KONYA FOOD AND AGRICULTURE UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPUTER ENGINEERING

CLASSIFICATION OF MULTIPLE SKIN DISEASES USING POTATO IMAGES

MASTER OF SCIENCE THESIS

SEDA KIZIL

KONYA

AUGUST 2019

KONYA FOOD AND AGRICULTURE UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

(MASTER'S THESIS)

CLASSIFICATION OF MULTIPLE SKIN DISEASES USING POTATO IMAGES

Seda KIZIL

Supervisor: Assist. Prof. Dr. Zafer ARICAN

Department of Computer Engineering

Meram- KONYA

August 2019

KONYA GIDA VE TARIM ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

(YÜKSEK LİSANS TEZİ)

PATATES GÖRÜNTÜLERİ KULLANARAK ÇOKLU DIŞ HASTALIKLARININ SINIFLANDIRILMASI

Seda KIZIL

Tez Danışmanı: Assist. Prof. Dr. Zafer ARICAN

Bilgisayar Mühendisliği Bölümü

Meram- KONYA

August 2019

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis of the degree of Master

………………………………

Assist. Prof. Dr. Zafer ARICAN (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis of the degree of Master

………………………………

Assist. Prof. Dr. Fatih NAR

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis of the degree of Master

………………………………

Prof. Dr. Sabri KOÇER

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis of the degree of Master

………………………………

Prof. Dr. Sencer BUZRUL

Director of the Institute of Natural and Applied Sciences

Approval of the thesis:

CLASSIFICATION OF MULTIPLE SKIN DISEASES USING POTATO IMAGES

This study titled "Classification of Multiple Skin Diseases Using Potato Images" and presented as thesis by **SEDA KIZIL** has been evaluated in compliance with the relevant provisions of KFAU Graduate Education and Training Regulation and KFAU Institute of Science Education and Training Direction and jury members written below have decided for the defence of this thesis and it has been declared by consensus that the candidate has succeeded in thesis defence examination dated.

Prof. Dr. Sencer BUZRUL

Director, **Institute of Natural and Applied Science, KGTU**

Assist. Prof. Dr. Arda SÖYLEV

Head of Department, **Computer Engineering, KGTU**

Assist. Prof. Dr. Zafer ARICAN

Supervisor, **Electrical-Electronics Engineering, KGTU**

Examining Committee Members:

Head: Prof. Dr. Sabri KOÇER

Computer Engineering, Necmettin Erbakan University

Member: Assist. Prof. Dr. Fatih NAR

Computer Engineering, KGTU

Member: Assist. Prof. Dr. Zafer ARICAN

Electrical-Electronics Engineering, KGTU

Date: 23.08.2019

ÖZET

PATATES GÖRÜNTÜLERİ KULLANARAK ÇOKLU DIŞ HASTALIKLARININ SINIFLANDIRILMASI

KIZIL, Seda
Yüksek Lisans Tezi, Bilgisayar Mühendisliği Bölümü
Tez Danışmanı: Dr. Öğr. Üyesi Zafer ARICAN
Ağustos, 2019 77 sayfa

Tarım ve tarıma dayalı sektörlerde kalite ve verimlilik diğer birçok sektörde olduğu gibi büyük öneme sahiptir. Tarım teknolojilerinin gelişmesi ile bu alanda yapılan çalışmalar da hız kazanmıştır. Küresel ısınma gibi faktörler, tarım ve tarım ürünlerini ciddi şekilde etkilemeye başlamıştır.

Tarımda meyve ve sebze hastalıklarının incelenmesi genellikle çıplak gözle yapılır. Bu hem zor hem de verimsiz bir yöntemdir. Hastalıkların tespit edilmesi için standarda ihtiyaç vardır.

Tarım ürünleri içinde patates, maliyet, birim alandan yüksek verim, yüksek besin değeri, sindirim kolaylığı, geniş kullanım alanı ve kolay yetişebilmesi özellikleri ile önemli bir yere sahiptir. Patates hastalıkları kalite ve verime zarar verir ve dolayısıyla bu durum üretici ve işletme kuruluşlarının kazancını azaltır.

Görüntü analizi ve makine öğrenme teknolojisi günümüzde yaygın olarak her alanda kullanılmaktadır. Bu tekniklerin tarıma ve tarıma dayalı endüstriler de kullanılması verimliliği arttırmakta ve hızlı çözümler sunmaktadır. Meyve ve sebzelerdeki hastalıkların tespit edilmesi için görüntü analizi ve makine öğrenme teknolojisinin kullanılması ile ilgili birçok çalışma yapılmıştır ve bu konuda çalışmalar halen devam etmektedir.

Bu çalışmada patateslerde yaygın olarak görülen üç hastalık için görüntü işleme teknolojisi kullanılarak hastalık tespiti ve sınıflandırması yapılmıştır. Önerilen yöntem %96 başarı oranı ile endüstriyel alanda kullanılabileceğini göstermiştir.

Anahtar Kelimeler: Görüntü işleme, destek vektör makineleri, patates dış hastalık tespiti ve sınıflandırılması

ABSTRACT

CLASSIFICATION OF MULTIPLE SKIN DISEASES USING POTATO IMAGES

KIZIL, Seda

MSc in Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Zafer ARICAN

August 2019, 77 pages

Quality and productivity have great importance in agriculture and agriculture-based sectors as in many other sectors. With the development of agricultural technologies, studies in this field have increased. Agriculture and agricultural products have been affected by for various reasons such as global warming severely.

Examination of fruit and vegetable diseases in agriculture are usually performed by the naked eye. This is both difficult and inefficient. Detection of diseases needs a standardization.

Potato is an important good among agricultural products with it cost advantage, high yield per unit area, high nutritional value, ease of digestion, wide area of use and easy to grow nature. However, potato diseases are harmful to quality and yield and reduce the earnings of producers and processing organizations.

Image analysis and machine learning methods are widely used in every field. The use of these techniques in agriculture and agriculture-based industries increases productivity and provides fast solutions. Many researches have been done to detect diseases in fruits and vegetables using machine learning technologies and there exist various ongoing studies.

In this study, disease detection and classification has been done for three common diseases in potatoes using image processing techniques. The proposed method has shown that it can be used in the industrial field with its success rate of 96%.

Key Words: Image processing, support vector machine, potato external disease detection and classification

ACKNOWLEDGEMENTS

Foremost, I would like to thank Konya Food and Agriculture University for accepting me to study in the Department of Computer Engineering. I thank my advisor Assist. Prof. Dr. Zafer ARICAN for the permanent support of my study and research. I am grateful for the devotion he showed during the research and writing of this thesis. I would like to thank Assoc. Prof. Dr. Şenol Zafer ERDOĞAN for his support. Lastly, I sincerely thank my family and all my friends for various kinds of encouragement and support.

TEXT OF OATH

I declare and honestly confirm that my study titled “Classification of Multiple Skin Diseases using Potato Images” and presented as Master’s Thesis has been written without applying to any assistance inconsistent with scientific ethics and traditions and all sources I have benefited from are listed in bibliography and I have benefited from these sources by means of making references.

23 /08/ 2019

Name SURNAME: Seda KIZIL

Signature:

YEMİN METNİ

Yüksek Lisans olarak sunduğum "Patates Görüntüleri Kullanarak Çoklu Dış Hastalıklarının Sınıflandırılması" adlı çalışmanın tarafımdan bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yararlandığım eserlerin bibliyografide gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

23/08/2019

Adı SOYADI: Seda KIZIL

İmza:

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations	Explanation
ANN	Artificial Neural Network
GSVM	Gauss Support Vector Machine
HOG	Histogram of Oriented Gradients
ID3	Decision Tree Algorithm
KNN	K- Nearest Neighborhood
LBP	Local Binary Pattern
LR	Logistic Regression
LSVM	Linear Support Vector Machine
MSVM	Multi Class Support Vector Machine
NN	Neural Network
OTSU	OTSU Algorithm
RBF	Radial Basis Function
RF	Random Forest
RGB	Red Green Blue
SURF	Speeded Up Robust Features
SVM	Support Vector Machines
TUIK	Turkish Statistical Institute

CHAPTER 1

INTRODUCTION

One of the most basic needs of humankind today is food. As humanity continues to exist, the food industry will always remain crucial. The basis of the food industry is agriculture. Nowadays, due to reasons such as wrong sowing, wrong irrigation, wrong fertilization, the desired yield cannot be obtained from the products in agriculture. In addition, global warming has a negative impact on the agricultural sector. Changing climate conditions affect agricultural products and require farmers to take precautions.

Potato is one of the most important products among the agricultural products. Also, potato is one of the most consumed foods in Turkey and the world. Potato was first introduced in Turkey in 1870 and has become the most consumed product as of today and started to be used in many sectors. For example, it has become an indispensable product of fast food sector and restaurants. Increasing demand requires automation in potato plant facilities. However, classification and quality control procedures are carried out manually in potato plant facilities. Besides, the potato industry is struggling with a lot of problems, for example inconsistent sorting, grading, human errors, increasing labor costs and production waste. In order to overcome such problems, factories began to show interest in machine learning systems.

Demand for quality and safe food products is increasing rapidly. The need to set these standards accurately, quickly and objectively continues to grow. Image processing provides an automated and cost-effective solution to meet these needs. Thus, image analysis techniques have various application areas in food industry.

According to Omid et al. (2010), the use of image processing and machine learning technology in agriculture and food has recently earned importance. Image analysis and machine learning systems are used in order to increase the market value of products and to comply with quality control and standards. Image processing technology can be successfully carried out, thanks to its ability to objectively determines the visual features of foodstuffs such as color density, color distribution, visual defects, size and shape.

There are some difficulties in the products produced and exported in our country and the products imported and consumed from foreign countries (Velioğlu, 2010). The problems in the potato industry are as follows:

- standardization in production,
- inadequacies in quality control,
- analysis is time consuming and labor dependent,
- problems caused by human factor in both production and quality control
- even the devices used in simple analysis are imported devices with high costs.

Computer vision algorithms are often used to identify and classify vegetable and fruit diseases and to classify them according to their size and weight (Pérez et al., 2017).

There are many studies on this subject using image processing in the literature. These studies are presented in the next chapter. In the third chapter, information about potatoes and potato diseases is given. In the fourth chapter, overview of image and pattern classification techniques is given and the used algorithms are explained. In the fifth chapter, the proposed algorithm for classification of potato diseases and underlying methods which are the subject of this thesis are explained. In the sixth chapter, the experiments and their results are given to confirm this thesis. In the seventh chapter, the results and the future studies are mentioned.

,

CHAPTER 2

LITERATURE REVIEW

Image analysis and machine learning technology are used in every field today. Its use in agriculture and agriculture-based sectors is important. Image-based analysis has advantage both in reducing human error and fast and safe analysis of high numbers of incoming products.

Many studies have been conducted on the detection of fruit diseases in agriculture by using image analysis techniques. For this purpose, various studies on variety of fruits have been done. Apricots, cherries, apples, cotton and potatoes can be given as examples of such fruits. When the studies on disease detection and classification are examined, it is observed that single type of disease is generally detected and classification is made on two classes together with healthy class. In this proposed approach, three external diseases were identified and four classes were classified with the healthy class. This study is the first of its kind due to multiple disease detection and classification using potato images.

Studies on vegetables and fruits, leaves and potatoes were examined separately and presented under three different sections.

2.1. Identification and Classification of Diseases in Fruits

The machine vision system developed using color images of cotton crops is a good example to the studies using image processing techniques in agriculture. Camargo and Smith identified diseased regions from cotton crops (Camargo and Smith, 2009). The attributes were extracted to identify the unique patterns in the image. The extracted attributes were classified using the Support Vector Machine (SVM). Tests were performed to determine the best classification model. As a result of the research, when the machine vision system is fed with the appropriate information, diseases can be correctly identified.

Anami et al, (Anami et al., 2011) studied plant diseases. In the study, a feature extraction algorithm was developed considering the characteristics of the fruits. For example, fruits such as mango, pomegranate, peanut can easily be identified using color characteristics. However, when the studies were examined with fruit diseases, it was determined that the fruits would not be identified by using only color characteristics. Therefore, a two-stage feature extraction process was performed using both color and texture attributes. After determining the features, the classification process was made by using artificial neural networks. Figure 2.1 shows examples to the images of diseased and normal fruits.

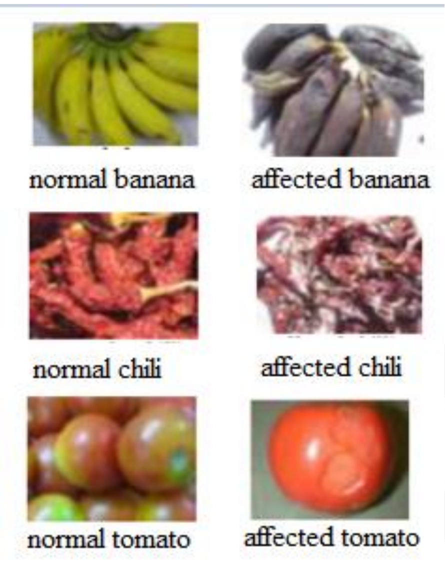


Figure 2.1: Diseased and normal images (Anami et al., 2011)

In Fruit Disease Detection and Classification (Sulakshana et al., 2017), apple diseases were identified. The approach has three basic steps namely image capture and pre-processing phase, feature extraction phase and classification phase. After image acquisition, unwanted background noise was removed, errors due to lighting and poor resolution have been resolved. Then K-Means clustering method was used for image segmentation. SVM was used for classification. Figure 2.2 shows the flow chart of the study.

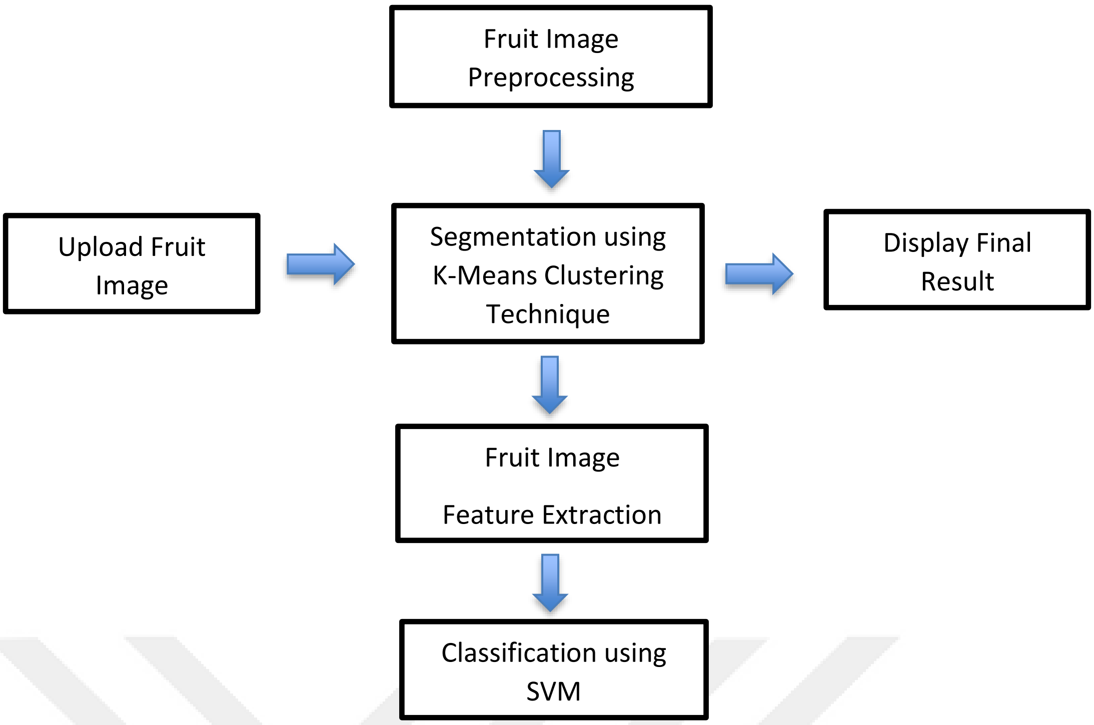
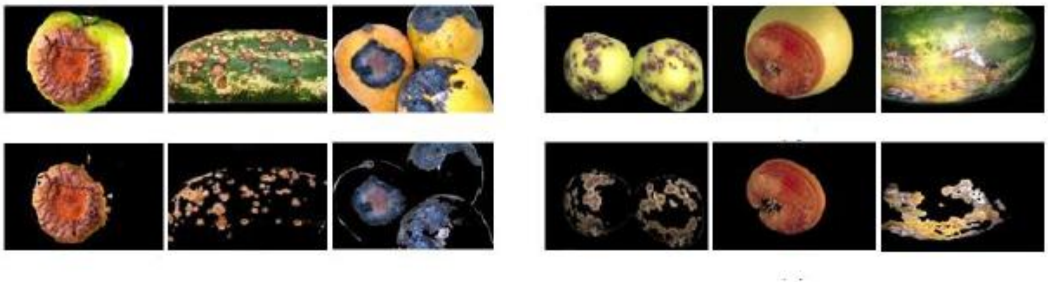


Figure 2.2: Proposed System Architecture (Sulakshana et al., 2017)

(Ranjit et al., 2016) to verify the proposed approach, collected 243 images for 34 disease classes from 10 different fruits, such as apple, banana, citrus, grapes. Each fruit species consists of different fruit diseases. K-Means and C-Means algorithms were used to detect stains and tissues caused by fruit diseases. (Figure 2.3) After the attributes were determined, the system was trained using the K-Nearest Neighborhood (KNN) classifier.

(a)



(b) (c)

Figure 2. 3: Identification and Classification of Fruit Diseases a) Sample images from dataset b) Image Segmentation using K-Means Clustering Algorithm (Before / After) c) Image segmentation using C-Means clustering algorithm (Before / After) (Ranjit et al,2016)

In (Dubey and Jalal, 2014) study, apple diseases are identified. In their study, classification was performed on four classes. Three common diseases of apple were identified in their studies. In the first stage of the study, segmentation was performed by using K-Means clustering technique. Local Binary Pattern (LBP) algorithm was used to extract features. Multiple SVM were used to classify the extracted attributes. As seen in Figure 2.4, although it was relatively successful to distinguish healthy apple samples, LBP features are not discriminative enough to classify apple diseases

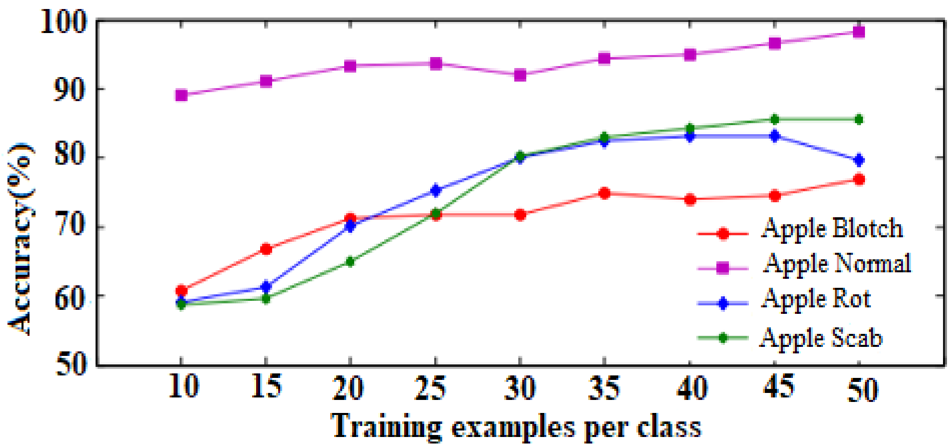


Figure 2.4: Accuracy per class for the LBP features in RGB color spaces using MSVM as a classifier (Dubey and Jalal, 2014)

2.2. Identification and Classification of Diseases on Leaves

In the study conducted by (Arivazhagan et al., 2013), diseases in plant leaves were identified. Various fruits and vegetables were used in the studies. These include beans, bananas, lemons, guava, breath, tomatoes and potatoes. Early blight and late blight with potato leaf disease were detected and 96% success was achieved using SVM classifier.

A system has been developed to detect diseases in grape leaves (Meunkaewjinda et al., 2008). In that a three-stage system stages were, first, grape leaf color segmentation, second, grape leaf disease segmentation and finally analysis and classification of diseases. The first step is a preprocessing module that divides irrelevant background into segments. After the preprocessing step, the attributes were removed and SVM was used for classification. As a result of the classification, the grape leaves were classified into three classes.

In another study (Mokhtar et al., 2015), tomato leaf diseases were identified. In the first step, real samples of diseased tomato leaves were collected and subjected to various pre-processing steps (image size adjustment and background noise removal). Then, characteristics of diseased areas in the leaf are extracted using Gabor wavelet transform. After the features were extracted, SVM were used for classification. Since the stains and tissues in the disease are similar, the SVM kernel model has been adopted to differentiate them. Cauchy, Invmult and Laplacian Kernel were used, and performances were compared. Finally, tomato leaves were divided into two classes as Powdery mildew or early blight. As a result of the experiments, they achieved 99.5% success.

In another study (Olsen et al., 2015), Situ leaf diseases were identified. Texture features were used to extract the features of the leaves. HOG algorithm was used to extract texture features. HOG algorithm provides a simple yet powerful approach to recognizing leaf texture. Four machine learning methods were used to classify leaf images. Logistic regression (LR), linear support vector machines (LSVM), Gauss support vector machines (GSVM) and ANN were used. The best results were obtained with Gaussian support vector machines with 94.72%.

In (Al-Hiary et al., 2011), Al- Hiary et al., studied leaf diseases and classification of these diseases. K-Means algorithm was used as the clustering algorithm. ANN were used to classify using texture feature sets. Their proposed method has been tested on five diseases affecting plants. (see Figure 2.5)

As a result, the study consists of three stages.

1) to identify the diseased area using the K-means clustering algorithm,

2) Color co-occurrence methodology was used to remove the tissue attributes affected by the disease on the leaf, and

3) ANN were used to determine and classify the disease.

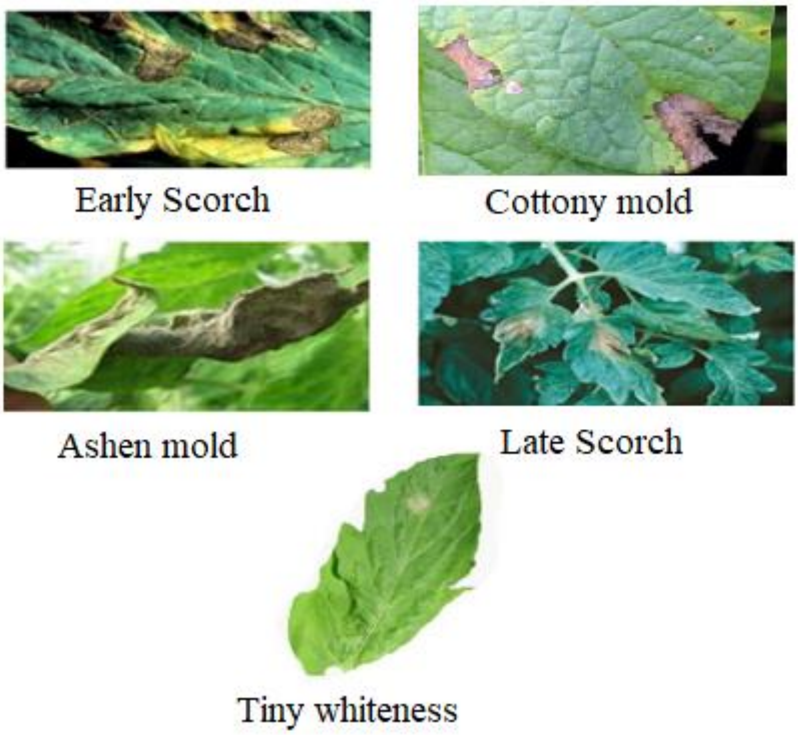


Figure 2.5: Leaf disease classes (Al-Hiary et al., 2011)

In (Sannakki et al., 2013) grape leaf diseases were identified in their study. In the first step, background noise was removed and all images were resized to the same size. Segmentation was performed using K-Means algorithm and ANN was used for classification. As a result of the classification, the grape leaves were successfully divided into two classes.

In (Pujari et al., 2016), plant diseases were identified. Two different classifiers namely SVM and ANN were used. Color and texture features on the leaf were used to detect plant diseases. The proposed method has been tested on six diseases. Namely; fungal, bacterial, viral, nematode, deficiency and normal (unaffected). A total of 900 pictures, 150 of which were from each class, were studied. It shows that SVM provides better classification accuracy compared to ANN (SVM: 92.17%, ANN: 87.48%).

2.3. Identification and Classification of Diseases on Potatoes

When the studies in the field of potatoes are examined, it has been observed that there are not enough studies about detection and classification of potato diseases. In the studies generally, a disease was detected. As a result of the classification, it was classified into two classes; such as Healthy and Unhealthy.

Common Scab of potato diseases affects the skin quality of potato tubers and reduces product quality. Product quality decrease has a significant effect on potato prices. (Nieto et al., 2011) proposed an objective method for detecting "Common Scab" disease on potato tubers using a hyperspectral imaging system tested by various experiments. Potatoes were obtained from some potato packaging companies harvested in 2009 in Spain, which belong to various degrees of Common Scab disease. Potato images were obtained in 320 spectral images (320 x 240 pixels) which were transferred to hyperspectral cubes in 900 nm and 1700 nm wavelength. The segmentation has 4 steps. First, the image was binarized using the Otsu algorithm, which calculates the optimal pairing threshold. In the second step, noises were removed by Gaussian blurring method and in the third step Blob analysis was applied. In the blob analysis, since the block with the largest area (except the background) is known to be a potato, this blob is selected, and the mask used to segment all images in the hyperspectral cube is created. SVM and RF (Random Forest) were used for classification. SVM (97.1%) algorithm performed better than RF (95.8%) algorithm. In terms of the use of hyperspectral images this study is unique. However, the identification and classification of a single potato disease is its weakness.

Segmentation was performed by Islam et al., (2017) on leaf images in order to reveal the characteristics of potato leaves. Disease detection and classification images belong to 3 disease classes namely, Late blight, Early blight and Healthy as shown in Figure 2.6. After the features were removed, the diseases were classified using multi-class SVM.

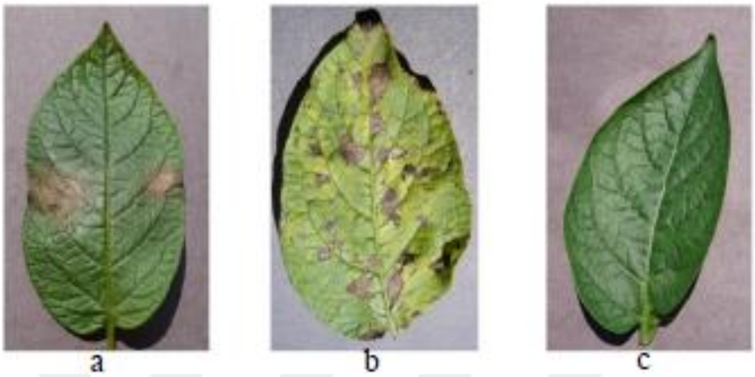


Figure 2.6: (a) Late Blight affected (b) Early Blight affected (c) Healthy (Islam et al., 2017)

(Razmjooy et al., 2012) proposed a hierarchical classification method for the quality assessment of faulty potatoes. In the developed system, sorting process was made according to the size of potatoes. The results proved that SVM was successful.

In a study proposed by (Samanta et al., 2012) external diseases in potato were detected. The images obtained were collected from different potato fields. Then, image segmentation was performed to identify the disease regions. After the segmentation, the attributes were grouped with the K-Means clustering algorithm. The disease was determined by histogram of the images.

2.4. Aim of the Thesis

Vegetable, fruit and plant leaf diseases detection using machine learning algorithms have been investigated. In these studies, one or more diseases were detected and classified. In studies with potatoes it has been observed that only one disease was detected. In addition, healthy potato class was added and classification was two classes.

The aim of this thesis is to identify and classify multiple diseases. The subject has not been seen that in the literature with potato diseases. To accomplish this, three common diseases of potatoes have been identified. These are BlackScurf, CommonScab and PowderyScab. The aim of the study is to develop a four-class classification algorithm. It is aimed to identify and classify diseases with one healthy and three diseased classes.

As a result of literature research, it is aimed to use SVM which gives the best results in the classification of vegetable and fruit diseases. As a result of multiple external disease classification in potato images, it is aimed to be used in industry with high success.

CHAPTER 3
POTATO AND POTATO DISEASES

3.1. General Information about Potato

Potato is the most commonly consumed food product after wheat, rice and corn in the world. Potatoes are important for humans as nutrition thanks to the protein, carbohydrate, minerals and vitamins it contains. Changing needs accelerated the research on agriculture. Especially today, changing climate and weather conditions have led people to do more research on agriculture (Çetiner, 2016).

Many crops today were brought to Europe by Columbus and his friends after the discovery of the New World in 1492. However, potatoes were not included in these products. The reason for this was that potatoes could not be discovered by the Spaniards until 1532, as they grew in the cool temperate climate of the Andes of South America. Until 1552, potatoes were not included in the literature. The exact time that potatoes arrive in Europe is not known (Hawkes and Ortega, 1993).

Potatoes are used almost all over the world thanks to its low price, high nutritional value, high yield per unit area, ease of digestion, wide usage area and growing in all kinds of climatic conditions. At the same time potatoes, spirits and starch are also important raw materials of the industry (Kargı, 2016).

The potato having approximately 150 years of history in Turkey, has become one of the most common agricultural products of the country. There is no precise information on how and where the potatoes brought to Turkey. According to Russian scientists, potatoes were first planted in the Sakarya River valley in the 1850s, near the Bosphorus and in Akova in the Adapazarı area. According to another source, came from Russia and the Caucasus to Turkey in the north in 1870 and it was grown in high plateaus in Eastern Anatolia and Black Sea region.

In parallel with agricultural production in Turkey, processing, marketing and consumption has become an important sector as well. Despite these developments, potato industry has not yet reached its full potential in Turkey (Çalışkan et al., 2010). According to the data of TUIK (Turkish Statistical Institute) in 2015, 2.312.234 tons of potato production is provided in Turkey from approximately 688.242 decares of farmland mostly growned in Niğde, Konya, Afyon, İzmir and Nevşehir (Table 3.1).

Table 3.1. Potato planted area and harvest amount for 2015 (Kargı, 2016)

City	Product	Year	Planted Area (Decare)	Harvested Area (Decare)	Production (Tons)	Yields (Kilogram)
Niğde	Potato	2015	227.466	227.466	674.773	2.966
Konya	Potato	2015	126.780	126.780	493.748	3.895
Afyon	Potato	2015	149.424	149.424	434.929	2.911
İzmir	Potato	2015	114.671	114.671	407.745	3.556
Nevşehir	Potato	2015	69.901	69.901	301.039	4.307

3.2. Types of Potato Diseases

Potato plants are exposed to many diseases and pests during their growth and development. Tuber yield varies significantly depending on the degree of this effect. When the diseases and pests are tackled in a timely manner and according to the technique, the yield will not be adversely affected as the possible damage will decrease significantly. Otherwise, potato diseases cause big losses. Potato roots and stems have various types of disease. Some of these bacteria are caused by soil structure and is effective in some of them. Table 3.2 shows 20 common diseases and their effects on potatoes.

In our study, Black Scurf, Common Scab and Powdery Scab diseases were identified. These three diseases are very dangerous due to their various effects and will be of great benefit if detected. For example, since Common Scab disease is an infectious disease, no crops should be planted within three years and seeds should not be separated from the collected soil.

Table 3.2: Summary of important aspects of 20 potato diseases and defects (Miller, 2006)

Disease	Location on Tuber	Spread within storage
CommonScab	External, general	No
PowderyScab	External, general	Yes
Rhizoctonia	External, general, must wash	No
SilverScurf	External, general, must wash	Yes
Bacterial soft rot	External, general; Internal, general	Yes
Blackleg	External, stem end; cut internal, stem end, and longitudinal	No
Early Blight	External, general; Internal, make shallow cuts through lesions	Yes
Freezing and Chilling	External; cut internal, stem end and cross section	No
Fusarium rot	External, general; Internal, cut through lesions	No
Late Blight	External, general; Internal, cut through lesions	Yes
Leak	External, general; Internal, cut around wounds and stem end	No
Mechanical Injury	External, general; Internal, cut through damage area	No
Pink rot	External, stem end eyes,lenticels; cut internal, turns pink	Yes
Ring rot	External, skin cracks; cut internal, near stem end	No
Root knot	External, general; internal, cut tangential	No
Black heart	Cut internal, longitudinal	No
Black spot	Cut internal, stem end half or on shoulder	No
Fusarium wilt	Cut internal, through stem end, only in xylem	No
Leafroll	Cut internal, cross section	No
Verticilium wilt	Cut internal, extends through vascular ring	No

3.2.1. Black Scurf

Rhizoctonia solani (Black Scurf) is a fungal disease of potatoes. Symptoms are root cancer, brown spots on potatoes, tuber growth disorders. Black or brown spots may be caused by soil or seed (Figure 3.1). Affected roots, stems and stolons show reddish brown necrotic patches called cankers. Root cancers cause plants to become stunted. The upper part of the leaves also causes rolling to have symptoms like blackleg.

Brown, slightly sunken lesions with distinct edges develop on the stem base and on stolons. Later a white collar can develop on the stems at soil level. The resulting pruning of the stem can lead to uneven emergence and gaps in crops (AHDB Potatoes, 2019).

Figure 3.1: BlackScurf diseased potato images (ResearchGate, 2019)

3.2.2. Common Scab

CommonScab disease in potato is caused by three types of viruses: S. scabies, S. acidstabs and S. turgidiscabies. Although these viruses are only observed in most potato fields and do not affect yield, they cause significant quality loss in tubers. This virus can be found in beets, radishes, rutabaga, turnips and parsnips, but these plants are rarely of economic importance. Symptoms of the diseases are limited to tubers which consist of lacquers and dark brown, circular or irregular lesions and are rough in tissue (Figure 3.2). The lesion of type depends on the type of potato, the tuber maturity in infection, the organic matter content of the soil, the strain of the pathogen and the environment (University of Massachusetts Amherts, 2013).

Figure 3.2: CommonScab diseased potato images (United Nations Economic Commission For Europe, 2019)

3.2.3. Powdery Scab

A soilborne pathogen called Spongospora subterranea causes PowderyScab, a disease that is one of the most important problems facing potato production in some regions. This virus is highly resistant to environmental stresses and can spread the disease on seed potatoes and in contaminated soil. In the last decade, important information has been gained about the biology of the pathogen and the epidemiology of the disease. However, there is no single effective control measure. When the risk of disease caused by the pathogen in the seed or soil can be accurately identified, effective dusty shell control can be improved (Merz and Falloon, 2009).

Figure 3.3: Powdery scab diseased potato images (Fera Science Ltd., 2019)

CHAPTER 4

IMAGE AND PATTERN CLASSIFICATION

Digital image processing can be used in all industrial applications. By using image processing techniques, digital image data can improve and object recognition can be achieved. Image processing consists of a series of operations. These operations start with the capture of the image and continue with the purpose of using different techniques. In parallel with the developing technology, the use of high-speed computers in image processing enables image processing to be used in many areas other than the existing areas (Samtaş and Gülesin, 2011).

After an image is digitally captured, multiple image processing methods can be used to extract features from it. Patterns are specific features of the image. Patterns obtained from the image should not be affected by rotation, shifting and scaling.

Pattern recognition is basically, the classification of data based on existing acquired knowledge or statistical information obtained from patterns and their representation. In a basic pattern recognition application, raw data is processed and converted to a form suitable for use by a machine. In the classification, a model is created using a set of training data. This model is tested with other data and assigned an appropriate class label.

Learning as part of classification problem can be defined as follows. It is to train a system using specific examples and adapt the results so that they can accurately classify when new data is unknown. The learning stage depends on how well the data supplied to the system performs. This is the most important step and it depends on the algorithms used during learning. The data sets are divided into two. One is the training set used to train the system and the other is the set used to test the trained system. The training set is used to create a model. It consists of a series of images used to train the system. The algorithms used during the training determine how the input images are associated with the output images. The system is trained by using training data sets with the specified algorithm. In general, 80% of data sets are used for training. Test data is used to test the system developed with training data. A set of data used to determine if the system is working correctly after training. 20% of the data in the data set is used for testing in general (sakilAnsari, 2019).

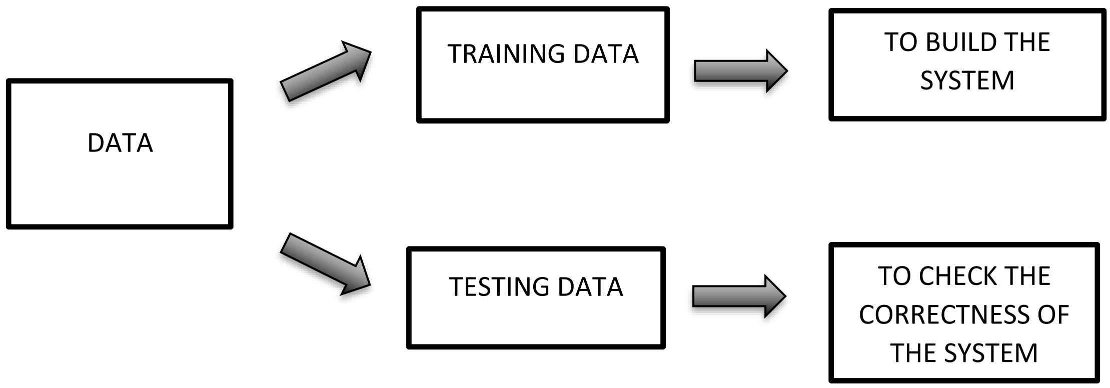


Figure 4.1: Pattern Recognition steps (sakilAnsari,2019)

4.1. Feature Extraction

Feature extraction is a low-level image processing method. Various features such as color, shape, size, texture can be extracted from the image. Feature extractor algorithms are used to extract these features used for object recognition and classification. Some of the most prominent feature extraction algorithms are; Speeded Up Robust Features (SURF), Histogram of Oriented Gradient (HOG), and Local Binary Pattern (LBP) (Naik and Patel, 2017).

Using feature descriptor algorithms, the image classification task has the following stages: Feature detection, feature description, and feature mapping.

- **Feature Detection**

The features extracted from the image are unique points and can be easily distinguished. For example, edges, stains, etc. Feature detection is to find whether a point can be selected as a property in particular. Feature detection is a fundamental step for successful image classification. That's why feature detectors are repeatable so that they can reliably find the same points of interest when viewing conditions change.

- **Feature Description**

After the features are detected, the region around each feature is removed and defined. Neighboring regions around each feature must be similarly defined. Feature identifiers are then calculated from each region. That is, a property vector is an n-dimensional vector that represents a feature based on some metrics in the display attributes.

- **Feature Matching**

Because property vectors represent objects, comparing two property vectors to perform image classification, two images are comparable to the similarity or difference. Essentially, there are two methods for comparing images: By measuring the distance between two property vectors or by measuring similarity. Such as, two features are compared by calculating the distance between two descriptors, the shorter the distance the greater the similarity. The commonly used distance measure is the Euclidean distance (Alwanin, 2014).

$$D\ (A,B)\ =\ \sqrt{\textstyle\sum_i (A(i)\ -\ B(i))^2} \tag{4.1}$$

4.1.1. Speeded Up Robust Features (SURF)

SURF (Bay et al., 2008) is a special algorithm used for the property descriptor. Basically, the SURF algorithm is used in image recording, object recognition, and classification applications. To find points of interest, SURF algorithm uses the Hessian blob detectors approximate integer value. Its property descriptor is based on the sum of the Haar wavelet response around the point of interest. Because of its invariance to scale, translation, illumination, contrast and rotation can easily find images taken in various conditions (Naik and Patel, 2017).

Some of the studies using SURF algorithm in food and agriculture are as follows. Pooja and Madival, (2016) used a SURF-based feature bag for the identification of food products. Pyramid compatible kernel approaches were used for classification and 86% classification rate was obtained. In the tissue analysis study conducted by Dailey et al. 2014, SVM was used as the classifier. As a result of the tests, 85% success was achieved for pineapple and 100% success for bitter melon. Another study using the SURF algorithm was done by Yogesh and Dubey in 2016. Defects in the fruit were identified in their proposed study. When the studies in this field are examined, SURF algorithm proved to be successful in disease detection.

4.1.2. Histogram of Oriented Gradients (HOG)

Histograms of Oriented Gradients (HOG) (Dalal et al., 2005) are feature descriptors used in object processing (Zulkifli, 2009). The HOG counts the occurrence of gradient orientation in local parts of the image. This method is like edge orientation histograms. It also scales the variable property transform identifiers and shape contexts. However, it is calculated in a dense grid of evenly spaced cells and uses overlapping local contrast normalization for improved accuracy.

The HOG algorithm is that the local object view and shape in an image can be defined by the distribution of intensity gradients or edge directions. For the application of these identifiers, the image is divided into small connected regions called cells.

A histogram of edge directions for each cell or edge orientations for pixels within the cell is generated. The union of these histograms then represents the identifier. For improved accuracy, a measure of intensity is calculated a larger region of the image to normalize local histograms. This value is used to normalize all the cells in the block, which are then called blocks. This normalization performs better in lighting changes or shade. Because the HOG descriptor operates on regional cells, the method does not affected by geometric and photometric transformations, excluding object orientation. The HOG concept is particularly attractive for classifications of leaves due to the nature of the regions with visible veins (Figure 4.2). These vessels are very lightening about the specific class that the leaf belongs to (Tsolakidis et al., 2014).

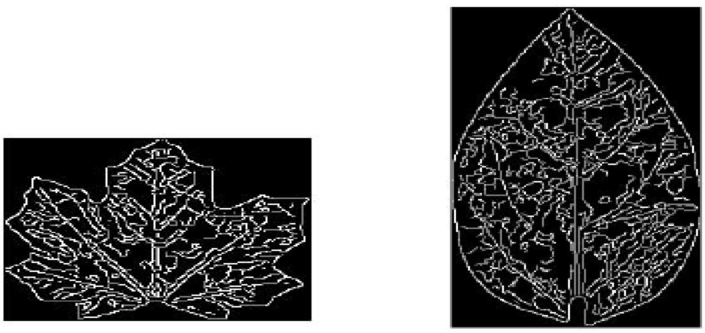


Figure 4.2: Veins of leaf (Tsolakidis et al., 2014)

In agriculture, the HOG algorithm is used in many areas for object preparation and recognition In 2014, proposed method (Tsolakidis et al., 2014) classified plant leaves which study HOG algorithm is combined with Zernike Moments. SVM was used for classification and Flavia and Swedish leaves were selected as data set.

In another study (Southgate, 2016) foods were divided into sections using the HOG algorithm. In other study (Olsen et al., 2015), HOG was used to extract traits from Situ leaves. The HOG attributes are calculated by taking the orientation histograms of the local zone edge density. Initially, edge gradients and orientations are computed for each pixel in the local region. The Sobel filters are used to achieve edge gradients and their orientations.

These local regions are divided into small areas called "cells". The size of each cell is 4 * 4 pixels. Histograms of 8 oriented edge gradients are calculated from each of these local cells. Then the total number of HOG attributes is 128 = 8 × (4 × 4) and creates a HOG property vector (Figure 4.3).

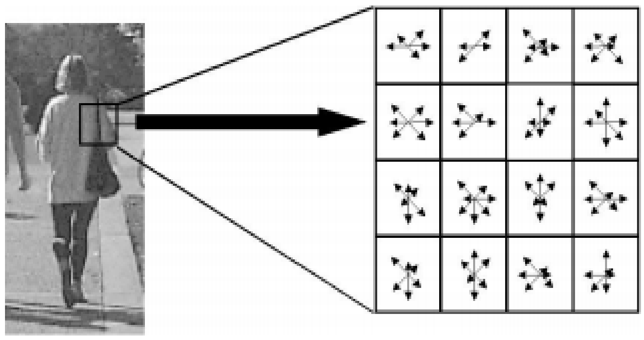


Figure 4.3: Extraction process of HOG feature (Kobayashi et al., 2007)

Minor changes to the windows position to refrain sudden changes in the identifier and less emphasis should be given to the distant gradients in the middle of the identifier. A Gauss weight function with an σ value equal to half the width of the identifier window is used to weigh the size of each pixel. The features extracted using the HOG algorithm are resistant to local geometric and photometric transformations with its histogram. Local spatial bins crowds tolerance to rotation and translation (Kobayashi et al., 2007).

4.1.3. Local Binary Pattern (LBP)

The LBP approach is a non-parametric method and a local structure formed by comparing neighboring pixels in the image with each other. This method, which has given very strong results in tissue analysis, is an effective method used in face analysis, image and video identification, environmental modeling, image tracking, motion analysis, biomedical and satellite images proposed by Ojala et al., the LBP algorithm is texture recognition very successful. This method is based on the assignment of binary values as a result of comparison of the image by selecting the pixel value in the middle of 3 × 3 windows as the threshold level. The generated binary number sequence is called the LBP code, which makes it possible to identify different types of attributes in the image. For example, edges, corners, light or dark areas, line regions. However, operators of 3 × 3 may not be effective to capture dominant features. Therefore, the length of the operator should be determined or assigned adaptively according to the problem. Instead, LBP definitions are made circularly. There are two parameters to be defined here. These are determined as the number of sample points P and the radius measure R of the symmetrical circular neighborhood.

LBP operator $LBP_{P,R}$ is created from the neighboring set of P pixels 2^P different output value is defined. In addition, if the binary LBP code obtained as a result of LBP has two or fewer passes, this is called uniform LBP. For example; 00000000, 00111000, 11100001 are uniform LBP codes. The LBP histogram also contains information about the distribution of local micro-patterns within the image (such as edges, bright areas, etc.). Thus, the characteristic of the image can be defined statistically (Kızrak, 2014).

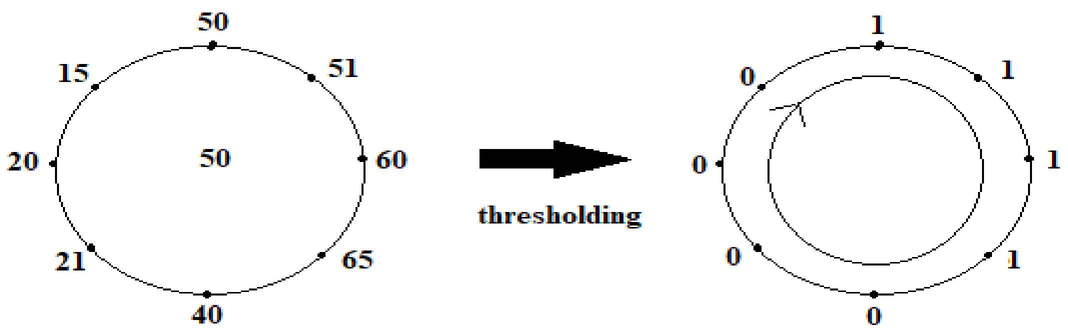


Figure 4.4: LBP Algorithm (Kızrak, 2014)

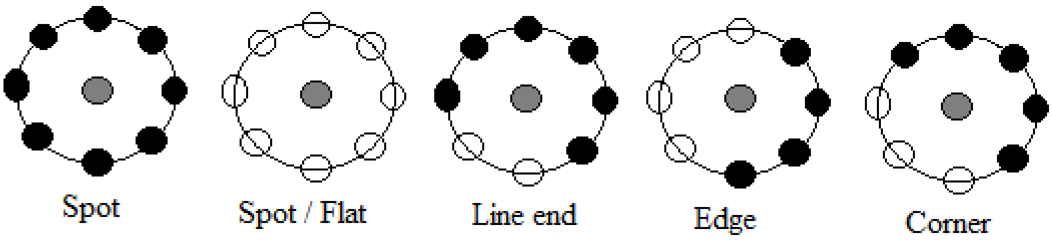


Figure 4.5: Meaning of LBP patterns in the image (Shan et al., 2009)

The LBP histogram computes micro-patterns over the entire image without location information. The image shape information is evenly distributed over small regions such as $R_0, R_1, \ldots\ldots R_m$ that the LBP histogram produces. The characteristics of the global shape and local texture of the image are subtracted from this histogram. There is a good trade-off between feature vector length and recognition performance (Kızrak, 2014).

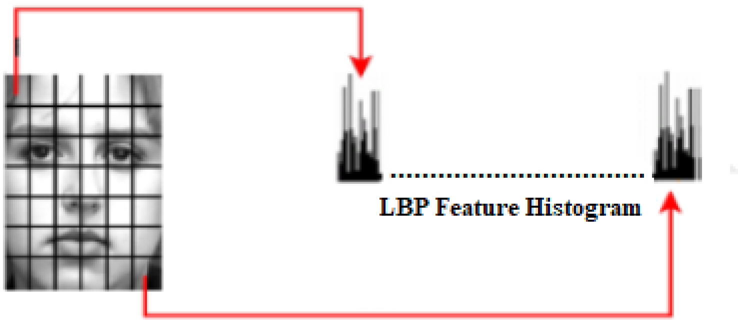


Figure 4.6: Obtaining histogram with LBP (Shan et al., 2009)

4.2. Bag of Visual Words Algorithm

The bag-of-words model is a most common method used in for object classification. The main idea of the algorithm is to measure every key point extracted from the image with visual words. Then each image is represented by a histogram of visual words. For this purpose, a clustering algorithm (i.e. k-means) to generate the visual words is used.

The concept of a “A visual bag of words” follows the same principle as "bag of words". The word bag model is a way of displaying text vectorially when modeling text with machine learning algorithms. The word order and position are not considered when doing this. Words are represented as a collection of key points in the text. Again, regardless of the order of these keywords, documents that share many same words are considered relevant to each other. Same approach applied to object categorization (Zhang et al., 2015).

Figure 4.7: Local Patches Identified from an image (Shah, 2012)

Key points are usually vectors in a higher dimensional space that is represented by local patches (Figure 4.7, Figure 4.8).

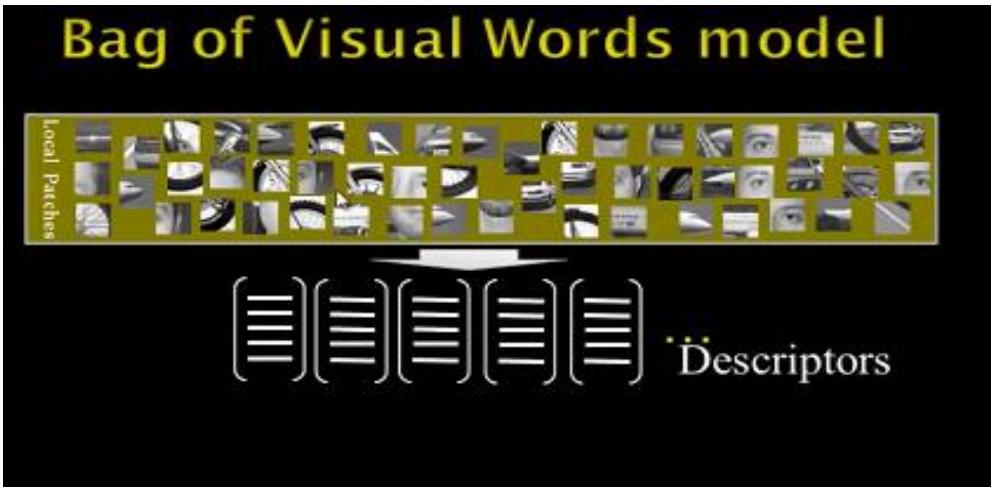


Figure 4.8: Key points

To efficiently handle these key points, each extracted key points is quantized into bag of visual words. Afterwards, each image is represented by a histogram of visual words (Figure 4.9).

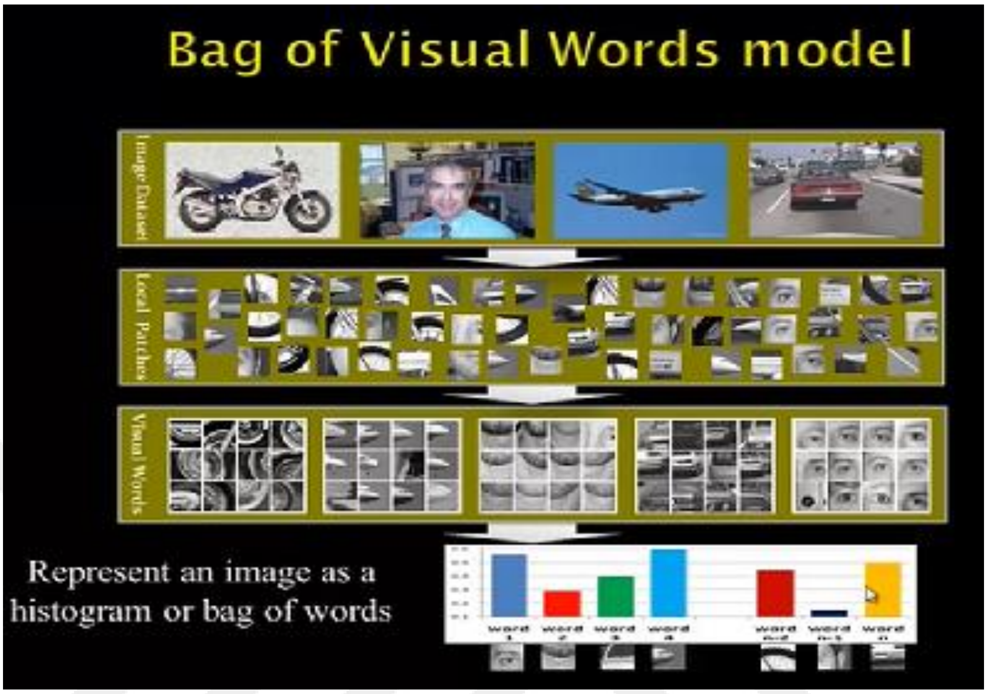


Figure 4.9: Portray an image as a histogram or bag of words

This is often referred to as word bag representation. And as a result, it transforms the object classification problem into a text classification problem. A clustering procedure (e.g., K-means) is often applied to group key points from all the training images into many clusters, with the center of each cluster corresponding to a different visual word (Figure 4.10).

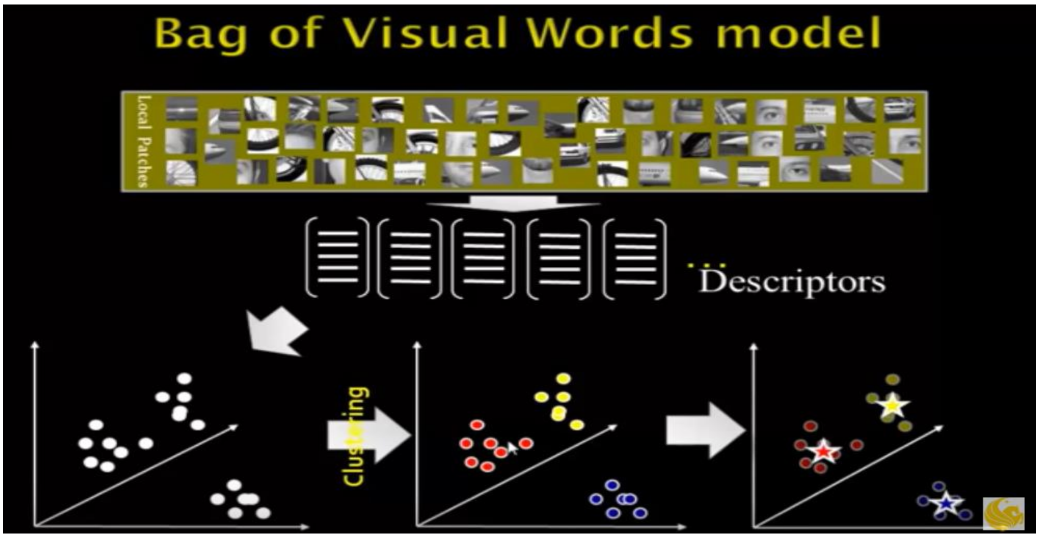


Figure 4.10: Bag of Words Clustering Procedure

4. 2. 1. K-Means Clustering Algorithm

The simplest form of clustering is to divide a specific set of clusters into discrete subsets and thus is a split clustering that aims to optimize specific clustering criteria. The most commonly used criterion calculates the square distance from the corresponding cluster center for each point and then the clustering error criterion that sums those distances for all points in the dataset. The K-Means algorithm is a popular clustering method that minimizes clustering failure. However, it is a local search procedure and its performance is largely due to initial starting conditions. Clustering is traditionally seen as an unsupervised method for data analysis (Likas et al., 2001).

K-means clustering is a widely used method for automatically dividing a dataset into K groups. K advances by selecting the first set center and then repeating them as follows:

1. Each instance d_i is assigned to the nearest cluster center.
2. Each cluster center has been updated to the average of C_j, the founding instances.

The algorithm joins when there is no other change in assigning instances to clusters. Clusters are initialized using randomly selected instances from the dataset. The datasets used can consist of numeric attributes or symbolic attributes. For numerical attributes, a Euclid distance measure is used, and the Hamming distance for symbolic attributes is calculated (Wagstaff and Rogers, 2001).

4.3. Classification Algorithms

The classification is to distribute data between the various classes on a data set. Classification algorithms learn this distribution pattern from the given training set and then try to classify it correctly on test data in which the class is not specific. The values that specify these classes on the dataset are called labels and are used to determine the class of data during both training and testing (Şeker, 2013).

Classification is generally carried out in accordance with a classification model. A classification model can be a function for which the target variable is calculated with the help of estimation variables. The relationship between the estimated variables and the target variables is determined by experiments on the data set. The data used in a classification process is basically separated from two: the first part of the data is the training data and the other is the test data.

The classification model was trained with training data and tested with test data. The classification success obtained in the test data gives the accuracy of the model. The training set contains the records from which the classification model will be obtained. The test set is used to determine the accuracy of a classification model. Approximately 70% of the analyzed data were used for training purposes and 30% for testing. However, sometimes these rates can be determined as 60% training data, 20% verification data, and 20% test data. The same data is used as both training and test data when the amount of data available is low. This method is called k-fold cross validation. The most commonly used method for K-fold cross validation is 10-fold cross validation, where 90% of the data is taken as training data and 10% of the data is taken as test data at each stage. After this process is done 10 times the average is given (Takcı, 2018).

The main classification algorithms are shown in Figure 4.11; Random Forest algorithm, ID3 (Decision Tree) Algorithm, K-Nearest Neighbor Algorithm, Naive Bayes Algorithm, SVM, ANN

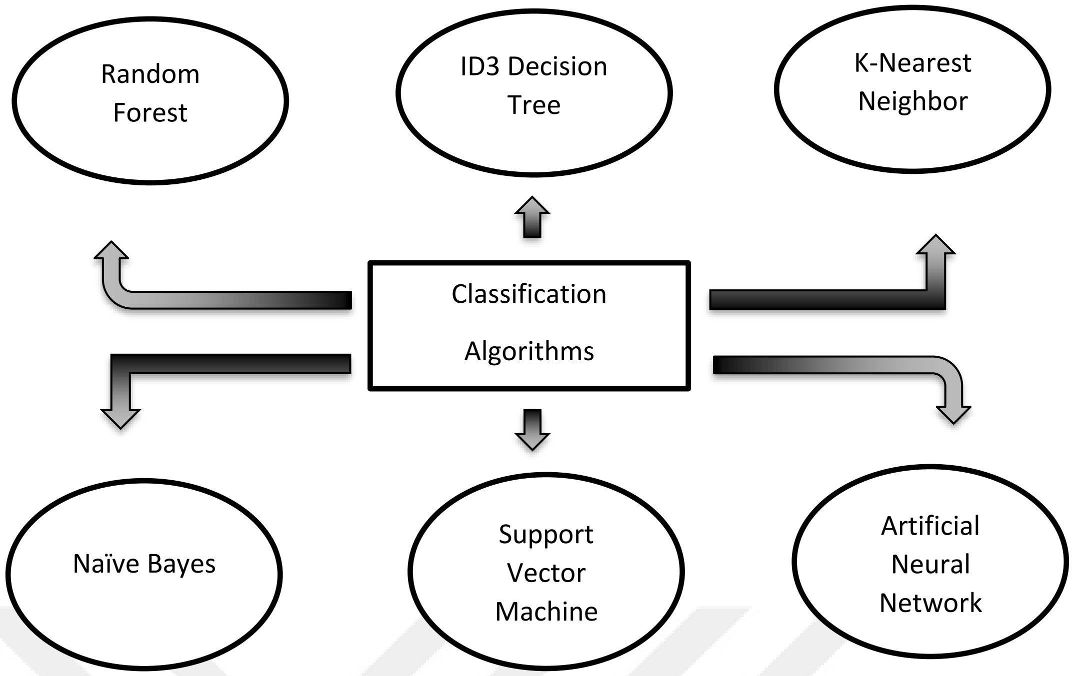


Figure 4.11: Classification Algorithm

Table 4.1: Advantages and Disadvantages of Classification Algorithm (Nikam, 2015)

ALGORITHM	FEATURES	LIMITATIONS
C 4.5 Algorithm	• Build Models can be easily interpreted. • Easy to implement • Can use both discrete and continuous values • Deals with noise.	• Small variation in data can lead to different decision trees. • Does not work very well on a small training data set. • Overfitting
ID3 Algorithm	• It produces the more accuracy result than the C4.5 algorithm. • Detection rate is increase and space consumption is reduced.	• Requires large searching time. • Sometimes it may generate very long rules which are very hard to prune. • Requires large amount of memory to store tree.
K- Nearest Neighbor Algorithm	• Classes need not be linearly separable. Algorithm • Zero cost of the learning process. • Sometimes it is Robust about noisy training data. • Well suited for multimodal classes.	• Time to find the nearest Neighbors in a large training data set can be excessive. • It is Sensitive to noisy or irrelevant attributes. • Performance of algorithm depends on the number of dimensions used.

Naive Bayes Algorithm	• Simple to implement. Algorithm • Great Computational efficiency and classification rate • It predicts accurate results for most of the classification and prediction problems.	• The precision of algorithm decreases if the amount of data is less. • For obtaining good results it requires a very large number of records.
Support Vector Machine Algorithm	• High accuracy. machine Algorithm • Work well even if data is not linearly separable in the base feature space.	• Speed and size requirement both in training and testing is more. • High complexity and extensive memory requirements for classification in many cases.
Artificial Neural Network Algorithm	• It is easy to use, with few parameters to adjust. • A neural network learns and reprogramming is not needed. • Easy to implement. • Applicable to a wide range of problems in real life.	• Requires high processing time if neural network is large. • Difficult to know how many neurons and layers are necessary. • Learning can be slow.

4.3.1. Artificial Neural Networks

Artificial Neural Networks (ANN) are computer systems that can learn events using human generated samples (examples of real brain functions) and determine how to react to events from the environment. Like the functional characteristics of the human brain, they are successfully applied in subjects such as learning, association, classification, generalization, feature identification, and optimization. ANN creates their own experiences with the information they obtain from the examples and then makes similar decisions on similar subjects (Altaş and Gülpınar, 2012).

The ANN consists of artificial cells that are hierarchically connected to each other and can operate in parallel. The main task of an ANN is to determine an output set that can correspond to an input set shown to it. In order to do this, the network is given the ability to generalize by learning with examples of the relevant event. This generalization sets the output sets corresponding to similar events.

It is a decision-making tool and a calculation method that can be used very effectively especially when there is no information about events but there are examples (Öztemel, 2006).

ANN is widely used in cases where there is a nonlinear, multidimensional, noisy, complex, inaccurate, incomplete data and especially in the absence of a mathematical model and algorithm for solving the problem. As with other programs, the data is not embedded in a database or within the program. The information is stored on the network and is difficult to disclose and interpret. On the other hand, in order to operate the ANN safely, they need to be trained and tested for their performance. They can only work with numerical information when entering data. ANN can self-organize and learn. One of the important advantages is the ability to process uncertain, incomplete information. The most important disadvantages are that the best solution cannot be guaranteed, and the network behavior cannot be explained because the determination of the appropriate network structure for the problem is usually done by trial and error (Öztemel, 2006; Haykin, 2009).

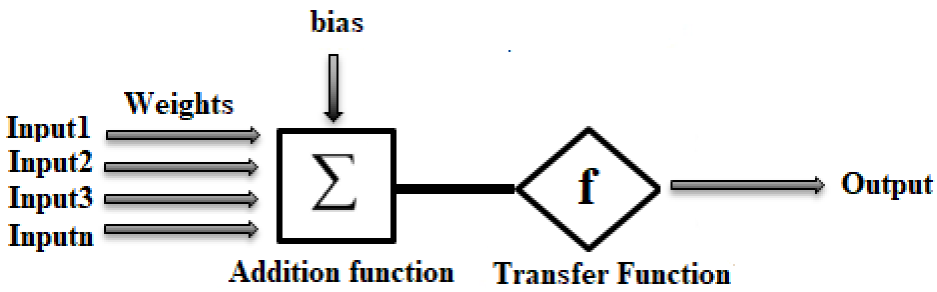


Figure 4.12: Basic ANN components

Figure 4.12 shows the simplified structure of the artificial nerve cell. Inputs show information from the outside world, weights indicate the importance of information coming to the cell and its effect on the artificial nerve cell. The sum function is used to calculate the net input to the cell. Bias is a constant and is also called the threshold value of the activation function. Finally, the transfer function performs the processing and output of the net input. In general, ANN are trained to produce an output corresponding to a input. Training of the network (adjustment of weight values) is performed by comparing the output with the target until the output of the network reaches the desired target. In its simplest definition, ANN are one of the best-known curve fitting techniques for complex mathematical connections. In the design of ANN, the transfer function and training function used are of great importance to obtain the best accuracy with the available data. The transfer functions frequently used in the design of ANN are given in Figure 4.13 (Velioğlu, 2010).

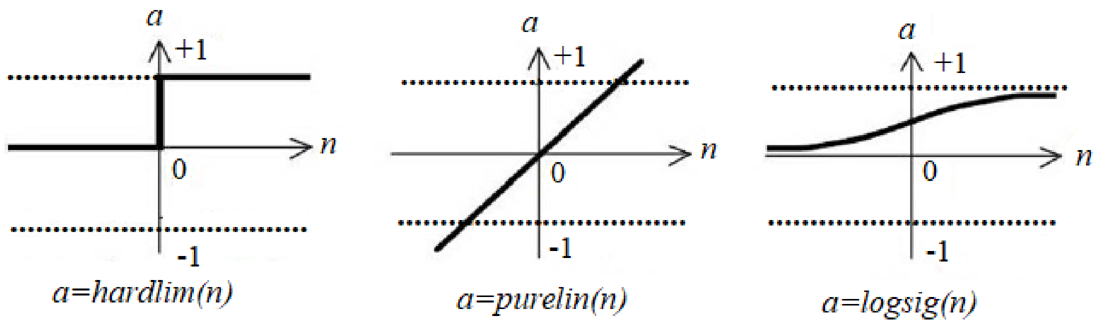


Figure 4.13: Transfer functions: (a) threshold value function, (b) linear function, (c) sigmoid function

Studies on ANN have started with single layer sensors. The most important feature of these sensors is that they divide the problem space into classes with a line or a plane. Once the inputs of the problem are multiplied by the weights and added, the class of the input is determined according to whether the value obtained is greater than or less than a threshold value. Classes are represented by the numbers 1 or 0. During learning, both the weights of the network and the weight value of the threshold unit are changed. The output of the threshold value unit is fixed and 1.

Multi-layer sensor networks most widely used ANN model, can solved many engineering problems (Figure 4.14). Multilayer networks have emerged as a result of studies to solve the XOR problem. These networks consist of 3 layers.

- **Input layer:** Retrieves information from the outside world.
- **Intermediate layers:** Process information from the input layer. It is possible to solve many problems with one intermediate layer. If the relationship between input / output of the problem to be learned by the network is not linear and the complexity increases, it can be used in more than one intermediate layer.
- **Output layer:** finds the output generated by the network for input from the input layer to the network by processing information from the intermediate layer. This output is transmitted to the outside world.

How many process elements should be in the input and output layers is determined by looking at the problem. There is no method to show the number of intermediate layers and the number of process elements in each intermediate layer. This is determined by trial and error. Each of the process elements in the input layer is depends on all the process elements. They depend on all the process elements in the output layer. The information flow proceeds from the input layer to the intermediate layer and from there to the output layer (Öztemel,2006).

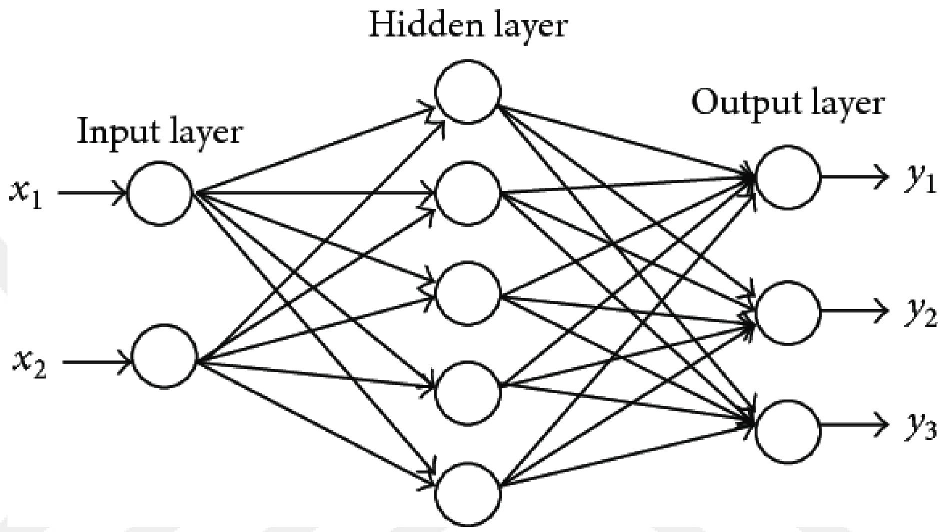


Figure 4.14: Multilayer Artificial Neural Networks

4.3.2. ID3 (Decision Tree) Algorithm

The decision tree approach is a method used to calculate approximate target functions and the learning function is represented by the decision tree. A decision tree is a descriptive and predictive model in the tree view. This model helps the decision maker decide which factors to consider when deciding and how each factor relates to the different outcomes of the decision in the past. Learning in decision trees, one of the most important classification tools algorithm is simple. The representation of self-knowledge is easily understood. Decision trees not only show decisions, but also include a description of decisions (Emel and Taşkın, 2005).

4.4. Support Vector Machines

4.4.1. Basic SVM Classification

Support vector machine is a non-parametric classification method based on statistical learning theory. SVM has been developed for binary classifications (Foody and Mathur, 2004). It is a method that uses the optimal algorithm to determine the boundary between classes in feature space. The method was originally designed for the classification of two-class linear data and then extented for the classification of multiclass and non-linear data. Basically, it is based on the principle of determining the hyperplane that can distinguish the two classes (Vapnik, 1995).

Advantages: (Ülgen, 2017)

- They are effective in high dimensional spaces.
- They are effective when there are less number of samples then the dimension of the feature space.
- The decision function uses limited training points. Therefore, it uses less memory.
- Versatile: Many different kernel functions can be used for the decision function.

SVMs are divided into two groups according to the linear separation and non-separation status of the data set.

4.4.2. Linear Support Vector Machines

In cases where samples of two classes are distributed linearly, SVM can be used for classification. In such a case, it is aimed to separate these two classes with the help of a decision function obtained by using educational data. The line that divides the data set into two is called the decision line. Although it is possible to draw infinite decision lines, it is important to determine the optimal decision line (Figure 4.15) (Ülgen, 2017).

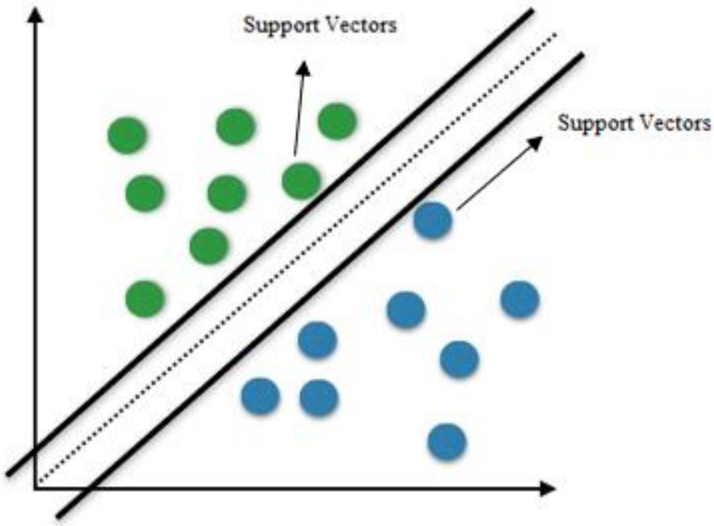


Figure 4.15: Support Vectors

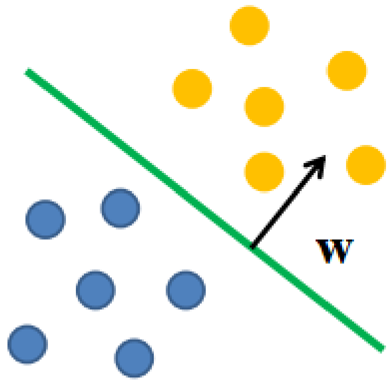


Figure 4.16: Linear SVM Two Linear Separable Classes W is weight vector

If we denote each feature point in feature space as X_i, $Y_i \epsilon \{1, -1\}$ the output representing the class to which the samples belong, and p $X \epsilon a^{P}$, high dimensional input vector; (X_i, Y_i) when a set of n volumes of pairs of S is given, it will best distinguish the samples from different classes,

$$w.x + b \tag{4.2}$$

It is the machine learning algorithm belonging to the class of supervised learning algorithms which helps to find linear hyperplane (Soman et al., 2011). Here, w, the normal of the hyperplane is also defined as the weight vector and b is constant.

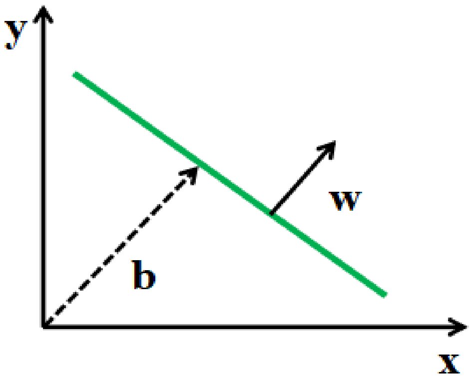


Figure 4.17: b constant value and w weight vector

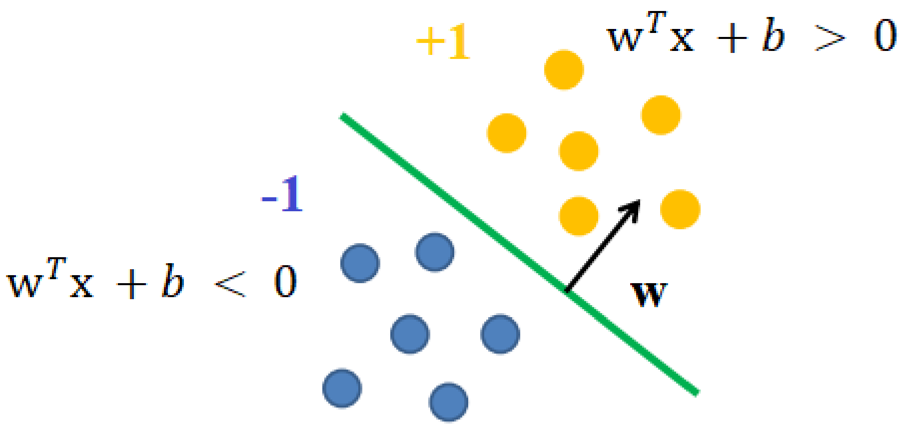


Figure 4.18: Class labels

The plane passing through the middle of the boundary planes and equidistant from both planes is defined as hyper plane. In Figure 4.18, (-1, +1) shows the class labels, *w* is the weight vector (normal of the hyperplane) and b is the constant value (Osuna et al., 1997).

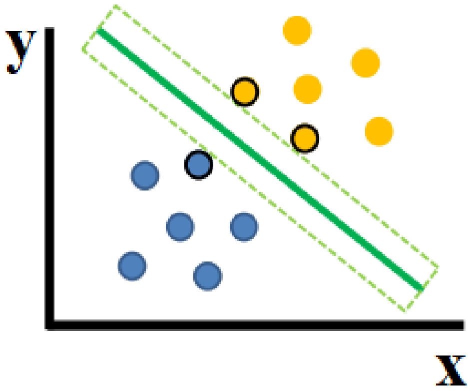


Figure 4.19: Boundary Planes

The planes on which the support vectors (Figure 4.19) are located and indicated by dashed lines are called boundary planes.

If the training data set can be separated linearly, the DVM tries to find the separation hyperplane with the largest boundary. In order to find this separation hyperplane, all the samples in the data set must provide the following inequalities (Soman et al., 2011).

$$f(x_i) = (w, x_i\ ) +\ b\ >= +1\ , y_i\ = +1 \tag{4.3}$$

$$f(x_i) = (w, x_i)\ +\ b\ <= -1\ ,\ y_i\ = -1 \tag{4.4}$$

Inequalities in (4.3) and (4.4) can be combined into a single inequality, as given in below equation.

$$y_i\ ((w, x_i)\ +\ b\ )\ - 1\ >=\ 0 \tag{4.5}$$

w is the norm of the normal plane w called the weight vector. Therefore, the distance to the hyperplane of the samples closest to the hyperplane must be equal to the norm of the weight vector that it has (Gunn, 1998; Schölkopf and Smola, 2002). Based on this theorem, the hyperplane that best separates the training examples is the plane that minimizes the following equation.

$$A\ (w)\ =\ \frac{1}{2}\ \|w\|^2 \tag{4.6}$$

Solving the optimization in below equation gives an optimal separation hyperplane is obtained to maximize the distance between the support vectors of the classes (Gunn, 1998; Cortes and Vapnik, 1995).

$$min\frac{1}{2}\|w\|^2 \quad \text{and} \quad y_i\,((w, x_i) + b) - 1 >= 0 \qquad (4.7)$$

4.4.3. Nonlinear Support Vector Machines

In practical applications however, it is not possible to linearly separate a dataset from the hyperplane most of the time. Nonlinear SVM are algorithms used when the data set cannot be separated by a linear function either fully or with a specific error. Therefore, the separation of classes is possible by estimating the separation curve. However, in practice it is difficult to estimate the curve. The geometric representation of the non-linear separation of the data set is given in Figure 4.20. (Ayhan and Erdoğmuş, 2014) In this case, the p-dimensional input vector x must be converted to the P-dimensional feature vector Φ. (Cortes and Vapnik, 1995).

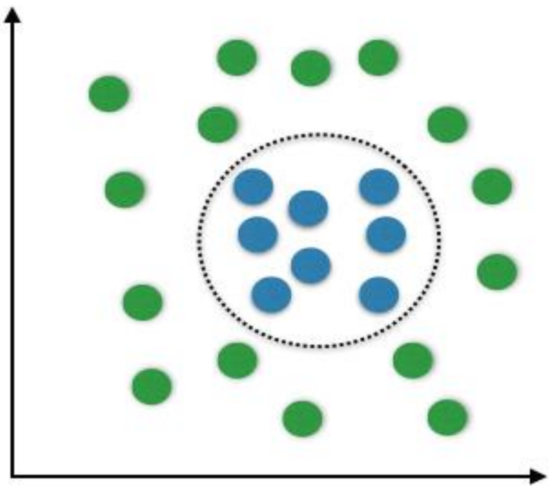


Figure 4.20: Two Class Problem of Linear Separation Case

For this conversion, the optimal separation plane must be defined in the property space. To achieve this, nonlinear mapping approach is used. Nonlinear mapping is an approach used to realize the linear separation of the original input space x into a higher dimensional *F* property space, a Hilbert space. Hilbert space is expressed as full inner product spaces with positive scalar products and elements of functions. The geometric explanation of the nonlinear mapping approach for a two-dimensional non-linearly separated data set is given in Figure 4.21. With the nonlinear mapping approach, the two-dimensional data set was moved to the three-dimensional property space to provide linear separation of the data set (Ayhan and Erdoğmuş, 2014).

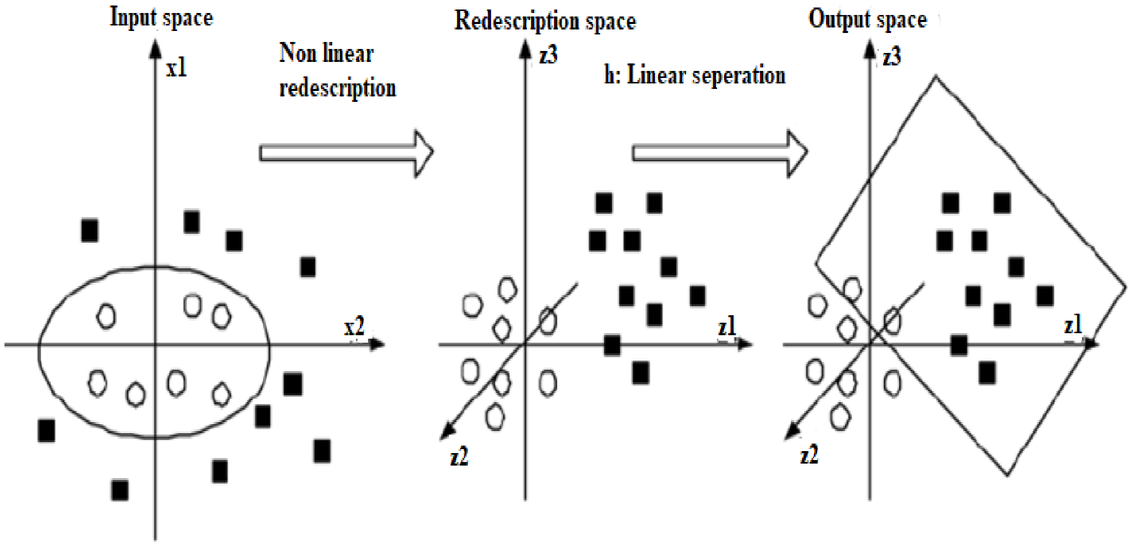


Figure 4.21: Example of Nonlinear Mapping Approach

In Figure 4.21, the property space of the functions, the two-dimensional input vector (X_1, X_2) and the three-dimensional property space (Z_1, Z_2, Z_3), is shown by the following equation.

$$A: R^2 -> R^3 \quad (x_{1,}\, x_{2,}) -> (z_{1,} z_{2,}, z_3) = (x_1^2, \sqrt{2\, x_1, x_2}\,, x_2^2) \tag{4.8}$$

Inner products of input vectors mapped in property space are obtained as in below equation (Soman et al., 2011).

$$= (x_2^2\, x'^2_2 + 2\, x_1\, x'_1\, x_2\, x'_2 + x_2^2\, x'^2_2)^{\,2} \tag{4.9}$$

Thus, the mapping process is carried out by elevating the data set from two-dimensional space to three-dimensional space (Soman et al., 2011).

As a result, the classifier decision function for the nonlinear SVM, depending on the separation hyperplane defined in the property space, is expressed by the following equation (Ülgen, 2017).

$$f(x) = sign\,((w', x_1) + b') = sign\,(\textstyle\sum_{i=1}^{a} y_1\, a_1 < Q(x_1, Q(x_1) >) \tag{4.10}$$

This mapping is called kernel. In statistical learning methods such as SVM, which kernel is used for which problem is an important problem. Appropriate kernel functions can lead to a significant increase in the generalization capability of the developed learning system (Barla et al., 2003).

The kernel function is an inner product in feature space, and is usually denoted as:

$$K\ (x, y)\ = <\ Q(x), Q(y) > \qquad (4.11)$$

In 4.11 K denotes the kernel function and Q denotes a non-linear mapping function.

Using a Kernel function, the algorithm can then be moved into a higher-dimension space without explicitly mapping the input points into this space. It is straightforward to apply the kernel trick thanks to the original formulation of SVM that consist of dot products. Resulting classifier, a hyperplane in the high dimensional feature space may be non-linear in the original input space (Souza, 2014).

$$\sum_{i=1}^{n} w_i\ (z_i, x) + b \longrightarrow \sum_{i=1}^{n} w_i\ (Q(z_i), Q(x))\ + b \longrightarrow \sum_{i=1}^{n} w_i\ \ k\,(z_i, x) + b \qquad (4.12)$$

Inner product *inner product applied to a* *Kernel function*

(possibly non- linerar) mapping Q.

Some of the most commonly used kernel methods: (Ülgen, 2017)

- Polynomial Kernel
- Radial Basis Kernel
- Histogram Intersection Kernel
- Chi-Square Kernel

A Kernel function converts training data. Thus, a nonlinear decision surface is transformed into a higher dimensional linear equation. If property data can be converted to a higher dimension, it is easier to separate classes with a linear function. The way to do this is to use the kernels functions (Crowley, 2016) (Figure 4.22).

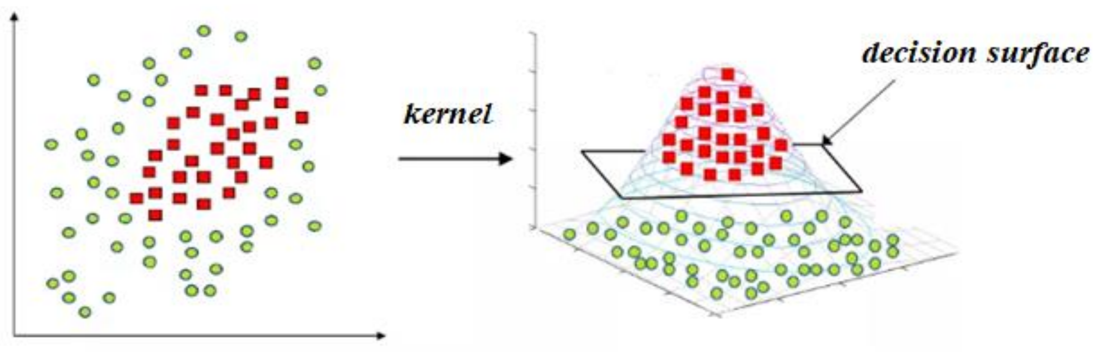


Figure 4.22: Separating classes using SVM kernel function (Ayan,2019)

4.4.3.1. Polynomial Kernel

The polynomial kernel is appropriate for problems where all training data is normalized. The corresponding kernel function is expressed by the following formula:

$$K\,(x, y) \;=\; (a\,x^{t}\,y \;+\; c)^{\,d} \tag{4.13}$$

Among the adjustable parameters, alpha is the slope, c is the term constant and d is the degree of polynomial (Güldoğan et al., 2017).

4.4.3.2. Radial Basis Kernel

It is seen that radial based kernel function is generally used in the applications in the literature because it gives better results. However, it is not known whether the radial-based kernel function is suitable for solving problems in every field. The performance of the kernel functions may vary according to data sets and problem. The Gaussian kernel is an example of a radial base function kernel. The corresponding kernel function is expressed by the following formula:

$$K\,(x, y) \;=\; exp\,(-\frac{\|x-y\|^2}{2a^2}) \tag{4.14}$$

As an alternative,

$$K\,(x, y) \;=\; exp\,(-\,y\,\|x - \;y\|^2) \tag{4.15}$$

is expressed by the formula.

In the case of overestimation, the exponential expression becomes approximately linear and high-dimensional projection begins to lose its non-linear structure. However, if the estimate is lower than the actual value, the function will not smooth and the decision limit will be extremely sensitive to noise values in the training data (Ayhan and Erdoğmuş, 2014).

4.4.3.3. Histogram Intersection Kernel

The histogram intersection algorithm uses the color information to know objects. The histogram intersection algorithm was suggested by Swain and Ballard its namely article “Color Indexing”. This algorithm is specially reliable when the color is a hard predictor of the object identity.

The histogram intersection does not need the accurate separation of the object from its background and it is robust to blocking objects in the foreground. A histogram is a graphical representation of the value distribution of a digital image. The value distribution is a way to represent the color appearance and in the HVS model it represents the saturation of a color. Histograms are invariant to translation and they change slowly under varied view angles, scales and in presence of occlusions.

Given the histogram *I* of the input image (camera frame) and the histogram *M* of the model (object frame), each one containing *n* bins, the intersection is defined as:

$$\sum_{j=1}^{n} min\,(I_j, M_j) \qquad (4.19)$$

The result of the intersection is the number of pixels from the model that have corresponding pixels of the same colors in the input image. To normalize the result between 0 and 1 divide it by the number of pixels in the model histogram:

$$\frac{\sum_{j=1}^{n} min(I_j, M_j)}{\sum_{j=1}^{n} M_j} \qquad (4.20)$$

When an unknown object image is given as input is compute the histogram intersection for all the stored models, the maximum value is the best match (Patacchiola, 2016)

In color images, the Histogram Intersection Kernel measures similarity between two color histograms. It is an algorithm unaffected by scale changes. It can be used successfully in non-segmented images as well (Barla et al., 2003). That intersection is larger when the two distribution are similar (Figure 4.23).

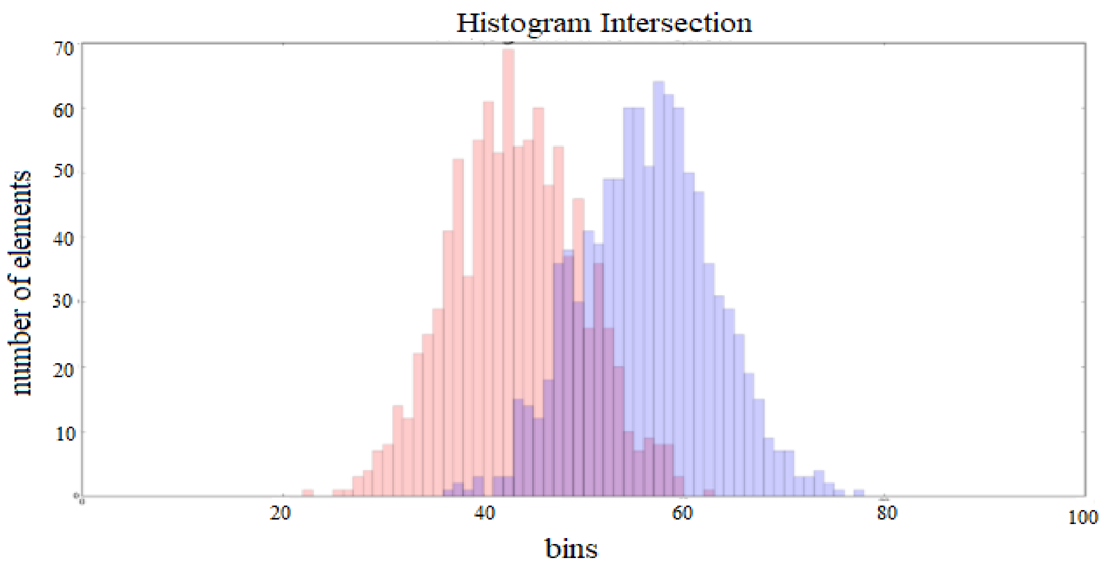


Figure 4.23: Histogram Intersection Kernel

4.4.3.4. Chi- Square Kernel

It is a kernel model commonly used in computer display applications. The Chi-Square Kernel (Maji et al., 2012) measures the distance between probability distributions. The Chi-Square Kernel is based on a comparison of the obtained values of the frequency of a class because it is divided by the expected frequency of the class. The chi-square Kernel is a commonly used algorithm for histogram-based image comparison, such as the Histogram kernel. Linear SVMs are preferred for efficiency reasons. However, non-linear kernels are used because they produce more accurate results in computer vision. Histogram Intersection Kernel and Chi-Square kernel kernels are commonly used to compare low-level histograms such as color and texture calculated on the image and to train the SVM classifier.

Chi- Squared Kernel:

$$K\,(x_i, x_j) \;=\; exp\,(-\,y\,\textstyle\sum_k \;\frac{(x_{ik} - x_{jk})^2}{x_{ik} + x_{jk}}) \qquad (4.21)$$

4.4.4. Multi Class SVM Classification

SVM are mainly used to separate data from two classes. In cases where there is more than one class Multi class SVM is used. When there are multiple classes, there are two types of approaches to distinguish between classes.

- Reduction of the problem into binary groups (One-to-one approach)
- Modeling the problem from a single group to all groups (One-to-all approach)

4.4.4.1. One-to-One Approach

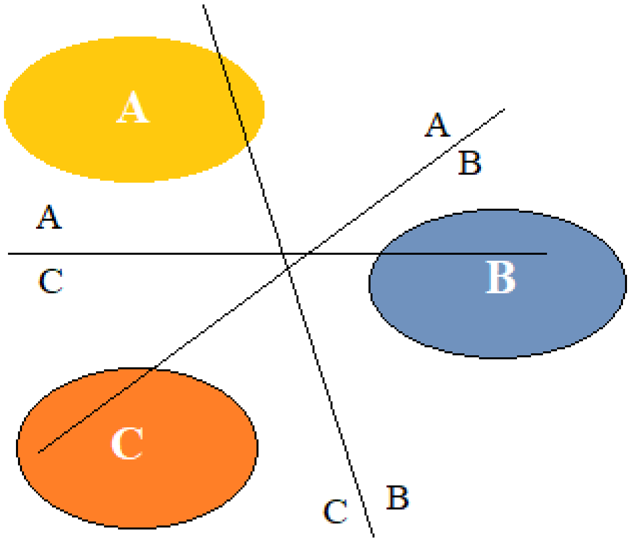


Figure 4.24: One- to- One Approach

In the above figure, a separate SVM was trained for both classes. During the classification of a new sample, each pair is questioned. That is, the AB, BC and AC pairs are questioned and whichever of these questions gets the most answers is put in this class.

Consider, for example, the results for the following queries:

AB-> A

BC-> B

AC-> A

In two of the above three queries, the result A was reached, so the new sample was considered closer to the A class.

4.4.4.2. One-to-Many Approach

According to the one-to-many approach used in multi-class classification, the problem turns into a "winner takes all". In this approach, a binary SVM classifier is trained for each class and check is such that the sample belongs to that class or not. To classify a test sample, this sample should belong to only one class. Below is the SVM extracted between three classes using this method.

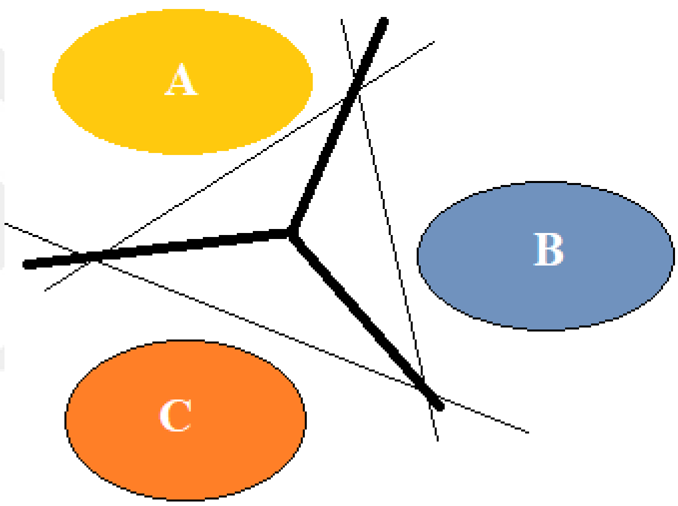


Figure 4.25: One-to-All Approach

As can be seen in the figure above, first three equations were found (fine lines in the figure). Then, by combining these line equations, a general SVM is used to differentiate the three classes (bold lines in the figure). For a sample to be marked as class A, SVM classifier for A vs others should select A, and SVM classifiers for B and C should select others.

CHAPTER 5

DISEASE CLASSIFICATION USING SUPPORT VECTOR MACHINES

Various scientific studies on the identification and classification of diseases in fruits and vegetables were examined using image analysis and machine learning in agriculture. These were mentioned in the literature section. The steps to be followed in this study are follows (Figure 5.1).

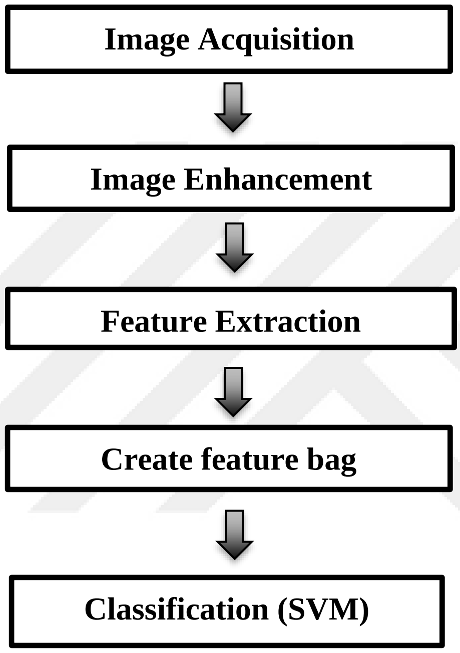


Figure 5.1: Steps in the classification of potato diseases

In this study, three diseases which are frequently observed in potatoes namely Common Scab (Streptomyces scabiei), Black Scurf (Rhizoctonia solani) and Powdery Scab (Spongospora subterranea) were examined.

The stains and tissues formed in these 3 diseases exhibit differences. In order to detect these stains and textures, feature extraction algorithms were applied, and classification was performed by using multiple class SVM. Figure 5.2 shows images of common diseases in potatoes.

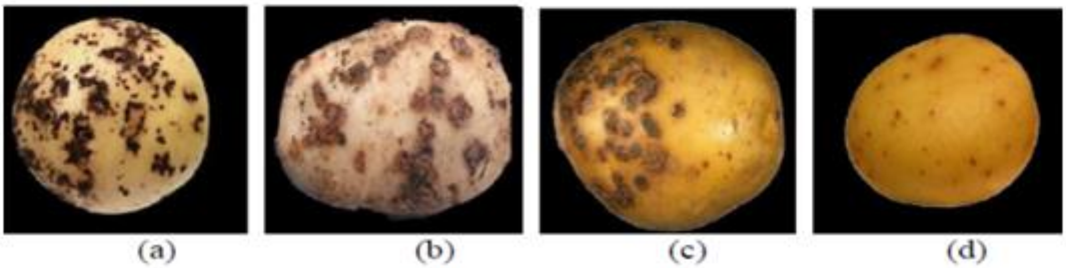


Figure 5.2: (a) Black scurf (b) Common Scab (c) Powdery Scab (d) Healthy potato

5.1. Image Acquisition

The images to be used in the study were obtained from previous studies that have worked with potatoes, from the website of the Turkish Ministry of Agriculture (Gıda Tarım ve Hayvancılık Bakanlığı, 2016) and the Internet. In this study, there are 4 classes including 3 diseased and 1 healthy. 100 images were obtained for each class. There are 400 images for a total of 4 classes. Since 400 images will not be enough for classification, data augmentation has been performed. For this purpose, each image is rotated at 60-degree angles to produce 6 different copies of an image. Thus, the total number of images was 2400. This augmentation by rotation also provides robustness to orientation changes. Figure 5.3 shows 60 degrees rotated examples of images of BlackScurf disease.

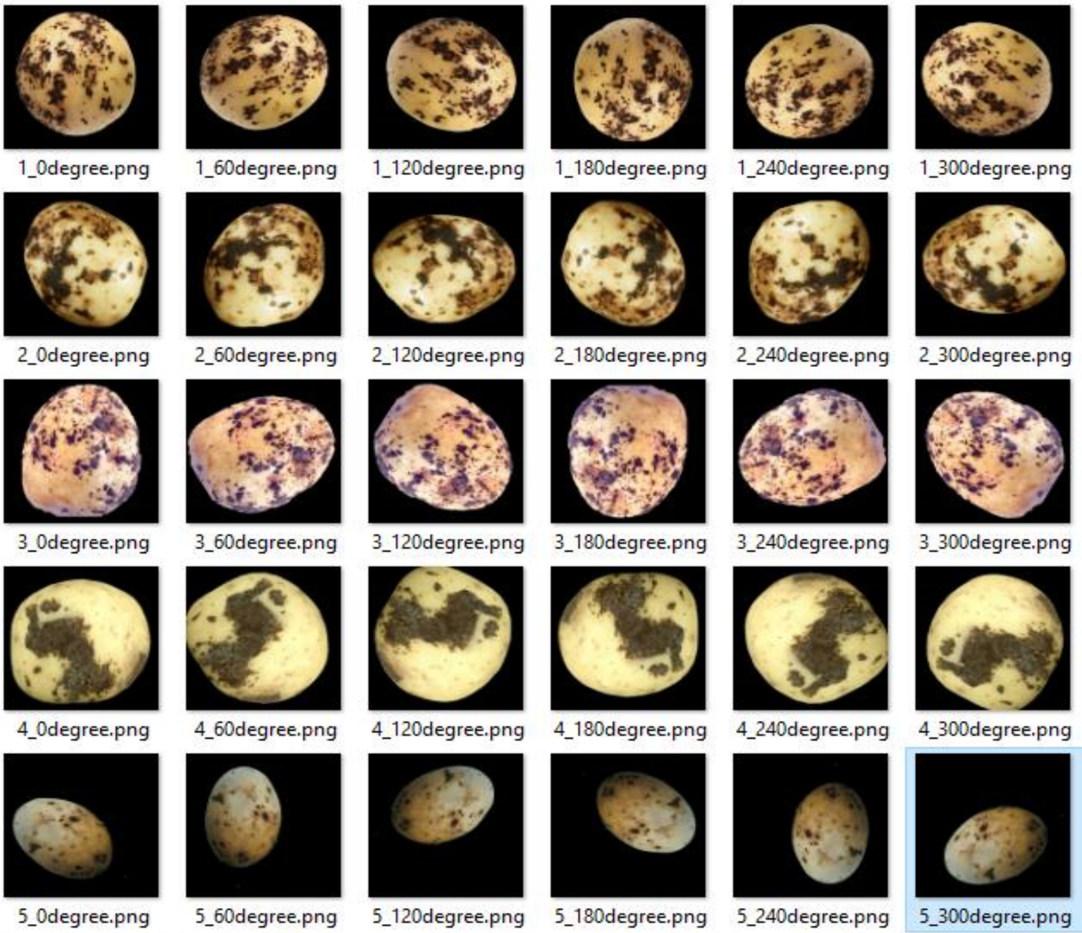


Figure 5.3: 60 degrees rotated samples of potato images of BlackScurf disease

5.2. Image Enhancement

After obtaining the images related to potatoes, image enhancement operations were performed. Since potato images were obtained from different sources, they all had to be converted to the same format. First, background noise was removed using OTSU algorithm (Açıl, 2017) and the images of the potatoes were distinguished. Then the images were resized to the same size.

5.2.1. Background separation

OTSU thresholding algorithm was used to subtract the potato image from the background. Threshold process is to make a black and white image according to a certain threshold value. In Otsu Threshold operation, this threshold value is determined automatically. As shown in Figure 5.4, a mask for the potato image is formed. The image was separated from the background by passing the potato through the mask.

Figure 5.4: Masked image of CommonScab disease

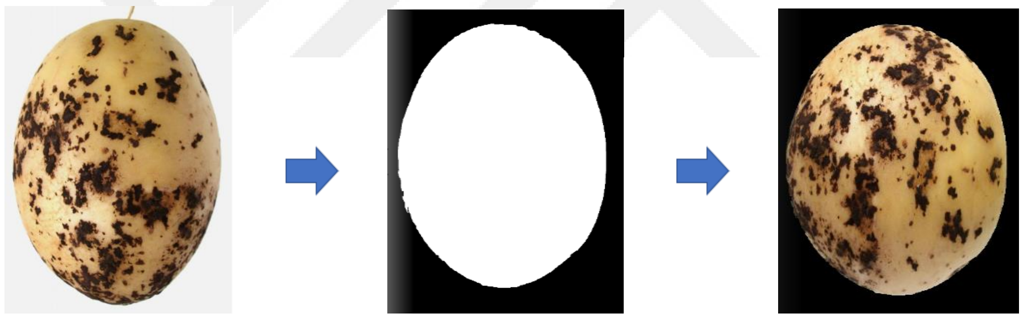


Figure 5.5: Masked image of Blackscurf disease

5.2.2. Resizing images

Since the pictures available on the internet are of different sizes, they are all brought to the same dimensions. For this purpose, it was sized in the dimensions of 400 × 400, 600 × 600, 800 × 800. The system was tested in all three dimensions. The best results are obtained on 600 × 600 images. The test results are mentioned in the Experiments and Results section.

5.3. Feature Extraction

Feature extraction is to identify unique patterns that may exist in the image. Feature extracting algorithms are used to detect identifiers in an image. The most commonly used are SURF (Speed up Robust Features) and HOG (Histogram of Oriented Gradients) algorithms. In order to apply the SURF and HOG algorithms to the images, the Accord.Net library [Accord.Net- Library] was used, which is commonly used for image processing in C# language.

Proposed system uses HOG feature and its compared to SURF. After the classification process, the success of the system, which was trained with the features extracted with HOG algorithm, was higher than SURF algorithm.

The performance of these algorithms affects the number of words selected during feature extraction. When the number of words is selected as 1500, the best performance is obtained from the algorithms. This behavior is explained in detail in next chapter.

After removing the features, classification process was performed by using SVM algorithm. As a result, when the features extracted using the HOG algorithm were classified using SVM, the performance of the system gave better results than the SURF algorithm. Comparative results of SURF and HOG algorithms will be given in Tests section.

5.4. Creating a feature bag using bag-of-words algorithm

Feature bag creation process is to group similar features after making meaningful features on the image. When the SURF and HOG algorithms are used to extract attributes, they can use a standard clustering algorithm based on Euclidean distances. Below are the images of BlackScurf (Figure 5.6) and Powdery scab (Figure 5.7) diseases, which are represented by using Bag of words algorithm.

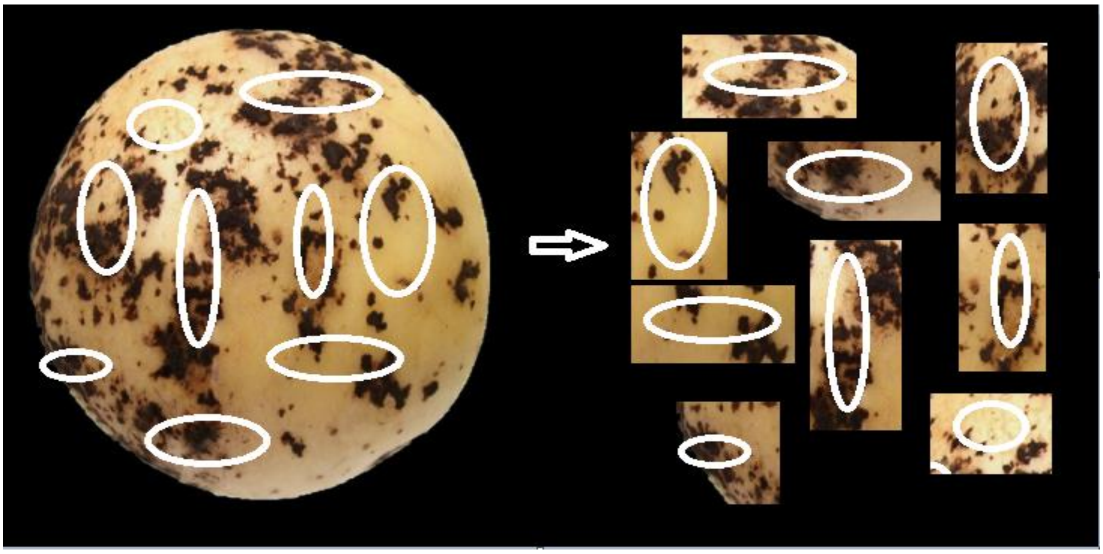


Figure 5.6: Black scurf disease image and Collection of Visual Words

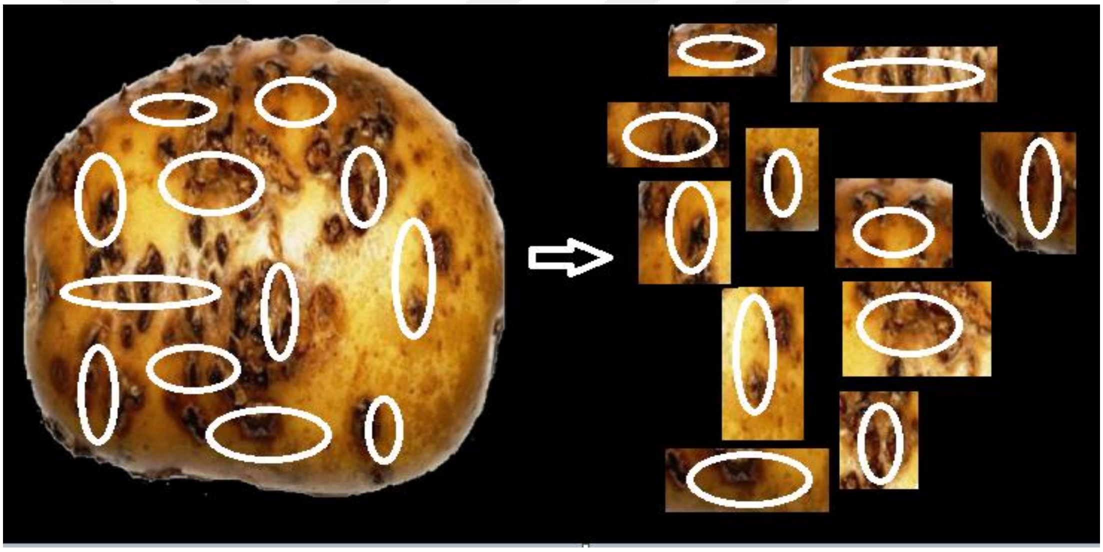


Figure 5.7: Powdery scab disease image and Collection of Visual Words

5.5. Multi-class Support Vector machine training and classification

In literature research in the second chapter, SVM has the best results in the classification of vegetable and fruit diseases. As a result of researches and experiments, it was decided to use SVM in this study. Because of there are multiple disease classes, one-to-all approaches were used for Multi-Class SVM classifier. In the classifier training, the SVM was trained by creating a histogram according to the frequency of the attributes in the attribute bag for all images.

SVM kernel method was used because the stain and tissue characteristics of the three common diseases were similar. Chi-Square and Histogram Intersection Kernel were used. The reason for using these two kernels are that they are widely used to compare low-level histograms, such as color and texture, calculated from images.

To verify the proposed classification method, various experiments are performed to test the success of the system. These are given in Chapter 6.

,

CHAPTER 6

EXPERIMENTS AND RESULTS

In this study, first, sample pictures for BlackScurf, CommonScab and PowderyScab diseases were collected. Potato pictures are obtained from some web sites that have been researched about potato cultivation, articles from national or international studies and documents prepared by the Turkey Ministry of Agriculture (Gıda Tarım ve Hayvancılık Bakanlığı, 2016). A total of 400 images were collected for four disease classes.

After the pictures are obtained, some pre-processing steps were performed. First, all pictures are resized to the same size (600 × 600). Background removal is performed to minimize the margin of error in feature extraction and classification after dimensioning. Then, the number of data was increased. To increase the number of data, each picture is rotated at 60 degrees to increase the total number of data to six times. After replication of the data, 2400 images for a total of 4 diseases; 1680 pictures (70%) were used for training and 720 pictures (30%) were used for testing. Aforge.Net library (Sert, 2017) is used in C# development environment to apply OTSU algorithm to image.

Various tests were performed to determine the accuracy of the developed system. To measure the effect of potato images on classification performance; 20 training data sets were created. Random images from 100 images were selected for each potato and 20 different training data sets were created.
The tests performed in this section are as follows.

1. Effect of number of images
2. Effect of number of features
3. SURF - HOG algorithms performance comparison
4. SVM kernel models performance measurement with HOG algorithm
5. SVM kernel models performance measurement with SURF algorithm
6. Classification success according to potato diseases
7. Comparison of sets according to classification results
8. Comparison of image sizes

6.1. Effect of number of images

In the first experiment, the effect of the number of images used on the success percentage was examined. Initially, 100 potato images were obtained for each disease. 25 of these images were used for the first test. A total of 100 potato images were tested.

Then, a total of 400 images were tested using the whole potato image. Last, the rotation was applied to the images. Images were rotated at 60, 120, 180, etc. degrees to increase the image.

The best classification results were obtained with 2400 images. During this test, HOG algorithm and Histogram Intersection Kernel was used. This test was made for 20 training sets and the results were obtained by averaging all of them. The following table (Table 6.1) shows the success percentages according to the number of images and the figure (Figure 6.1) plotted according to this table

Table 6.1: Effect of number of images

Number of images	**Success percentage**
100	40%
400	45%
800	55%
1200	60%
1600	75%
2000	85%
2400	**95%**
2800	90%
3200	80%
3600	75%

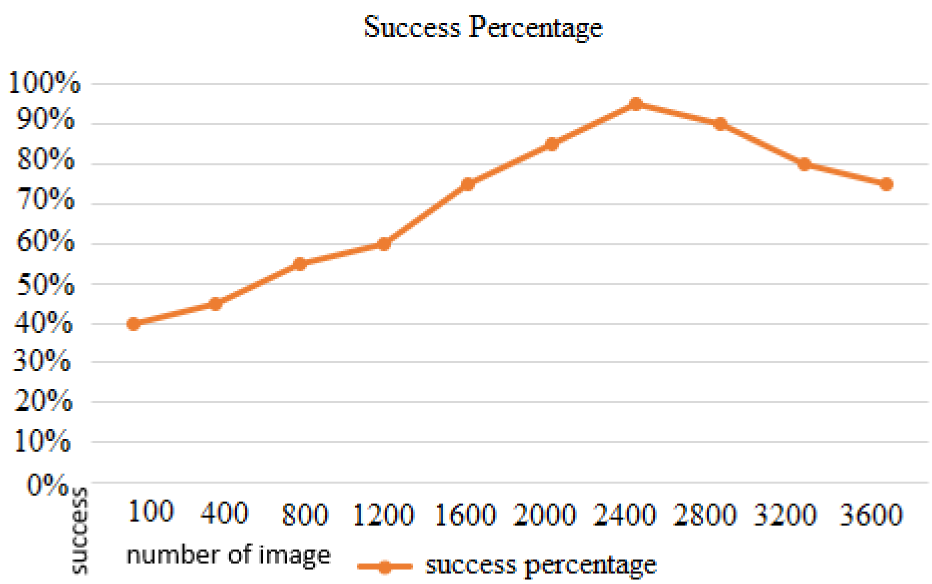


Figure 6.1: Graph drawn according to Table 6.1.

As can be seen from the above graph, when the number of potato images is low, the system does not provide enough learning, while the best results are obtained in 2400 images. When there are more than 2400 images, the success of the system has decreased due to excessive learning.

6.2. Effect of number of features

In the second experiment, the effect of the number of features on the performance of the proposed approach was investigated. Similar features extracted from the image are grouped using the Bag of words algorithm, while the number of attributes determined affects the classification performance. The number of attributes is used to measure the effect of the Bag of Words algorithm on the success of the system. As a result of the experiments, the best results were obtained when 1500 attributes were selected. During this test, HOG algorithm and Histogram Intersection Kernel was used. This test was made for 20 training sets and the results were obtained by averaging all of them. The following table (Table 6.2) shows the success percentages according to the number of features and the figure (Figure 6.2) plotted according to this table.

Table 6.2: Effect of number of features

Number of features	**Success percentage**
100	80%
200	82%
300	83%
400	85%
500	87%
600	89%
700	90%
800	90%
1000	91%
1200	92%
1500	**96%**
1800	95%
2000	95%
2500	94%
3000	94%
3500	93%
4000	90%
4500	88%
5000	85%

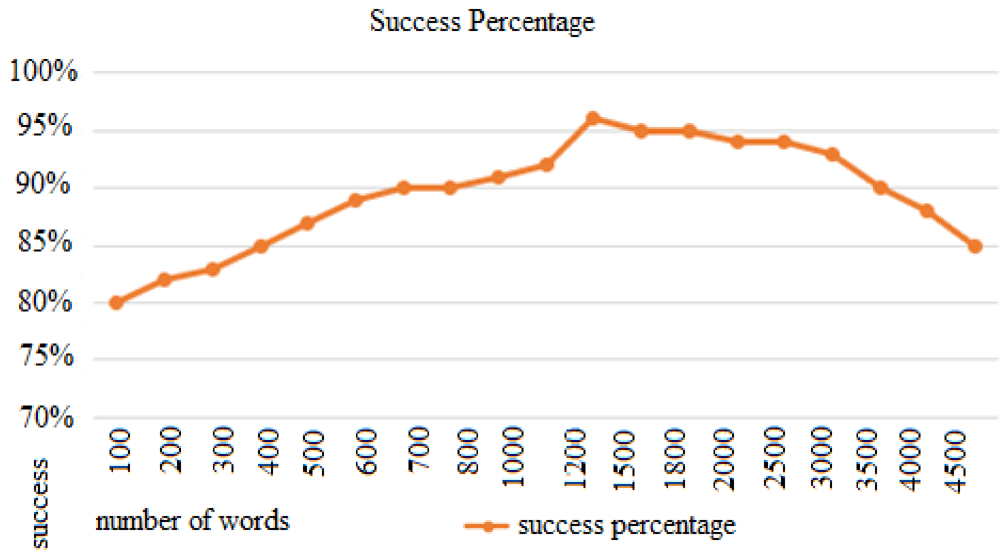


Figure 6.2: Graph drawn according to Table 6.2

As shown in the figure above, when the number of attributes to be grouped is selected less, the performance of the system is very low. When the number of attributes is increased, the performance of the system increases, but increasing after a certain value decreases the performance of the system.

6.3. SURF - HOG algorithms performance comparison

In the third experiment, the effect of feature extractor algorithms were investigated on classification performance. As shown in Figure 6.3, when the system is trained with the features extracted using the HOG algorithm, the success rate of classification is high. The reason that the HOG algorithm performs better than the SURF algorithm is that it can more accurately define tissue attributes on diseased potatoes with gradient directions. Since the SURF algorithm collects the gradient values, it loses some direction information. During this experiment, the number of features is 1500, the number of images is 2400, Histogram Intersection Kernel was used as SVM kernel model.

Table 6.3: Success percentages SURF- HOG algorithm

Set	SURF	HOG
Set1	84%	93%
Set2	85%	96%
Set3	83%	94%
Set4	83%	94%
Set5	83%	94%
Set6	88%	96%
Set7	86%	94%
Set8	85%	95%
Set9	84%	94%
Set10	88%	96%
Set11	83%	93%
Set12	85%	95%
Set13	90%	97%
Set14	87%	94%
Set15	88%	96%
Set16	83%	94%
Set17	84%	94%
Set18	83%	94%
Set19	88%	96%
Set20	90%	97%

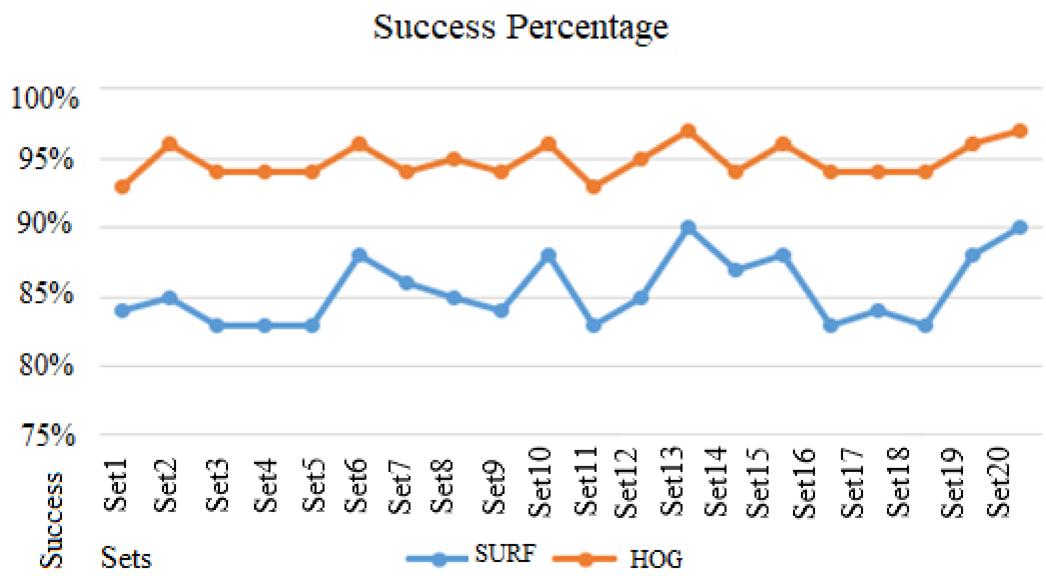


Figure 6.3: Graph drawn according to Table 6.3

6.4. SVM kernel models performance measurement with HOG algorithm

In the fourth experiment, the selected kernel function for the SVM was tested. Due to the similarity of the stains and textures on potatoes, a kernel model was needed due to the possible overlap of class sample regions between classes. Chi Square and Histogram Intersection Kernel performances were investigated. As shown in Figure 6.4, when the Histogram Intersection Kernel is used as a kernel model, classification success is high. During this test HOG algorithm was used. The number of features is 1500 and the number of images is 2400.

Table 6.4: SVM Kernel Models Performance with HOG algorithm

Set	Histogram Intersection Kernel	Chi-Square Kernel
Set1	93%	89%
Set2	96%	90%
Set3	94%	88%
Set4	92%	87%

Set5	94%	88%
Set6	96%	90%
Set7	94%	87%
Set8	95%	89%
Set9	94%	88%
Set10	96%	90%
Set11	93%	87%
Set12	95%	89%
Set13	97%	91%
Set14	94%	88%
Set15	96%	90%
Set16	94%	87%
Set17	94%	88%
Set18	94%	87%
Set19	96%	90%
Set20	97%	91%

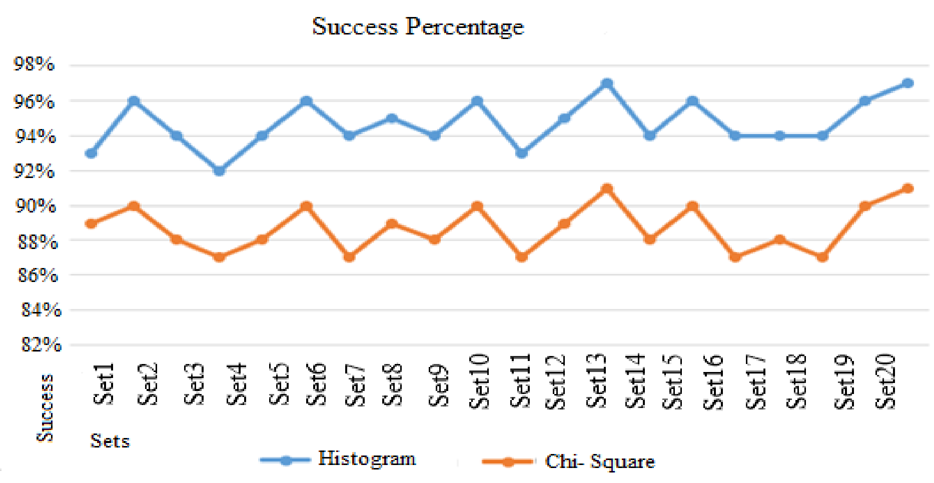


Figure 6.4: Graph drawn according to Table 6.4

The reason why the classification performance of the Histogram Intersection kernel is high, extract features using HOG algorithm. Because, since both approaches were histogram based, they supported each other and showed a harmonious performance.

6.5. SVM kernel models performance measurement with SURF algorithm

The difference from the fourth experiment is that SURF is used as a feature extractor algorithm to measure the success of the selected model for the SVM. As shown in Figure 6.5, Histogram Intersection Kernel performed better than Chi Square Kernel. During this test SURF algorithm was used. The number of features is 1500 and the number of images is 2400.

Table 6.5: SVM Kernel Models Performance with SURF algorithm

Set	Histogram Intersection Kernel	Chi - Square Kernel
Set1	84%	65%
Set2	88%	70%
Set3	83%	68%
Set4	81%	65%
Set5	83%	68%
Set6	88%	75%
Set7	86%	70%
Set8	85%	73%
Set9	84%	69%
Set10	88%	75%
Set11	83%	70%
Set12	85%	73%
Set13	90%	78%

Set14	87%	75%
Set15	88%	76%
Set16	83%	70%
Set17	84%	73%
Set18	83%	73%
Set19	88%	75%
Set20	90%	78%

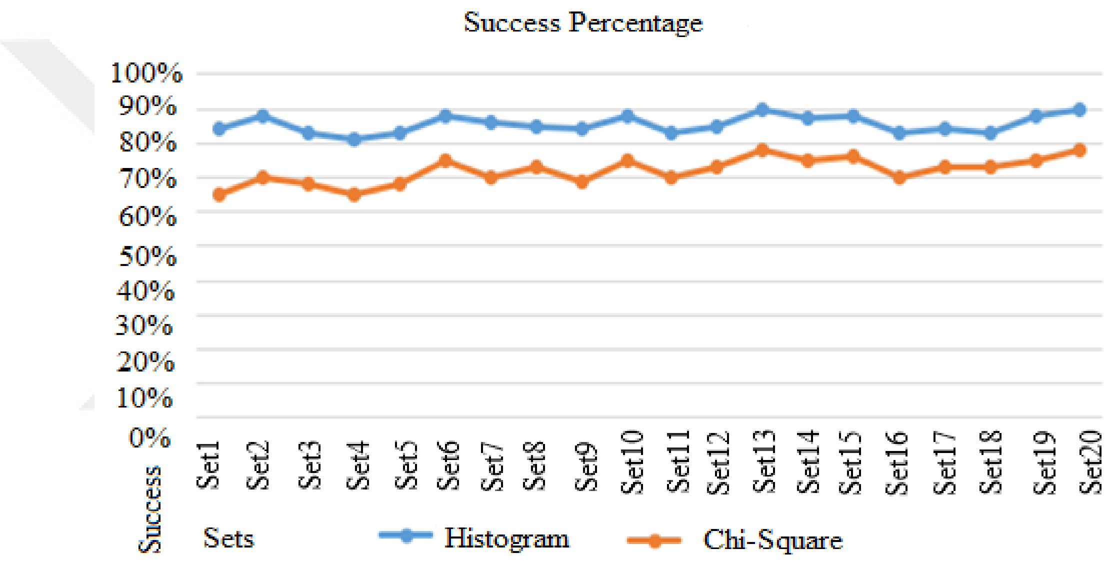


Figure 6.5: Graph drawn according to Table 6.5

Histogram Intersection Kernel showed good results with SURF algorithm. However, the HOG-Histogram Intersection Kernel pair showed better results than the SURF-Histogram Intersection Kernel pair. Accordingly, the features extracted using the SURF algorithm did not perform well when using both kernels. This is because the HOG-Histogram Intersection algorithms have similar features that complement each other.

The reason for the high performance of the HOG algorithm is the use of Histogram Intersection Kernel as the kernel for classification. Because both algorithms work histogram based. HOG algorithm is used to extract the attributes from the image. The Histogram Intersection Kernel has the best results when comparing similar histograms when classification. When Table 6.4 and Table 6.5 are combined, Table 6.6 is obtained. As a result of combining feature extraction algorithms and kernel functions, the best results were obtained in the Histogram Intersection Kernel-HOG pair.

Accordingly, the HOG algorithm and Histogram Intersection kernel pair is very effective in detecting stains on potatoes and has proved to perform very well in classification.

Table 6.6: Combination of feature extraction algorithms and kernel functions

Set	Histogram-SURF	Chi - Square-SURF	Histogram-HOG	Chi - Square-HOG
Set1	84%	65%	93%	89%
Set2	88%	70%	96%	90%
Set3	83%	68%	94%	88%
Set4	81%	65%	92%	87%
Set5	83%	68%	94%	88%
Set6	88%	75%	96%	90%
Set7	86%	70%	94%	87%
Set8	85%	73%	95%	89%
Set9	84%	69%	94%	88%
Set10	88%	75%	96%	90%
Set11	83%	70%	93%	87%
Set12	85%	73%	95%	89%
Set13	90%	78%	97%	91%
Set14	87%	75%	94%	88%
Set15	88%	76%	96%	90%
Set16	83%	70%	94%	87%
Set17	84%	73%	94%	88%
Set18	83%	73%	94%	87%
Set19	88%	75%	96%	90%
Set20	90%	78%	97%	91%

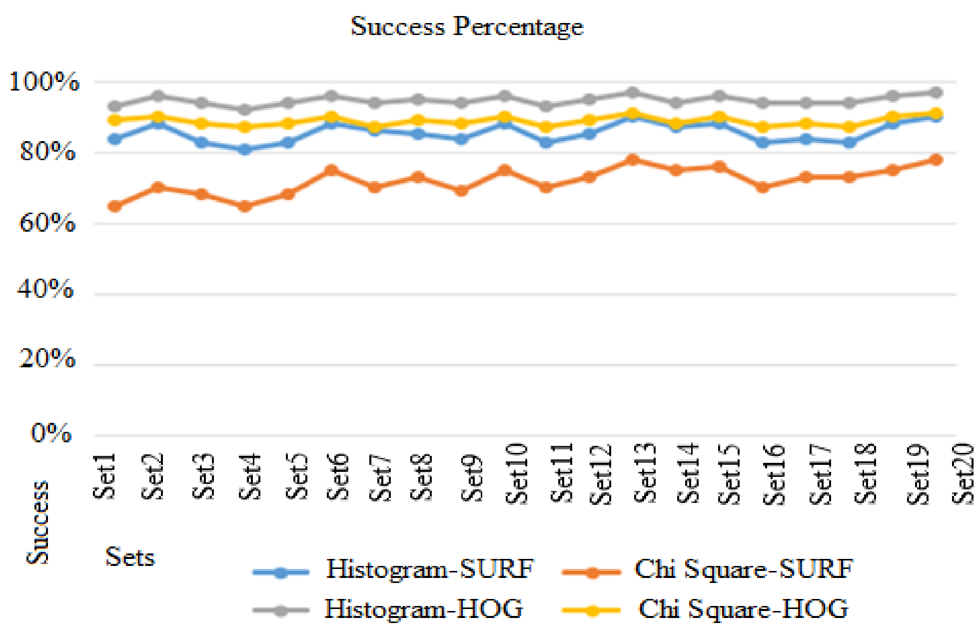


Figure 6.6: Graph drawn according to Table 6.6

6.6. Classification success according to potato diseases

In the sixth experiment, the success of classification according to potato diseases is examined. While the best results were obtained in the healthy class, Commonscab diseases classification success is lower than the others.

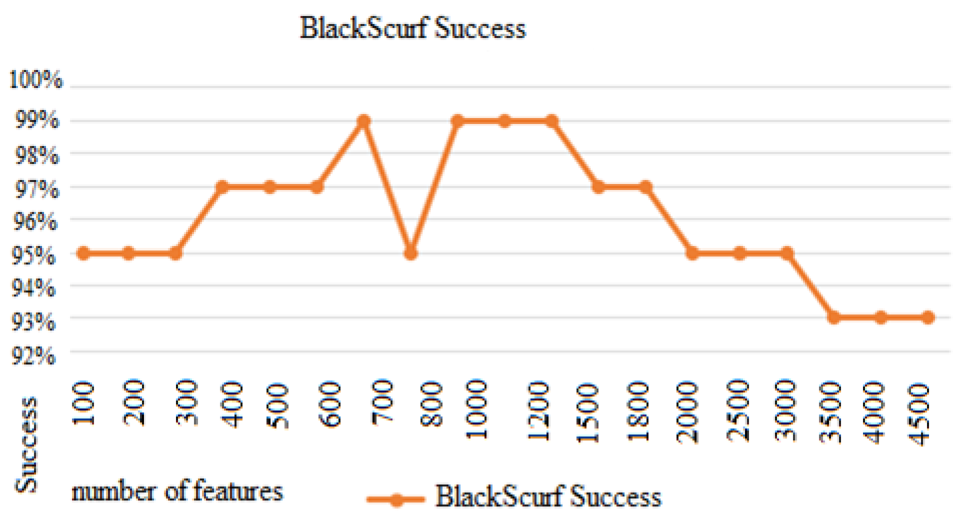


Figure 6.7: Black scurf Class Success

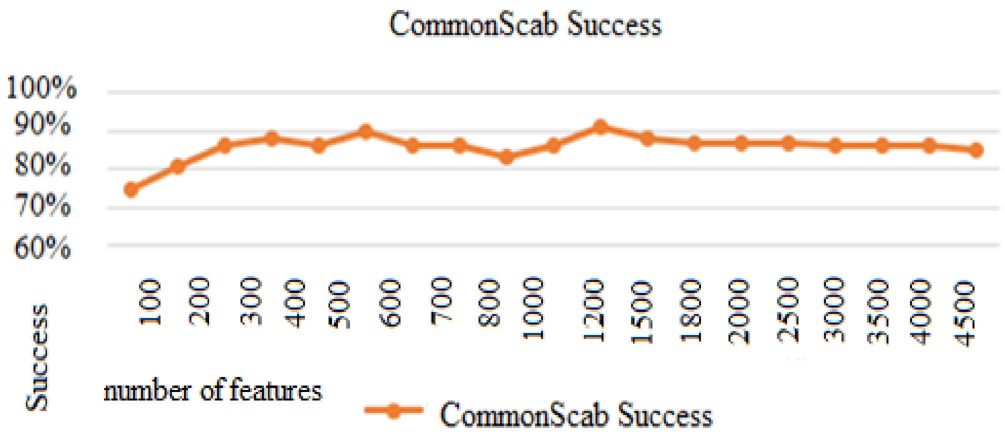


Figure 6.8: CommonScab Class Success

The classification success of BlackScurf and CommonScab diseases is shown in the above figures. Since the number of features is between 1000-1500, classification success is high. This is because the bag of words algorithm used to group similar features has good results in these feature numbers.

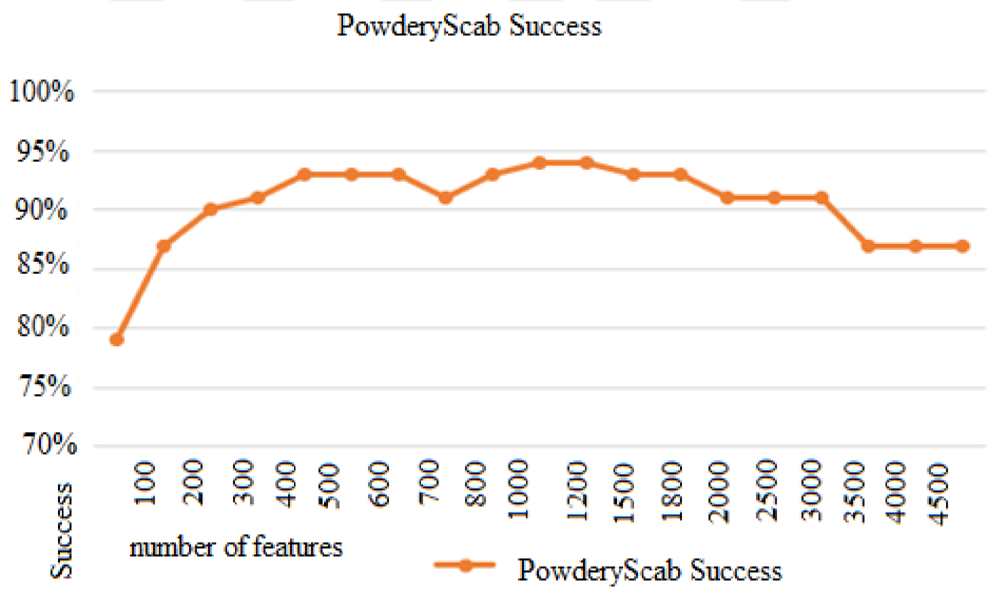


Figure 6.9: PowderyScab Class Success

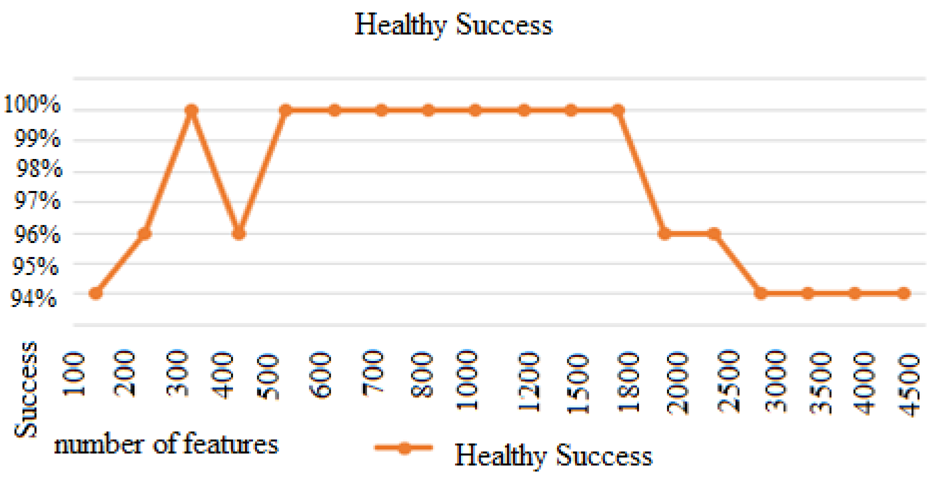


Figure 6.10: Healthy Class Success

The classification success of PowderyScab and Healthy classes is shown in the figures above. Since the number of features is between 1000-1500, classification success is high. 100% success was achieved when the number of features for the healthy class was between 500-1800.

The percentages of achievement according to the number of attributes of the four classes are shown in Table 6.7. The figure formed according to this table is shown in Figure 6.11.

Table 6.7: Percentage of achievement by attribute number of four classes

Number of Features	BlackScurf	CommonScab	PowderyScab	Healthy
100	96%	75%	79%	97%
200	96%	81%	87%	97%
300	96%	86%	90%	98%
400	97%	88%	91%	100%
500	97%	86%	93%	98%
600	97%	90%	93%	100%
700	98%	86%	93%	100%
800	96%	86%	91%	100%
1000	98%	83%	93%	100%
1200	98%	86%	94%	100%
1500	98%	91%	94%	100%
1800	97%	88%	93%	100%
2000	97%	87%	93%	100%

2500	96%	87%	91%	98%
3000	96%	87%	91%	98%
3500	96%	86%	91%	97%
4000	95%	86%	87%	97%
4500	95%	86%	87%	97%
5000	95%	85%	87%	97%

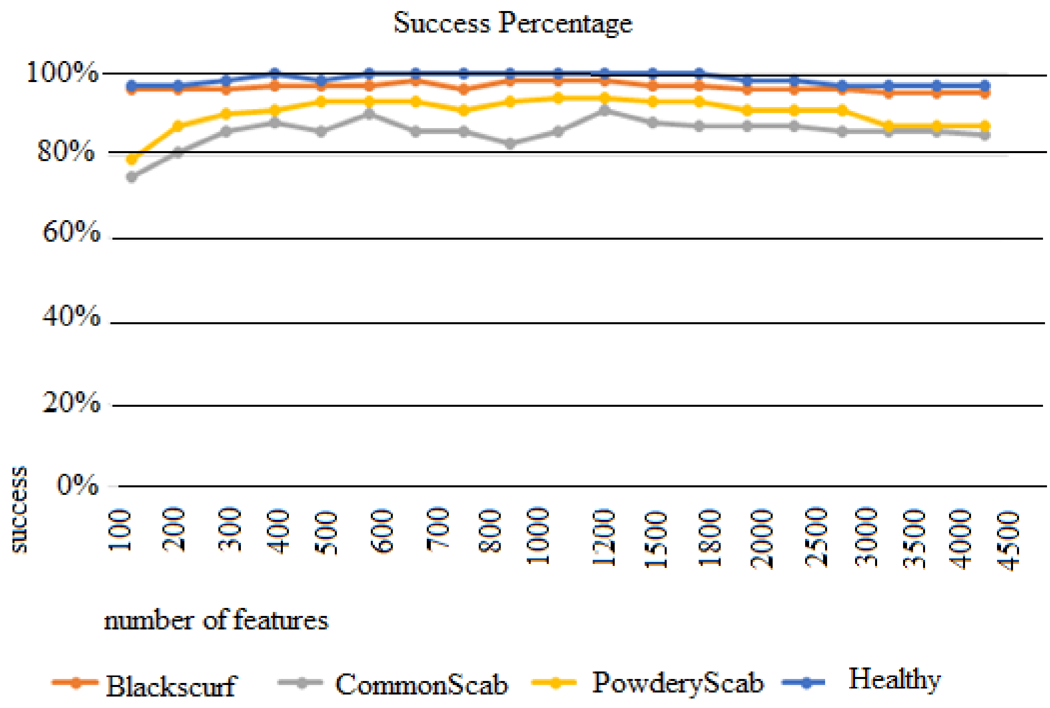


Figure 6.11: Graph drawn according to Table 6.7

Finally, a confusion matrix was formed to observe the wrong class selection of the classification method. According to the results given in Table 1, there was no wrong transition between healthy and diseased classes, while the most wrong transition occurred for CommonScab disease. If the developed system is used for the classification of diseased and healthy potatoes, it can achieve a 100% successful classification process. When the other studies in this area are examined, the success rate is generally 97% in the classifications made with two classes. In this respect, it has been superior to other studies.

According to the table, for example, the test pictures of Blackscurf disease in the first row were detected 99% correctly and 1% of them were found wrong by comminution with Commonscab disease. Likewise, 92% of the test pictures of Commonscab disease on line 2 were detected correctly, while 3% were compared to Blackscurf disease and 5% to PowderyScab disease. During these results, the HOG attribute was obtained using the Histogram Intersection kernel model SVM for 1500 attribute numbers.

Table 6.8: Confusion Matrix

Percent	BLACK SCURF	COMMON SCAB	POWDERY SCAB	HEALTHY
BLACK SCURF	99%	1%	0	0
COMMON SCAB	3%	92%	5%	0
POWDERY SCAB	2%	3%	95%	0
HEALTHY	0	0	0	100

For example, the images below belong to CommonScab disease, but the system mistakenly classified PowderyScab.

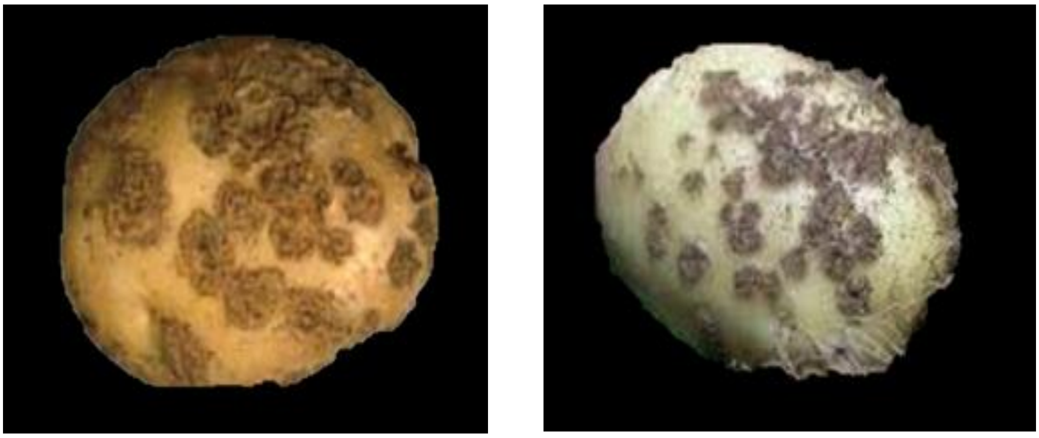


Figure 6.12: Misclassified Common Scab Images as Powdery Scab

Figure 6.13: PowderyScab sample images

6.7. Comparison of sets according to classification results

Randomly generated sets were obtained by random mixing of 100 images found for each potato disease. The effect of these clusters on the success of classification is as shown in the graph. Since all potato images were random in the generated clusters, the system was prevented from learning using the same potato images. The purpose of creating clusters is to increase the durability of the system by using different images and different clusters during the training and testing phase.

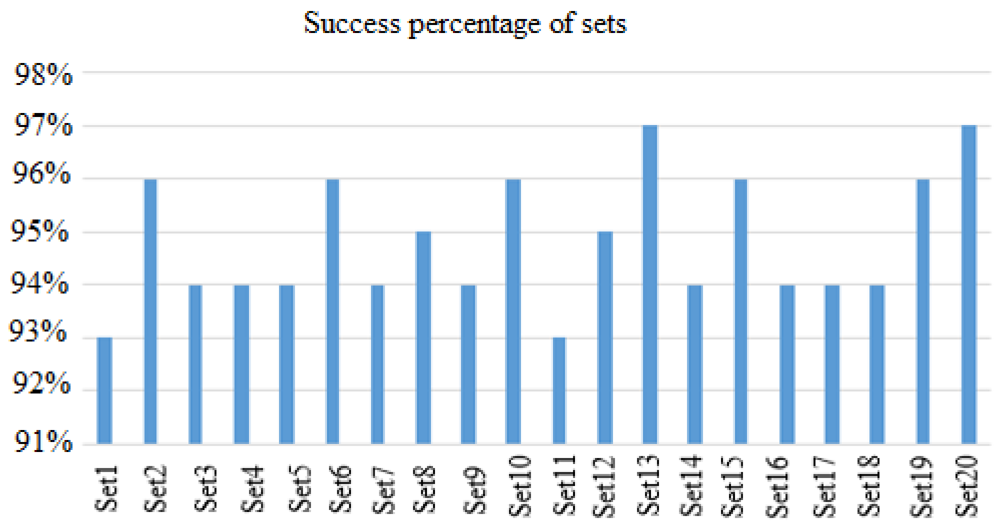


Figure 6.14: Comparison of sets

6.8. Comparison of image sizes

Since the images obtained from the internet were of different sizes, they were all sized to be the same size. For this purpose, it is sized 400 x 400, 600 x 600, 800 x 800. The system was tested in all three dimensions. The best results are obtained on 600 × 600 images.

400x400 information and texture properties are reduced. Disease tissue and patterns cannot be clearly distinguished because the images are small. At 800x800, the edges become blurred by interpolation, and the disease becomes more difficult to distinguish because the tissues and patterns are enlarged. Therefore, the extracted features remain weak to differentiate the disease.

The best of 600x600 is the average size of the images obtained in this area. Disease stains and tissues can be clearly distinguished. The clearness of the symptoms of the disease ensures the correct removal of the features. Therefore, the classification performance is high.

Table 6.9: Comparison of image sizes

Set	400*400	600*600	800*800
Set1	75%	93%	80%
Set2	78%	96%	83%
Set3	76%	94%	81%
Set4	76%	94%	81%
Set5	76%	94%	81%
Set6	78%	96%	83%
Set7	76%	94%	81%
Set8	77%	95%	82%
Set9	76%	94%	81%
Set10	78%	96%	83%
Set11	75%	93%	80%
Set12	77%	95%	82%
Set13	79%	97%	85%
Set14	76%	94%	81%
Set15	78%	96%	83%
Set16	76%	94%	83%
Set17	76%	94%	81%
Set18	76%	94%	81%
Set19	78%	96%	83%
Set20	79%	97%	85%

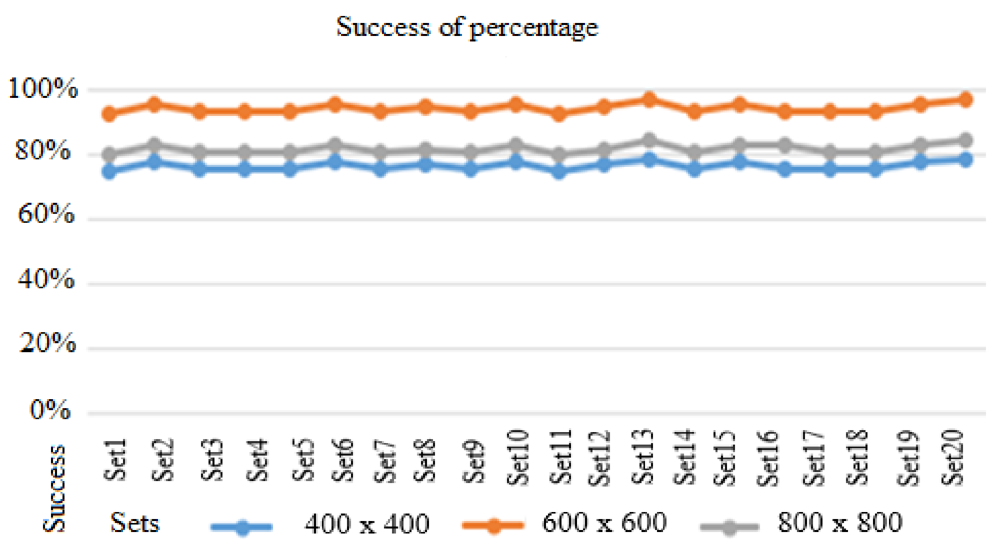


Figure 6.15: Success percentage of dimensions

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this thesis an approach was proposed to identify and classify potato external diseases using multi-class SVM. In the proposed method, three common diseases of BlackScurf, CommonScab and PowderyScab diseases were investigated. The system was trained according to stain and tissue characteristics caused by these diseases. In the light of literature research on potato diseases, this thesis, which has been identified and classified more than one potato disease, is the first in its field.

This thesis has been proved to be successful by achieving 96% success rate as a result of various experiments. It has also achieved 100% success in distinguishing healthy and diseased potatoes. The algorithms used to achieve these successful results have great impact. The compatibility of HOG algorithm and Histogram Intersection Kernel methods increased system performance. Since both approaches were histogram-based, successful results were obtained. In this study, three potato external diseases was detected and classified successfully and HOG-Histogram Intersection Kernel compatibility was found.

The developed method can be used in a possible automation system to perform automatic sorting of potatoes prior to fabrication and consumption. As future work, it is planned to identify and classify internal diseases by increasing the number of potato diseases. Deep Learning technique is planned to be used in the classification process. There may also be more than one disease sign on a potato. These are also planned to be identified in the next study. The proposed method has been developed for the analysis of potatoes brought from the field in the industrial environment. As future work, an improvement is planned in the field for farmers to detect diseases immediately after harvest.

BIBLIOGRAPHY

Accord.Net-Library, http://accord-framework.net/, (Access Date: 14.08.2019)

Açıl, S., 2017, "Görüntü İşleme: Otsu Thresholding", https://medium.com/@sddkal/g%C3%B6r%C3%BCnt%C3%BCi%CC%87%C5%9Fleme-otsu-thresholding-3f20e1c846ff, (Access Date: 14.08.2019)

Agriculture and Horticulture Development Board, "Black Scurf and Stem Canker", https://potatoes.ahdb.org.uk/media-gallery/detail/13214/2636 (Access Date: 14.08.2019)

Al-Hiary, H., Bani-Ahmad, S., Reyalat, M., Braik. M., and ALRahamneh, Z., 2011, Fast and Accurate Detection and Classification of Plant Diseases, *International Journal of Computer Applications,* 17(1): 31-38

Altaş, D., Gülpınar, V., 2012, Karar Ağaçları ve Yapay Sinir Ağlarının Sınıflandırma Performanslarının Karşılaştırılması: Avrupa Birliği Örneği, *Trakya University Journal of Social Science,* 14(1): 1-22

Alwanin, R., 2014, An Approach to Image Classification Based on SURF Descriptors and Color Histograms, Master Thesis, *University of Manchester-Faculty of Engineering and Physical Sciences,* 92p. (Unpublished)

Anami, B.S., Pujari, J.D., and Yakkundimath, R., 2011, Identification and Classification of Normal and Affected Agriculture/horticulture Produce Based on Combined Color and Texture Feature Extraction, *International Journal of Computer Applications in Engineering Sciences,* 1(3): 356-360

Arivazhagan, S., Shebiah, R.N., Ananthi, S., and Varthini, S.V., 2013, Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features, *Agricultural Engineering International,* 15(1): 211-217

Ayhan, S., and Erdoğmuş, Ş., 2014, Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi, *Eskişehir Osmangazi Üniversitesi IIBF Dergisi,* 9(1): 175- 198

Ayan, B., 2019, "Sosyal Ağlarda Makine Öğrenmesi 11 [Kernel SVM]", http://bugraayan.com/2019/04/17/sosyal-aglarda-makine-ogrenmesi-11-kernel-svm/, (Access Date: 14.08.2019)

Barla, A., Odone, F., and Verri, A., 2003, Histogram Intersection Kernel for Image Classification, *Proceedings / ICIP International Conference on Image Processing,* 2: 513-516

Bay, H., Ess, A., Tuytelaars, T., and Gool V., 2008, Speeded-Up Robust Features (SURF), *Computer Vision and Image Understanding,* 110(3): 346-359

Camargo, A., and Smith, J.S., 2009, Image pattern classification for the identification of disease-causing agents in plants, *Computers and Electronics in Agriculture,* 66: 121-125

Çalışkan, M.E., Onaran, H., and Arıoğlu. H., 2010, Overview of the Turkish Potato Sector: Challenges, Achievements and Expectations, *Potato Research,* 53(4): 255-266

Cortes, C., and Vapnik, V., 1995. Support vector networks, Machine Learning, 1-25.

Crowley, J.L., 2016, Kernel Functions and Support Vector Machines, *Intelligent Systems: Reasoning and Recognition, Second Semester Lesson 5*

Çetiner. S., 2016, Tarihsel süreçte patates tarımı ve kültür, Sabancı Üniversitesi, 10p.

Dailey, M.N., Chaivivatrakul, S., 2014, Texture-based fruit detection, *Precision Agriculture,* 15(6): 662-683

Dalal, N., and Triggs, B., 2005, Histograms of Oriented Gradients for Human Detection, *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR),* 886-893

Dubey, S.R., and Jalal, A.S., 2012, Adapted Approach for Fruit Disease Identification using Images, *International Journal of Computer Vision and Image Processing (IJCVIP),* 3: 44-58

Emel, G.G., Taşkın, Ç., 2005, Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması, *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi,* 6(2): 221-239

Fera Science Ltd, "Powdery Scab (Spongospora subterranea)", https://www.fera.co.uk/catalog/product/view/id/65/s/powdery-scab/ (Access Date: 14.08.2019)

Foody, G.M., and Mathur, A., 2004, Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification, *Remote Sensing of Environment*, 93(1-2), 107–117.

Gıda Tarım ve Hayvancılık Bakanlığı, 2016, "Patates Hastalık ve Zararlıları ile Mücadele", https://www.tarimorman.gov.tr/GKGM/Belgeler/Bitki%20Sa%C4%9Fl%C4%B1%C4%9F%C4%B1%20Hizmetleri/hastalik_zararlilari_ile_m%C3%BCcadele_dokumanlari/patates.pdf, (Access Date: 14.08.2019)

Gunn, S. R., 1998, Support vector machines for classification and regression. Technical Report, Faculty of Engineering, *Science and Mathematics, School of Electronics and Computer Science.*

Güldoğan, E., Çolak, C., and Taşdelen, B., 2017, Çeşitli Çekirdek Fonksiyonları ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama, Doctoral Thesis, *İnönü Üniversitesi ve Mersin Üniversitesi Biyoistatistik Ve Tıp Bilişimi Anabilim Dalı Ortak Doktora Tezi*

Haykin, S., 2009, Neural Networks and Learning Machines, Third edition, Pearson Education, 34-35

Hawkes, J.G., and Ortega, J.F., 1993, The early history of the potato in Europe, *International Journal of Plant Breeding,* 70(1-2): 1-7

Huang, D., Shan, C., Ardabilian, M., Wang, Y., and Chen, L., 2011, Local Binary Patterns and Its Application to Facial Image Analysis: A Survey, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 41(6): 765-781

Islam, M., Dinh, A., Wahid. K., and Bhowmik, P., 2017, Detection of Potato Diseases Using Image Segmentation and Multiclass Support Vector Machine, *IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 30-33

Kargı, A.B., 2016, "2015 TÜİK Verilerine Göre Türkiye'de Tarım Ürünlerinin Dağılışı", http://www.cografyadefterim.com/kpss-cografya/2015-tuik-verilerine-gore-turkiyede-tarim-urunlerinin-dagilisi.html (Access Date: 14.08.2019)

Kızrak, M.A., 2014, Local Binary Pattern Yöntemi ile Yüz İfadelerinin Tanınması, Şekil Tanıma Proje Raporu

Kobayashi, T., Hidaka, A., and Kurita, T., 2008, Selection of Histograms of Oriented Gradients Features for Pedestrian Detection, International Conference on Neural Information Processing Neural Information Processing, pages 598-607

Likas, A., Vlassis, N., and Verbeek, J., 2001, The global k-means clustering algorithm, *Pattern Recognition*, 36(2): 451-461

Maji, S., Berg, A.C., and Malik, J., 2012, Efficient Classification for Additive Kernel SVMs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 66-77

Merz.U., and Falloon, R.E., 2009, Review: Powdery Scab of Potato—Increased Knowledge of Pathogen Biology and Disease Epidemiology for Effective Disease Management, *Potato Research*, 52: 17-37

Meunkaewjinda, A., Kumsawat, P., Attakitmongcol, K., and Srikaew, A., 2008, Grape leaf disease detection from color imagery using hybrid intelligent system, *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 1: 513-516

Miller, J., Olsen, N., and Nolte, P., 2006, Diagnosis & Management of Potato Storage Diseases, *University of Idaho College of Agricultural and Life Sciences*, 8p.

Mokhtar, U., Ali, M.A.S., Hassenian, A.E., and Hefny, H., 2015, Tomato leaves diseases detection approach based on support vector machines, *International Computer Engineering Conference (ICENCO)*, pages 246-250

Naik, S., and Patel, B., 2017, Machine Vision based Fruit Classification and Grading - A Review, *International Journal of Computer Applications*, 170(9): 22-34

Nieto, A.D., Formella, A., Carrion, P., Fernandez, E.V., and Delgado. M.F., 2011, Common Scab Detection on Potatoes Using an Infrared Hyperspectral Imaging System, *International Conference on Image Analysis and Processing (ICIAP)*, 6979: 303-312

Nikam., S.S., 2015, A Comparative Study of Classification Techniques in Data Mining Algorithms, *Oriental Journal of Computer Science& Technology*, 8(1): 13-19

Ojala, T., Pietikäinen, M., and Mäenpää, T., 2002, Multiresolution Grayscale and Rotation Invariant Texture Classification with Local Binary Patterns, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24: 971-987

Olsen,A., Han, S., Calvert, B., Ridd,P., and Kenny, O., 2015, In Situ Leaf Classification Using Histograms of Oriented Gradients, *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*

Omid, M., Khojastehnazhand, M., and Tabatabaeefar, A., 2010, Estimating volume and mass of citrus fruits by image processing technique, *Journal of Food Engineering,* 100(2): 315-321

Osuna, E.E., Freund, R., and Girosi, F., 1997, Support Vector Machines: Training and Applications, *A.I. Massachusetts Institute of Technology and Artificial Intelligence Laboratory*, Massachusetts.

Öztemel, E., 2006, Yapay sinir ağları. Papatya Yayıncılık, İstanbul.

Patacchiola., M., 2016, "The Simplest Classifier: Histogram Comparison", https://mpatacchiola.github.io/blog/2016/11/12/the-simplest-classifier histogram-intersection.html (Access Date: 28.08.2019)

Pérez, D.S., Bromberg, F., and Diaz, C.A., 2017, Image classification for detection of winter grapevine buds in natural conditions using scale-invariant features transform, bag of features and support vector machines, *Computers and Electronics in Agriculture*, 135: 81-95

Pooja, H., Madival, S.A., 2016, Food Recognition and Calorie Extraction using Bag-of- SURF and Spatial Pyramid Matching Methods, *International Journal of Computer Science and Mobile Computing,* 5(5): 387-393

Pujari, J.D., Yakkundimath, R., and Byadgi. A.S., 2016, SVM and ANN Based Classification of Plant Diseases Using Feature Reduction Technique, *International Journal of Interactive Multimedia and Artificial Intelligence,* 3: 6-14

Ranjit, K.N., Chethan, H.K., and Naveena, C., 2016, Identification and Classification of Fruit Diseases, *Int. Journal of Engineering Research and Application,* 6(7): 11-14

Razmjooy.N., Mousavi, B.S., and Soleymani.F., 2012, A real-time mathematical computer method for potato inspection using machine vision, *Computers & Mathematics with Applications,* 63(1): 268-279

ResearchGate, “Symptoms of black scurf on tuber of potato”, https://www.researchgate.net/figure/Symptoms-of-black-scurf-on-tuber-of potato_fig3_287996397 (Access Date: 14.08.2019)

sakilAnsari, 2019, “Pattern Recognition | Introduction”, https://www.geeksforgeeks.org/pattern-recognition-introduction/,(Access Date: 14.08.2019)

Samanta.D., Chaudhury, P.P., and Ghosh.A., 2012, Scab Diseases Detection of Potato using Image Processing, *International Journal of Computer Trends and Technology,* 3(1): 110-113

Samtaş, G., and Gülesin, M., 2011, Sayısal Görüntü İşleme ve Farklı Alanlardaki Uygulamaları, *Electronic Journal of Vocational Colleges,* 85-97

Sannakki, S.S., Rajpurohit, V.S., Nargund, V.B., and Kulkarni, P., 2013, Diagnosis and Classification of Grape Leaf Diseases using Neural Networks, *Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*

Sert, B., 2017 "(Ders 07) Aforge.Net ile Otsu Threshold (C#)", http://www.eprogramci.com/aforge-net-ile-otsu-threshold-c/, (Access Date: 14.08.2019)

Schölkopf, B., and Smola A.J., 2002, Learning with Kernels, *MIT Press*, 626p.

Shah, M., 2012, "Lecture 17: Bag-of-Features (Bag-of-Words)", https://www.youtube.com/watch?v=iGZpJZhqEME (Access Date: 14.08.2019)

Shan, C., Gong, S., and McOwan, P.W. 2009, Facial expression recognition based on Local Binary Patterns: A comprehensive study, Image and Vision Computing, 27(6): 803-816

Soman, K.P., Loganathan, R., and Ajay, V., 2011, Machine learning with SVM and other kernel methods. PHI Learning Pvt. Ltd., 486p.

Southgate, J., 2016, Food Classification to provide Nutritional Information, Master Thesis, University of the Western Cape Computer Science, 48p.

Souza, C., 2010, "Kernel Functions for Machine Learning Applications", http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/, (Access Date: 14.08.2019)

Souza, C., 2014, "Handwriting Recognition Revisited: Kernel Support Vector Machines", https://www.codeproject.com/Articles/106583/Handwriting-Recognition-Revisited-Kernel-Support-V, (Access Date: 14.08.2019)

Sulakshana, A.G., Kalyani, S.D., Monali, K.W., and Priyanka, R.D., 2017, Fruit Disease Detection and Classification, *International Research Journal of Engineering and Technology (IRJET)*, 04(12): 1151-1154

Şeker, Ş.E., 2008, "Çok sınıflı DVM (Multiclass SVM)", http://bilgisayarkavramlari.sadievrenseker.com/2008/12/01/cok-sinifli-dvm-multiclass-svm/, (Access Date: 14.08.2019)

Şeker, Ş.E., 2013, "Sınıflandırma (Classification)", http://bilgisayarkavramlari.sadievrenseker.com/2013/03/31/siniflandirma-classification/, (Access Date: 14.08.2019)

Takcı, H., 2018, "Sınıflandırma algoritmaları (Classification algorithms)", https://verimadencisi.blogspot.com/2018/03/onemli-bir-veri-madenciligi-teknigi.html, (Access Date: 14.08.2019)

Tsolakidis, D.G., Kosmopoulos, D.U., and Papadourakis, G., 2014, Plant Leaf Recognition Using Zernike Moments and Histogram of Oriented Gradients, *Hellenic Conference on Artificial Intelligence Artificial Intelligence: Methods and Applications,* 8445: 406-417

United Nations Economic Commission for Europe, "List of Diseases and Pests", https://www.unece.org/fileadmin/DAM/trade/agr/standard/potatoes/ListOfDiseasesAndPests.htm (Access Date: 14.08.2019)

University of Massachusetts Amherts, The Center for Agriculture, Food and the Environment, 2013, “Potato, Scab”, https://ag.umass.edu/vegetable/fact-sheets/potato-scab (Access Date: 14.08.2019)

Ülgen, K.E., 2017, “Makine Öğrenimi Bölüm-4 (Destek Vektör Makineleri)”, https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi-b%C3%B6l%C3%BCm-4-destek-vekt%C3%B6r-makineleri-2f8010824054, (Access Date: 14.08.2019)

Üstüner, M., Şanlı, F.B., Balçık, F.B., Esetlili, M.T., 2013, Destek Vektör Makineleri Tekniği ile Sınıflandırma: Rapideye Örneği, *Türkiye Ulusal Fotogrametri ve Uzaktan Algılama Birliği VII. Teknik Sempozyumu*

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. New York: Springer-Verlag.

Velioğlu, H.M., 2010, Görüntü İşleme Teknolojisi ve Yapay Sinir Ağlarının Gıda Ürünlerinde Bazı Kalite Özelliklerinin Belirlenmesinde Kullanılması, *Doktora Tezi, Namık Kemal Üniversitesi Fen Bilimleri Enstitüsü*, 90p. (Unpublished)

Wagstaff. K., Rogers. S., and Schroedl. S., 2001, Constrained K-means clustering with background knowledge, *Proceedings of the Eighteenth International Conference on Machine Learning*, 577-584.

Yogesh and Dubey., A.K., 2016, Fruit defect detection based on speeded up robust feature technique, *International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO),* 590-594

Zhang, Y., Jin, R., and Zhou, Z.H., 2015, Understanding Bag-of-Words Model: A Statistical Framework, *International Journal of Machine Learning and Cybernetics*, 1(1-4): 43-52

Zulkifli, Z., 2009, Plant leaf identification using moment invariants & general regression neural network, Master Thesis, Universiti Teknologi Malaysia, Faculty of Computer Science and Information Systems.