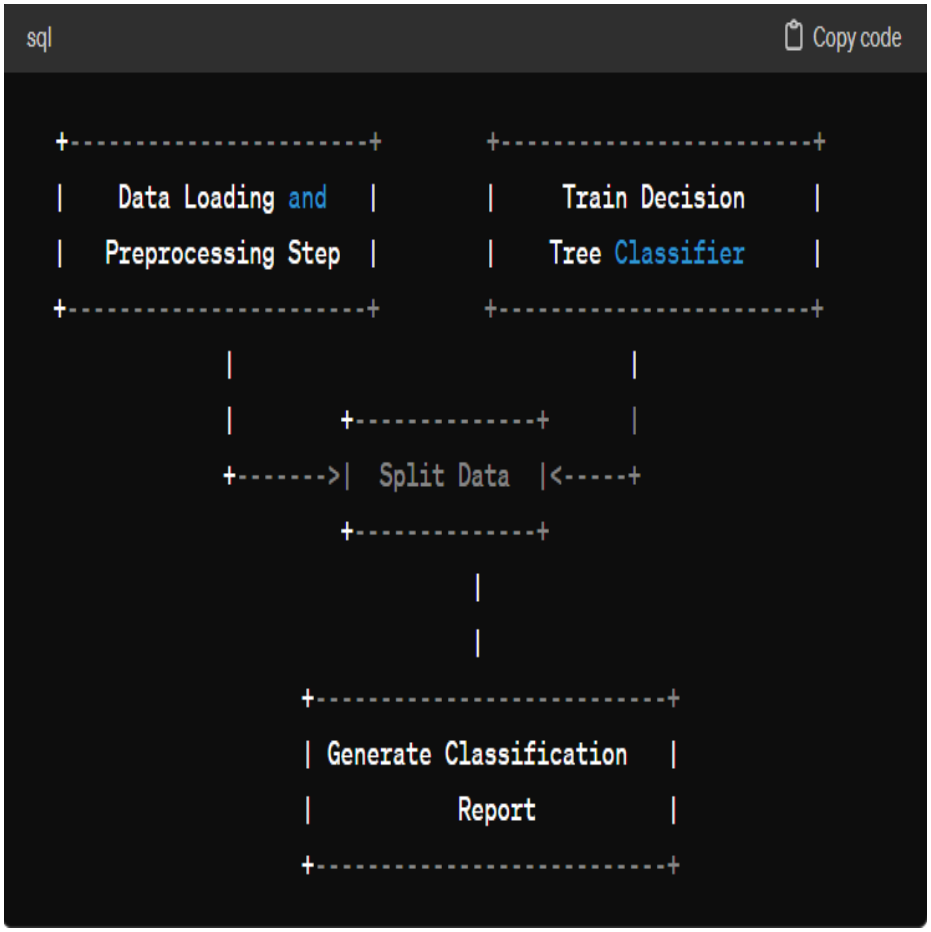


Fig. 4: Model Implementation Workflow

6.1 Data Pre-processing

A function was implemented for the purpose of data pre-processing. The data samples were labeled images. So, The function named image_pre-processing takes two arguments: data_directory and category_name. The images are pre-processed from a specified category(target class) directory within the given data_directory which is the parent directory of the category_name directory. Fig 3 shows the images of 5 particular classes of cassava Plants. The function loads images and converts them to numpy array after resizing them to a fixed size (224,224). After selecting the data samples for a particular target class, TensorFlow was used to extract features from the image. The folder name of the dataset was used to label the features of each sample. .

6.2 Model Training and Tuning

Libraries used in the code:

- os
- numpy
- matplotlib.pyplot
- sklearn.model_selection
- sklearn.metrics
- imblearn.over_sampling
- tensorflow

The dataset was split into training and testing sets using train_test_split of scikit learn. The ratio of training and testing data was 80/20. The shape of the dataset between 5 models varies as the number of samples in each class is different. The decision tree was selected to train the model. Here the model is trained using the features extracted from one category. For each category, the model produced an accuracy of 1.

6.3 Model Evaluation

Classification report is a popular evaluation technique for Machine learning models. In this project metrics known as Precision, Recall, F1-Score, Support, and Accuracy were observed closely to understand the model performance. High precision, Recall, and F1-Score indicate the strength of the model. Figure 5,6,7,8,9 shows the classification report of models trained using the features of cgm,cbsd,cbb,cmd, healthy respectively. The models showed outstanding results in all metrics.

Classification Report:

	precision	recall	f1-score	support
cgm	1.00	1.00	1.00	155
accuracy			1.00	155
macro avg	1.00	1.00	1.00	155
weighted avg	1.00	1.00	1.00	155

Fig. 5: cgm

```
Classification Report:
              precision    recall  f1-score   support

        cbsd       1.00      1.00      1.00       289

    accuracy                           1.00       289
   macro avg       1.00      1.00      1.00       289
weighted avg       1.00      1.00      1.00       289
```

Fig. 6: cbsd

```
Classification Report:
              precision    recall  f1-score   support

         cbb       1.00      1.00      1.00        94

    accuracy                           1.00        94
   macro avg       1.00      1.00      1.00        94
weighted avg       1.00      1.00      1.00        94
```

Fig. 7: cbb

```
Classification Report:
              precision    recall  f1-score   support

         cmd       1.00      1.00      1.00       532

    accuracy                           1.00       532
   macro avg       1.00      1.00      1.00       532
weighted avg       1.00      1.00      1.00       532
```

Fig. 8: cmd

```
Classification Report:
              precision    recall  f1-score   support

     healthy       1.00      1.00      1.00        64

    accuracy                           1.00        64
   macro avg       1.00      1.00      1.00        64
weighted avg       1.00      1.00      1.00        64
```

Fig. 9: healthy

7 Experiments and Results

To validate the model's performance several types of experiments were performed. A paper using the same dataset gain an accuracy of 93In their study they used Contrast Limited Adaptive Histogram Equalization(CLAHE) as an image processing technique and Convolutional Neural Network (CNN) to classify the dataset into various classes. The challenge of the strategy was that the dataset was highly imbalanced. To check the validity of their claim the data was plotted using matplotlib. Their claim was correct. Fig 10 shows the data distribution in each class.

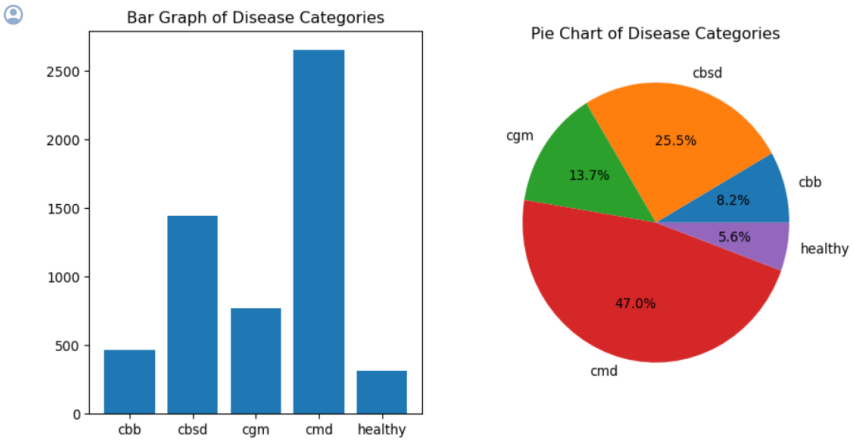


Fig. 10: Data Distribution per class

7.1 Experiment 1

The implementation technique of the paper was followed to train a CNN model with CLAHE pre-processing technique. Unfortunately, this replica showed 43% accuracy.

7.2 Experiment 2

Image augmentation technique of TensorFlow was applied to the features. The CNN model trained using this strategy showed a 54% result

7.3 Experiment 3

CLAHE as image pre-processing, Image augmentation technique of TensorFlow was applied to the features. The CNN model trained using this strategy showed 46% accuracy

7.4 Experiment 4

Dimensional reduction technique like Principle Component Analysis (PCA) was utilized in this experiment. The model logistic regression trained using PCA-enhanced features showed 24% accuracy.

The following table shows the accuracy value of different experiment performed while building this project, along with the accuracy of the project.

Table 1: Comparison of ML Techniques and Accuracy

ML Techniques	Accuracy
TensorFlow + CLAHE + CNN	43%
TensorFlow + Image Augmentation + CNN	54%
TensorFlow + CLAHE + Image Augmentation + CNN	46%
TensorFlow + tSNE/PCA + CNN	<= 25%
This Project	100%

8 Statement about responsible AI

To harness the power of AI in the way of greater good and advancement there are some principles that are followed in practice. A responsible AI refers to the ethical practice of artificial Intelligence technologies with consideration for social and cultural impact. Responsible AI includes the practice of implementation of AI in such a way that it respects human rights while addressing potential risks. To ensure responsible AI practices authorities take measures such as ethical consideration, fairness and bias mitigation, Transparency and Accountability, consent, and data protection.

9 Conclusion

The model achieved 100% accuracy after several experimental attempts. This is a remarkable achievement that shows the model's strength to accurately predict target variables across all observations. Several research studies were observed to build the model. The feature engineering using the TensorFlow library alone contributed highly to the outcome of the model. The model performs best for datasets with labeled values. But in the case of imbalanced data, it doesn't show promising results. There is scope for improvement in the case of the imbalanced dataset. The project will show promising results if the data is properly labeled with experts.

Acknowledgements

I would like to express thanks to Dr. Michael Beck and Professor Sheela Ramanna for their support and guidance during the project.

I am also grateful to my colleagues for their valuable comments and insights about the project.

References

1. Fabrizio Balducci, Donato Impedovo, and Giuseppe Pirlo. Machine learning applications on agricultural datasets for smart farm enhancement. volume 6, page 38. MDPI, 2018.
2. Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Rubby Aworka, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Moez Krichen, and Charles Lebon Mberi Kimpolo. Crops yield prediction based on machine learning models: Case of west african countries. volume 2, page 100049. Elsevier, 2022.
3. Timnit Gebru ErnestMwebaze. Cassava disease classification. Kaggle, 2019.
4. Emily J McCallum, Ravi B Anjanappa, and Wilhelm Gruissem. Tackling agriculturally relevant diseases in the staple crop cassava (manihot esculenta). volume 38, pages 50–58. Elsevier, 2017.
5. GAOGD Sambasivam and Geoffrey Duncan Opiyo. A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. volume 22, pages 27–34. Elsevier, 2021.