VIETNAM NATIONAL UNIVERSITY OF HOCHIMINH CITY

THE INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING



**APPLYING FEDERATED LEARNING TO BUILD PLANT DISEASE RECOGNIZING AI MODEL**

By

LE TRIEU LONG

A thesis submitted to the School of Computer Science and Engineering

in partial fulfillment of the requirements for the degree of   
Bachelor of Information Technology/Computer Science/Computer Engineering

Ho Chi Minh City, Vietnam  
Year

**APPLYING FEDERATED LEARNING TO BUILD PLANT DISEASE RECOGNIZING AI MODEL**

APPROVED BY: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ,   
Dr. Vo Thi Luu Phuong.   
\*(*Type Committee names beneath lines*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
(*Typed Committee name here*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
(*Typed Committee name here*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
(*Typed Committee name here*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
(*Typed Committee name here*)

THESIS COMMITTEE   
(Whichever applies)

# ACKNOWLEGMENTS

I’m extremely grateful to the knowledges gathered from 3 years study at International University. Lecturers, friends have done a massive role of my field learning process since the very beginning steps.

With special thanks to Ph.D. Vo Thi Luu Phuong who guide me at IoT Research Club and in this pre-thesis project as an Inspirable Advisor.

*Note: Paper A4, Top: 2.5cm; Bottom: 2cm; Left: 3cm; Right: 2cm*

# TABLE OF CONTENTS

[ACKNOWLEGMENTS 3](#_Toc514144956)

[TABLE OF CONTENTS 4](#_Toc514144957)

[LIST OF FIGURES 6](#_Toc514144958)

[LIST OF TABLES 7](#_Toc514144959)

[ABSTRACT 8](#_Toc514144960)

[CHAPTER 1 9](#_Toc514144961)

[INTRODUCTION 9](#_Toc514144962)

[1.1. Background 9](#_Toc514144963)

[1.2. Problem Statement 9](#_Toc514144964)

[1.3. Scope and Objectives 9](#_Toc514144965)

[1.4. Assumption and Solution 9](#_Toc514144966)

[1.5. Structure of thesis 10](#_Toc514144967)

[CHAPTER 2 11](#_Toc514144968)

[LITURATURE REVIEW/RELATED WORK 11](#_Toc514144969)

[2.1. Review 1 **Error! Bookmark not defined.**](#_Toc514144970)

[2.1.1. Sub review 1 11](#_Toc514144971)

[2.2. Review 2 **Error! Bookmark not defined.**](#_Toc514144972)

[2.2.1. Sub review 2 13](#_Toc514144973)

[CHAPTER 3 14](#_Toc514144974)

[METHODOLOGY 14](#_Toc514144975)

[3.1. Overview 14](#_Toc514144976)

[3.2. User requirement analysis 15](#_Toc514144977)

[3.2.1. Sub 1 15](#_Toc514144978)

[3.3. System Design 15](#_Toc514144979)

[3.3.1. Datadase design 15](#_Toc514144980)

[3.3.2. User Interface design 15](#_Toc514144981)

[3.4. Abc 15](#_Toc514144982)

[CHAPTER 4 15](#_Toc514144983)

[IMPLEMENT AND RESULTS 15](#_Toc514144984)

[4.1. Implement 16](#_Toc514144985)

[4.1.1. Abc **Error! Bookmark not defined.**](#_Toc514144986)

[4.2. Results 16](#_Toc514144987)

[4.2.1. Result 1 16](#_Toc514144988)

[4.2.2. Result 2 16](#_Toc514144989)

[CHAPTER 5 17](#_Toc514144990)

[EXPERIMENTAL RESULT 17](#_Toc514144991)

[5.1. Dataset Description 17](#_Toc514144992)

[5.2. Evaluation 17](#_Toc514144993)

[5.3. Illustration 17](#_Toc514144994)

[CHAPTER 6 18](#_Toc514144995)

[CONCLUSION AND FUTURE WORK 18](#_Toc514144996)

[6.1. Conclusion 18](#_Toc514144997)

[6.2. Future work 18](#_Toc514144998)

[REFERENCES 19](#_Toc514144999)

[APPENDIX 20](#_Toc514145000)

# LIST OF FIGURES

*Figure 1.1: Abc*

# LIST OF TABLES

*Table 3.1: Cde*

# ABSTRACT

Abc…

# CHAPTER 1

# INTRODUCTION

## Background

Nowadays, with the increasing world population, food is one of the main problems humans must deal with for sustainable development. Agriculture becomes more significant in many countries to handle food security. Plant disease always is one of the most dangerous “enemies”, therefore, the ability to recognize disease is a crucial key to keeping a healthy crop.

## Problem Statement

There are many factors could affect the result of each planting season, for instances:

methods used, quality of the seeds, disaster… and the most dangerous to the crops from the very beginning of the agriculture are diseases. Diseases on plants could be recognize by observing the symptom showed, like on leafs

## Scope and Objectives

While in past, the symptom is recognized by farmers using most experience or knowledges gather from verbal and books, today the scientist developed the computer vision model using computer to recognize disease given the pictures of leaf which is more efficient and easier to use. The model needs to be feeded by large dataset provided from the real farms to have the best accuracy, however collect images from the farms do not want to public the images of the problems occuring on their plants and collecting images from farms also requires massive bandwidth and memory to process.

We will build an AI model trained by federated learning (FL) framework to recognize plant diseases. The application can provide the farmer ability to recognize and share information about plant diseases. It should have the following features:

- Friendly UI.

- Preserve the privacy of farmers’ data. The data is not collected to a server to train the AI model.

All requirements meet these features of the AI model trained by federated learning framework.

## Assumption and Solution

In the project presentation, there are assumptions to commit. The mobile application was only made as to present the result model trained by federated learning algorithm and the UI design is not an focusing criteria.

The quality of model is evaluation in the laboratory and not made as production, therefore, application could not work as expected in many situations.

## Structure of thesis

The Pre-thesis structured in 6 chapters are organized as follows:

* **Chapter 1:** Introducing
* **Chapter 2:** Literature review
* **Chapter 3:** Methodology: include user requirement analysis, system design, database design and user Interface design. It will also have activity diagram and use case diagram.
* **Chapter 4:** Implement and Result in view of a work can be considered as a dedicated one or not through practical success.
* **Chapter 5:** Experimental result
* **Chapter 6:** Conclusion and Future work is the summary of processes in the pre-thesis.

# CHAPTER 2

# LITURATURE REVIEW/RELATED WORK

## Plant Disease dataset

Many research done to make an plant disease recognization with many approaches like SVM (Support Vector Machines), K-Nearest Neighbours (KNN), ConvNet models like) InceptionNet, VGGNet, etc.

Angie K. Reyes et al training CNN on soybean images draw from PlantVillage Dataset. Furthermore, augmentation, dropout layer and L2 regularization are also experimented to figure out which is suitable to get the highest result. The output was 99.32% with imbalanced data, augmentation and dropout layer combined. [6].

R. Sujatha et al using many models from Machine Learning to Deep Learning to compares performance of them on plant leaf disease dataset [7]

## Federated Learning Algorithm

There was much research talk about federated training algorithm since Google introduced the first paper relevant [4] applied in many fields

Ryan McDonald et al introduced distributed strategies for perceptron training as solution for big and complex data. The model was trained on [5]. H. Brendan McMahan et al published Communication-Efficient Learning of Deep Networks from Decentralized Data [3] paper in which introduced fedAvg (Federated Averaging) algorithm in communication process to deal with unbalanced and non-IID dataset. SGD used as an optimization algorithm applied on MNIST CNN, SHAKESPEARE LSTM, CIFAR 10. Training with various models: a multi-layer perceptron, two different convolutional NNs, a two-layer character LSTM, and a large-scale word-level LSTM in different conditions. The output result that the fedAvg algorithm outperformed fedSGD in all assumption conditions with the loss decrease faster and the test accuracy is better though. However, these researchs are just applied experimentally on toys dataset. Therefore, we would try to test them in a real dataset

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

### Sub review 2

Abc...

…

# CHAPTER 3

# METHODOLOGY

## Overview

## ML Pipeline

Chosen Mobilenet V3 as the main model because of it advantages on size, which is suitable for the devices. Commonly, there are basic steps to fitting a model: data preprocessing, model training, model evaluation and validation.

Diagram

Description automatically generated

Common AI model training process

Federated learning algorithm, the process become more complex. T

Diagram

Description automatically generated

Federated Algorithm round process

To stimulate the real situation, N clients are using as substitute to each farm being trouble with one type of plant diseas and there is one server that process and hold the main model parameters. Each client has its S size dataset included images of size N in size of 4000 x 2672 px and the model provided by the server. There are connections between clients and server to send and receive parameters of model.

* + N must larger than the number of classes in all datasets own by clients
  + S is the number unified and chosen by either server or clients

Federated training with fedAvg algorithm used in the research to construct a central model from the parameters received from clients that can recognize all the data images the clients have.

## User requirement analysis

Abc...

### Sub 1

Abc...

## System Design

Abc….

### Datadase design

### User Interface design

## Abc

# CHAPTER 4

# IMPLEMENT AND RESULTS

## Implement

Abc…

* + 1. Model Training
       1. Pytorch
       2. Jupyter Notebook
    2. Mobile Application

## Results

Experiments is built by criteria :

+ balance of data (non-iid , iid )

+ User per round: 15, 25, 50 ,100

+ Number user per round: 15, 25 ,50, 100

### Result 1

Abc…

### Result 2

Abc…

# CHAPTER 5

# EXPERIMENTAL RESULT

## Dataset Description

The Plant Panthology – FCGV8 is the dataset made for the FGVC8 (The Eight Workshop on Fine-Grained Visual Categorization) Plant Pathology 2021 Challenge competition []. Contain 23,249 multi-class and 1,703 multi-label Apple leaf’s images in RGB. The multi-class dataset contains images among 5 disease categories including Cedar apple rust, Powdery mildew, Apple scab, Fogeye leaf spot, Complex and Healthy.

|  |  |  |  |
| --- | --- | --- | --- |
| Foliar diseases | Total images | Training dataset | Test dataset |
| Apple scab | 6034 | 4827 | 1207 |
| Frogeye leaf spot | 3973 | 3183 | 796 |
| Ceder apple rust | 2329 | 1863 | 466 |
| Powdery mildew | 1513s | 1210 | 303 |
| Complex | 3610 | 2888 | 722 |
| Healthy | 5784 | 4631 | 1158 |
| Total | 23249 | 18602 | 4652 |

Table \_. Total number of multi-class data

Chart, pie chart

Description automatically generated

Figure \_. The raw dataset labels percentile.

Deal the huge amount of data, we will take a sample of 6000 images from the multi-class dataset while keeping almost the same distribution of labels in environment.

Chart, pie chart

Description automatically generated

Figure \_. 6000 images sample dataset labels percentile

## Evaluation

* + 1. Introduction

The effective of training a common model depend on many parameters (learning rate, batch size, epoch, ...). However, optimizing a model using a federated learning algorithm we have also considered some specific parameters (number client per round, local epoch, distribution of number per client). In this experiment, we expect to figure out the relation of these parameters and data distribution with the accuracy of the model. The result is expected to be used for hyperparameters tuning in federated training processes.

Experiment divided into 3 cases with different number clients used for training process per round (5/100, 15/100, 50/100) while other parameters is By analyzing the change in number epoch needed to converge, the maximum accuracy, minimum loss, we can know whether clients per round can result better. In each case, we tried to perform 3 cases of label distribution:  1-2 labels per client, 3 labels per client and not limited (random) the log can show which circumstance is better and the consonance with the number of clients per round.

* + 1. Experiment Cases
       1. Non-IID dataset
       2. IID dataset

## Illustration

Abc…

# CHAPTER 6

# CONCLUSION AND FUTURE WORK

## Conclusion

Abc...

## Future work

Abc…

# REFERENCES

1. MARTY HALL, LARRY BROWN “Core Servlets and JavaServer Pages™ - Volume 1 Core Technologies”, Second Edition, 2003.
2. MARTY HALL, LARRY BROWN “Core Web Programming”, Second Edition, 2001.
3. DAVE CRANE, ERIC PASCARILLO, DARREN JAMES, “Ajax in Action”, 2006.
4. JAMES L. WEAVER, KEVIN MUKHAR, AND JIM CRUME, “Beginning J2EE 1.4: From Novice to Professional”, 2004.
5. ?

# APPENDIX