الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
جامعة عمار ثليجي الاغواط
AMAR TELIDJI LAGHOUAT UNIVERSITY
كلية العلوم
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS AND INFORMATICS

Master Thesis

Domaine : Mathematics and Computer science

Filiere : Computer science

Option :Informatics and Decision System

realized by :

Mahfoudi Moussa Annaceur

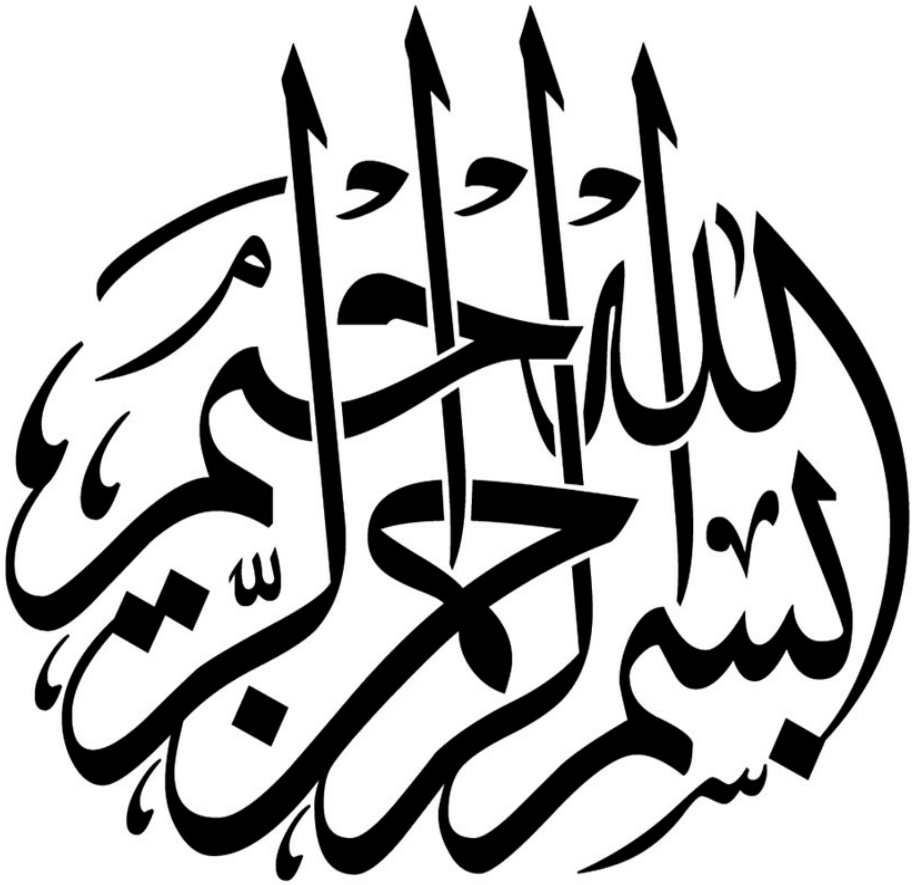
Theme

Predction Disease using Machine Learning

Publicly supported on 30-06-2022 before the jury composed of :

Dr Hamini Narjes	**MC(A)**	**President**
Dr Bouzidi Mohamed	**MC(A)**	**Examiner**
Pr Cherroun Hadda	***MC(A)***	***Supervisor***

Academic Year 2021/2022


بسم الله الرحمن الرحيم

ACKNOWLEDGEMENT

First of all, I thank Allah, our founder, for giving me the strength, the will and the courage to do this modest work and i would thank my parents with all my respect and love ,as a sign of affection ,gratitude for all their encouragement and sacrifice throughout my studies,I would like to thank my coach **Professor CHERROOUN Hadda** for her supervision, her availability and to provide all the material and moral possibilitie and her precious advice during the elaboration of this work and also thank **Dr Gouluma Younes** who has allocated his time to us to add lessons on the machine learning that were not included in our formation.And Special Thank to my colleague **SACI Abedlhamid** for motivated me to finish this work . all thanks also to the members of the jury who agreed to judge and assist in the defense of this work. I would also like to express our gratitude to all the teachers in the Department of Informatics for their scientific and educational qualifications. I would like to warmly thank all my loved ones and all those who from far and wide,have brought me their solicitude to accomplish this work.

Abstract

The economy of countries depends on agriculture to meet the needs of the population, achieve self-sufficiency and export of peasant products to other countries, but there are some problems that threaten this sector, including fungal diseases and pests .These disease are mainly caused by weather and climate changes. In this work we investigate the usage of machine learning to predict these diseases and learn their relationship with the weather changes .Our work is apart of a large smart agriculture project that aims gathering dataset on agriculture using Internet of things . This dataset en composes many weather features and it is annotated manually by an expert on targeted diseases .we focus the deployed ML algorithm is Random forest and SVM. The generated models are achieved for disease prediction.

keywords : smartagriculture,disease prediction, ,SVM,Forest,Internet of things,Machine Learning

Résumé

L'économie des pays dépend de l'agriculture pour répondre aux besoins de la population, atteindre l'autosuffisance et exporter des produits paysans vers d'autres pays, mais il y a certains problèmes qui menacent ce secteur, y compris les maladies fongiques et les parasites . Ces maladies sont principalement causées par les conditions météorologiques et les changements climatiques. Dans ce travail, nous étudions l'utilisation de l'apprentissage automatique pour prédire ces maladies et apprendre leur relation avec les changements climatiques . Notre travail fait partie d'un grand projet d'agriculture intelligente qui vise à recueillir des données sur l'agriculture à l'aide de l'Internet des objets . Ce jeu de données compose de nombreuses caractéristiques météorologiques et il est annoté manuellement par un expert sur les maladies ciblées . nous concentrons l'algorithme ML déployé est Random forest et SVM. Les modèles générés sont obtenus pour la prédiction des maladies.

mots clé : smartagriculture,disease prediction, ,SVM,Forest,Internet of things,Machine Learning

ملخص

يعتمد اقتصاد البلدان على الزراعة لتلبية احتياجات السكان وتحقيق الاكتفاء الذاتي وتصدير منتجات الفلاحين إلى بلدان أخرى، ولكن هناك بعض المشاكل التي تهدد هذا القطاع، بما في ذلك الأمراض الفطرية والآفات. هذا المرض ناتج بشكل أساسي عن تغير الطقس والمناخ. في هذا العمل، نحقق في استخدام التعلم الآلي للتنبؤ بهذه الأمراض ومعرفة علاقتها بتغيرات الطقس. عملنا بصرف النظر عن مشروع زراعي ذكي كبير يهدف إلى جمع مجموعة بيانات عن الزراعة باستخدام إنترنت الأشياء. تؤلف مجموعة البيانات هذه العديد من ميزات الطقس ويتم تعليقها يدويًا من قبل خبير في الأمراض المستهدفة. نركز على خوارزمية پی المنتشرة وهي غندم ذرست و صطپ. يتم تحقيق النماذج المتولدة للتنبؤ بالأمراض.

TABLE DES MATIÈRES

1 Introduction **2**

2 Generalities **7**

2.1 Smart Agriculture 7

2.2 Application in Smart Agriculture 7

2.2.1 Diseases Detection 7

2.2.2 Weather Management in Green House 9

2.2.3 Precision Irrigation 9

2.3 Plant Disease 9

2.4 Types of Plant Disease 10

2.5 Weather parameters 10

2.6 Problem of Plant Disease Prediction 11

2.7 Conclusion 12

3 Machine Learning **13**

3.1 Introduction 13

3.2 Machine Learning 13

3.2.1 Supervised Learning 14

3.2.2 Unsupervised Learning 15

3.3 Reinforcement Learning 16

3.4 Machine Learning Algorithms 17

3.4.1 Logistic Regression 17

3.4.2 Support Vector Machine(SVM) 18

3.4.3 Decision Tree . . . 19
3.4.4 Random Forest . . . 20
3.5 Deep Learning . . . 21
3.6 Principle of Artificial Neural Networks . . . 22
3.6.1 Types of Deep Neural Networks . . . 23
3.7 Applications of Deep Learning . . . 23
3.7.1 Translation and Language Processing . . . 23
3.7.2 Assistant Drive Cars . . . 23
3.7.3 Healthcare . . . 24
3.8 Conclusion . . . 25

4 Related Work **26**
4.1 Diseases prediction . . . 26
4.2 Classification and Symptoms Visualization . . . 28
4.3 Disease Recognition . . . 29
4.4 Conclusion . . . 31

5 Proposed Approach **32**
5.1 Global Overview . . . 32
5.2 Metrics . . . 33
5.2.1 Precision . . . 34
5.3 Dataset Collection and Annotation . . . 35
5.3.1 Pre-processing . . . 36
5.4 Experiments . . . 37
5.4.1 Used Tools . . . 37
5.4.2 Implementation . . . 38
5.4.3 Results . . . 40
5.5 Conclusion . . . 42

TABLE DES FIGURES

1.1 Global View of project *Dz-SmartFarm* 3

3.1 types of regression [1] 15
3.2 clustering [2] 16
3.3 logistic regression [3] 17
3.4 Support vector Machine [4] 18
3.5 rondom forrst algorithem [5] 20
3.6 ANN discription [6] 21
3.7 ANN discription [6] 22
3.8 healthcare using deep learning [7] 24

4.1 Reviewed Related Work 27

5.1 Overview of the ML-Based Plant Disease Prediction System 33
5.2 A Sample of the Data set 36
5.3 Import the dataset script 38
5.4 Defining the Forest Model 39
5.5 SVM model script 39
5.6 Test of the predicting model Script 40
5.7 Performances of the Generated Models in term of accuracy 40

LISTE DES TABLEAUX

4.1 accuracies obtained using CNNs training differnet datasets. 30
4.2 summary of related work. 31

5.1 Information about the Datasets. 36
5.2 SVM-based model Vs Random Forest model performances. 41

Abbreviation

IOT : Internet Of things

ML : Machine Learning

SVM :Support Vector Machine

AI :Artificial Intelligence

RL :Reinforcement Learning

DL :Deep Learning

ANN :Artificial Neural Network

CNN :Conventional Neural Network

RNN :Recurrent Neural Network

CHAPITRE 1

INTRODUCTION

Context

Healthy, sustainable and inclusive food systems are critical to achieve country's development goals. Agricultural development is one of the most powerful tools to end extreme poverty, boost shared prosperity, and feed a projected 9.7 billion people by 2050. Growth in the agriculture sector is two to four times more effective in raising incomes compared to other sectors. Agriculture is also crucial to economic growth. in 2018, it accounted for more than 25% of global gross domestic product (GDP) in some least developing countries [1].

However agriculture-driven growth, poverty reduction, and food security are at risk : Multiple shocks – from COVID-19 related disruptions to extreme weather, pests and conflicts – are impacting food systems, resulting in higher food prices and growing hunger.

Our research work is part of a larger project called Dz-Smart-Farm that aims harnessing the Algerian Agriculture by means of usage and deployment of advanced Information Technologies. This latter includes Internet of Things, Machine Learning and its advanced techniques with sustainable energies.

The Project functionalities are illustrated in Figure 1.1. Its grantees are essentially farmers, agriculture actors : politics, leaders and researches in smart agriculture domains .

The project aims to monitor crop health through different analysis, the main ones are :

1. Deployment of optimized IoT networks that can gather practical Data,

1. https://www.worldbank.org/en/topic/agriculture/overview#1 seen on 08/05/2022

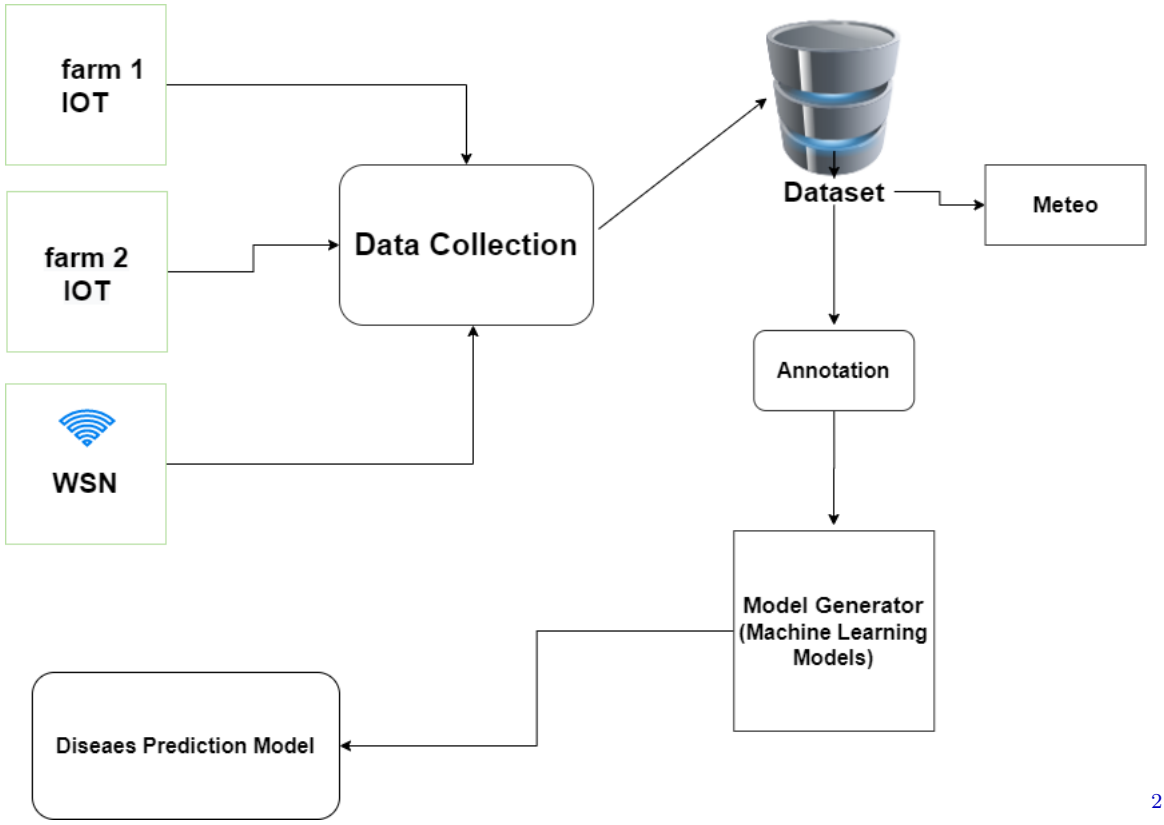


FIGURE 1.1 – Global View of project *Dz-SmartFarm*

2. Creation of many largest Datasets from real in-sit environment(plant images,weather).
3. Providing farms by real practical information in order to well manage their corps outcomes,
4. Prediction of diseases,
5. Remote farms Monitoring,
6. Provide effective decision support tools for decision makers.

Our contribution in this substantial project is the involvement in the generation of Machine Learning based models that can predict plant diseases using the collected datasets on weather in sit .

Algeria is focusing on improving the peasant sector and setting large amounts in order to meet its markets and aspire to export agricultural products, such as agricultural products, but there are obstacles that threaten agriculture and the growth of some plants, including diverse terrain.

In fact, Algeria's climate changes from one region to another and peasants grow crops according to the climate of their regions and, to the extent possible, trying to avoid diseases affecting their crops.

In addition, the peasant according to us budget places pesticides, inculcates some crops in green houses and supplies the land with the mineral salts that plants need to grow well. However, some of the above-mentioned are regarded as obstructionists. Pesticides kill insects, but they negatively affect the acidity of the ground.

Weather change also affects plants, especially sand winds in the desert and snow in the north, sometimes kills the plant and temperature change affects plants that need water. Heat leads to evaporation. This may not take into account the desire of some plants.

.

Plant Disease Problem

Plant disease can be defined in many ways, but one of the simplest definitions describes disease as any condition in a plant caused by living and nonliving agents that interferes with its normal growth and development. Diseases or plant health problems can impact plants in many ways since all parts of a plant can be affected including flowers, leaves, fruits, seeds, stems, branches, growing tips, and roots.

Many different factors can cause plant health problems. These factors can be divided into two groups based on whether they are living or nonliving. Nonliving disease agents, often called antic agents, include factors such as environmental stress or cultural care. Living disease agents, called bio-tic agents or plant pathogens, include microorganisms such as fungi and bacteria. Both abiotic and biotic agents will be described in greater detail in the section.

.

Objectives and Motivations

The main objective of this master thesis is to carry out the agriculture using machine learning with data set collected by IOT , So the goal is to predict the decease caused by the weather effect.

Thesis Structure

This manuscript is divided into 4 chapter

The first chapter, we react the generalities about the agriculture field and some concept such as the Application of smart agriculture and some kind of disease, this dissertation has also the following chapters :

The chapter 2, is presented the literature review made, describing the state-of-art : Machine learning and deep learning the main existing techniques. The related works done in these subject are also presented in the third chapter.

the chapter 4 the implementation stage developed during the thesis is explained.

To finalize, finishes the dissertation by presenting the final remarks and future work.

CHAPITRE 2

GENERALITIES

In this chapter we present the agriculture and weather changes in Algeria and we describe the smart agriculture and their application ,in addition to we explain some types of plant diseases to give a pick in our subject .

2.1 Smart Agriculture

Smart agriculture is a relatively new word, and the majority of farmers are unaware of what it entails. The use of technology such as the Internet of Things, sensors, positioning systems, robotics, and artificial intelligence on a farm is referred to as smart agriculture. The ultimate objective is to improve the quality and quantity of crops while reducing the amount of human work required [**?**].

2.2 Application in Smart Agriculture

There are many field and application that can benefit from smart agriculture technologies :deseases detection ,weather management in green houses ,managing and controlling irrigation ...Here after we details some of thier :

2.2.1 Diseases Detection

In this application we use the AI to analyse the plant to discover some diseases can threat it by using image processing and deep learning to predict diseases ,In addition to

using climate data to analyse and detect diseases .

Detection and identification of diseases in crops could be realized via both direct and indirect methods. Direct detection of diseases includes molecular and serological methods that could be used for high-throughput analysis when large numbers of samples need to be analyzed. In these methods, the disease causing pathogens such as bacteria, fungi and viruses are directly detected to provide accurate identification of the disease/pathogen. On the other hand, indirect methods identify the plant diseases through various parameters such as morphological change, temperature change, transpiration rate change and volatile organic compounds released by infected plants [8]. .

2.2.2 Weather Management in Green House

Weather control refers to keep the desired values of the green house inside parameters as temperature, light, humidity, and CO2 concentration in order to provide optimal conditions for the plants. With respect to the energy cost, the temperature in greenhouse should keep around a desired level. Lower accuracy requires higher set point in temperature, which results in more energy consumption . The temperature can be increased by heating of the air and, at the same time, the use of thermal screens may decrease the energy consumption during nighttime periods. The light control is mainly based on the requirement of the plant growing age. The control of humidity provides to avoid the extremely high or low air humidity values. Concerning to energy saving avoiding the high humidity is more important. The control of humidity can be achieved by heating and/or by ventilation [9].

Advantages of Greenhouse Climate Control

- Reduced risk for molds and diseases.
- Improved plant growth, yeald and quality.
- Accurate control of greenhouse climate systems.
- Saved energy cost.

2.2.3 Precision Irrigation

Precision irrigation is a unique sustainable agricultural approach that allows the application of water and nutrients to the plant at the right time and place and in small measured doses in order to provide it with optimal growing conditions [10]. Today the technologies help to develop this approach to estimated the water by using for example ; smart rubber and Arrow drip-per . This precise tools are used to deliver water directly to the root. It consists of arrow adapter, rubber tubing, and water arrow drip-per.

2.3 Plant Disease

In general, a plant becomes diseased when it is continuously disturbed by some causal agent that results in an abnormal physiological process that disrupts the plant's normal structure, growth, function, or other activities. This interference with one or more of a plant's essential physiological or biochemical systems elicits characteristic pathological conditions or symptoms [8].

2.4 Types of Plant Disease

A large spectre of disease can effect plants :

Leaf Spot : dead, discolored, or injured areas of tissue that usually have distinct margins ; spots often appear on leaves or fruit.

Blight : rapid yellowing, browning, collapse, and death of leaves, shoots, stems, flowers, or the entire plant.

Chlorosis : yellowing of leaves and stems that are normally green.

Necrosis : browning or blackening of areas on a plant indicative of the death of plant cells.

Wilt : loss of turgor or drooping of leaves, shoots, or the entire plant due to lack of water.

Distortion : twisting or other abnormal traits of leaves, stems, and shoots.

Mosaic : uneven pattern of yellow, light green, or dark green, usually on leaves.

Canker : dead area on a stem or branch ; can be sunken, swollen, or discolored and is usually distinguished from adjacent healthy tissues by color.

Rot : breakdown and decay of plant tissue, often used to describe conditions in roots and fruit.

Dieback : death of the tips of leaves, shoots, and stems ; failure of branches to develop, especially in the spring.

Witches' broom : abnormal proliferation of shoots from the same point on a plant resulting in a bushy, broom-like appearance.

Gall : a swelling or abnormal growth of plant tissues ; can develop on leaves, stems, and roots.

Stunt :abnormally small-sized plant parts due to the failure of those plant parts to grow to full size ; often used to describe an entire plant [11].

2.5 Weather parameters

The elements of weather are the character atmospheric variables that paintings collectively to create distinctive weather situations and set up climate styles. The number one elements include temperature, air pressure, wind, humidity, precipitation, visibility, clouds, and sunlight publicity.And the change of these parameter has a negative impact

in the agriculture field .

2.6 Problem of Plant Disease Prediction

The planet disease prediction can help the farmers to avoid the threat and improve the corp by learning from the previous expert . This is to overcome diseases that threaten the crops and use pesticides before the disease occurs . In order to address these problems we can use machine learning to anticipate some dealing with some diseases. .

2.7 Conclusion

In this chapter we learned some preliminaries about the smart farm and techniques used in irrigation and protecting plants from the green house method and mentioned some diseases and weather parameters that affects the condition of the plant. In addition, we learned about the problem of predicting diseases .

CHAPITRE 3

MACHINE LEARNING

In this chapter we present some concepts about AI , machine learning and deep learning and talking about the neural networks and their types .

3.1 Introduction

Human beings endeavor to organize their tasks and teach them to the computer so that they do not repeat them by presenting their expertise and knowledge to the machine in order to perform these tasks that are routine and new and the machine learns through practice and expertise in the field. Traditionally we try to solve some type of problems by defined algorithms and implement them and keep for the computer to execute these instructions, and we prefer to introduce and deployment of artificial intelligence, that computers can solve problems like humans . So what's the technology used in our study : plants disease predication in agriculture ? And how to bypass this disease by using machine learning ?

3.2 Machine Learning

Machine learning methods are based on teaching computers to learn from dataset and have an experience instead of being explicitly programmed to do so.Data , Models and training are the element of machine learning , The models are trained to find patterns and correlations in large data sets and to make a decisions and predictions. Machine

learning applications improve with use and become more accurate the more data they have access to[12]. Applications of machine learning are become all around us like in healthcare, industry, military... .

There are four learning models That can be used in this domain : supervised, unsupervised, semi-supervised, or reinforcement.

3.2.1 Supervised Learning

it's one of the most used methods in ML, it takes both training data and its associated output during the training process. The major goal of supervised learning methods is to learn the association between input training data and their labels. For this it performs multiple training data instances Supervised algorithms are called supervised because the machine learning model learns from data samples with the knowledge of the output [13].

Some of the well-known supervised machine learning algorithms are : KNN (k-nearest neighbors), Decision tree, Logistic Regression, and Random Forest.

we can divide supervised learning algorithms in following two classes :

Classification : Classification-based tasks predict the categorical output responses or labels for the given data. This output, which will belong to a specific discrete category, is based on what our ML model learns in the training phase.

Types Of Learners In Classification

1. **Lazy Learners**– Lazy learners simply store the training data and wait until a testing data appears. The classification is done using the most related data in the stored training data. They have more predicting time compared to eager learners. Eg – k-nearest neighbor, case-based reasoning.

1. **Eager Learners**– Eager learners construct a classification model based on the given training data before getting data for predictions. It must be able to commit to a single hypothesis that will work for the entire space. Due to this, they take a lot of time in training and less time for a prediction.Eg – Decision Tree, Naive Bayes, Artificial Neural Networks [14].

Regression : Regression-based to predict the continues numerical output responses or labels for the given training data. This output will also be based on what our ML model learns in the training phase. For example : to make decision about tow thinks [15] .

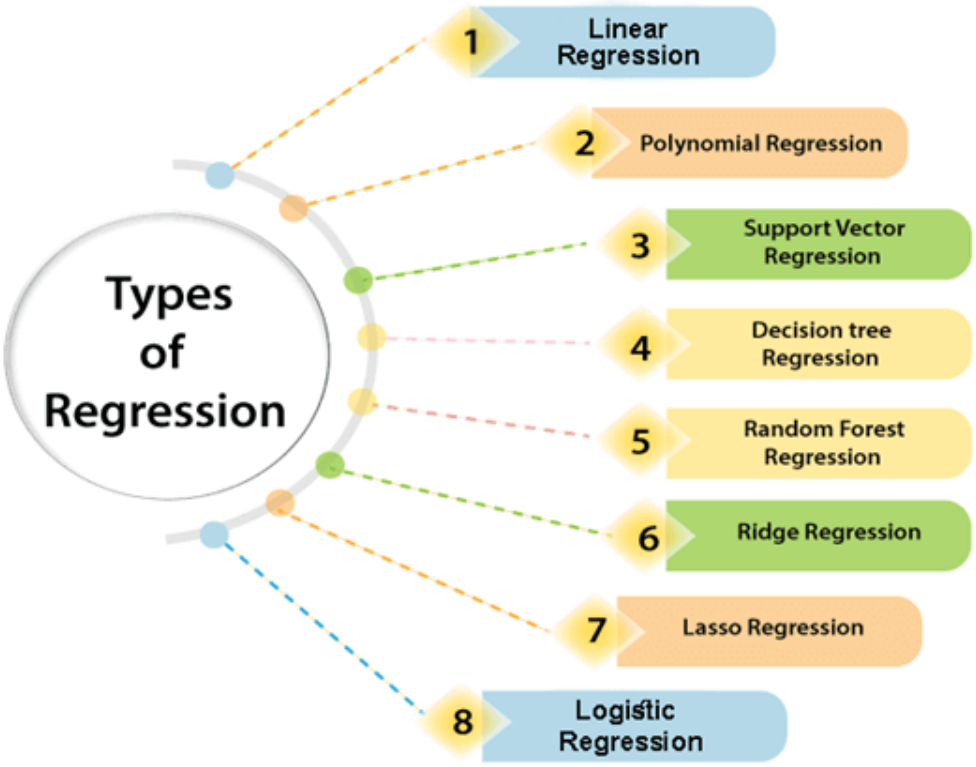


FIGURE 3.1 – types of regression [1]

3.2.2 Unsupervised Learning

In unsupervised machine learning, we use a learning algorithm to discover unknown patterns in unlabeled datasets.

This is contrary to supervised machine learning that uses human-labeled data. Unsupervised learning algorithms use unstructured data that's grouped based on similarities and patterns.

Clustering Analysis

Clustering is the process of dividing uncategorized data into similar groups or clusters. This process ensures that similar data points are identified and grouped. Clustering algorithms is key in the processing of data and identification of groups (natural clusters).

The following image shows an example of how clustering works.

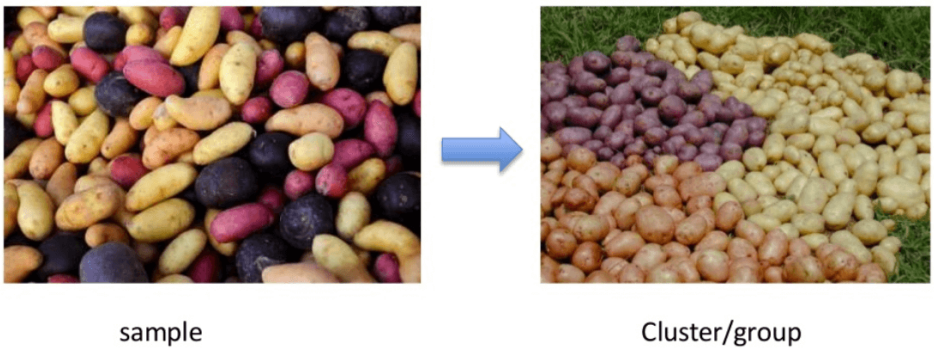


FIGURE 3.2 – clustering [2]

3.3 Reinforcement Learning

Reinforcement Learning (RL) is the technological know-how of choice making. It is ready studying the ideal conduct in an surroundings to gain most praise. This most beneficial behavior is found out via interactions with the environment and observations of ways it responds, much like youngsters exploring the sector round them and studying the movements that assist them acquire a intention.

This discovery manner is comparable to an ordeal-and-mistakes seek. The first-rate of movements is measured by using not simply the instantaneous reward they return, however additionally the not on time reward they might fetch [16].

3.4 Machine Learning Algorithms

3.4.1 Logistic Regression

It is the proper regression evaluation to conduct when the structured variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive evaluation. Logistic regression is used to describe facts and to explain the relationship among one structured binary variable and one or extra nominal, ordinal or ratio-stage impartial variables.

Sometimes logistic regressions are tough to interpret ; the Intellectus Statistics tool without problems permits you to behavior the evaluation, then in simple English interprets the output [15].

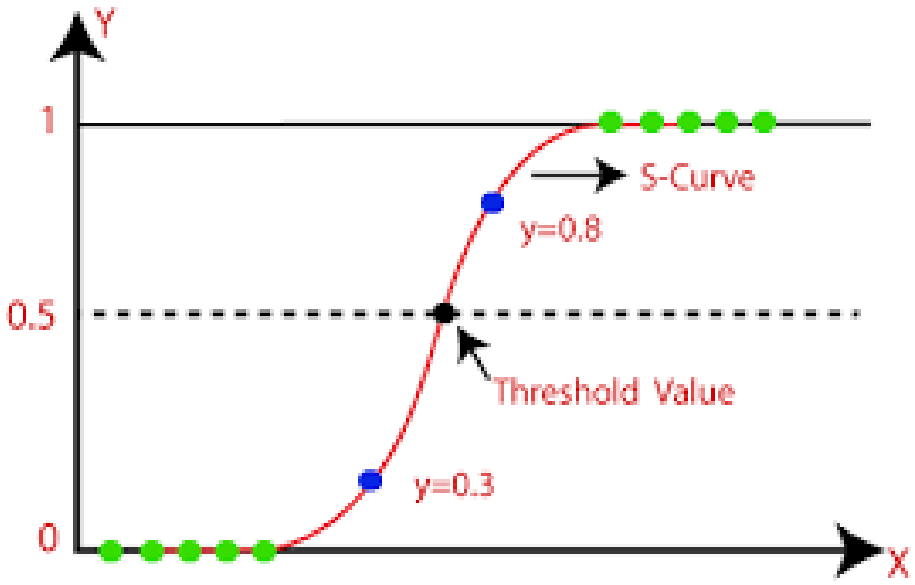


FIGURE 3.3 – logistic regression [3]

3.4.2 Support Vector Machine(SVM)

is a supervised system learning set of rules that can be used for both category or regression challenges. However, it's far commonly utilized in class troubles. In the SVM set of rules, we plot each data object as a factor in n-dimensional area with the price of each characteristic being the price of a specific coordinate. Then, we perform class by locating the hyper-aircraft that differentiates the 2 instructions thoroughly (look at the below snapshot) [17].

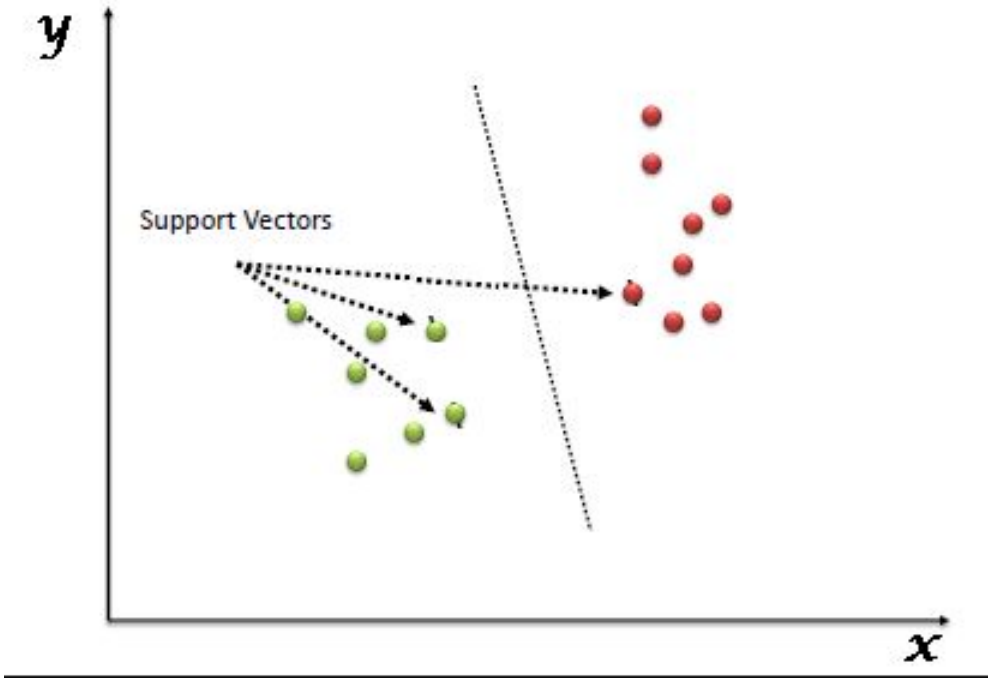


FIGURE 3.4 – Support vector Machine [4]

3.4.3 Decision Tree

A selection tree is a supervised getting to know approach that has a pre-defined target variable and is most often used in classification problems. This tree may be implemented to both specific or continuous input & output variables. The education technique resembles a go with the flow chart, with each internal (non-leaf) node a check of an characteristic, each department is the final results of that take a look at, and every leaf (terminal) node includes a class label. The uppermost node within the tree is known as the foundation node.

In the selection method, the sample (population) is break up into or more sub-populations units of maximal, that's decided via the most tremendous splitter or differentiation inside the enter variables.

The closing goal is to create a predictive model that may take observations about a pattern (the branches) and make accurate conclusions about the sample's target fee (the leaves).

Types of Decision Trees :

The predominant families of selection timber are described by means of characteristic :

Classification Trees : Used are expecting the magnificence to which the facts sample belongs.

Regression Tree : Used when the outcome isn't a classifier, but alternatively an actual range.

Some approaches assemble multiple choice bushes, or ensembles, to resolve unique issues. A few not unusual examples :

Boosted trees – Used to educate times that have been formerly incorrectly modeled. For example, AdaBoost. This works for both regression and class problems.

Bootstrap aggregated decision trees – Used for classifying information that's hard to label by means of using repeated sampling and constructing a consensus prediction [18]

3.4.4 Random Forest

Random Forest, like its call implies, consists of a massive variety of man or woman choice timber that perform as an ensemble. Each individual tree inside the random wooded area spits out a category prediction and the class with the maximum votes turns into our version's prediction (see determine under) [19].

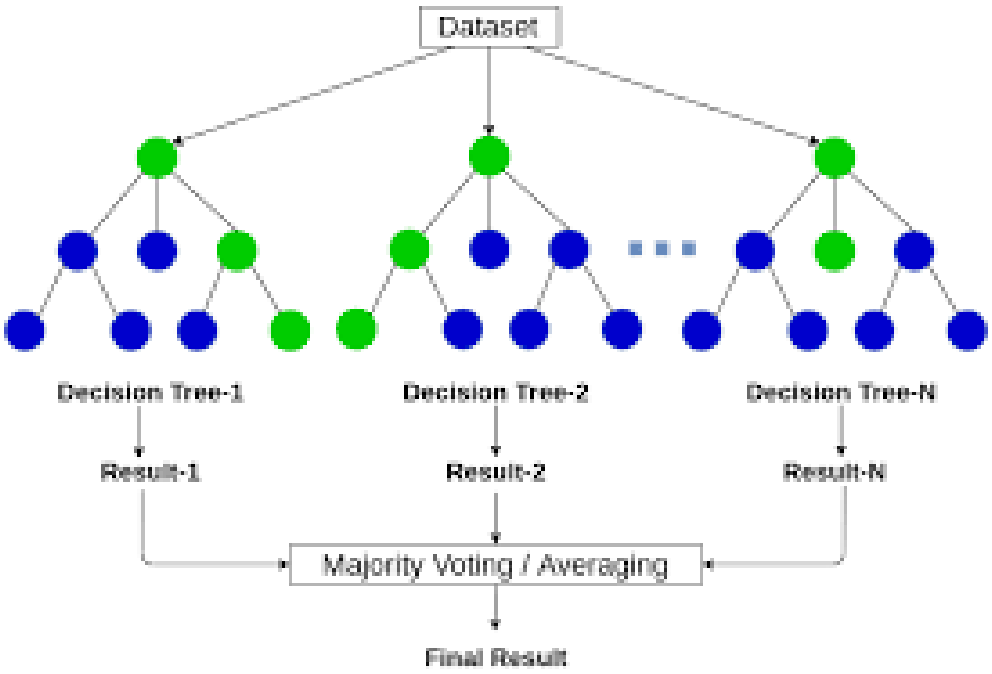


FIGURE 3.5 – rondom forrst algorithem [5]

3.5 Deep Learning

Deep learning is a sort of synthetic intelligence, derived from the system Learning that changed into superior with the intention of making algorithms succesful examine and enhance independently, in assessment to programming wherein the device virtually executes to the letter of predetermined rules..

Artificial Neural Network is a computer system stimulated by way of the functioning of the human brain to research. This is a ramification of Deep Learning era, which is itself part of the Artificial Intelligence sub-class and Machine Learning. This kind of community is described by a set of layers of neurons that are strongly interconnected

A network of neurons is made of a set of successive layers, every of which takes its enter from the outputs of the preceding one. That this set is fully connected. Each layer is a set of neurons that have no connection between them and receive digital information from neighbouring neurons. The layer set consists of an input layer that reads the enter values, an output layer that gives the system effects and between these two hides one with numerous so-known as hidden layers concerned inside the switch. As shown within the following Figure .

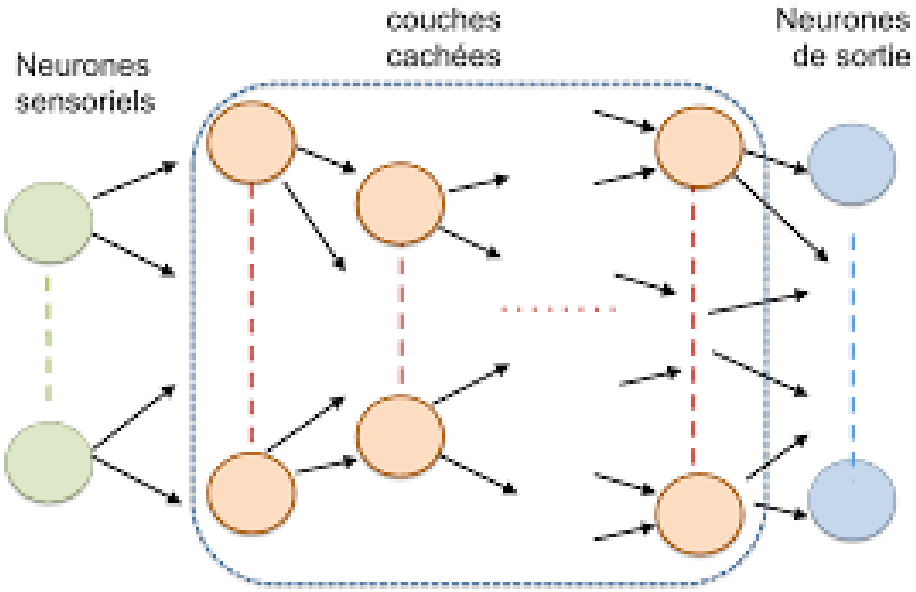


FIGURE 3.6 – ANN discription [6]

3.6 Principle of Artificial Neural Networks

Let's now depict the key actions that were made at various points during the use of this NN to define the principle of how a neural network functions.

Each neuron has a weight coefficient that is allocated to it.

Each inputted value may be increased using this coefficient, and all outcomes would then receive an addition feature. A property like this is referred to as an aggregate characteristic.

The sum received will then be compared with a predetermined threshold when we apply an activation characteristic. The neuron ceases to prompt if the sum is below the threshold, which frequently occurs when the sum is negative, indicating that the data are uninteresting. The neuron activates and transfers the data to the next layer neuron (hidden layers) until the remaining statistics reach the remaining neuron if this sum is above the threshold (typically in the scenario where the sum is effective). The final outcome (output) will then be available.

This principle is illustrated in the following Figure .

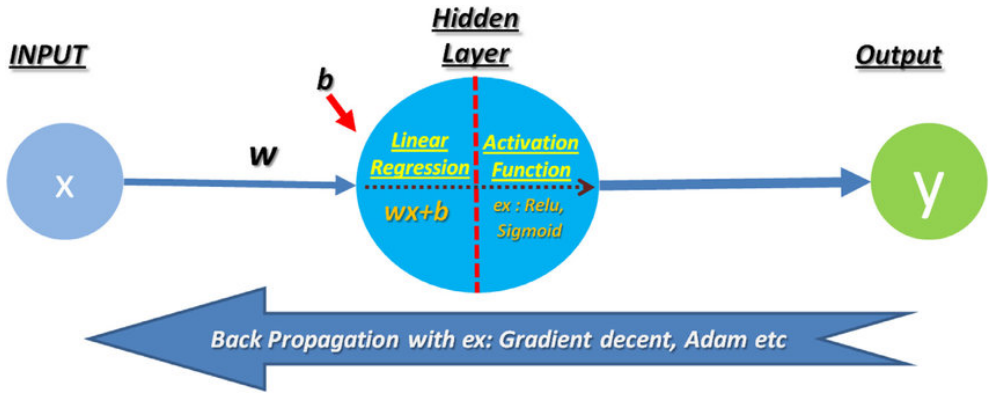


FIGURE 3.7 – ANN discription [6]

3.6.1 Types of Deep Neural Networks

Neural Networks referred to as "feed-forward" (to propagation before)

That is the handiest variation and the statistics moves simplest in a single route, it is going at once through the input to the processing nodes (hidden layers) and then to the outputs, with out a cycle or loop within the network.

Recurrent Neural Network (RNN)

This mode of studying is a bit extra complex, they have got at least one cycle in their structures, they shop the outcomes produced by using the processing nodes and feed the model with these effects. Its using in : automated pattern recognition, computerized speech translation...

Conventional Neural Network(CNN)

He way of operating is inspired through a biological method that is the visual cortex of animals. The statistics to be processed bypass via a mess of layers (from the input layer to the output layer).

This form of network applies in special fields : vision computing , photo and video popularity, natural language processing...

3.7 Applications of Deep Learning

The deep learning have applied in a lot of domain which is :

3.7.1 Translation and Language Processing

Translation is based in neural network from large data to matching the most commune words and it using NLP to translated ,orthography : like google research give their users suggestion to help us for research using BERT data set .

3.7.2 Assistant Drive Cars

In these application of deep-learning use the CNN to control the stats of the driver to protect them from accident .and organize the roads are obstructed by traffic accidents.

This technology protect people because most accidents lead to disasters.

3.7.3 Healthcare

One of the application of Deep learning is healthcare .most diseases are scanned .It's the role of identifying object and it is a feature of CNN by selection the dimensions of viruses like : SARS .

FIGURE 3.8 – healthcare using deep learning [7]

3.8 Conclusion

In this chapter, we first looked at the representation of machine learning as a wellknown idea and its different types. Then we delved into one among its kinds, deep learning to know, and at this degree we mentioned the perception of neural networks, on which deep studying and its different types are based. We have followed up with the operating principle of this kind and its one of a kind regions of application..

CHAPITRE 4

RELATED WORK

In this chapter we will present some previous work related to the use of artificial intelligence in agriculture. Researchers used many ways to detect diseases and gave them different results.

For the purpose of our study on the topic, we have reviewed some research paper that have dealt with the forecasting plant diseases and the way to prevent them.

Figure 4.1 illustrates the reviewed work according to their targeted task. In fact .....

4.1 Diseases prediction

Priyanka et al. [20] have proposed an approach for the Prediction of Potato Late Blight Disease based upon weather parameters using Artificial Neural Network. They have used Weather data that were obtained from field sites in Pantnagar(India). The weather parameters used are -maximum temperature, minimum temperature, maximum humidity, minimum humidity, and rainfall.

The accuracy of model depends on training dataset. They are split the dataset into subgroups, training and testing dataset.The dataset contains weather information from the year 2011-2015. It contains 88 rows in total. The dataset is divided such that there are 3 cases for testing the model. The Researches in this work reached divided their dataset into many combinations (so-called observations) :

58 observations in training set and 30 observations in testing set 44 observations in training set and 44 observations in testing set 30 observations in training set and 58

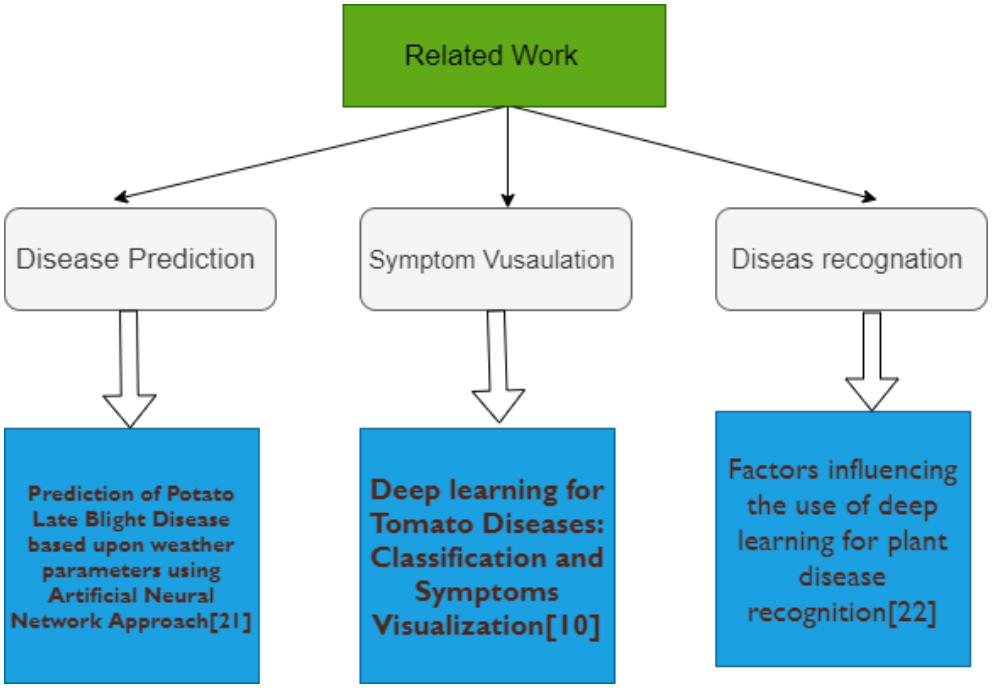


FIGURE 4.1 – Reviewed Related Work

observations in testing set.

The entire process of disease prediction can be broken down into several steps ranging from : i) data acquisition, ii)data normalization, iii) training to disease manifestation prediction.

Their prediction process acts in steps as follows :

— Step 1 : Collect weather data for disease prediction.

— Step 2 :Normalize the data if needed between values 0-1

— Step 3 :Create an ANN model

— Step 4 :Divide the dataset into training and testing data. The dataset is divided 3 times such that the training and testing dataset is as follows

— Step 5 :Choose the activation function-Sigmoid, TanH, and ReLu

— Step 6 :Training of network by using the dataset

— Step 7 :Prediction of manifestation of diseases. The output will be in form of 3 classes-didn't occur, occurred with less severity and occurred with high severity

— Step 8 :Evaluation of model for the given dataset

Output

The output is in form of three classes : a) 0 for disease did not occur b) 1 for disease occurred but with less severity c) 2 for disease occurred but with high severity.

Results

The model was tested for different activation function as well as different training and testing set. The maximum accuracy achieved is 90.909%. This model proves to be useful for prediction of Late Blight with the help of weather data and this make timely intervention so as to reduce the crop loss. Combining aspects of ANN and well established statistical tools may offer a more flexible option for the future.

4.2 Classification and Symptoms Visualization

The tomato product in Algeria is classified among the most widely consumed crops, and Algeria produces thousands of tons of it, but there are diseases that threaten this crop, and the state is making efforts to bypass these diseases by modern technological including deep learning.

In their work Brahimi[21] proposed an approach that uses Deep Learning for classifying plant diseases, focusing on images of leaves. This study presents two main contributions in plant disease classification : 1) Improvement in classification pipeline using deep models. and 2) Detection of disease symptoms in the infected leaf

The main phases of their approach are :

1. Pre-training phase : in this phase they train deep architectures on a large dataset like ImageNet using powerful machines. The objective of this phase is the initialization of network weights for the next phase.
2. Training (fine-tuning) : the fine-tune output network from the first phase using images of tomato leaves affected by nine diseases. At this stage, they used a small dataset compared to the pre-training dataset.
3. Disease classication : in this mode, the user takes a picture of a leaf and use the produced deep model to determine which disease affect tomato plant.
4. Symptoms detection and visualization : after the disease classification in the precedent mode, the user can visualize the regions that characterise the disease. This

symptoms visualization method helps the inexperienced user by giving them more information about the disease mechanism.

They used Pre-trained Models for building a classifier for plant diseases without hand-crafted features, we use deep learning approach, especially convolutional neural networks (CNN). CNNs are directly trained using raw images. As a result, the proposed system learns to extract features from data without the intervention of human in features engineering.

The deployed **Dataset** is published in goodfellow. This dataset is open access repository of images published online that contain more 50000 images for leaves. They extract only images of tomatoes leaves. Their experiments demonstrate the benefit of using pre-trained model, especially if the number of examples used in training is not very large, unlike that used in disease classification context.

4.3 Disease Recognition

Barbedo [22] has investigated the factors Influencing the Use of Deep Learning for plant Disease Recognition. In fact, he provides an in-depth analysis about the main factors that affect the performance of deep learning-based tools for plant disease recognition under realistic conditions. The goal was to provide some guidelines to make the investigation of deep learning-based methods for disease recognition more thorough and realistic.

The database used in the experiments is a subset of an $50,000$ images of 171 diseases affecting 21 plant species. It encompasses only images of corn diseases were used in the context of this work. He used some rules were applied for division consistency : a) images were manually blacked out prior to the subdivision ; b) healthy tissue occupied at least 20% of the cropping area ; c) isolated symptoms were taken individually ; d) clustered symptoms were taken as a group ; e) widespread symptoms were taken both as a whole and divided into homogeneous regions.

Transfer learning was applied to a pretrained CNN (GoogLeNet) using the Neural Network Toolbox provided by Matlab. At first, three different CNNs were trained : the first using the original unprocessed images, the second using whole images with background manually removed, the third using the subdivided images. In each case, 80% of the samples were used for training and 20% for validation. A fourth CNN was trained with a reduced

Dataset	Training Samples	Accuracy
Orginal	1,584	76%
Background removed	1,584	79%
Subdivided(full)	1,6005	89%
Subdivided(Reduced)	1,584	81%

TABLE 4.1 – accuracies obtained using CNNs training differnet datasets.

version of the training dataset containing subdivided images, so it would match the size of the original training dataset.

Training datasets were augmented using operations of rotation, mirroring, addition of Gaussian noise, brightness adjustment and contrast adjustment. As a result, the size of the training set was increased 12-fold and overfitting problems were reduced . The final accuracies for each network were obtained using a 10-fold cross-validation approach. The characteristics of the images that were misclassified by the CNNs were carefully analysed, and nine factors were identified as having the most impact on the results.

They reported that the relative differences between the four trained neural networks are important indicators of some of the main factors that affect the effectiveness of CNNs for plant disease recognition.

The accuracies abtained are between 79% and 89%.

work	Approach used	Accuracy
Disease Prediction[21]	ANN	82%
Symptom visualization[10]	CNN	84%
Disease recognition[22]	CNN	89%

TABLE 4.2 – summary of related work.

To summarize all the previous reviews that we've seen in this chapter we have to cover these important points : diseases prediction , symptom detection and we see the factors influencing the use of deep learning for Plant disease recognition. The table 4.2 summarize this works.

4.4 Conclusion

In this chapter we have reviewed some work that have dealt with the application of Machine learning and Deep learning as support for some smart agriculture applications mainly dedicated to and Symptom detection disease prediction. The lessons learned from this modest literature review numerous. First, the dataset are chosen to train the models. Furthermore, the Accuracy of such models used and compared all the models.

CHAPITRE 5

PROPOSED APPROACH

In this chapter, the proposed solution for disease prediction within its implementation and data-based forecasting is then described, along with the methodology and instructions for doing so.

5.1 Global Overview

Figure 3.1 illustrates the overall predict system.let us mention that the weather data are collected by means of Internet of Thing (IoT) network **??**. In fact in this dataset, the sensed data are essentially related to weather and agriculture parameters such Temperature, humidity, soil moister, pressure,wind. . ..

In order to predict plant disease ,we rely on machine learning based solution

The main Machine Learning methods investigated are *SVM* and *Random Forest*. We have projected to use *Deep Learning* unfortunately it was delayed until we gather enough real data to be more efficient.

These prediction models are narrowed on dataset collected from a farm in Laghouat city. These models aim to predict some diseases in order to protect crops and develop their production.

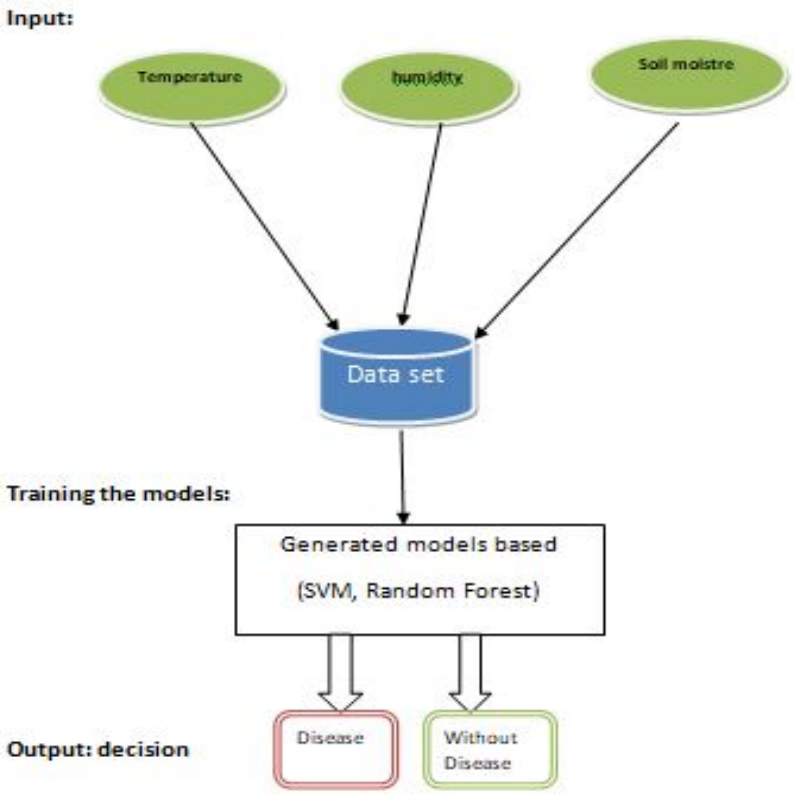


FIGURE 5.1 – Overview of the ML-Based Plant Disease Prediction System

5.2 Metrics

When evaluating the performance of a classifier, the used metrics are crucial. For instance the Confusion matrix is used to describe classifier prediction performance in two-class problems. By comparing the actual and expected values of all samples, The confusion matrix visualizes properly or erroneously predicted samples of each class. It's especially beneficial in supervised learning, when each sample has a target class assigned to it. In addition there are other important metrics which are described bellow.

Confusion Matrix

The confusion matrix is a specific table visualizing the performance of the classifier. Usually, in the field of machine learning a confusion matrix is refer to as the error matrix. An image region is said to be positive or negative, depending on the data type. Furthermore, a decision for the detected result can be either correct (true) or incorrect (false).

Therefore, the decision will be one of four possible categories : true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The correct decision is the diagonal of the confusion matrix [23].

— **True Positives (TP) :** These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also Yes.

— **True Negatives (TN) :**These are the correctly predicted negative values which means that the value of actual class is no and the value of predicted class is also No.

False positives (FP) and False negative (FN) are occur when our actual class contradicts with the predicted class.

— **False Positives(FB) :** When actual class is no and the predicted class is yes.

tem **False Negatives (FN) :** When actual class is yes but the predicted class is No.

Accuracy

The most common statistic used in data mining is accuracy. It's calculated as the percentage of accurate decisions made out of the total number of decisions made. However, accuracy does not always work in every situation. When there is an imbalance in class, accuracy is no longer a good measure since it may be biased toward the majority class and does not reflect minority class achievement. By classifying all data as negative cases, a classifier can achieve high accuracy.

$$Accuracy = \frac{TP + Tn}{TP + FP + FN + TN}$$

5.2.1 Precision

Precision is defined as the proportion of positive cases accurately identified in outcomes (also known as Positive Predictive Value).

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

Recall

The ratio of true observations that the model will classify as true is known as recall, sensitivity, or true positive rate (TPR). The recall values range from 0 to 1, indicating no

true positives and no false negatives. Low recall indicates a high number of false negatives.

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

$F1$ score

$F1$ score is defined as weighted average of Precision and Recall.

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

5.3 Dataset Collection and Annotation

As mentioned bellow, We are collecting the data by an IoT network and store them into an *CSV* file which serves as input in our application to train our model for prepare it to the prediction.

The parameters captured by the IoT are namely : temperature, humidity and soil moister in a farm at Laghouat. For the other parameters (Wind Speed, Rain), we have acquired them from the weather local station (N24) at Laghouat.

For the annotation task, we have been assisted by an agriculture expert who give us some information about the targeted disease namely :

— **Beatrice** : this is disease is caused mainly by the combination of three weather parameters : wind of sand and high temperature and humidity. In fact, When a wind of sand fights intensively it causes some wounds on the leaves and legs of the plant. If simultaneously a high temperature with intense humidity occur this leads to appearance of rot on the wounds. This fact affects negatively the plant.

— **Meldio** disease destroys the crop. It is caused by the fungus Phytophthora infestans. It Attacks all the top parts of the plant (leaves, stems, flowers and fruits). It can cause significant damage that could completely damage the crop. Caused by high temperature and humidity and usually appears in Spring and infects potatoes, tomatoes and onions etc.

	Shape	Size
Data	(1025,5 line)	65 KO

TABLE 5.1 – Information about the Datasets.

This table show details about the dataset used Figure 5.2 shows a sample of our dataset with the related manual annotation.

1	temperature	humidity	soil-moistur	ventsable	target
2	25.7000008	40	0	0	0
3	25.7000008	39	0	1	1
4	25.7000008	38	0	0	0
5	25.7000008	33	0	1	1
6	25.7000008	42	0	0	0
7	25.7000008	45	0	0	1

FIGURE 5.2 – A Sample of the Data set

5.3.1 Pre-processing

In order to model the phenomenon described above, we must prepare the data so it will be able to be introduced into the Machine Learning model that will be described below. The input data of the model will be transformed from their original values to values which varies between 0 and 1, and this to allow the model to converge. the output values that are the cell identifier for each observation, will be prepared in another way, due to the fact that the model is a classification model. The output values are tow classes 0 to no disease 1 to disease.

5.4 Experiments

In this section, all methods and tools used to make the planet disease prediction are being reported and discussed. At the beginning, tools that are utilized in the research are described in details. After that, the implementation process started with data generation, prepossessing and generated models are presented with a short summary at the end of the section.

5.4.1 Used Tools

The used tools are mainly :

Google Colab

Colaboratory, or Colab for short, is a product of Google Research. It allows anyone to write and run any Python code from a browser, and is especially useful for machine learning, data analysis, and education. Technically, Colab is a hosted Jupyter notebook service that requires no setup and provides free access to computing resources, including GPUs. Colab is used extensively in the machine learning community with applications including [1] :

— Developing and training neural networks

— Experimenting with TPU

— Disseminating AI research

Python Python is an interpreter, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development and used as a scripting or glue language to connect existing components. Python's simple, easy-to-learn syntax emphasizes readability and, therefore, reduces program maintenance costs. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed [2]

Python libraries

The libraries used in our work :

— **Pandas** is a library for working with and manipulating tabular style data. In many

1. https://colab.researchgoogle.com(seenon2022-05-01)
2. https://www.python.org/doc/essayes/blurb(seen on 2022-04-30)

ways, you can think of it as a replacement for a spreadsheet, only it is much more powerful.

— **Numpy** is a python library, intended to handle multidimensional matrices or arrays and mathematical functions.

— **Matplotlib** is a python library, it is used for plotting and visualizing data as a graph.

— **SKlearn**

Sklearn is the most usable and resilient machine learning library in Python. It offers a variety of fast tools for machine learning and statistical modeling, including as classification, regression, clustering, and dimensional reduction, through a Python consistency interface.

— **Seaborn** SeaBorn is a matplotlib-based Python data visualization package. It has a high-level interface for creating visually appealing and instructive statistics visuals.

5.4.2 Implementation

Hereafter, some skech of our implementation are reported

Importing the Dataset

Import the dataset as CSV file and convert it to Dataframe describing in this figure .

FIGURE 5.3 – Import the dataset script

Creating the dataset testing and training and splitting the dataset into 20% for testing and 80% for training.

Machine Learning Models

We invoke the machine learning models to train our data to prepare them to the prediction and measure the accuracy and compare their performances.

The Random forest models the following script shows the random forest model training after a tuning phase

```
[9] from sklearn.ensemble import RandomForestClassifier
    rf=RandomForestClassifier(n_estimators=3,random_state=1)
    rf.fit(X_train.T,y_train.T)

    acc_test=rf.score(X_test.T,y_test.T)*100
    acc=rf.score(X_train.T,y_train.T)*100
    accuracies['Random Forest']=acc
    print("Random Forest Algorithm Accuracy Score :{:.2f}%".format(acc))
```

FIGURE 5.4 – Defining the Forest Model

Support vector machine models

Hereafter, we are defining the SVM model. A tuning phase is first invoked. In fact we have calculated the hyper parameters *Gamma* and *C* value by using the *grid searchCV* algorithm to choose the best values. The following Figure reports the related script.

```
from sklearn.svm import *
from sklearn.model_selection import GridSearchCV

# defining parameter range
param_grid = {'C': [0.1, 1, 10, 100, 1000],
      'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
      'kernel': ['rbf']}

grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3)
# fitting the model for grid search
grid.fit(X_train.T, y_train)
# print best parameter after tuning
print(grid.best_params_)
```

FIGURE 5.5 – SVM model script

Predicting Model We deploy this generated model by giving the test values captured in the farm (temperature, humidity, wind) to predict the targeted disease is present or not. This scripted by the following code.

```
y_head_rf=rf.predict(x)
print(y_head_rf)
```

```
[0 1 0 1 1 1 1 0 1 0 0 1 1 0 0 0 1 1 1 1 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 0 0
```

FIGURE 5.6 – Test of the predicting model Script

5.4.3 Results

In this section, we summarize some results of the models used in our approach. The we compare the models on the gathered datasets in the Figure bellow we compare the accuracy of the tow models.

The reached results in term of accuracy for both predicting model are promising more than 80%. However the SVM-based model beats the Random Forest-based one.

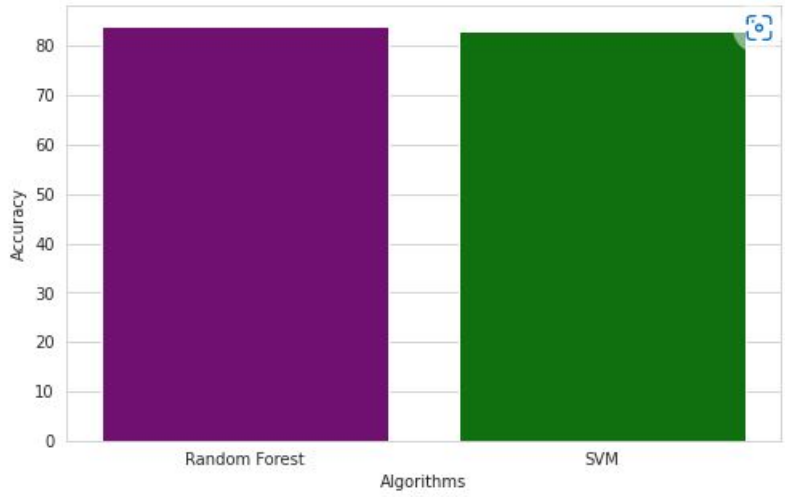


FIGURE 5.7 – Performances of the Generated Models in term of accuracy

Table 5.2 reports the performances in term of the four used metrics. These results are also very suitable for the problem of disease prediction based on the used forecasting

Models	SVM (%)	Random Forest (%)
Accuracy	83	84
Recall	66	59
Precision	71	60
F1 Score	68	59

TABLE 5.2 – SVM-based model Vs Random Forest model performances.

dataset. We can notice that Forest reaches a highest accuracy (84%) comparing to SVM, which had (83%) accuracy, Meanwhile, the recall ,precision and F1 score with 66% , 71%,68% respectively are highest then Recall(59%),Precision(60%)and F1 score (59%) of the Forest One can ask why get this result is due to two facts. First the size of the dataset is the modest. Second, the annotations are done manually according to an empirical thresholds.

5.5 Conclusion

In this chapter, We have described the proposed approach, its implementation and the experimental phase. In addition, we have demonstrated that our Machine learning based models using the SVM and Random Forest and the dataset collected by the IoT network in the farm are very suitable for the prediction of diseases. In fact, we get a prediction with 83% and 84% accuracy by using SVM and Random Forest respectively to detect disease.

CONCLUSION AND PERSPECTIVES

In this work we have investigated using Machine learning based approach to reach a Smart-agriculture application. In fact, Plant diseases adversely affected the quality and productivity of the Algerian crop but by collecting data weather values by Inter of Things and applying machine learning algorithms we reached some promising result in order to predict some of the fungal diseases caused by weather change.

In this project we collected information and created dataset thanks to an IoT network and the Meteorological Center in the city of Laghouat. The annotation are performed manually thanks to a farming expert who gave us information about some fungal diseases and their correlation with weather data. This helps us to be able to train both ML model (Random forest and SVM).

This smart-agriculture application is performed to make farmers aware of the need to prevent them and avoid their dangers, especially plants that are not resistant such as tomatoes, potatoes and grapes.

This project is in its beginnings. It can be promoted using perspectives :

— **Image extension** : The user can predict the type of disease that affects his crop by adding another deep learning branch. It can take a picture of a plant paper and the disease is predicted through images of symptoms stored in the database.

— In addition, we can add a camera robot that can automatically inspect the farm by using computing Vision and CNN.

BIBLIOGRAPHIE

[1] type of regression | Aidoc Blog. https://www.javatpoint.com/regression-analysis-in-machine-learning/(seen on 2022-02-10).

[2] Maha Sharkas Stephen Marshall Dina A. Ragab1, 2 and Jinchang Ren. The ingredients of machine learning. *Applied Artificial Intelligence*, 2017.

[3] Mrs. Soja Rani Swathi. Prediction of crop yield and cost by finding best accuracy using machine learning approach. *Applied Artificial Intelligence*, page 53, 2021.

[4] Support Vector Machine. https://academy.hsoub.com/questions-support-vector-machines/(seen on 2022-0212).

[5] Classification of machine learning algorithms . https://www.engineersgarage.com/machine-learning-algorithms-classification/(seen on 2022-02-15).

[6] Schematic of showing the principle of artificial neural network. | Download Scientific Diagram. https://www.researchgate.net/figure/Schematic-of-showing-the-principle-of-artificial-neural-network_fig3_347821040(seen on 2022-03-10).

[7] Deep Learning in Healthcare and Radiology | Aidoc Blog. https://www.aidoc.com/blog/deep-learning-in-healthcare/(seen on 2022-03-10).

[8] plant disease - definitions of plant disease | britannica. https://www.britannica.com/science/plant-disease/Definitions-of-plant-disease(seen on 2022-02-10).

[9] Istvan Farkas. *Greenhouse, Climate Control*, pages 338–339. Springer Netherlands, Dordrecht, 20. https://ondo.io/company/news/011.

[10] Netafim. Precision irrigation. https://www.netafim.com/en/precision-Irrigation/#:~:text=Precision20irrigation.

[11] Plant health problem. https://portal.ct.gov/CAES/Fact-Sheets/Plant-Pathology/Plant-Health-Problems (seen on 2022-02-10).

[12] What is machine learning ? | Definition, types, and examples | SAP Insights. https://www.sap.com/insights/what-is-machine-learning.html(seen on 2022-02-10).

[13] Why is Visible Learning more effective ? https://www.tutorialspoint.com/what-is-supervised-learning(seen on 2022-02-10).

[14] Clustering in Unsupervised Machine Learning | Engineering Education (EngEd) Program | Section. https://www.section.io/engineering-education/clustering-in-unsupervised-ml/(seen on 2022-03-10).

[15] Logistic Regression in Machine Learning - Javatpoint. https://www.javatpoint.com/logistic-regression-in-machine-learning(seen on 2022-03-10).

[16] What is Reinforcement Learning ? – Overview of How it Works | Synopsys. https://www.synopsys.com/ai/what-is-reinforcement-learning.html(seen on 2022-03-10).

[17] SVM | Support Vector Machine Algorithm in Machine Learning. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/(seen on 2022-03-10).

[18] Decision Tree Definition | DeepAI. https://deepai.org/machine-learning-glossary-and-terms/decision-tree(seen on 2022-03-10).

[19] Understanding Random Forest. How the Algorithm Works and Why it Is... | by Tony Yiu | Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2(seen on 2022-03-10).

[20] Priyanka Sharma, BK Singh, and RP Singh. Prediction of potato late blight disease based upon weather parameters using artificial neural network approach. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–13. IEEE, 2018.

[21] Mohammed Brahimi, Kamel Boukhalfa, and Abdelouahab Moussaoui. Deep learning for tomato diseases : Classification and symptoms visualization. *Applied Artificial Intelligence*, 31(4) :299–315, 2017.

[22] Jayme G.A. Barbedo. Factors influencing the use of deep learning for plant disease recognition. *Biosystems Engineering*, 172 :84–91, 2018.

[23] Maha Sharkas Stephen Marshall Dina A. Ragab1, 2 and Jinchang Ren. Breast cancer detection using deep convolutional neural networks and support vector machines. *Applied Artificial Intelligence*, page 9, 2017.