A GUAVA LEAF DISEASE DETECTION BY MACHINE LEARNING

BY

MD. RADOANUL HAQUE
ID: 172-15-9878

SAMIUL ISLAM
ID: 172-15-9738

NISHAT ANJUM MAMATA
ID: 172-15-9640

This Report Presented in Partial Fulfillment of the Requirements for

The Degree of Bachelor of Science in Computer Science and Engineering

Supervised By
Dr. Md. Ismail Jabiullah

Professor

Department of CSE

Daffodil International University

Co Supervised By

Assistant Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

1 JUNE 2021

APPROVAL

This project titled "**GUAVA LEAF DISEASE DETECTION BY MACHINE LEARNING**.", submitted by **MD. RADOANUL HAQUE, SAMIUL ISLAM, and NISHAT ANJUM MAMATA** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as suitable for the partial completion of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents. The presentation has been held on December 2020.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Azmoon

Nazmun Nessa Moon **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Aniruddha Rakshit

Aniruddha Rakshit **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this thesis has been done by us under the supervision of **Dr. Md. Ismail Jabiullah, Professor Lecturer (Senior Scale), Department of CSE,** and co-supervision of **Ahmed Al Maruf, Sr. Lecturer, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Md. Ismail Jabiullah
Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Ahmed Al Marouf
Sr. Lecturer Department of CSE
Daffodil International University

Submitted by:

Radoanul Haque

Md. Radoanul Haque
ID: 172-15-9878
Department of CSE
Daffodil International University

Samiul Islam

Samiul Islam
ID: 172-15-9738
Department of CSE
Daffodil International University

Nishat Anjum

Nishat Anjum Mamata
ID: 172-15-9640
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing makes it possible to complete the final thesis successfully.

We really grateful and wish profound gratitude to Dr. Md. Ismail Jabiullah , Lecture Dept. of Computer Science and Engineering, Daffodil International University, Dhaka. Deep knowledge & keen interest of our supervisor in the field of “Machine Learning and Deep Learning “to carry out this thesis. His endless patience, academic guidance, constant back-up, constant and energetic supervision, positive analysis, important advice, reading many lower draft and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Prof. Dr. Touhid Bhuiyan** , Professor, and Head, Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

ABSTRACT

Fruit diagnosis and early identification The production of healthy fruit industry is more critical for plant diseases. Farmers' general monitoring system can take time, costly and often incorrect. This paper offers an overview of target recognition through grouping of numerous images and machine learning methods for the guava leaf disease detection. Our system has been developed based on machine learning algorithm. In this work rust, white fly, leaf spot and sound disease has been detected.

For the whole method, a number of machine learning programs (MLs) were used, such as Scikit-learn, Pandas, Matploatlib, Numpy. In the pre-processing of images, we have also used Scikit-learn to implement algorithms. In order to check the validity of our work we use five separate K-Nearest Neighbour(KNN), Vector Support (SVM), Tree Classifier Decisions and the Random Forest. Naive Bayes. The most effective algorithm. This five algorithms were studied. Finally, this high-precision algorithm detects guava leaf disease.

TABLE OF CONTENTS

CONTENTS **PAGE NO**

CHAPTER

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

In tropical and subtropical regions of the world, Guava is one of the most common fruit. It was found that Guava has an advantage for people with Asthma, hypertension, oral ulcers, scurvy, congestion in the lungs, bacterial infections etc. In order to meet the minimum daily needs of its people Bangladesh produces less than 30 percent of the fruit required. Any 80% of families in the world eat fewer fruits than the minimum daily requirement that is recommended. As a result, the pervasive dietary shortcomings of vitamins A and C, and other nutrients, weaken humans. The condition can be improved by cultivating guava fruit trees and the cultivation of guava that can meet the needs of Bangladeshi residents. [1] The object detection of these systems shifts from customer-server technologies to user classification techniques, where cameras are used for input/output front ends. Once a computer capability is created for a camera system. That is why we have developed an information system that identifies the object automatically and defines the guava leaf disease. ML allows the machine to learn and use artificial intelligence (artificial intelligence) automatically to maximize knowledge execution without special programming. We describe the image with the ML method in the image data set to construct our model. ML libraries have been used including Pandas, Matploatlib, Numpy, Pandas, Numpy, Matplotlib, Tensorflow, Pytorch, etc. It also uses the ML library. In the pre-processing of the image we use TensorFlow and Scikit-learn. We used five common classifiers and in-depth neural network study algorithms to assess the validity of our work. In order to determine the highest precision algorithm, the six algorithms have been tested. Initially for our model, we cover two categories to identify and classify the image. We have used the best precision generated algorithm for image classification to perfectly identify the Guava Leaf Disease. On average, our definition can be helpful for farmer.

1.2 Motivation

Our beautiful world is the ecosystem for all kinds of people. Any of these are born with the blessing that they can all have their attributes and virtues. Guava is one of the favorite fruit of people of Bangladesh. There are lots of guava garden available in our country. As a result, production is also so much in our country. But the problem is when guava leaf is affected by any virus or disease this effect is also flow to the fruit. As a result, yield is decreased. If it is possible to predict guava leaf disease in early stage, then the amount of losses will be reduced. This phenomenon motivated us to do our work.

1.3 Problem Definition

Problems need to be described in the original state of the system and boundary conditions that describe physical situation at the boundaries with mathematical restrictions. The main boundary of our work is data collection. Because of pandemic situation we were unable to more data. we collected only 200 data for our research. After collected our data we need to convert image to pixel array for this purpose we used CV2 library. Then we store all data as pickle file. For training we used 200 data. In our work color is a very important factor because when guava leaf affected by any virus or disease then it will be yellow color. And sound leaf color in green. So we work with color full image that means color channel is 3(red, green, blue). After data preprocessing we used Machine Learning algorithm because our dataset is small, so we did not use deep learning algorithm. Then we calculate accuracy of each algorithm and we selected best algorithm that is produced highest accuracy.

1.4 Research Questions

- Can the program detect the image of diseased leaf?
- How detecting image can be helpful to farmer?
- Can the detection accurately define the image?
- Will the online model be deployed?
- How do people assist with this work?

1.5 Research Methodology

In this section we will discuss the collection of test results, data pre-processing, classification, algorithm section, algorithm implementation and evaluation of algorithms. The effectiveness of the model is discussed at the end of this chapter.

1.6 Research Objectives

The main objective of our research is to help farmer, who cultivate guava. Guava leaves disease is a very common phenomenon in our country. For this guava yield rate decreased. So we tried to build an intelligence system that can help farmer to detect guava leaves disease quickly and more effectively. As a result, they can save their time and money. In future we will make an android app that can automatically detect the guava leaf disease.

1.7 Research Layout

The report will have appeared as regards:

Chapter 1 Give this study a review. This first part is an important step in the initial analysis. In addition, the following chapter outlines the motivations for us to do such analysis. The definition of the problem is the most important aspect of this chapter. This segment is also included in the issue of research and the challenge.

Chapter 2 Consisting of the context analysis, give the relevant work in this area a brief overview. Notable work with machine learning is mentioned here, in particular prediction work

Chapter 3 This chapter enhances the mathematical techniques of this work. This is the analysis methodology; Furthermore, this chapter illustrates Machine Learning procedural approaches.

Chapter 4 Is about the results assessment. The results of the analysis contained in the table. s related to the evaluation of the result. That means it refers to the result or outcome of the research

Chapter 5 Part of the investigation's ending. The model output is seen in this section. This segment also demonstrates the exactness of the relation. The web implementation component of the model and performance is also included in this section. The chapter concludes by examining the work's shortcomings. The future work has been encoded.

1.8 Expected Outcome

- Our system will help the farmer
- Farmer’s time and cost will be reduced
- Easy to use for detecting guava leaf disease
- We will make an intelligence system that also can suggest medicine

1.9 Summary

In this chapter we discussed about introduction of our work. We discussed research objective, motivation. The main objective of our work to help the farmer. As they faced different problem for cultivation guava we motivated to mitigate this problem.

CHAPTER 2

BACKGROUND

2.1 Introduction

Computer education for prediction was conducted in several trials. Prediction is one of Machine Learning's most recurrent applications. There have been enormous experiments to forecast the of leaf disease of different tree to address the corresponding solution. This experiments concentrated on problems and used various machine learning algorithms to solve the problem. This chapter provides an overview of the related work carried out efficiently by some specialists in the aforesaid sector.

2.2 Related Works

For the artist, type, material and year of creation of the artwork, T. Mensink et. al. [2] suggested a detection scheme. The 112,030 pictures depicted in the public exhibition of artworks of the Rijksmuseum in Amsterdam, Netherlands. The accuracy of the device was measured according to the accuracy of the weighted frequency. For the calculation of categorization of artworks, the Main Average Precision (MAP) algorithm has been used. Accuracy for the labeling of materials was also evaluated with Chart. You have encoded the images with modern Fisher vectors (FV). They used 70 percent data from the data collection for classification and regression models.

O. Bimber et. al. [3] have identified its handheld museum guidance device Phone Guide. The methodology was studied in order to attain theoretically feasible identification rates under realistic conditions using an adaptive picture classification. The captured images derived global color characteristics for object detection and used a 3-layer neural network. Once the closest neighbor process is used to identify different key frames, data base photos contain intermediate relationships close to the key frames. The job of the customer was to update the mobile image classification system with 93.47 percent accuracy.

J. S. Hare et. al. [4] A new technique has been developed to capture content-based image data and recognize images objects that have been noisily corrupted and converted through the imagery system. They also used a space vector model to efficiently index any image. After that, a two-step rating procedure is used to calculate the correct picture. The algorithm is particularly immune from variations in queries and images from the database. Lower Invariant Scale Transform function (SIFT) Descriptor13 is used in implementation of this procedure. The pictures are reconstructed into a vector on the basis of the frequencies of "visual" terms in the image. 850 images from Nat image collections are used in data sets. There were 850 images from the national museum's photo collections. 200 randomly chosen pictures were used to verify the computer.

T.-Y. Lin et. at. [5] The authors have developed feature pyramids, using the multiscale hierarchy of deep convolutional networks at a marginal extra cost A high-level semantic feature map has been developed on all levels using a top-down architecture with side-links. The architecture of the Pyramid Network (FPN) has significantly improved with the feature extractor. Without bells and whistles, the COCO recognition benchmark obtained the high-end single model data. The system is based on a low-resolution design, semanthropically influential components, semanthropic features, top-down path and side connections.

To use fast pyramids for work, P. Dollár et. al. [6] Three different visual recognition technologies have been updated by Mehedi. For identifying the footpath and objects, the results were checked. Their technique is broadly applicable to multiscale analysis-including vision algorithms. The Caltech, INRIA, TUD-Brussels and ETH databases and PASCAL VOC are their strategies. You can measure them on broad spectrum images, but it does not work on images from the narrow band of spectrum. They also showed the ability to identify the Aggregated Channel Functionality in football (ACF). ACF uses a consistent scale curve, histogram and LUV color channel.

C. Szegedy et. al. [7] Tried to distinguish objects accurately by using the strength of the DNNs. They also designed a method for building a binary mask of the bounding object. DNN-subject to regression. You demonstrate that the DNN regression can learn classification features and gather geometric data. After removing masks from multiple items in one set with DNN regression, a small group of huge subwindows would be fitted with a DNN tracker. It uses one network to forecast four half of the case in the object box mask and four networks. The method has been validated with 5000 images in 20 classes in the Pascal Visual Object challenge (VOC) 2007 dataset. Test- and assessment system c of VOC2012

X. F. Hermida et. al. [8] Many methods of image processing, including adaptive threshold detection and angle detection were used in the proposed Braille text recognition system using Optical Character Recognizing Philosophy (OCR). The areas with two adaptive thresholds have been transformed by a method in black and white regions. To calculate the limits, the luminance histogram was used. The machine also observed falsified scores and lost points.

G. Sainarayanan et. al. [9] aimed at developing a technology capable of helping blind people by the detection of obstacles. The visual sensor is present in the universe and the system is made up of an SBPS, a headgear vision sensor, and a pair of stereo earphones. It is then evaluated with fuzzy clustering algorithms using an image treatment scheme in real time. The picture then becomes a detailed organized stereo-acoustic pattern and is converted into the stereo earphones. An object recognition from the background is checked with a FUZZI LVQ (FLVQ).

The proposal to automatically recognize jackfruit from digital images by Md Tarek Habib et al. [10]. Most of them were 3 kinds of pests, red Rhizopus, pink and leaf fungus. As a dataset, they used 480 color images. Bicubic interpolation to resize an image to a default height. Histogram equalization helped to stretch the comparison between pictures in segmentation. Segmenting using the k-means clustering approach to distinguish disease-free components from disease-affected components. Finally, the nine highly pronounced classifiers have been classified. There are logistics, SVMs, Random Tree, RIPPRR, KNN, Naive Bayes, BPN, and CPN. These are logistics regressions. There are logistics, SVMs, Random Tree, RIPPRR, KNN, Naive Bayes, BPN, and CPN. These are logistics regressions. Random forest has the maximum accuracy of 92.5% and KNN has the lowest accuracy of 70.42% of these classificatory.

2.3 Research Summary

The above analysis in various research teams shows that every day research on the processing of images has been carried out in recent times. There were also some positive effects on this argument. Although there are not enough funds, it is hopeful that this area will become more resourceful every day after a single day.

2.4 Challenges

The data sets are the main barriers to this function. We need efficient inputs but not enough proven methods to optimize them. Another challenge is to have insufficient services in this region.

3.5 Summary

In this chapter we discussed about related researches. Prediction is one of the most recurring applications for machine learning. Huge tests have been conducted to anticipate leaf diseases of different trees to address the appropriate remedy.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The methodology relates to the study project's overall approach and reasoning. It includes learning the techniques and theory or concepts used in your field to create a strategy that suits your goals. The approach chapter can clarify design decisions by showing that the chosen approaches and techniques are best suited to research purposes and objectives and can yield genuine and reliable results. We have followed a 7-stage module to construct our platform. Figure 3.1 shows our module phase.

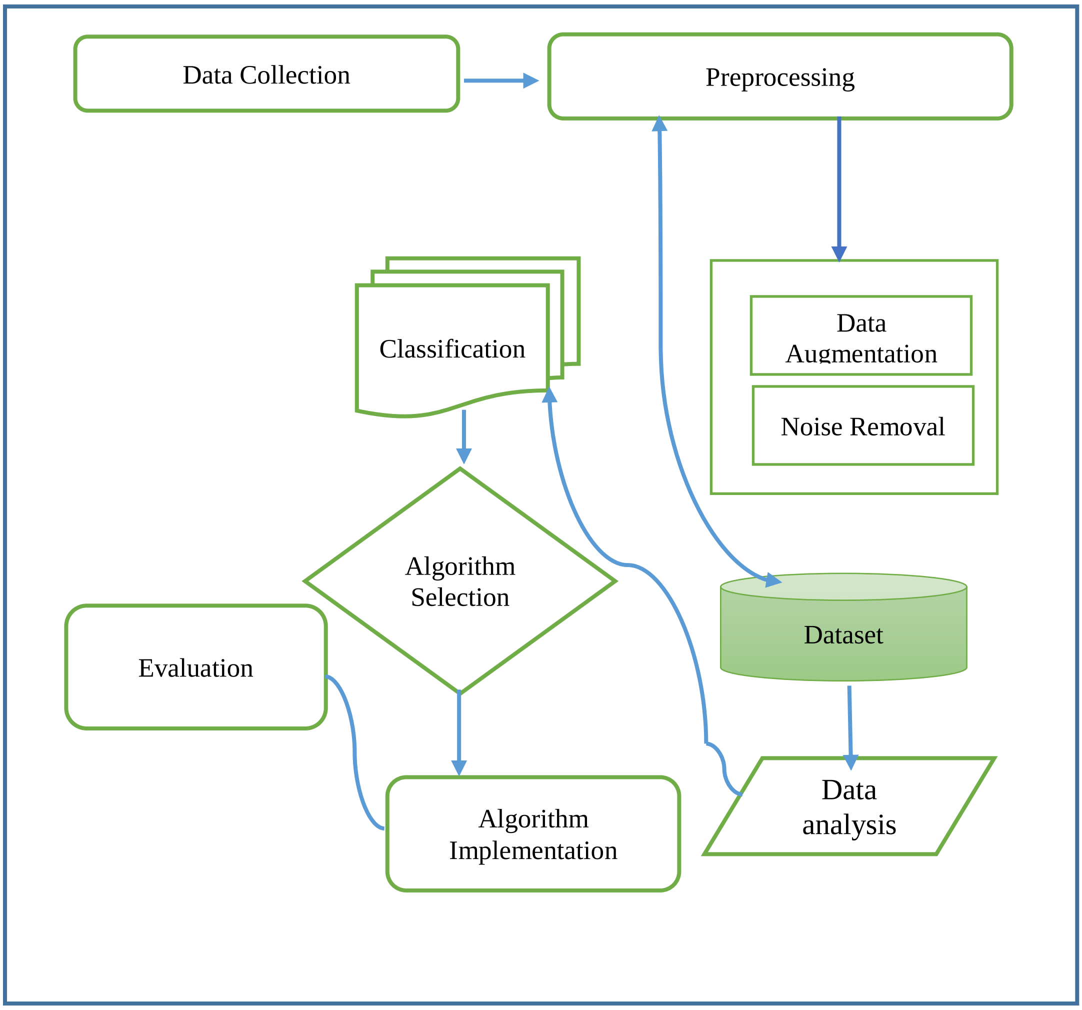


Figure 3.1: Methodology diagram

3.2 Data collection:

A dataset consists of a selection of raw statistics and analysis material. We collected our required dataset from guava garden. But due to corona pandemic we could not visit more garden. So we collected only 200 original picture.

3.3 Pre-Processing

For various machine learning algorithms, the process of converting raw data into pixel array formats may be used. For the preparation of our model, we used Tensor Flow Keras. It is a library for free and open source data flow and differentiable programming. Our data set was divided into two sections. In the first part, our model was educated. About 400 photographs were taken. The second part included 129 pictures for testing purposes. We processed them in full color pictures before we translate all the raw images into 224 pixels in height and width.

3.4 Dataset

We must transform the original when building our dataset. The picture was then transformed into the pixel array. We stored this pixel array into pickle file.

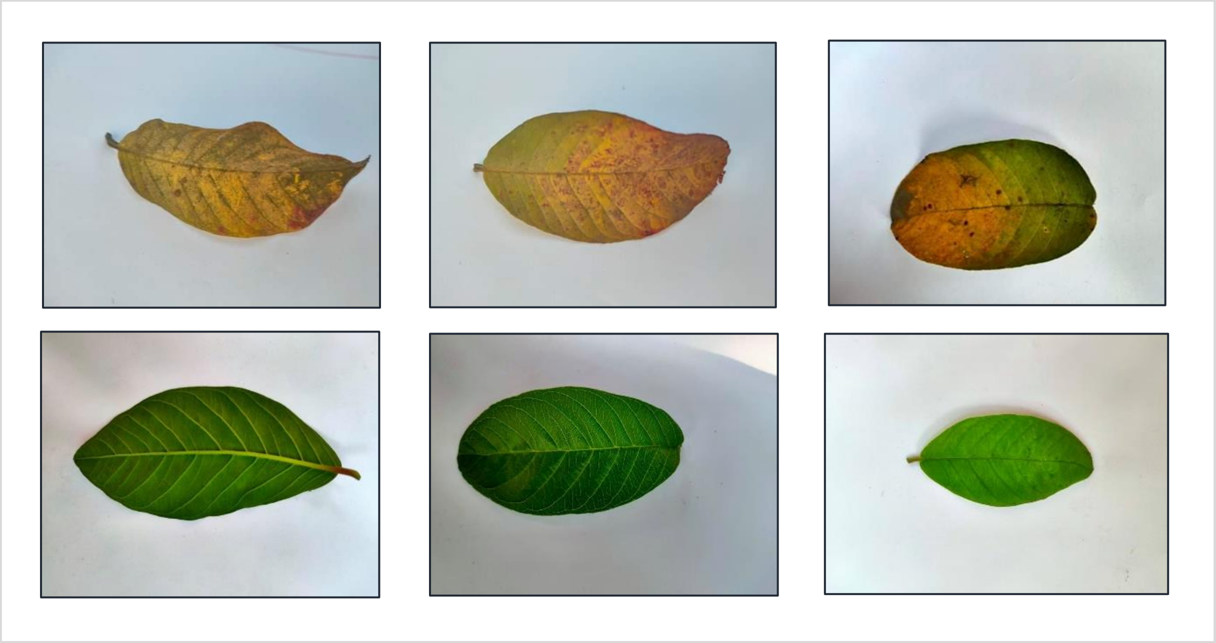


Figure 3.2 Sample of dataset

Figure 3.3 represents the sample of our dataset. our dataset first columns represent the white fly leaf spot and rust disease and second row represents the sound leaf. We can see that leaf spot diseased leaf is yellow colored whitefly yellow and black, rust is also yellow and green wand sound leaf is only green color. That means color is an important fact in our work. we described the color effect in next section.

3.5 Data Analysis

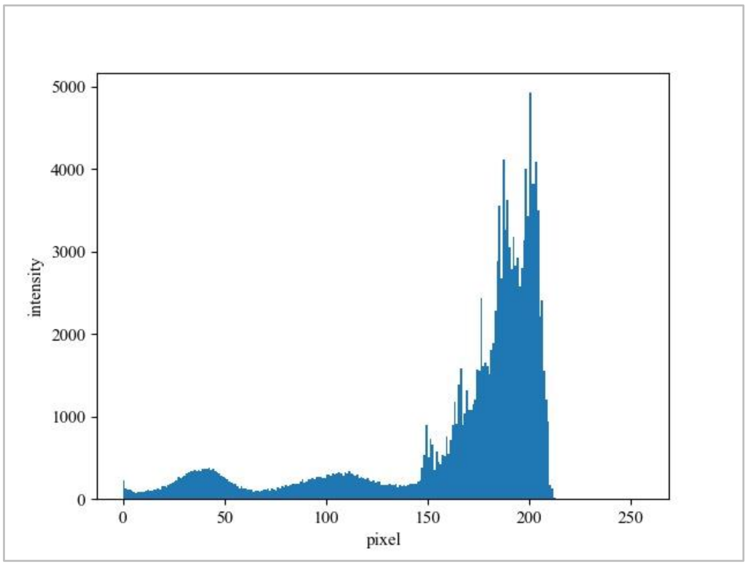


Figure 3.3. histogram of leafspot disease

Data analysis is very important part of any research. By data analysis we can found internal meaning of any dataset.[17] figure 3.3 represents the histogram of leafspot disease. Here Y axis represents the intensity of pixels and X axis represents the pixel. We said that color is an important thing in our research. Form this pixel vs intensitiy graph we can see that most number of pixel is 150 to 200 out of 0 to 255.

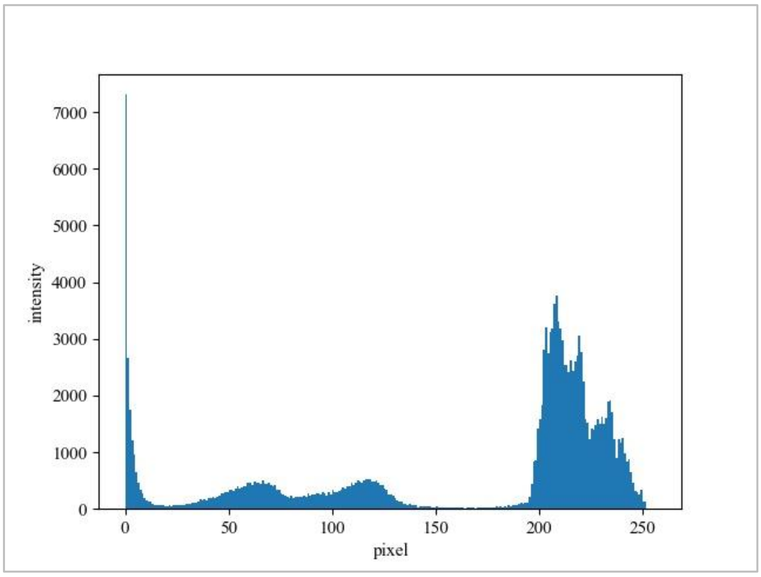


Figure 3.4. Histogram of sound leaf.

Figure 3.4 represents the histogram of sound leaf. Same as figure 3.3 X axis represents the pixels of image and Y axis represents the Intensity of pixel. Actually we used white color background for both images. That’s why the intensity of pixel number 0 to 150 are close to each other. But we noticed that from pixel number 151 to 255 are different of two type of images. This pixel vs intensity figure we can see that for sound leaf pixel number 200 to 255 are high out of 0 to 255. On the other hand, most number of pixel is 150 to 200 out of 0 to 255 for leaf spot disease.

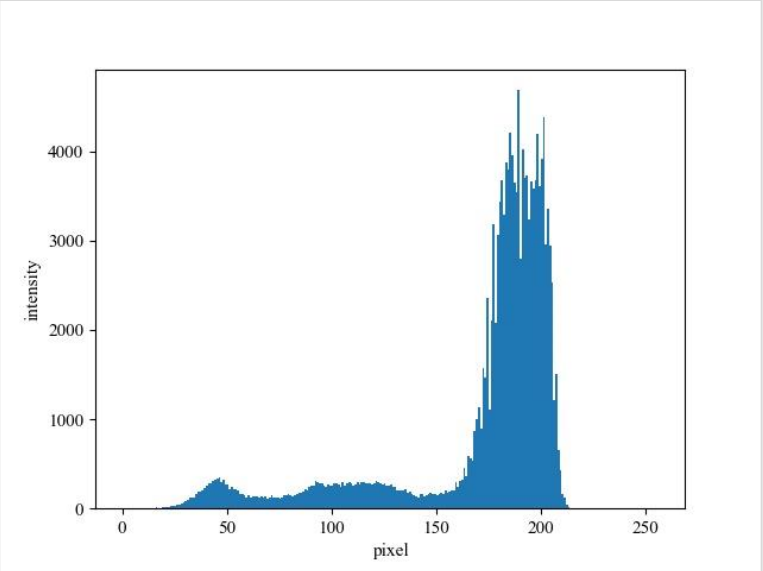


Figure 3.5. Histogram of white fly disease.

Figure 3.5 represents the histogram of white fly disease. Here Y axis represents the intensity of pixels and X axis represents the pixel. We said that color is an important thing in our research. Form this pixel vs intensitiy graph we can see from this graph like leaf spot disease the most pixel is exist on 150 to200 but in this case some pixel is competetevily less then leaf sopt disease.

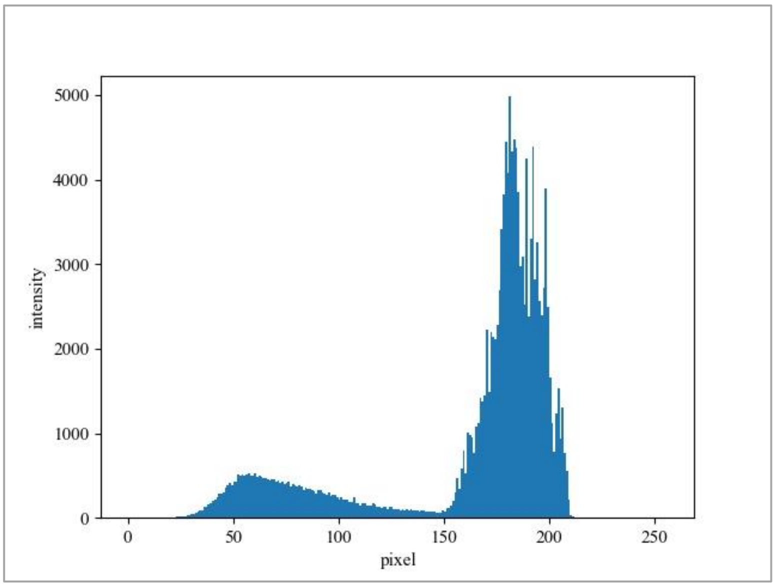


Figure 3.6. Histogram of rust disease.

Fgure 3.6 represents the histogram of rust disease. Here Y axis represents the intensity of pixels and X axis represents the pixel. Form this pixel vs intensitiy graph we can see that most number of plexel is exit on 150 to 180. From pixel analysis we can express that leaf sopt, white fly, and rust disease contains more pixel between 150 to 255.

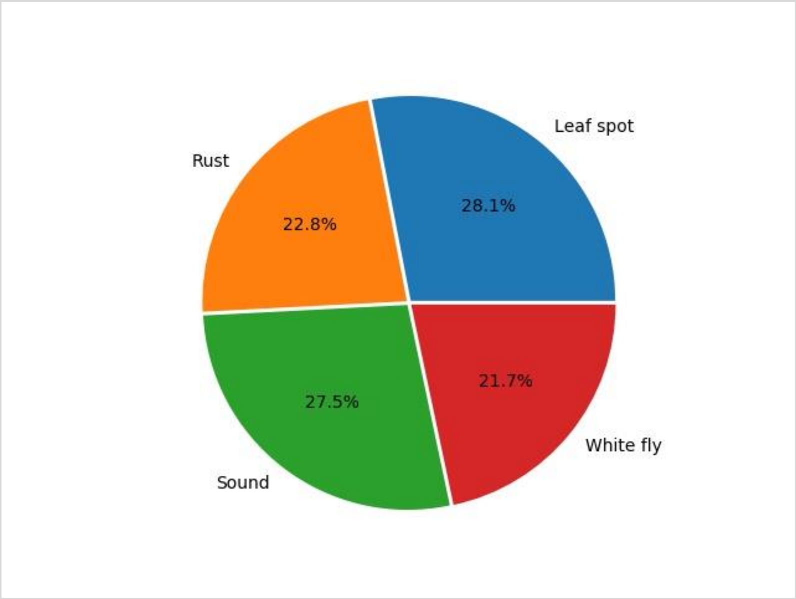


Figure 3.7. Classification of dataset

Figure 3.7 represents the classification of our dataset. Our work is a binaray classification. The four classes are Leaf Spot, Rust, Sound, White Fly. Our total dataset contains 549 images. Leaf spot contains 28.1%, Rust contains 22.8%, Sound contains 27.5% and white fly contains 21.7%. We tried to keep all amount closely, as a result all percentage is between 20 to 30%

3.6 Algorithm Selection

The right classification of the images depends on our model. In this case, we have selected five mainstream classification algorithms. All the algorithms which we used were KNN, SVM, Random Forest, Naive Bayes and the decision book. We checked all output algorithms for the best in our model training.

3.7 Algorithm Implementation

Table 3.1 represents the parameter usage of different algorithm. We used different parameter to implement each model. for KNN we use number of neighbor 2 and random state is 0. For Decision Tree we also use random state 0 for SVM algorithm we use linear kernel and random state 42. For Naïve Bays we use GaussianNB. We use for Random Forest algorithm we use 100 n estimator. We implement our model by using this parameter. We choose best parameter value that is produced highest accuracy.

Table: 3.1 Parameter Usage

Algorithms	Details
KNN	n_neighbors=3, p=2,random_state=0
Decision Tree	random_sate=0
SVM	Kernel=linear, random_state = 42
Naïve Bayes	GaussianNB
Random Forest	n_estimators=100

3.8 Evaluation

Evaluation systematically determines the value, value and importance of a topic to characterize an evaluation or research of a program while some merely consider evaluation to be associated with application research. [10] Our best performed model is Random Forest. For both accuracy and different score, it produced best performance among all the algorithms. So we decided to use this algorithm for real life prediction. Figure 3.8 represents the evaluation of our research. In this graph yellow color represents the predicted and blue color represents the real value. For evaluation of our model we used 33 images of four classes. For leaf spot disease real image was 8 and our model also predict 8 images with no error. For rust and white fly we used 6 real images and our system also predict accurately. For sound images we used 13 picture and our system predict 12 sound images that means it is only one error among 33 images.

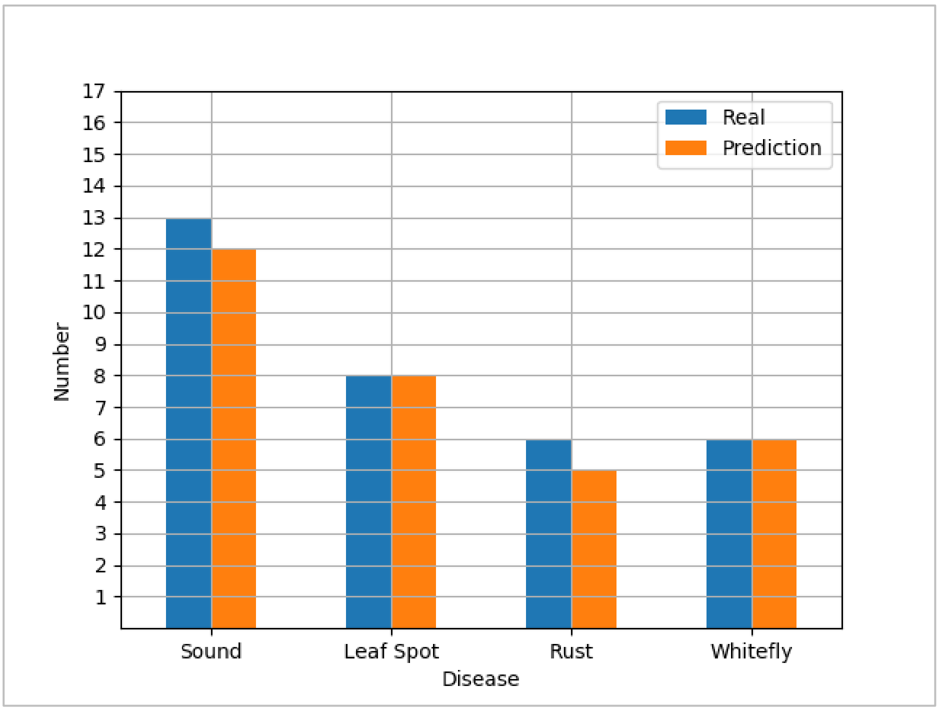


Figure 3.8 Real Vs Predicted Comparison

The identification of the most appropriate output criterion for the classification of imbalanced datasets is a big problem. A few examples have demonstrated in previous work that imbalance may have a significant effect on the importance and sense of precision and many other famous success metrics. [11]

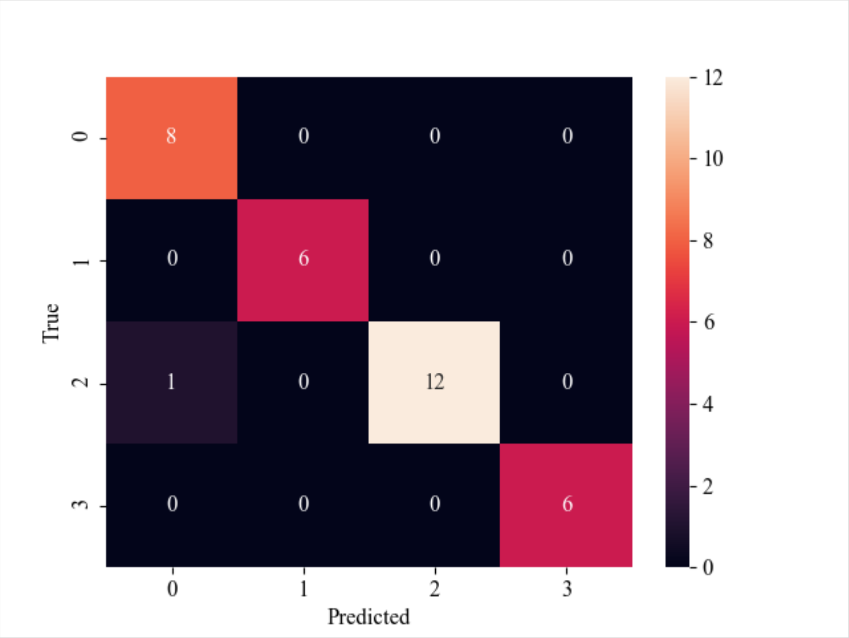


Figure 3.9 Confusion matrix

Figure 3.7 represents the confusion matrix of our prediction results. Here 0, 1, 2, 3 represents leaf spot, rust, sound, white fly. This confusion matrix shows that for sound images our system predict one of them is rust. But originally it was sound. It is an error and percentage is very low. So we can say that our model is very good for detecting guava leaf diseases.

3.11 Summary

In this chapter we have discussed the methodology approach of our research. We discussed all steps individually. The more important things are colors and we discussed the color effect of our work by color analysis using histogram. We also showed that how we implemented our work and which parameter is used in our model. At last we showed the evaluation process of our work.

CHAPTER 4

RESULT ANALYSIS

4.1 Introduction

This section relates mostly in the scientific analysis of objective proof and test results. When we analyze the subject, what is the outcome analysis first? Without interpretation and without analysis, the section on implications should be prepared. The guide can also be found in the University Papers section. The results will be published and the test will be shown. We have already seen various algorithms and we would explain which algorithms in five algorithms are stronger. As a parameter for data calculation, we also have selected accuracy, reminder and f1.

4.2 Experimental Result

Table 4.1 shows the accuracy table. We used 30 to 70 percent research results to figure out which things work well. Yellow boxes indicate the maximum accuracy of the test percentage for each algorithm. The majority of algorithms are conducted below 30% of the test outcome, as seen in this table. Only KNN algorithm produced highest accuracy using 40% test data. Our lowest accuracy is 49.70% achieved by Naïve Bayes Algorithm. And the highest accuracy is 92.12% that is achieved by Random Forest algorithm that is represented by red rectangular box. Not only 30% data usage rate but also for 40, 50, 60 and 70 Random Forest algorithm produced best performance.

Table 4.1 Accuracy table

Test Data usage rate	Algorithms				
	KNN	*Decision Tree*	*Naïve Bayes*	*Random Forest*	*SVM*
30%	83.03%	72.73%	49.70%	92.12%	87.88%
40%	80.45%	77.73%	47.73%	90.00%	88.18%
50%	80.00%	72.36%	50.91%	86.18%	88.00%
60%	76.97%	72.12%	53.03%	86.36%	86.67%
70%	76.88%	72.99%	52.99%	83.38%	84.42%

Table 4.2 Different score matrix

Score Matrix	Algorithms				
	KNN	*Naive Bayes*	*Decision tree*	*Random Forest*	*SVM*
F1 Score	0.83	0.55	0.78	0.92	0.88
Recall	0.83	0.55	0.78	0.92	0.88
Precision	0.84	0.62	0.80	0.93	0.89

Beside accuracy calculation we also calculate f1 score recall, precision and specificity score. This score calculated by using only 30% data usage rate. We used 30% because most of algorithm performed better by using this percentage. Table 4.2 represents different score matrix. From this graph we can see that most of algorithm produce better for every score, without Naïve Bayes algorithm another important think is Random Forest algorithm produced better score among all the algorithms. For every score matrix Random produced better. The recall and specificity scores of algorithms. This analysis tells us Random Forest fitted data so better than other algorithms. So we have decided to use Random Forest algorithm for prediction.

4.2.1 KNN

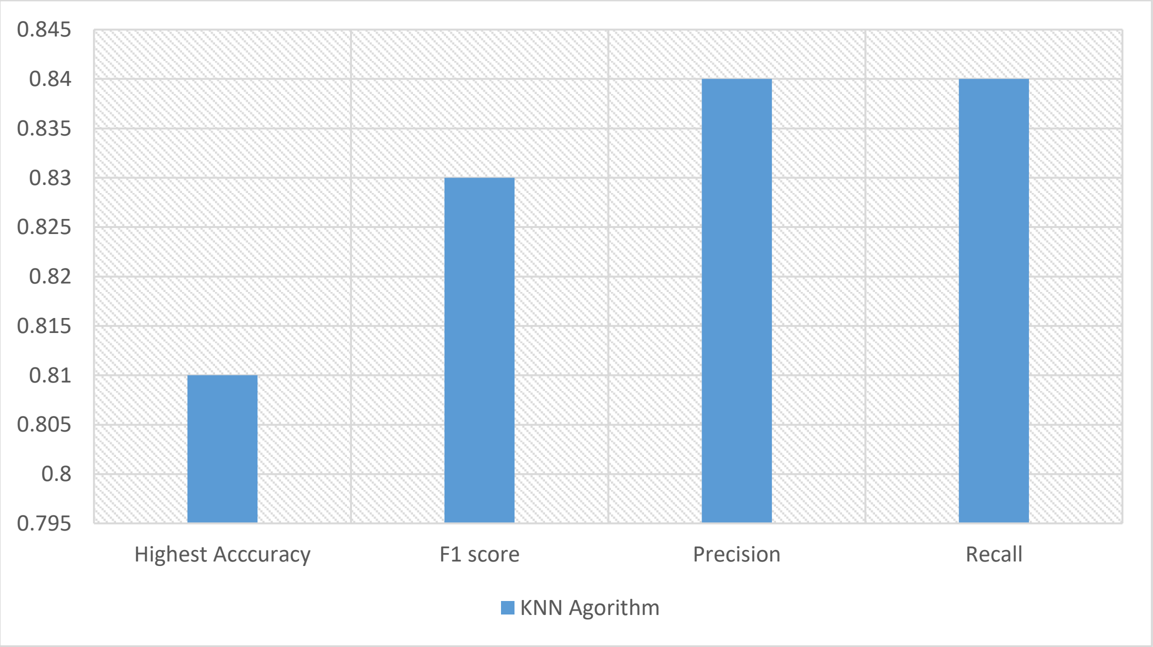


Figure 4.1 Different Score comparison graph of KNN

KNN is simple to use and the most popular machine study algorithm is suboptimal. [13] A KNN algorithm, which is easy to use, is the most common machine study algorithm. Figure 4.1 Represents the scores for all parameters. We had an accuracy rate of 81% for KNN model, while the data rate is 40%. The f1 and precision score is 0.83 and recall score is 0.84. we can see from Figure 4.1 all of the score are very good and close to each other. The KNN algorithm does not contain a plan but for the whole array of data, known as the training data package. While data can be conserved by various ways, kd trees constitute the most used data structure of the KNN algorithm. It helps to immediately and effectively look to balance the latest patterns. The training data is curated and modified as long as new data are produced.

4.2.2 Decision Tree

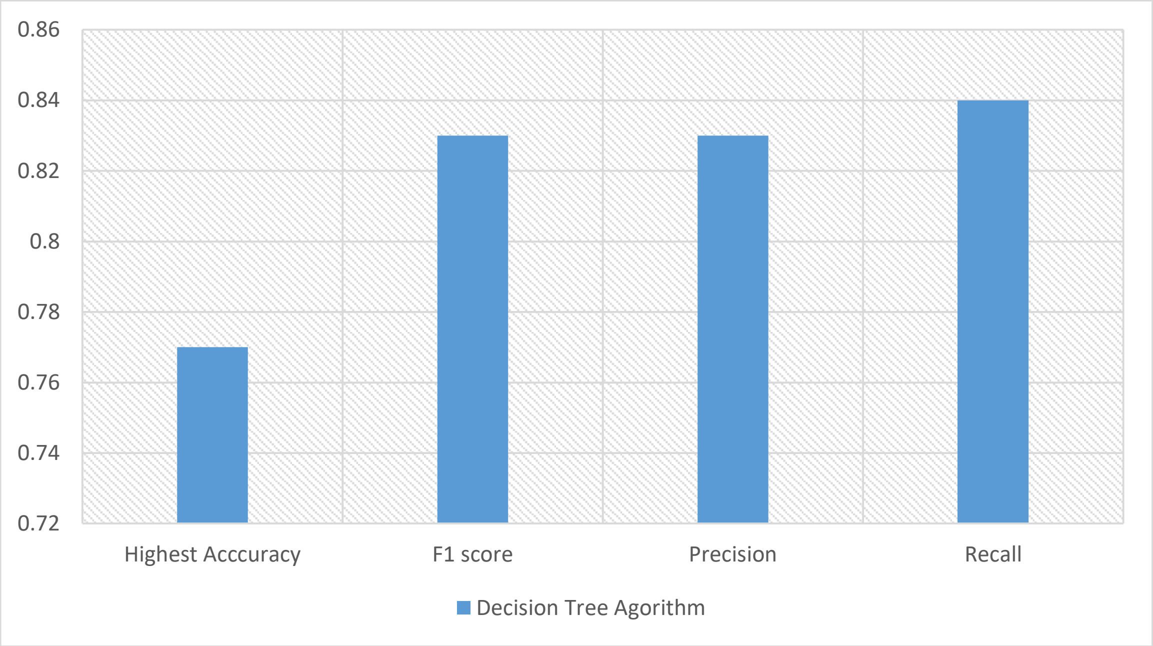


Figure 4.2: Different Score comparison graph of Decision Tree.

Decision Tree is a hateful algorithm, which divides each node properly using local knowledge. One of the consequences is that with the divisional variables a stronger tree may be changed. Trees are very versatile and proven to be limited in their interactions. The downside is that the tone knew the results, called a high variance. There are also large gaps in overfitting, with tree predictions too positive. There are impressive outcomes and a complex dataset in the decision tree. [16] One consequence is that by changing the division variables, a stronger tree can be created. Trees are highly versatile, regarded as low distortion in their interactions. This is the adverse effect of Trees, known as low distortion, in their very high degree of flexibility. The downside is that they learn the sound of the performance, known as high variance. High variances often lead to overfitting, with too positive predictions made by the tree. Figure 4.2 represents that the highest accuracy of Decision tree algorithm is 72.73% using 30% test data. and the f1 and precision score are 0.78 and 0.78 correspondingly. Recall score is 0.80 We can see that decision tree also produced very good performance.

4.2.3 SVM

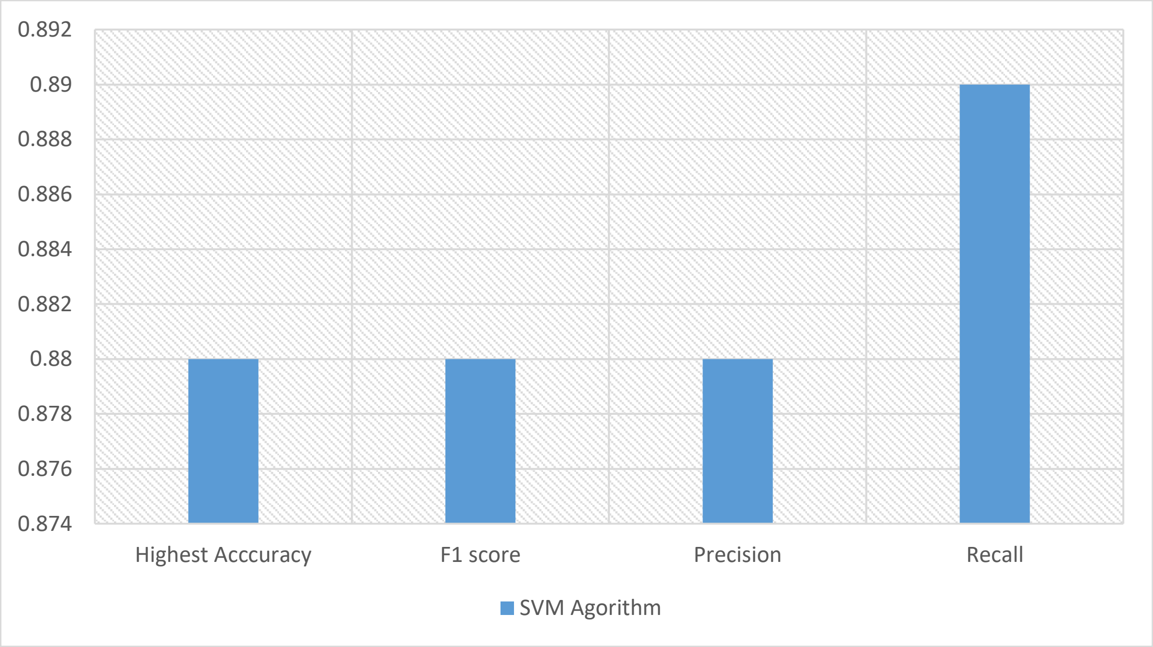


Figure 4.3: Different Score comparison graph of SVM

"Support Vector Machine" (SVM) is a master-training algorithm that can also be used for regression and classification. However, it is mostly contained in questions of description. Each data object in the SVM Algorithm is labeled as a point in n-dimensional space by the value of a basic teamwork Then we categorize the plane which makes very good differences between the two classes. Help vectors are primarily autonomous monitoring coordinates Figure 4.3. The after figure. The SVM is a border that separates the most between the two parties. We found second best performance from SVM algorithm. The highest accuracy was 88% using 30% testing data usage rate. The f1 score is 0.88 precision score is 0. The recall score is 0.88. From figure 4.4. we can see that all score is better and close to each other.

4.2.4 Random Forest

Random forest is a versatility and easy to use algorithm that generates an excellent result, even with hyper parameters, most of the time. It is also a common algorithm because it is simple and varied. For classification and regression functions, it can be used. We shall learn in this article how the RFAl does, what is different and how the other algorithms are used. It builds a "forest" with decision-making trees, usually "sacking" qualified. The fundamental theory of the box method is that a combination of learning styles can boost the final outcome. Classification and regression tasks may be used by random forest. [14] In our classification task random forest generates 92.12% accuracy the precision and specificity scores are 0.92 as like SVM algorithm. Figure 4.4 represents among all other algorithm this algorithm produced better performance. Random forest algorithm in the purpose of prediction.

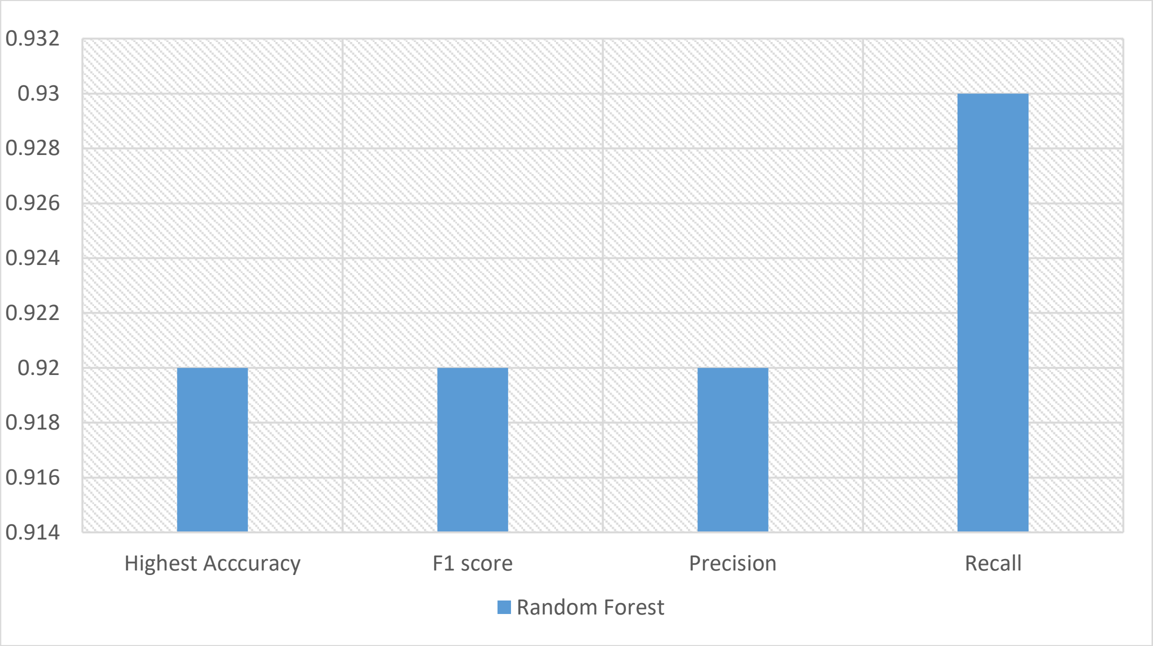


Figure 4.4: Random Forest Score Comparison

4.2.4 Naïve Bays

Logistic regression is one of the most popular algorithms for machine learning, and is a supervised method of learning. The category dependent variable is estimated from a series of independent variables. It is used. The value of a category dependent variable is predicted by logistic regression. The outcome must be categorical or unobtrusive. This may be "Yes" or "No," "0" or "1" or "fact," etc. but it contains probabilistic values of between 0 and 1. Logical regression, except how it was used, is much like the linear regression. Linear regression is used to solve regression problems and logistic regression is used for problem resolution. Linear regression is used to resolve regression problems and logistic regression is used to solve classification problems. [15] the highest accuracy of naïve bayes algorithm is 49%. Figure 4.5 represents F1 sore is 0.66 recall precision scores are correspondingly 0.55, 0.55, 0.62. We noticed that Naive Bayes generate lowest accuracy among all the algorithms as well as other scores is also lower than all algorithms.

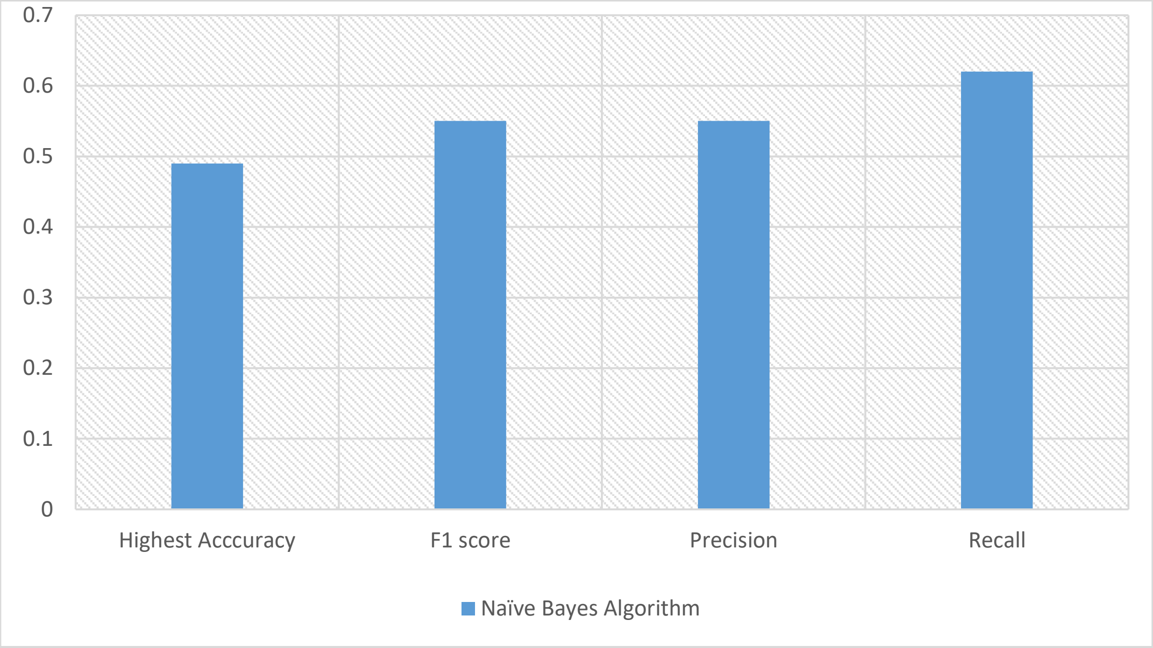


Figure 4.5: Naïve Bayes Score Comparison

4.3 Probability Calculation

In our work we also calculated probability distribution by our best performed algorithm. As random forest provided us best performance from all other algorithm. So we calculated probability of each disease by this algorithm. Every result is shown in below:

4.3.1 probability calculation for leaf spot disease

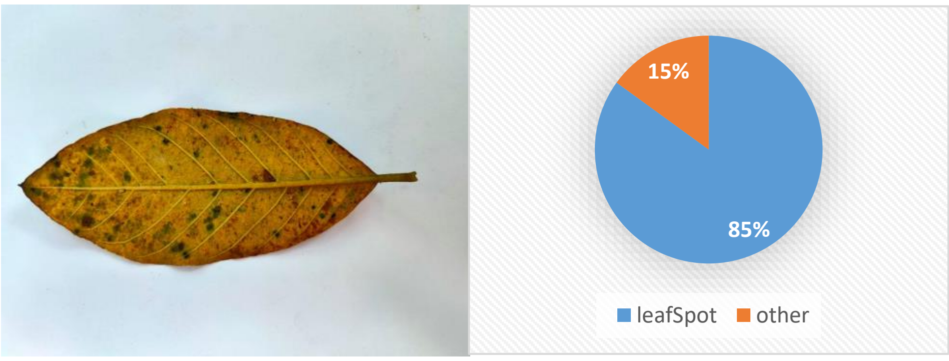


Figure 4.6: percentage of leafspot disease

Figure 4.6 represents percentage of leaf spot disease. Form this graph we can see that our system detects 85% leafspot disease. And 15% other disease.

4.3.2 Probability Calculation for rust disease

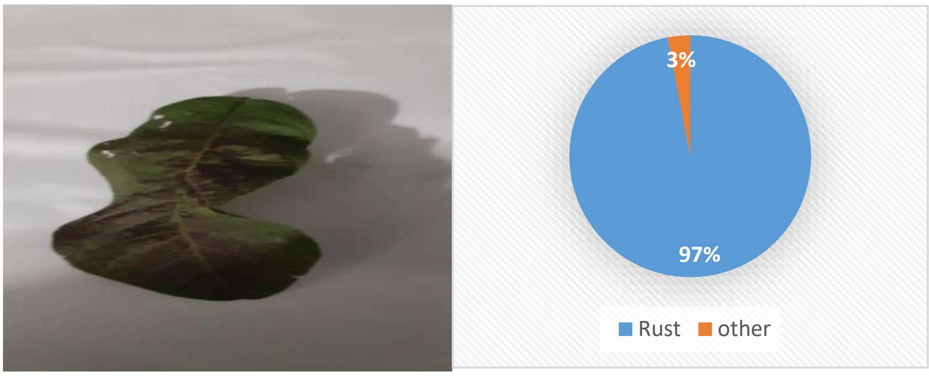


Figure 4.7: percentage of rust disease

Figure 4.7 represents percentage of leaf rust disease. Form this graph we can see that our system detects 97% rust disease. And only 3% other disease.

4.3.3 Probability calculation for sound

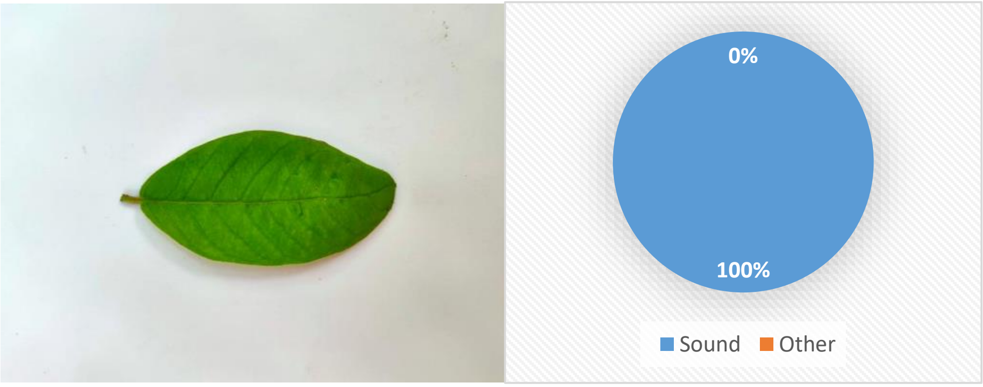


Figure 4.8: percentage of sound

Figure 4.8 represents percentage of sound leaf. Our system can detect 100% sound leaf. There is no error here.

4.3.4 Probability Calculation for Whitefly

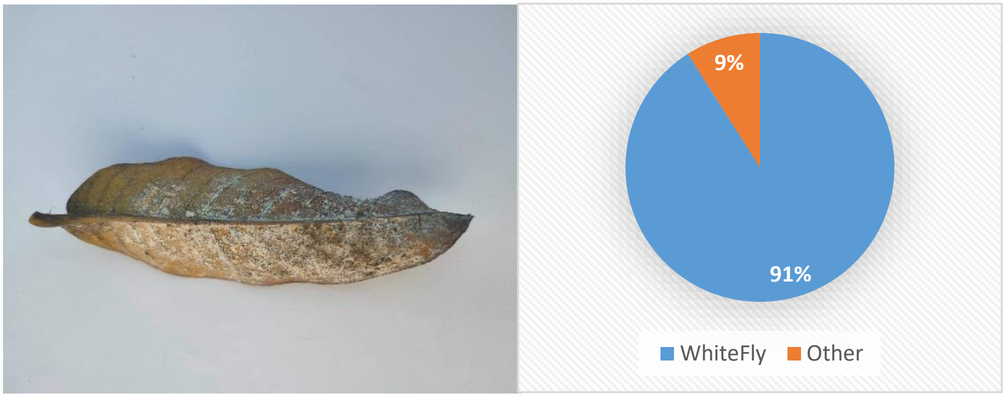


Figure 4.9: percentage of whitefly disease

Figure 4.9 represents percentage of whitefly disease. Our system can detect 91% sound leaf. And only 9% other disease.

4.4 Summary

In this chapter we discussed about our results. We showed different score matrix and their performing percentages by graphically. We can see that every algorithm performed very good on different percentage of data usage rate.

CHAPTER 5

SUMMARY, CONCLUSION AND FUTURE WORK

5.1 Summary of the Study

It does not suspect that many research activities, especially on image recognition, take place in image processing. Recent research is expanding on this topic, while the consequences of such a number of works are a cataclysmic shift in our computer life. We get several noteworthy applications in real life for the value of such research works. However, such research on guava leaf disease in Bangladesh is not very severe. It is a matter of great concern. However, several researchers from different countries have started to investigate this area. We actually detect the diseases in guava in our study. we actually detect the leaf spot disease in guava leaf.

5.2 Conclusion

In our work, every algorithm works carefully. We studied the efficiency of five algorithms and found the best algorithm for detecting guava leaf disease.

Finally, we have a model from which this disease can be detected. We used many computer algorithms in our work. We used various methods for measurement. We have consistency, f1 score accuracy, and recall to select the right algorithm. We picked this algorithm for prediction generation, which gives the highest accuracy and f1 score. We encountered many challenges in doing this study. Data collection was the biggest challenge. We do not gather further data as a pandemic condition. 200 initial datasets are collected. In the pre-processing state, we gathered 200 original data and increased it to a range of 400. We tried to develop a model that can predict the disease of guava leaf after the preprocessing. We hope that the farmer will help with our model.

5.3 Recommendations

There are a few remarkable suggestions for this is given bellow:

- Increasing the performance of data analysis and producing improved results
- Improvements will also improve the efficiency of data collection. Create a more complex algorithm with higher data.
- Continuity fits best in the field to collect data. If this research is extended in the world, it is stronger.

5.4 Future Work

Following are the guidelines for further production:

- In future work on our dataset, will use advanced algorithms.
- Bangladesh by adding additional towns in order to expand the data volume. Our model will be more reliably assessed if we incorporate all other towns or rural areas in our region.
- We will use deeper technologies to obtain proper know-how from our data if it is feasible to gather large amounts of data.
- Last but not least we will try building an intellectual infrastructure that will allow users to detect guava leaf disease

REFERENCE

[1] Islam, Md & Rahman, Md Mushfiqur & Alam, Khondoker & Naher, Nazmun & Hoque, Sanzida & Alam, Md. (2015). Bacterial leaf blight of guava saplings at Dhaka, Gazipur, Barisal and Khagrachori districts of Bangladesh. 2411-6610.

[2] T. Mensink and J. Van Gemert, "The rijksmuseum challenge: Museum-centered visual recognition," in Proceedings of International Conference on Multimedia Retrieval, 2014, pp. 451-454.

[3] O. Bimber and E. Bruns, "PhoneGuide: Adaptive image classification for mobile museum guidance," in 2011 International Symposium on Ubiquitous Virtual Reality, 2011, pp. 1-4: IEEE.

[4] J. S. Hare and P. H. Lewis, "Content-based image retrieval using a mobile device as a novel interface," in Storage and Retrieval Methods and Applications for Multimedia 2005, 2005, vol. 5682, pp. 64-75: International Society for Optics and Photonics.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[6] P. Dollár, R. Appel, S. Belongie, P. J. I. t. o. p. a. Perona, and m. intelligence, "Fast feature pyramids for object detection," vol. 36, no. 8, pp. 1532-1545, 2014.

[7] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in Advances in neural information processing systems, 2013, pp. 2553-2561.

[8] X. F. Hermida, A. C. Rodriguez, and F. M. J. P. o. I.-B. Rodriguez, "A Braille OCR for Blind People," 1996.

[9] G. Sainarayanan, R. Nagarajan, and S. J. A. S. C. Yaacob, "Fuzzy image processing scheme for autonomous navigation of human blind," vol. 7, no. 1, pp. 257-264, 2007.

[10] Md. Tarek Habib, Md. Jueal Mia, Mohammad Shorif Uddin, Farruk Ahmed "An in-depth exploration of automated jackfruit disease recognition". In Journal of King Saud University –Computer and Information Sciences, 25 April 2020

[11] ‘Evaluation’ https://en.wikipedia.org/wiki/Evaluation

[12] Amalia Luque, Alejandro Carrasco, Alejandro Martín, Ana de las Heras,The impact of class imbalance in classification performance metrics based on the binary confusion matrix,Pattern Recognition, Volume 91,2019,Pages 216-231,ISSN 0031-3203,https://doi.org/10.1016/j.patcog.2019.02.023.

[13] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. J. N. Zhang,"Efficient kNN classification algorithm for big data," vol. 195,pp. 143-148, 2016.

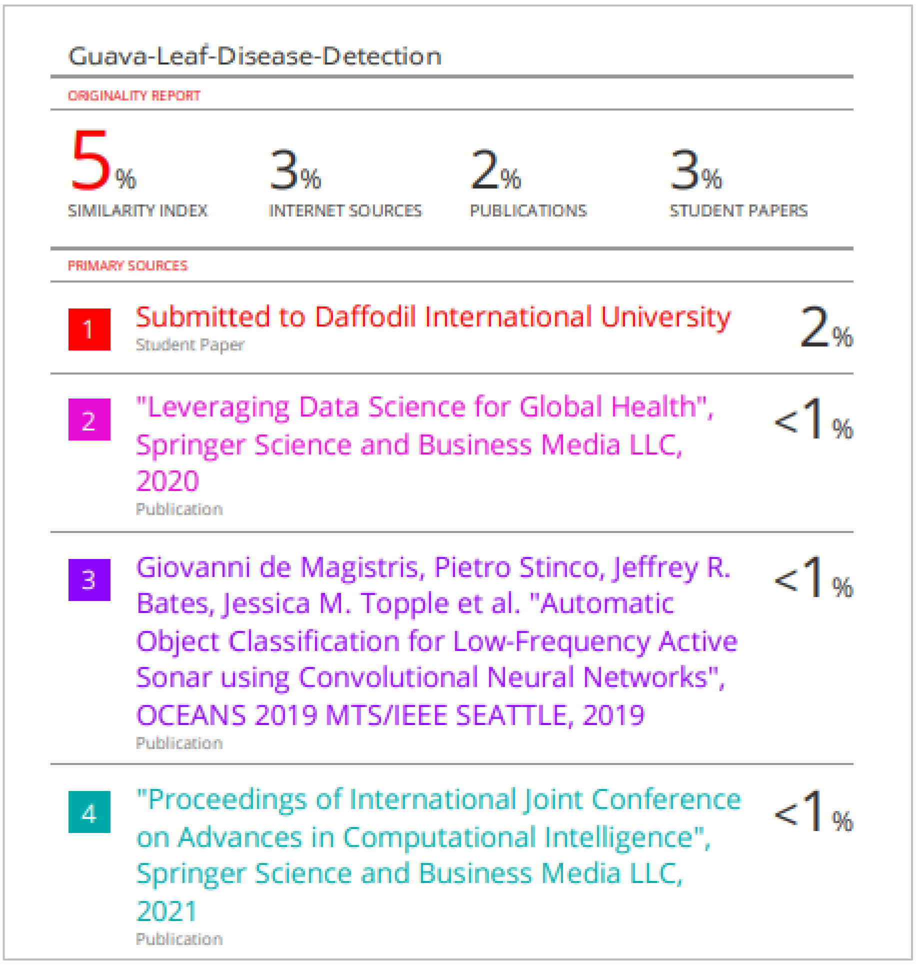
[14] L. H. Simon Bernard, and S´ebastien Adam, "Influence of Hyperparameters on Random Forest Accuracy," Springer, 2009.

[15] H. J. I. J. o. P. R. Zhang and A. Intelligence, "Exploring conditions for the optimality of naive Bayes," vol. 19, no. 02, pp.183-198, 2005.

[16] S. R. Safavian, D. J. I. t. o. s. Landgrebe, man,, and cybernetics, "A survey of decision tree classifier methodology," vol. 21, no.3, pp. 660-674, 1991.

[17] M. M. Hasan, M. T. Zahara, M. M. Sykot, A. U. Nur, M. Saifuzzaman and R. Hafiz, "Ascertaining the Fluctuation of Rice Price in Bangladesh Using Machine Learning Approach," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225468.

PLAGIARISM REPORT



Guava-Leaf-Disease-Detection

ORIGINALITY REPORT

5%	3%	2%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	"Leveraging Data Science for Global Health", Springer Science and Business Media LLC, 2020 Publication	<1%
3	Giovanni de Magistris, Pietro Stinco, Jeffrey R. Bates, Jessica M. Topple et al. "Automatic Object Classification for Low-Frequency Active Sonar using Convolutional Neural Networks", OCEANS 2019 MTS/IEEE SEATTLE, 2019 Publication	<1%
4	"Proceedings of International Joint Conference on Advances in Computational Intelligence", Springer Science and Business Media LLC, 2021 Publication	<1%