Journal of Theoretical and Applied Information Technology
30th September 2021. Vol.99. No 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# A PREDICTION OLIVE DISEASES USING MACHINE LEARNING MODELS, DECISION TREE AND NAÏVE BAYES MODELS

[1]JAFAR DRDSH, [2]DERAR ELEYAN, [3]AMNA ELEYAN

[1,2] Department of Applied Computing, Technical University-Kadoorie, Tulkarem, Palestine [3]Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BH, United Kingdom,

E-mail: [1] j.h.drdsh1@students.ptuk.edu.ps, [2]d.eleyan@ptuk.edu.ps,[3] a.eleyan@mmu.ac.uk

**ABSTRACT**

Machine Learning Models as Decision Tree and Naïve Bayes 'NB' models are widely used to predict diseases. This paper has used these models to predict olive diseases. It relies on image processing of the olive leaves and predict the type of disease according to the information gathered from different images. Where 2000 images were collected of an olive leaf which is affected by the disease and healthy. The results show that the accuracy of prediction in the decision tree model is 97% and in the NB model it has reached 80%.

This idea was inspired by an idea found in disease prediction, such as the Agrobase application, which analyzes and processes images, and then returns to the associated database to analyze and compare the results, and then gives the prediction result.

In this research, we focus on the accuracy of prediction and image analysis, where we highlighted the most deadly disease in olives, olive leaf spot disease, so its color was analyzed and open cv was adopted in the Python language.

**Keywords:** *Olive Diseases, Machine Learning Models, Decision Tree, Naïve Bayes models, Olive Spot Diseases*

## 1. INTRODUCTION

The olive tree has a special sanctity in some religions such as Islam. Muslims care about the olive tree because it is blessed by God, and for other religions, the olive tree has a special sanctity. Everyone agrees on the importance of this tree very much.

Olives, also called Olea Europaea, have a place for sorting Olea. This tree lives about 400-600 years. Where this tree was found in the eastern Mediterranean 7000 years ago. The olive tree is the main and only source of olive oil production, which has amazing features [8].

The importance of the olive tree can be deduced from its spread in 64 countries. The production of olive trees increased permanently from 1965 to 2017. In 2017, the yield of olive oil production reaches 20 million tons, and the cultivated area increased to 10 million hectares. This study is for 64 countries, and olive tree diseases caused a decrease in production from 31,453 hectares in 1961 to 19,319 hectares in 2017, all due to the death of a large number of fruitful and productive olive trees [8].

At the global level, the productivity of olive trees decreased. For example, in Spain it decreased to 44% of the annual olive harvest [9] compared to previous figures. In a subsequent study [9], studies predict the disappearance of 20,000 olive trees in the coming decades, and in Italy the olive yield decreased to 57% compared to last year [10].

Because of Xylella fastidiosa, a specific type of plant bacteria, there are serious risks to olive trees [11], in Italy there are 11 million olive trees threatened with extinction because there is no treatment until this time [12].

The olive tree is affected by many diseases. Among the diseases of olives are microorganisms, olive knot, bacterial spot, bacterial leaf curse, minute leaf spot, plant fly, olive leaf spot, leaf shape, natural olive oil fly, olive bark creeping creeping, olive borer, olive moth, nitrogen and calcium Magnesium, zinc deficiency, and a lot of multiple diseases. Neofabria infection and peacock leaf spot are among the diseases that appear in the form of a spot on olive leaves, as this disease is contagious and is transmitted from one tree to another during the season of its spread.

There's a allocate of progress inside the field of plant ailment revelation utilizing picture taking care of strategies. But since disease is an outstandingly changing term the recognizing components will persistently alter and be from a endless super-set. So to initiate a common and working for all sort of disease calculation will ceaselessly be a challenge.

In this research paper, two objectives will be discussed, the first is to provide high accuracy in detecting olive diseases, and the second is to reduce the potential material losses when trying to detect the disease through the accuracy in identifying olive diseases early.

An evolutionary model methodology was taken after in this paper, this demonstrate recommends breaking down of work into littler chunks and prioritizing them. The demonstrate permits for changing necessities as well as all work in broken down into viable work chunks, see figure 1 that shows the evolutionary model Phases.

## 2. LITERATURE REVIEW

One of the critical foliar maladies influencing olive trees is peacock spot disease caused by Cycloconium oleaginum, moreover known as olive leaf spot and bird's-eye spot. Seriously, contaminated trees appear defoliation of takes off coming about in destitute twig development and destitute natural product set, and extraordinary harm of manor. Contamination is regularly related to tall stickiness and winter conditions (cool and moo light), where tall temperatures limit spore germination and development. Indications of infection begin as dirty blotches create green dark circular spots 0.1 to 0.5 inch in distance across with a swoon yellow corona around the spot. More injuries are created within the lower portion of the tree. Takes off may drop rashly and twig passing may happen due to defoliation [1].

Olive leaf spot (OLS), a disease that appears on olive leaves, has several names, one of which is known as peacock spot or bird's eye spot, is caused by the biotroph parasitic pathogen Fusicladium oleagineum (syn. Spilocaea oleaginea, Cycloconium oleagineum), agreeing to the late proposed utilize of the Sort Fusicladium rather than Venturia for those species which show as it were anamorphic arrange. It is one of the foremost vital contagious infections that influence olive trees, and in cases of extreme diseases might cause abdicate misfortunes of around 20%. The infection causes unmistakable injuries basically on the upper surfaces of the clears out, which are at first subtle dingy blotches, but afterwards create into the sloppy green to nearly dark circular spots encompassed by a yellow corona. Petioles, natural products and stems are too helpless, but seldom show injuries. The tainted clears out drop rashly, and defoliation influences the vegetative and regenerative development of olive trees negatively [2].

OLS disease can happen at any time of the year, but ordinarily amid late harvest time to early summer, on the off chance that natural conditions are positive. In hot dry climate conditions, conidia stay practical but inert on contaminated clears out and begin to grow early in winter. Conidium generation is ideal at 15oC and/or temperatures extending from 2 to 25oC and tall stickiness (85%). Conidia of S. oleaginous are scattered by rain sprinkle or wind-borne water beads [3].

Previous scientific papers in this field were searched for several techniques to analyze the problem of olive leaf spot disease, one of the techniques that were investigated was to determine the number of spots on the olive leaf and classify the level of tree disease based on the number of spots [13]. And in another research, the relationship between disease spot on olive leaf with the area around the spot on olive leaf, and determination of disease behavior based on analysis using image processing and MATLAB program and k mean in machine learning. The first research paper has weakness in it that it depends on the number of spots on the olive leaf to determine the percentage of disease, and this matter leads to inaccuracy in determining the percentage of disease. It gives a disease rate of 33%, as for weakness in the second research paper, it did not search except to determine the behavior of the disease and did not search for its definition and its percentage on olive leaves [8]. As

for us in this research paper, we will work on an accurate technique in determining the seriousness of olive leaf disease, despite many researches, but we will present this technique in which solutions to the weaknesses of previous research, because the olive leaf will be analyzed accurately in terms of time and taking into account the external influences and factors that affect On the accuracy of the result and the exception of every property that is not relevant to the prediction.

OpenCV is defined as an open source vision package for PC with more than 500 methods for image and video analysis, fabricated object analysis, restorative imaging, security, computer software, camera development, stereoscopic vision and computerized thinking. Since its inception in 1999, it has been developed because the vision for engineers about computer vision based on the C language is simplified and misuses multi-core processors. Interpretation 1.0 was pushed in a few 006 and a big moment occurred in 2009 with the speed of OpenCV 2 that anticipated fundamental changes, particularly the modern C++ interface.

Work has been done to reduce the scale of vital lines of code for visual sensitivity coding as well as to reduce standard programming errors such as memory leakage (through the task of modified information and deallocation) that will appear once OpenCV is misused in C. More recently, methods have been developed in an OpenCV stack based on C++ language interfaces. There is active progress in communication for python ruby mat lab and others.

The OpenCV supported in Python was used to develop the algorithm underlying this paper.

Naive Bayes Classifier is considered as one of Statistical Bayesian Classifier and it is known with its simplicity. It suppose that all variables participate towards classification and mutually correlated, so it called Naive. Such assumption called class conditional independence. Idiot's Bayes, Independence Bayes, and Simple Bayes are other interesting calls. They have the ability to predict the probabilities of class membership, the probability of a given data item is an example for a specific class label. The presence of a specific feature (attribute) or its absence of a class has no relation with any other feature presence or its absence when the class variable is given at Naive Bayes classifier [5].
Bayesian Theorem is the basis of Naive Bayes Classifier technique and it is applied at high dimensionality of the inputs [3]. Bayes Theorem is the basis of Bayesian classification as shown below:

X is a data sample that its class label is not known and H is some hypothesis. While class C may include the data sample X. Bayes theorem is using P(C), P(X), and P(X|C) for calculating the posterior probability P(C|X).

Where P(C|X) is the target class posterior probability.

P(X|C) is the likelihood and is the predictor probability of the given class.

P(C) is called the class prior probability.

P(X) is the predictor prior probability of the class. Explain as in equation 1.

A decision tree structured by three main parts. These parts concluded as root, internal and leaf nodes. The internal nodes indicate a test conditions for a feature (attribute).

Each branch states the test condition result. Each terminal node (or leaf node) had its class label. The root node is considered as the topmost node. The decision tree divide and conquer the data. In the decision tree, each path compose a decision rule. In general, it exercises a greedy process start from top to bottom. The technique of decision tree classification has two phases. First phase is the tree construction, while the second phase is tree clipping, the decision Tree algorithm is one of the supervised learning algorithms family. The supervised learning algorithms do not deal with regression and classification problems , but decision tree algorithm solve such problems.

Creating a training model is the goal of decision tree. The training model is used to predict the class or to predict the target variable value by attending learning simple decision rules concluded from previous data ( training data).

We begin with the tree root for predicting a class label In Decision Trees.
The root attribute values are compared with the attribute-recorded value. Based on comparison, corresponding to that value we follow the branch and move to the following node.

1) Decision Trees Types
The type of target variable is the basis of classification of decision trees. So it has the below types:

1. Categorical Variable Decision Tree: the main property for this decision Tree is to have a categorical target variable.
2. Continuous Variable Decision Tree: the main property for this decision Tree is to have a continuous target variable. [5].

A splitting criterion is used in Decision-tree algorithms. This criterion separate a node from a tree and aim is to reduce the imperfection of a node. The rate for each predictor variable provided by splitting criteria. To select variables and keep it in the model, it should have the best splitting criterion rate. Information Gain, the most common and substantial splitting criteria are Gini index and Gain ratio.

Decision tree algorithms include classification tree. Regression tree is included too. These trees are named CART. CART use predictor variables to split data sets to subsets. The root nodes will be created by splitting the subsets repeatedly. The CART algorithm exercise in two stages. First stage is creating a binary division. The second stage is pruning a tree with respect to cost-complicity. Further, the algorithm select the best variable by using the Gini impurity index. Equation 2 measures the impurity of Gini index.
where Pi is the probability that one of the registers in D. It belongs to class Ci. Pi is estimated by | Ci, D | / | D |. The amount is found in categories of m. The first variable, in a decision tree, is the leading factor. While other variables assorted in order of importance. The root node is the variable that split the entire population.

The splitting will be done with the highest information gain. In popular data mining methods, the data is divided into two parts. These two parts are the training data and test dat. The training data set is approximately 70% of the data, while the test data set is about 30% of the data. The model was created on a training data set [6].

The DT algorithm in machine learning is considered one of the best algorithms, but almost the first in this field, due to the accuracy of the knowledge that results from its use for prediction.

## 3. LIMITATIONS

OLS is a disease that multiplies and spreads in high humidity, and this is the biggest determinant during the process of writing this scientific paper. Leaves were investigated in areas of high humidity with shade at the bottoms of valleys, and the largest possible number of olive leaves infected with the disease were collected from those places.

## 4. FIGURES

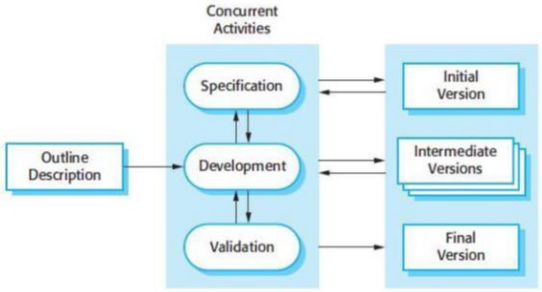


*Figure 1 Evolutionary methodology*

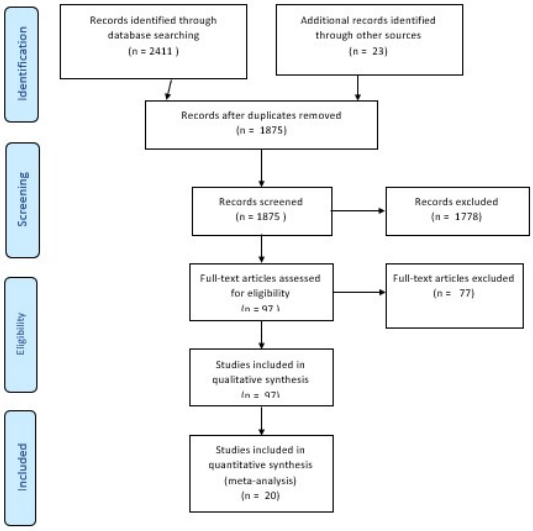


*Figure 2 PRISMA diagram*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

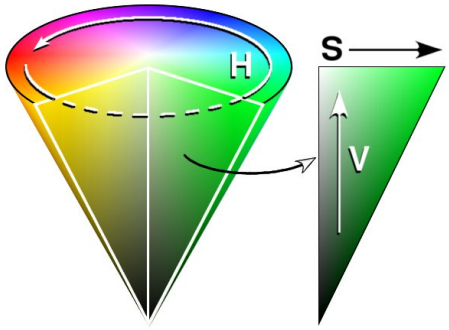


*Figure 3 HSV color*

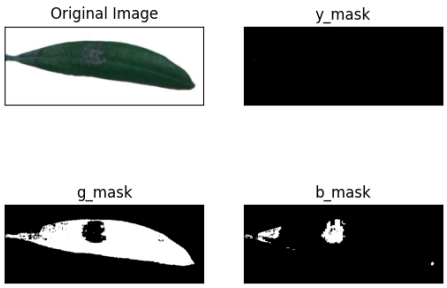


*Figure 4 Image processing*

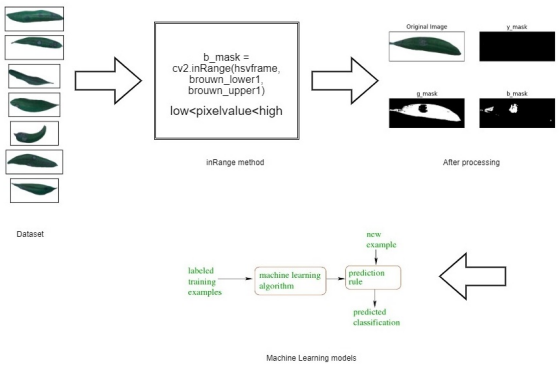


*Figure 5 Flow chart*

*Figure 6 Level 0*

*Figure 7 Level 1*

*Figure 8 Level 2*

*Figure 10 Level 4*

*Figure 9 Level 3*

| ID | Y_% | G_% | B_% | Level of disease |
|---|---|---|---|---|
| 0 | 0.013129 | 53.68439 | 46.30248 | Level 2 |
| 1 | 0.014902 | 24.34429 | 75.64081 | Level 4 |
| 2 | 0.005439 | 64.76475 | 35.22981 | Level 2 |
| 3 | 0.451653 | 5.584677 | 93.96367 | Level 4 |
| 4 | 0 | 26.70306 | 73.29694 | Level 3 |
| 5 | 0.001857 | 32.81045 | 67.18769 | Level 3 |
| 6 | 12.72838 | 16.95183 | 70.31979 | Level 3 |
| 7 | 0.015229 | 66.07578 | 33.90899 | Level 2 |
| 8 | 0.022608 | 46.93382 | 53.04358 | Level 3 |
| 9 | 0.846204 | 84.85869 | 14.2951 | Level 1 |

*Figure 11 Dataset*

## 5. EQUATIONS

$$P(C|X) = \frac{P(X|C).P(C)}{P(x)} \quad (1)$$

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (2)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

## 6. METHODOLOGY

As shown in the PRISMA diagram (Figure 2), this methodology was followed in the study of the research literature.

In this research, a specific technique was identified that was used to analyze and classify the level of olive leaf spot disease, which appears on the olive leaf in the case of leaf spot.

The spot of the disease is often in a circular shape, but sometimes it takes an irregular shape. Pictures of olive leaves were collected from several local farms and a group on the Internet, an infected group of 2000 leaves and another uninfected group also reached 2000 healthy leaves.

A program has been developed using the Python language and using the open cv image processing library to analyze the images and find out the percentage of the spot from the total image area. The cv2.inRange () method was used to determine the stain percentage in the olive leaf

```
frame = cv2.imread(image_Path)
        hsvframe = cv2.cvtColor(frame,
cv2.COLOR_BGR2HSV)

        yellow_lower = np.array([20, 100, 100])
        yellow_upper = np.array([30, 255,
255])
        y_mask = cv2.inRange(hsvframe,
yellow_lower, yellow_upper)

        green_lower = np.array([20, 0, 0])
        green_upper = np.array([120, 100,
100])
        g_mask = cv2.inRange(hsvframe,
green_lower1, green_upper)

        brouwn_lower = np.array([20, 10, 0])
        brouwn_upper = np.array([35, 255,
88])
        b_mask = cv2.inRange(hsvframe,
brouwn_lower1, brouwn_upper)
```

Lower and upper refer to the boundary of the disease spot region, meaning a pixel within this boundary is set to 255 and any pixels outside this boundary are set to 0 and in this way, the disease area and its area are determined due to the image [7].

hsvframe is the image of the olive leaf that is analyzed to determine the percentage of the disease on it, as it was converted from RGB to HSV

Any captured image consisting of a matrix of RGB colors, and the R, G and B components of the object's color in an image is a digital matrix that depends on the amount of light falling on the object, and because of the similarity and abundance of these colors and the large number of images, it is difficult to distinguish the characteristics of the image with their qualities for each object alone. To make the distinction easy, the matrix has been converted to HSV space format. The HSV color space describes colors in terms of hue, saturation, and value. Therefore, if image qualities play an important role, the HSV color model is usually preferred over the RGB model. Describes the HSV model where this formula is similar to the human eye in color perception. In general, RGB stands for primary colors, while HSV is the color of an object's attributes.

brightness [7]. Explanation of HSV color space, Explain as in Figure 3

After determining the percentage of the area of the affected area from the total area of the olive leaf, the determination was made on all the olive leaves that were collected, Explain as in Figure 4

Where the data was collected in an Excel file after analyzing the images and analyzing their colors to prepare them to be reliable data to make a prediction with them as shown in Figure 11

Machine learning methods were used to predict the percentage of disease based on the results of analyzing the images, the proportions extracted from it, where the Decision Tree and Naïve Bayes models were used to predict the proportion of disease on the olive leaf, where the disease level was divided into five levels, less than 1 healthy level,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

from 1 to 25% of the second level of injury, and from 26 to 50% of the third level, and from 51 to 75% of the fourth level, and from 76 to 100% of the fifth.

```
def build_tree(self):
    self.data = pd.read_excel(r'dataset-1.xlsx')
    self.df = pd.DataFrame(self.data)
    df2 = pd.DataFrame(self.df,
columns=['Y_%', 'G_%', 'B_%', 'Level of
disease'])
    self.df.fillna(self.df.mean(), inplace=True)

    X = self.df[['Y_%', 'G_%', 'B_%']]
    y = self.df['Level of disease']
    y.fillna('level 0', inplace=True)


    one_hot_data = pd.get_dummies(X)

    clf = DecisionTreeClassifier()
    clf_train = clf.fit(one_hot_data, y)
```

The lines of code above represent the building of the Decision Tree model.

```
def build_NB(self):

    print('Step 1: Load and Pre-Process The
Data .....')
    candidates = pd.read_excel(r'dataset-
1.xlsx')
    df = pd.DataFrame(candidates,
columns=['Y_%', 'G_%', 'B_%', 'Level of
disease'])
    print(df)
    print(' ')

    print('Step 2: Subset The Data to build
model .....')
    X = df[['Y_%', 'G_%', 'B_%']]
    y = df['Level of disease']
    print(' ')

    print('Step 3: Split The Data Into Train
And Test Sets .....')
    X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.25,
random_state=0)
    print(type(X_train))
    print(type(y_train))
    print(' ')

    print('Step 4: Build A MultinomialNB
Classifier .....')
    MultiNB = MultinomialNB()
    MultiNB.fit(X_train, y_train)
    print(1)

    print('Step 5: Prediction: use the model to
predict for X_test and compare the result of
prediction with y_test to measure the accuracy
.....')
    print(' ')
    y_expect = y_test
    y_predict = MultiNB.predict(X_test)
```

The lines of code above represent the building of the Naïve Bayes model.

## 7. RESULTS

Through this technique of a Prediction Olive Diseases using Machine Learning Models, Decision Tree and Naïve Bayes models, the results of determining the area of injury to the olive leaf and its percentage from the total area of the leaf can be seen, where the peacock disease spot is separated from the rest of the olive leaf, and then the results of the image analysis and processing are purified, Explain as in Figure 5 and Figures 6, 7, 8, 9, 10.
As is the case in the above picture, it is clear how the disease is concentrated in the diseased olive leaves and the healthy leaf is devoid of spots, where the first level is the leaf was healthy from any disease, and in the remaining four levels the disparity of disease between the leaves is cleared
As the percentage of accurate prediction of disease in olive leaves through machine learning models reached 97%.

**print('The accuracy of MultinomialNB Classifier:', accuracy_score(y_expect, y_predict))**

The lines of code above represent the building of the accuracy.

Thus, a Decision Tree was built, as shown in the Figure 11 below

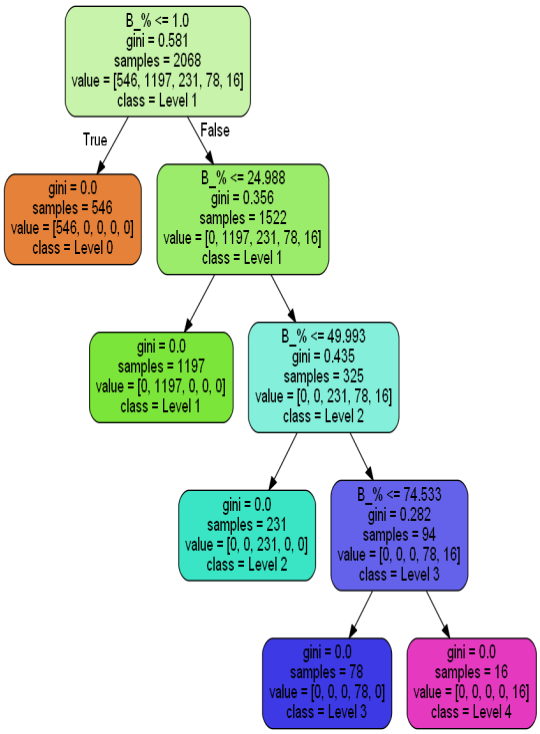


*Figure* 11 *Decision Tree*

**dot_data = tree.export_graphviz(clf_train,out_file=None, feature_names=list(one_hot_data.columns.values),**

**class_names=['level0','level 1', 'level 2', 'level 3', 'level 4', ''], rounded=True,filled=True)**

**graph = pydotplus.graph_from_dot_data(dot_data)**

**graph.write_png("Levelofdisease.png")**

The above lines of code were used to build a Decision Tree as an image that appears for the researcher to be able to take information and analyze it according to the results that appear in the accuracy of the models and in the prediction results.

After showing results, the two objectives of this research are achieved, the first is to provide high accuracy in detecting olive diseases, and the second is to reduce the potential material losses when trying to detect the disease through the accuracy in identifying early olive diseases.

## 8. CONCLUTION

In this research paper, an accurate technique is presented in predicting the percentage of olive leaf spot disease, with a 97% accuracy rate in the machine learning models used, as the language of this technique is of high accuracy after isolating all the irrelevant factors and analyzing the factor relevant to the disease directly, where the irrelevant factors were represented in color. The yellow-green that appears on the olive leaf, as these colors are not relevant to olive leaf spot disease, and the factors relevant to the disease are represented in brown.
In the proposed technique, the color ratio is determined on the olive leaf, which gives a better result than determining the number of spots on the olive leaf.

## 9. FUTURE WORKS

Many tree diseases will be studied and researched so that this technology expands to include many diseases until it reaches high accuracy in predicting diseases and provides the best research in this field for every researcher or technician who wants to develop an application in this field, and research will also be done on diseases of plants, small shrubs or The annual plant so that its diseases can be detected using highly efficient techniques, which contribute to raising agricultural yields in the world.

By building an algorithm with a unique technology and an application on smart phones linked to a guided imaging plane, so that the crop and field are completely photographed to predict the level of disease in the crop, this is at the level of olive trees as well as for the rest of the plants.

**ACKNOWLEDGMENT:**

The authors wish to thank Palestine Technical University-Kadoorie (PTUK) for supporting this research work as part of PTUK research fund.

**REFRENCES:**

[1] Mahmoud, K., Khalaf, l., Nasser, B., Morouj, Z., Biological control of olive leaf spot (peacock spot disease) caused by Cycloconium oleaginum (Spilocea oleaginea). Academicjournals Vol. 2(6). 9(2010)
[2] George T., Anastasios S., George S., Laminarin Induces Defense Responses and Efficiently Controls Olive Leaf Spot Disease in Olive. mdpi/journal/molecules. 1(2021)
[3] Nawaf A., Mazen S., Visible/Near infrared (VIS/NIR) spectroscopy and multivariate data analysis (MVDA) for identification and quantification of olive leaf spot (OLS) disease. Palestine Technical University Research Journal, 2014, 2(1)
[4] Chaitanya S., A Survey Paper on Satellite Image Using OpenCV Library overHadoop Framework. MAT journals. 2018
[5] Sayali D., Channe, H., Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064. 2014
[6] Maryam, T., Mohammad, T., Sara, S., Parichehr, H., Ali, R., Habibollah, E., Ali, T., Gordon, A., Ferns,. Mohsen, M., Majid, G. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. Elsevier. 2017
[7] Mohammed, A., Noor, A., Mohammed, A. Detecting Crows on Sowed Crop Fields using Simplistic Image processing Techniques by OpenCV in comparison with TensorFlow Image Detection API. IJRASET. 2020
[8] Aditya, S., Rajveer, S., Olive spot Disease Detection and Classification using Analysis of leaf image texture. Science direct. 2019
[9] Vilar, J., (2019). Accessed on 1-05-2019. URL: https://www.oliveoiltimes.com/wp-content/uploads/2019/06/salvemos-el-buen-aceite.pdf.
[10] GRANITTO, Y.,Olivetimes. URL: https://www.oliveoiltimes.com/olive-oil-business/italian-olive-oil-production-falls-to-record-lows/66852.
[11] Abbott, A., . Accessed on 1-05-2019. URL: https://www.nature.com/articles/d41586-018-07389-8.
[12] Borunda, A.,. Accessed on 1-05-2019. URL: https://www.nationalgeographic.com/science/2018/08/italy-olive-trees-dying-xylella/.
[13] Mokhled S., Al-Tarawneh. An Empirical Investigation of Olive Leave Spot Disease Using Auto-Cropping Segmentation and Fuzzy C-Means Classification. World Applied Sciences Journal 23 (9): 1207-1211, 2013