T. R.
VAN YUZUNCU YIL UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF ELECTRIC AND ELECTRONIC ENGINEERING

PREDICTION OF COLORECTAL CANCER USING SUPPORT VECTOR MACHINES ALGORITHM

M.Sc. THESIS

PREPARED BY: Bawar Safwat Hussein MATEEN
SUPERVISOR: Assoc. Prof. Dr. Rıdvan SARAÇOĞLU

VAN-2020

T. R.
VAN YUZUNCU YIL UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF ELECTRIC AND ELECTRONIC ENGINEERING

PREDICTION OF COLORECTAL CANCER USING SUPPORT VECTOR MACHINES ALGORITHM

M.Sc. THESIS

PREPARED BY: Bawar Safwat Hussein MATEEN

This study was supported by Scientific Research Studys Coordination Unit of Van Yuzuncu Yil University with study no:…….

VAN-2020

ACCEPTANCE and APPROVAL PAGE

This thesis entitled "Prediction of Colorectal Cancer Using Support Vector Machines Algorithm" presented by Bawar Safwat Hussein MATEEN under supervision of Assoc. Prof. Dr. Rıdvan SARAÇOĞLU in the department of Electrical and Electronics Engineering has been accepted as a M. Sc. thesis according to Legislations of Graduate Higher Education on 06/01/2020 with unanimity of votes members of jury.

Chair: Assoc. Prof. Dr. Emre ÇOMAK Signature:

Member:Assoc. Prof. Dr. Rıdvan SARAÇOĞLU Signature:

Member:Assist. Prof. Dr. A. Oğuz KIZILÇAY Signature:

This thesis has been approved by the committee of The Institute of Natural and Applied Science on ..17.../..01.../2020. with decision number .2020/4-1

Signature

Director of Institute

Prof. Dr. Ahmet ÇELİKSOY
Enstitü Müdürü

THESIS STATEMENT

All information presented in the thesis obtained in the frame of ethical behavior and academic rules. In addition all kinds of information that does not belong to me have been cited appropriately in the thesis prepared by the thesis writing rules.

Bawar S. MATEEN

ABSTRACT

PREDICTION OF COLORECTAL CANCER USING SUPPORT VECTOR MACHINES ALGORITHM

MATEEM, Bawar
MSc. Thesis Electric and Electronic Engineering
Supervisor: Assoc. Prof. Dr. Rıdvan SARAÇOĞLU
January 2020, 69 pages

Colorectal cancer is a common type of cancer disease that begins in colon or rectum. It is the second most common cancer type of cancer among females, and the third among males. Colorectal cancer causes the death of thousands of people around the world every year. The early detection and prediction of colorectal cancer increases the success chances of colorectal cancer treatment, since colorectal cancer is localized, curable and needs less cost for the treatments in its early stages. This study tried to predict and detect colorectal cancer by using support vector machines algorithm. And that by the implementation on a dataset that is based on the participants' lifestyle, which is the first time to predict colorectal cancer by this type of datasets. This dataset included 22 information about each participant. The results show that colorectal cancer is predictable with a high accuracy by this type of dataset using support vector machines algorithm. And these results can be more accurate if the information of more colorectal cancer patients were collected, and if the data were recorded in the local area of the research to work with better performence for prediction of colorctal cancer in local people.

Keywords: Artificial Intelligence, Colorectal Cancer, Kernel Functions, Support Vector Machines.

ÖZET

KOLOREKTAL KANSERIN DESTEK VEKTÖR MAKINELERI ALGORITMASI ILE TAHMINI

MATEEN, Bawar
Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı
Tez Danışmanı: Doç. Dr. Rıdvan SARAÇOĞLU
Ocak 2020, 69 sayfa

Kolorektal kanser, kolon veya rektumda başlayan yaygın bir kanser hastalığı türüdür. Kadınlar arasında en sık görülen ikinci kanser türü, erkekler arasında ise üçüncü kanser türüdür. Kolorektal kanser her yıl dünya çapında binlerce insanın ölümüne neden olmaktadır. Kolorektal kanser lokalize, tedavi edilebilir olduğundan ve erken evrelerinde tedavi için daha az maliyete ihtiyaç duyduğundan, erken teşhis ve tahmini, kolorektal kanser tedavisinin başarı şansını arttırır. Bu çalışmada, destek vektör makineleri algoritması kullanılarak kolorektal kanser tahmin ve tespit edilmeye çalışılmıştır. Katılımcıların yaşam tarzına dayanan bir veri kümesinin kullanılmasıyla, bu tür veri kümeleriyle kolorektal kanseri tahmin edilmiştir. Bu veri seti her katılımcı hakkında 22 bilgi içermektedir. Sonuçlar kanserin, destek vektör makineleri algoritması kullanılarak bu tür bir veri kümesi ile yüksek bir doğrulukla öngörülebilir olduğunu göstermektedir. Eğer daha fazla kolorektal kanser hastasının bilgisi toplanırsa ve veriler bölgesel olarak toplanıp yine bölgesel modellemeler yapılırsa daha iyi performansla çalışıp daha doğru sonuçlar elde edilebilecektir.

Anahtar Kelimeler: Çekirdek Fonksiyonları, Destek Vektör Makineleri, Kolerktal Kanser, Yapay Zeka.

ACKNOWLEDGMENT

Foremost, I want to thank Allah Almighty for the wisdom he bestowed upon me, the strength, peace of my mind, and good health in order to finish this study.

My special thanks to my dear supervisor Assoc. Prof. Dr. Rıdvan SARAÇOĞLU for his help, advices, suggestions, and effort that he gave me during more than two years.

I would like to express my special gratitude to Prof. Dr. Naci GENÇ, Assoc. Prof. Dr. Muzaffer ATEŞ, and Assoc. Prof. Dr. Murat CANAYAZ for teaching and helping me whenever I needed.

Special thanks to all my teachers in University of Duhok and everyone who had taught during the years of my study, without them I were not able to reach these achievements. Special thanks to Mr. Muhammad Guhdar who was behind the idea of the study, and was ready for helping every time I needed help.

Thanks to my friends who were patient with me, helped me and prayed for me to reach my goals.

Great thanks to my father Mr. Safwat Berwari who supported and motivated me to get better in every side of the life, and he is a part of every good thing happening to me, Thanks to my mother deprived herself of sleep every day to wake me up, and gave me love and compassion, and patient over the alienation for me, my great grandfather, my sister and my brothers, my uncles and all the wonderful members of my family, and all of my relatives, I apologize because my words do not fulfill their right.

2020

Bawar Safwat Hussein MATEEN

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

SYMBOLS AND ABBREVIATIONS

Some symbols and abbreviations used in this study are presented below, along with descriptions.

Abbreviations	Description
SVM	Support Vector Machines
AI	Artificial Intelligence
ML	Machine learning
CRC	Colorectal Cancer
RBF	Radial Basis Function
YYU	Yuzuncu Yil University
TR	Turkish Republic

LIST OF APPENDIX

Page

1. INTRODUCTION

Nowadays the new technology has become an important part of our life, and it made the most of human processes easier than before. Also, the new technological inventions had converted many actions that were impossible to be done by human to possible and easy things. Due to the effectiveness and effortlessness of using technology, people are trying to use technology in various new fields that they did not reach before, this technology is called Artificial Intelligence.

On the other hand, health is one of the most important things that people think about it, because it is relevant with human life and death. Because of that, people have given a major importance to their health, to have a good healthy life. So, human beings had searched for the causes that damage their lovely life, and they spend time, money and effort to know, treat, diagnose and predict the diseases that are dangerous to their health.

Since the technology toke its place in our life, it has been used frequently for the medical purposes. One of medical fields that technology helped in considerably is the prediction of the diseases. Researchers had written numerous research papers to predict the diseases by using Artificial Intelligence algorithms especially the diseases that are too dangerous and that cause death usually, such as cancer, which is one of the most dangerous diseases that the humanity is afraid of.

This research is a new part of the combination between health care and technology, because an efficient algorithm which is Support Vector Machine (SVM) is used to predict a common type of cancer that is colorectal cancer. The previous process is done by using a dataset that is based on information of a big number of participants, who are healthy or patients of colorectal cancer, and this dataset will include the information about the participants' lifestyle especially those information that are related to the risk factors of colorectal cancer. This research is the first study that predict colorectal cancer using lifestyle based dataset using SVM algorithm which is an

artificial intelligence algorithm that had been proved to be an effective algorithm with several types of datasets.

1.1. Aim of the Study

The main purpose of this study is the prediction of colorectal cancer disease, and that by SVM algorithm, which trains on a dataset and by the comparison of the new data with the data that it trained on, the algorithm decides that the new data belongs to which class. Also, the study tries to identify the best situation of the dataset and kernels' parameters, in such a way that gives the best accuracy of the classification and prediction.

1.2. Study Importance

Colorectal cancer is one of the most common types of cancer all over the world (Society, 2019), this disease is the second most common type of cancer among females and the third among males, and it diagnosed 1.2 million new cases over the world and it caused 608,700 deaths only in 2008 (Jemal, et al., 2011). The early detection and prediction of colorectal cancer is crucial since it increases the success chances of the treatment, because cancer in general and colorectal cancer especially is localized, curable and needs less cost for the treatments in its early stages (Etzioni, et al., 2003), but if the disease takes a long period of time it may spread to other parts of the body, and it may cause new cancer types in the patient. On the other hand, using artificial intelligence classification algorithms especially SVM algorithm is an ideal way to classify big datasets which are not able to be classified by the human. So it is considerable that using this preferable tool for prediction of a dangerous disease such as colorectal cancer is a very important study.

In addition to the previous reason, the study is important because it will use a dataset that is related to lifestyle of the participants and risk factors of colorectal cancer,

which is a new type of datasets that had been used to predict colorectal cancer by artificial intelligence algorithms.

1. LITERATURE REVIEW

Cancer is a common disease, that people are doing every possible thing for its diagnosis and treatment, and because of the effectiveness of SVM classification algorithm, it has been used commonly in prediction and classification of Cancer disease, and most of the articles that are written in this subject used DNA Micro-array dataset, but few of them are related to colorectal cancer and most of them are about other types of cancer, by taking a look at these articles these are noticed:

Yu et al. (2004) used Artificial Neural Network (ANN) and SVM in their study to differentiate between 55 Colorectal Cancer (CRC) patient's serum samples, 35 Colorectal Adenoma (CRA) patient's serum samples, and 92 Healthy Persons (HP) serum samples. Their program could separate between CRC and CRA with a specificity of 83%, sensitivity of 89% and positive predictive value (PPV) of 89%. And it can separate between CRC and HP with a specificity of 92%, sensitivity of 89% and PPV of 86%.

Li et al. (2014) study's aim was the detection of Colorectal Cancer (CRC) with near-infrared Raman spectroscopy and feature selection techniques. The related nucleic acids, lipids, and proteins of tissues are identified with the ant colony optimization (ACO) and SVM. This study provided an accuracy of diagnosis of 93.2% for identifying CRC from normal Raman spectroscopy.

Kominami et al. (2016) developed a real-time image recognition system that can predict colorectal lesions and solve some problems with the recommendations of The Preservation and Incorporation of Valuable Endoscopic Innovations (PIVI) committee of the American Society for Gastrointestinal Endoscopy (ASGE). The results showed that the accuracy between endoscopic and their system diagnosis was 97.5%.

Duan and Rajapakse (2004) tried to solve two problems of cancer classification with mass spectrometry data, using Support Vector Machine Recursive Feature Elimination (SVM-RFE). Their study showed that SVM-RFE could select a good peaks subset with much better performance and prediction accuracy than T-statistics.

Polat and Güneş (2007) used Least Square Support Vector Machine (LS-SVM) on Wisconsin Diagnostic Breast Cancer (WDBC) dataset which was derived from a group of images using Fine Needle Aspiration (biopsies) of the breast. They reached a high classification accuracy which was 98.53%.

Sewak et al. (2007) used Support Vector Machines (SVMs) with linear, polynomial and RBF kernel functions were trained using WDBC dataset. Their system obtained an accuracy of 99%.

Hong and Cho (2008) proposed a method which SVM is generate with One-vs.-Rest (OVR) scheme and probabilistically ordered using naïve Bayes classifiers (NBs). This method produced an accuracy that is higher than conventional methods.

Subashini et al. (2009) in their study compared between polynomial kernel of SVM and Radial Basis Function Neural Network (RBFNN). They used Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The results showed that RBFNN outperformed the polynomial kernel of SVM for the correct classification of tumors.

Akay (2009) studied breast cancer diagnosis by a SVM based method that was combined with feature selection, he also used Wisconsin Diagnostic Breast Cancer (WDBC) dataset that is used commonly by the researchers who used machine learning for breast cancer diagnosis. The results showed that the accuracy of this method was 99.51% which is a very good accuracy ratio.

Yu et al. (2011) in their study used SVM to determine best prognostic model from combinations of 14 pathway-related markers (p53, APC, p21ras, E-cadherin, endothelin-B receptor, Shp2, ADCY-2, SPARCL1, neuroligin1, hsp27, mmp-9, MAPK, MSH2 and rho) from the patients. The results showed that seven of the previous markers (SPARCL1, Shp2, MSH2, E-cadherin, p53, ADCY-2 and MAPK) were related to the prognosis and clinical pathological features of the CRC patients.

Maulik and Chakraborty (2014) in their study they aimed to predict scheme that combines fuzzy preference based rough set (FPRS) method for gene selection with semi supervised Support Vector Machines (SVMs). They noticed as a result that their method can achieve success and is biologically related to cancer diagnosis and drug discovery.

Machhale et al. (2015) in their study they used SVM, K- Nearest Neighbor (KNN) and Hybrid Classifier (SVM-KNN) to classify 50 Magnetic Resonance Imaging (MRI) samples, the results showed that the hybrid classification method (SVM-KNN) reached the highest classification accuracy which was 98%.

Wang and Huang (2010) in their study the results of tumor marker detection for gastric cancer, lung cancer and colorectal cancer were collected. Then SVM with best kernel function were used on it. Some diagnostic classifiers were validated like combined diagnosis test, logistic regression and decision tree. The results showed that the accuracy of SVM classifier among 4 kinds of classifiers was high.

However, the articles that used DNA microarray had developed using SVM for classifying microarray data, like what happened in these articles:

Guyon et al. (2002) used SVM based on Recursive Feature Elimination (RFE). They demonstrated experimentally that their techniques classified better and more biologically relevant genes to cancer.

Rifkin et al. (2003) In their paper they applied a methodology that combines class specific (one vs. all) binary Support Vector Machines (SVMs) to the diagnosis of common tumors using DNA microarray data from tumor samples that included 14 of the most common cancer types. The results was accurate by 78%.

Chen (2003) made a system based on Genetic Algorithm (GA) and SVM, and he combined bootstrap methods to GA to overcome training sample small size problem. The results showed that this method could find genes that separate between normal and cancer cells.

Peng et al. (2003) had combined Genetic Algorithm (GA) with SVM so they reached an accuracy of 87.93% for the eight-class and 85.19% for the fourteen-class cancer classification, so these results outperformed the results from previous published methods.

Liu (2004) had used SVM on gene expression profiles of lung cancer, prostate cancer, and colon cancer samples, instead of other previous research that used machine learning with bioinformatics because the researchers most use a large training set which take a lot of time, and also it may not represent the real world data distribution.

Zhu and Hastie (2004) used Penalized Logistic Regression (PLR) instead of SVM because SVM has a weakness that it does not provide any estimate of the underlying probability. So they got a result that PLR performance is similar to SVM but PLR provided an estimate of the underlying probability.

Wang and Makedon (2004) had compared performance of Relief-F algorithm with other filtering methods like Information Gain, Gain Ratio, and x2-statistic, to SVM and K-nearest neighbors (K-NN). The results showed that the results of Relief-F algorithm comparable with SVM and K-NN.

Statnikov et al. (2005) tried Multi-category Support Vector Machine (MC-SVM) for cancer diagnosis from gene expression data. Their result showed that MC-SVM performed better than other algorithms so they constructed a software system that named Gene Expression Model Selector (GEMS) for this process.

Duan et al. (2005) had proposed a selection method that is using a backward elimination procedure similar to support vector machine recursive feature elimination (SVM-RFE). But in this method at each step it is computing the feature ranking score. That gives more accurate results than the original SVM-RFE.

Tang et al. (2005) had used a hybrid algorithm which was Granular Support Vector Machines - Recursive Feature Elimination (GSVM-RFE), to select the gene that is related to Prostate Cancer. They saw that this algorithm is much more accurate than Support Vector Machines - Recursive Feature Elimination (SVM-RFE). And also the method extracted a “perfect” gene subset with 17 genes with 100% accuracy.

Shen and Tan (2005) in their study the researchers presented the use of Penalized Logistic Regression to classify cancer, and they combined two Dimensional Reduction methods with the Penalized Logistic Regression so accuracy and speed are enhanced. They also compared their results with other two machine learning methods; Support Vector Machine and Least-Squares Regression and the results of their method was at least equal or better than the other two methods.

Chu and Wang (2005) had used SVM in their study for cancer classification with microarray data. They used some dimensionality reduction methods like Fisher Ratio, Class-separability measure, T-test, and Principal Component Analysis (PCA) for the

selection of the genes, they said that they could reach to the same accuracy of previous published results but with less features.

Mao et al. (2005) had proposed two classifiers; Fuzzy Support Vector Machine (FSVM) and Binary Classification Tree based on SVM. They applied these techniques for analyzing breast cancer data, small round blue-cell tumors, and acute leukemia data. They noticed that FSVM Based on Support Vector Machine Recursive Feature Elimination (SVM-RFE) can find most important genes that affect certain cancer types with high accuracy.

Wang et al. (2005) used a correlation-based feature selector combined with other machine learning algorithm like SVM. They showed that they can obtain at least as good results as previous published results on acute leukemia, and diffuse large B-cell lymphoma. They also showed that they can select relevant genes with high confidence.

Liu et al. (2005) in their study they joined Genetic Algorithm (GA) with All Paired (AP) SVM methods for multiclass cancer categorization. They saw that this method can select the relevant gene to cancer with the best performance of classification until the research date.

Zhang et al. (2006) developed a Recursive Support Vector Machine (R-SVM) algorithm to choose important genes for data with noise. They compared its performance to Recursive Feature Elimination Support Vector Machine (SVM-RFE). They noticed that these properties can improve over SVM-RFE by 5% to 20%.

Duan et al. (2007) in their study they proposed a multiclass classification method for gene selection. This method was selecting genes in backward eliminate on and it was computing its ranking scores at each step from weight vectors analysis of multiple two-class linear SVM classifiers from One-vs.-One (OVO) or One-vs.-All (OVA) breakdown of the classification problem of multi-class. The study results proved the effectiveness of this method to select a set of genes to have a good accuracy in the classification.

Lanza et al. (2007) in their research they have investigate 23 colon cancer samples that was characterized using Microsatellite Stability (MSS) and 16 samples by High Microsatellite Instability (MSI-H) for genome-wide expression of microRNA

(miRNA) and RNA, they suggested that mRNA/miRNA expression signature combination may improve bio-molecular classification of cancer disease.

Zhou and Tuck (2007) in their study they showed a family of four extensions to Multiclass Support Vector Machine - Recursive Feature Elimination to solve the problem of gene selection. They considered all classes simultaneously during the stages of gene selection, they have proposed some extensions identify genes that lead to more accuracy in the classification.

Alba et al. (2007) in their study they compared between two hybrid algorithms; Particle Swarm Optimization (PSO) with SVM, and Genetic Algorithm (GA) with SVM to classify high dimensional Microarray. They evaluate a new PSO version called Geometric PSO by using a binary representation in Hamming space. Another important contribution was discover of new good results identifying significant in the variety development of these types of cancer disease (breast, colon, leukemia, lung ovarian, and prostate).

Statnikov and Aliferis (2008) had compared Random Forests (RF) to SVM, they worked on 18 datasets and they noticed that SVM outperform RF often by a large margin.

Lin et al. (2007) had worked on datasets from New Zealand and Germany to predict Colorectal Cancer (CRC) they investigate that using different micro-arrays could be validated application to the alternate series of patients, the prediction rate for New Zealand were 77% and for Germany were 84%.

Wang et al. (2007) had divided a large and complex dataset that contained 14 cancer types into a group of small binary classification problems and after that they applied the divide-and-conquer approach to every classification problem. They used Fuzzy Neural Network (FNN) and SVM. As a result they obtained accuracy that was comparable to previous results but with only 28 genes instead of 16,063 genes. So their method can reduce the number of genes for a good diagnosis.

Del Rio et al. (2007) had developed a program to select discriminatory genes by the significance analysis of microarrays algorithm and classify the results using SVM. As a result they reached a total accuracy of 95%.

Anand and Suangthan (2009) in their study they used One-vs.-All Support Vector Machine (OVA-SVM) to select relative genes with cancer, and they used three different estimating probability methods from decision value. They noticed that the most consistent one was Platt's approach while isotonic approach was better for the datasets with unequal proportion of samples in different classes. The three probability methods affected on the improvement of the accuracy of the classification.

Mishra and Sahu (2011) in their study the genes are clustered by k-means and then Signal-to-Noise Ratio SNR ranking is used to get the features with the highest ranking from each cluster and then given to two classifiers like SVM and k-NN, then the genes are ranked using SNR and then also it is classified again. This experimental reached an accuracy of 99.3%.

Li et al. (2012) had obtained gene expression profiles of 55 early stage primary CRCs, 56 late stage primary CRCs, and 34 metastatic CRCs. They developed a novel gene selection algorithm Support Vector Machine Recursive Feature Elimination (SVM-RFE) by incorporating T-statistic to make (SVM-T-RFE). And they achieved an accuracy of classification 100%. And they compared (SVM-T-RFE) with (SVM-RFE) performance and they saw that (SVM-T-RFE) is more accurate than (SVM-RFE) for prediction using less than or equal number of selected genes.

Nikumbh et al. (2012) had used two techniques that are hybrid from Biogeography-based Optimization (BBO) and Random Forests (BBO – RF) and BBO – Support Vector Machines (SVM) with gene ranking as a heuristic for microarray gene expression analysis. Their results show that genes selected by (BBO-RF) and (BBO-SVM) are more accurate from previously reported algorithms.

2. MATERIALS AND METHODS

3.1. Cancer

During our lives, our bodies' cells divide healthily and replace other cells in an under controlled fashion, but when these cells divide and multiply in an uncontrolled way this causes the well-known disease that is cancer (Chang, 2018).

Cancer cells can grow everywhere in our bodies, because in our bodies trillions of cells divide everywhere. When cancer happens the natural process of division and growing up of the new cells breaks down, due to the survival of the damaged and the old cells that must die and be replaced, so the new cells grow up with no need of it, and these cells can divide continuously until it becomes a tumor. However, some types of cancer do not form solid tumors such as leukemias.

Cancer tumors are malignant, this means that it can invade other parts of the body. Also there is another way that cancer can transfer to other body parts which is through the blood or the lymph system, when cancer cells break off and travel to far organs from the first cancer tumor (What Is Cancer?, 2015).

What are the differences between cancer cells and normal cells?

1. Cancer cells are not as specialized as normal cells.
2. The ability of cancer cells to ignore the stopping signals which the normal cells respond to.
3. Cancer cells can evade the body immune system and hide from it, unlike the normal cells (What Is Cancer?, 2015).

Cancer major types

Cancer has more than 100 types which divided into the major groups below;

1. carcinoma: Which are the most common cancer with about 80% of cancer cases, that include colorectal, skin, breast, pancreas, and lung cancers, which are malignant growths in the tissue lining the organs of the body.
2. Sarcoma: Which is the cancer that appear in bone, fat, cartilage, blood vessels, muscle or other connective or soft tissues, this type is not common comparatively.
3. Leukemia: It is the cancer that starts in blood-forming tissue, like the bone marrow, and causes producing a big amounts of abnormal cells of blood.
4. Lymphoma: is the cancer of lymphocytes.
5. Melanoma: is the most dangerous skin cancer which that make skin-pigments.

3.1. Colorectal Cancer

If the cancer begins in the colon it will be called colon cancer, and if it begins in the rectum it will be called as rectal cancer. Those two types of cancer are known as colorectal cancer.

The trend of the new cases of colorectal cancer and the colorectal cancer deaths are decreasing yearly in adults who are 55 years or older. However, the number of colorectal cancer new cases in adults who are younger than 55 years is increasing slightly in recent years.

The colon is a digestive system part in our bodies. This systems responsibility is to remove and process foods to take the nutrients that the body needs, and pass out the unnecessary materials from the body. The last part of the digestive system is the large intestine, which begins with the colon (also called large bowel), and finishes with rectum and anal canal. The colon is approximately 5 feet long. And rectum with anal canal length is 6-8 inches. (Colorectal Cancer Prevention (PDQ)–Patient Version, 2019)

3.2.1. How colorectal cancer begins?

The majority of colorectal cancers begin as a growth in the inner lining of the rectum or the colon, which is called polyps. The polyps have different types, over time (usually after many years) some of them are able to convert to cancer. However, not any type of polyps will change into cancer. This changing ability depends on the polyp type. The 2 major polyps' types are:

Adenomatous polyps (adenomas): This type often becomes cancer, and due to that it called pre-cancerous condition.

Hyperplastic and inflammatory polyps: These types of polyps are more common, but they are safer from changing to cancer.

Also, there are another causes that make the polyps more dangerous to contain cancer and increase colorectal cancer's risk, and these factors are:

The size of the polyp if it was bigger than 1 cm.

If there were several polyps (more than 2).

There is another pre-cancerous condition which is called dysplasia. If this condition is found when it is removed the polyp will be more likely to change into cancer (Colorectal Cancer Prevention (PDQ)–Patient Version, 2019).

3.1.2. How does colorectal cancer spread?

Over time, cancer polyps can grow up in the internal wall of colon or rectum, which includes several layers. Colorectal cancer's beginning is in the mucosa which is the innermost layer and it can enlarge through the other layers.

During the cancer staying in the colon or rectum layers, its cells can transfer into the vessels of blood or lymph, so it can reach anywhere in the body.

The colorectal cancer stage depends on how the cancer grows deeply in the colon or the rectum wall (Colorectal Cancer Prevention (PDQ)–Patient Version, 2019).

3.1.3. Types of colorectal cancer

- Adenocarcinomas: This type is the major type of colorectal cancer, because nearly 96% of colorectal cancers are adenocarcinomas. This type of cancer begins in mucus cells which are responsible to lubricate the intestine inside. The doctors mean this type of colorectal cancer almost every time they talk about colorectal cancer. There are some adenocarcinoma's sub-types that might have a worse prognosis than the other sub-types, like mucinous and signet ring.
- Carcinoid tumors: This type of tumors begin from a specific hormone-making cells in the colon and rectum.
- Gastrointestinal stromal tumors (GISTs): These tumors begin Cajal interstitial cells in the colon's wall. This type of tumors can grow among the digestive tract, but this type is not found so often in colon. However, not all of them are cancer.
- Lymphomas: These cancers are the immune system cells' cancer. Their majority begin in lymph nodes, but their ability to begin in the rectum or colon make them studied as a type of colorectal cancer.
- Sarcomas: this is a rare type of colorectal cancer, which is able to begin in the connective tissues of the colon or the rectum wall, such as blood vessels or muscle layers (Colorectal Cancer Prevention (PDQ)–Patient Version, 2019).

3.2. Artificial Intelligence (AI)

"Artificial intelligence" is a term that was initiated 63 years before known in 1956. There are several definitions for artificial intelligence, one of them is "The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings" (Copeland, 2019). Another definition is the one that says "Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems, these processes include learning, reasoning, and self-correction" (Kumar, Artificial Intelligence: Definition, Types, Examples, Technologies, 2018).

3.3.1. Artificial intelligence subfields

Artificial intelligence has many subfields, and day by day the developments of AI make new fields that artificial intelligence works in, some of artificial intelligence subfields will be shown below briefly:

1. Machine Learning (ML): In this method the target will be defined, but the machine by itself will search for the steps until reaching the target by the training. For example, if somebody want to identify two simple things for a child, like apple and orange, he must show him some samples of each type and the child will differentiate between them by himself in the future, this is how machine learning works (Kumar, Artificial Intelligence: Definition, Types, Examples, Technologies, 2018).
2. Neural networks: they are some machine learning methods that use interconnected units that operate on the inputted data like the neurons, and by passing information through these units neural network finds a connection or make a meaning for this information.
3. Deep learning: Deep learning uses enormous neural networks that have many processing units, and due to the developments of the computing power and new techniques of training it is able to learn complex patterns from a huge data.
4. Internet of things: It is a system that connect mechanical and digital devices, animals, and people together. AI allow us to make these systems (Artificial intelligence, 2019).
5. Natural Language Processing (NLP): It is defined as the automatic software operations on natural language, such as text and speech. One of the familiar uses of this method is the email spam detection that has developed considerably in the mail system that is used nowadays.
6. Vision: It is possible to say that it is the field that make the machine see. This process is done by capturing pictures using cameras, and convert it to digital signals

that can be processed by the machine. Its best results are usually achieved by using machine learning so these both fields are interlinked.

7. Robotics: It is the field that focuses on the design and producing robots. Robots are often used to do some difficult or dangerous tasks for humans. Everyone can see the role of robots in many processes around as such as car assembly lines, medical uses, office cleaning, food serving, and many other processes.
8. Autonomous Vehicles: This is an area of AI that is under focus considerably nowadays. And there are many new types of vehicles that behave by itself like cars, buses, and trains. Also, many autonomous weapons are made last years that made a revolution in wars' tactics (Kumar, Artificial Intelligence: Definition, Types, Examples, Technologies, 2018).

Advanced algorithms: Many algorithms are being improved and combined together for the purpose of analyzing big data rapidly. These algorithms make it possible to predict rare events (Artificial intelligence, 2019).

3.3. Machine Learning

ML is a type of algorithms that make the applications more accurate in the prediction of outcomes without a previous programming. The basic idea of machine learning is building such algorithms that can take input data and by using statistical analyzing to predict the outputted (Rouse, 2018).

Machine learning makes it possible to analyze huge data quantities. Since it delivers more accurate results in a faster way to recognize gainful commercial opportunities the risks, it might need additional resources and time for training well, but merging machine learning artificial intelligence and other new technologies could make machine learning more effective in managing big amounts of data (What is Machine Learning? A definition, 2017).

Machine Learning is a part of artificial intelligence where the machine learns things by itself if it had sufficient data and enough computation capability. In this field

machine learns and gets experience by some thinking processes like what is done by every human beings, In a situation that it uses data as an experience and the computational power as thinking capability. Machine learning is the most used and usefulway to reach artificial intelligence because of the discoveries that has been done by (Kumar, Artificial Intelligence vs Machine Learning, 2018).

3.4.1. Machine learning algorithms' categories

- Supervised machine learning algorithms: These algorithms can use what it learned before to predict the new events using labeled examples. After sufficient training, these systems can predict the labels of new outputs. Also, they can compare the outputs with the outputs that are labeled correctly to make better results. An example for this type of algorithms is support vector machine.
- Unsupervised machine learning algorithms: These algorithms are used when there was no labeled or classified data to train on. In other words, this type is used if inputted data was without variables for the output data. So these algorithms can -+describe data structures from an unlabeled data, k-mean and K-NN (k nearest neighbors) are examples of unsupervised machine learning algorithms.
- Semi-supervised machine learning algorithms: This type stands between supervised learning algorithms and unsupervised learning algorithms, because these algorithms use labeled and unlabeled data in the training process. This method can improve the accuracy of learning in the systems that use it.
- Reinforcement machine learning algorithms: This is a type of algorithms, which have the ability to learn by the interacting with their environment by actions' producing and discovering errors or rewards. A simple feedback of rewards that known as reinforcement signal is needed for the system to know the best action this is known as the reinforcement signal (What is Machine Learning? A definition, 2017).

3.4. Classification Algorithms

Classification algorithms are methods that categorize data to classes, these algorithms train on some data to predict the class of another information. There are many classification algorithms and some of them are explained briefly below:

- Support vector machines (SVM): Support vector machine is a method that tries to build a hyper-plane or a set of them to separate between data classes, and it tries to achieve a hyper-plane that is the furthest from data-points of each class, and the distance between the nearest data-point and the hyper-plane is named "margin", so the classification errors' rate is lower in the big margin hyper-plane.

  Kernel functions: Kernel is a set of mathematical functions that works on the inputted data to transform it to the form that is required, and it has several kinds such as; linear, polynomial, RBF and sigmoid, but in this study three types of kernels were used, which are; Linear, Polynomial and Gaussian RBF.
- Decision Tree: Decision Tree is a learning algorithm that predicts the value of a target point based on other points' values which the algorithm trained on it, it uses a tree-like structure for the training consequence.
- Nearest neighbors: Nearest neighbors is an algorithm that predict the value of a point depending on the values of the closest points to it.
- Neural Network: Neural Networks are a group of algorithm that act like the human mind to recognize patterns, these algorithm take the input data as input layer and work on this data and transfer it to other layers until giving us the output layer.
- AdaBoost: AdaBoost stands for "Adaptive Boosting", AdaBoost is an ensemble classifier which means it had been made up of several classification algorithms, and it combines the results of these weak algorithms to give a powerful result.

3.5. Support Vector Machines (SVM) in Details

Firstly, there are some basics that are needed to be known to understand SVM method well, and to know the process that it works by. One of the words that it is used in this subject is “vector”.

3.6.1. Vector

Vector is a mathematical object that can be drawn as a narrow like (Figure 3.1):

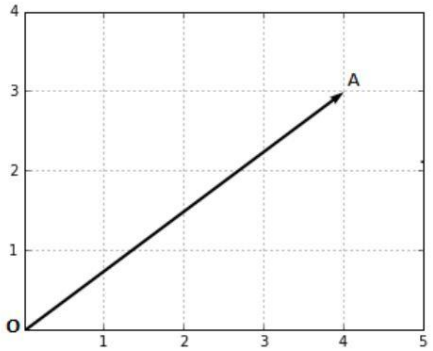

Figure 2.1. Vector.

This narrow begins from the origin point ‘O’ to the point ‘A’, so it could be written

$$\overrightarrow{OA} = (4, 3) \quad (3.1)$$

The vectors have two things, which are the length and the direction as in (Figure 3.2):

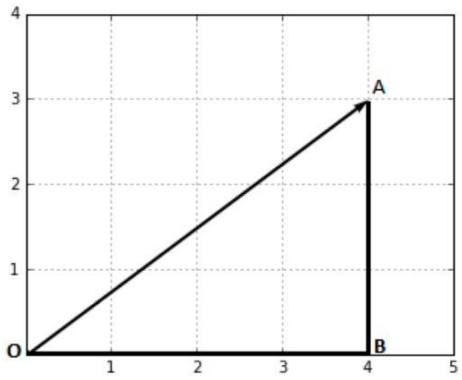

Figure 2.2. The magnitude of this vector is the length of the segment OA.

Vector's length $\|OA\|$ can be calculated by this equation

$$OA^2 = OB^2 + AB^2 \tag{3.2}$$

$$OA^2 = 4^2 + 3^2 \tag{3.3}$$

$$OA^2 = 25 \tag{3.4}$$

$$OA = \sqrt{25} \tag{3.5}$$

$$\|OA\| = OA = 5 \tag{3.6}$$

And the direction of the vector can be calculated by the way below:

$$w = \left(\frac{u1}{\|u\|}, \frac{u2}{\|u\|}\right) = (\cos\theta, \cos\alpha) \tag{3.7}$$

3.5.2. Linear Separability

There are some classification problems that it can be separated linearly, if these problems were in one dimension the separator will be a point, the two dimension separator must be a line, and three dimension problem will be separated by a plane as shown in (Figure 3.3), (Figure 3.4) and (Figure 3.5) below:

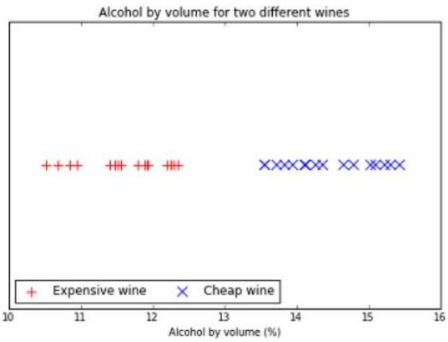

Figure 2.3. An example of linearly separable data.

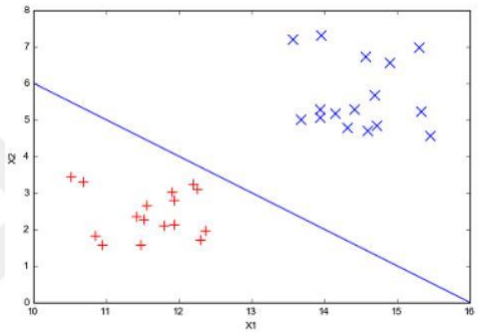

Figure 2.4. Data separated by a line.

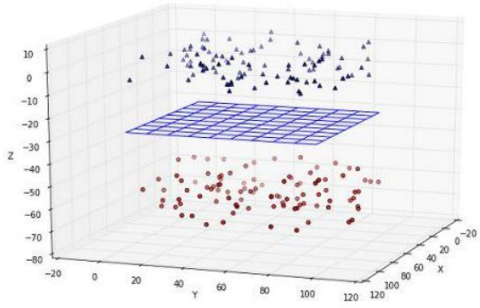

Figure 2.5. Data separated by a plane.

However there are some problem that are cannot be separated between their classes linearly, like the examples in (Figure 3.6), (Figure 3.7) and (Figure 3.8):

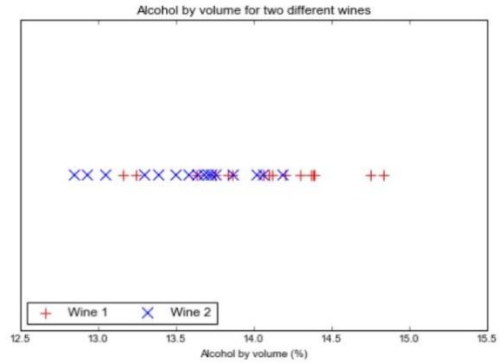

Figure 2.6. Non-linearly separated data in 1D.

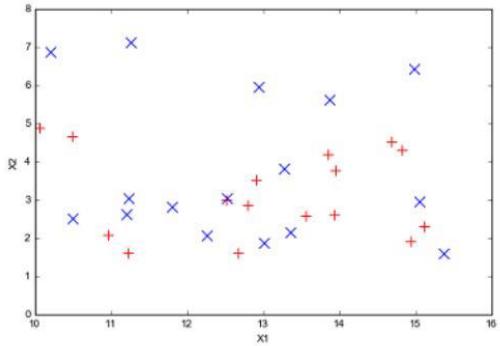

Figure 2.7. Non-linearly separated data in 2D.

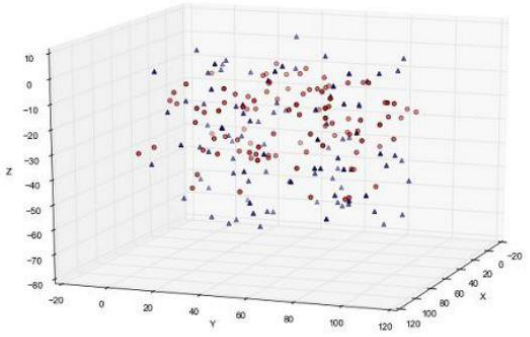

Figure 2.8. Non-linearly separated data in 3D.

3.5.3. Hyperplanes

Hyperplanes are the boundaries that linearly divide the n-dimensional data points in two component, so the data points fall on either side of the hyperplane can be attributed to different classes (Gandhi, 2018).

It is known that a line can be represented by this equation:

$$y = ax + b \tag{3.8}$$

$$ax - y + b = 0 \tag{3.9}$$

Assume that $x = x1$ and $y = x2$ so

$$ax1 - x2 + b = 0 \tag{3.10}$$

If two vectors were defined as $X = (x1, x2)$ and $W = (a, -1)$ the equation can be represented by the way below (know that $W \cdot X$ is the dot product of W and X):

$$W \cdot X + b = 0 \tag{3.11}$$

This is the equation of the hyperplanes of any dimensions, and it is the dimension that used for the vectors at the same time.

For better understanding the equation of hyperplanes two vectors will be defined which are $W = (w0, w1)$ and $X = (x, y)$, so it could be defined in this way:

$$W \cdot X + b = 0 \tag{3.12}$$

This is equal to:

$$w0x + w1y + b = 0 \tag{3.13}$$

$$w1y = -w0x - b \tag{3.14}$$

$$y = \frac{-w0}{w1}x - \frac{b}{w1} \tag{3.15}$$

If a and c were defined as:

$$a = -\frac{w0}{w1} \tag{3.16}$$

$$c = -\frac{b}{w1} \tag{3.17}$$

$$y = ax + c \tag{3.18}$$

3.5.4. How to classify data by a hyperplane?

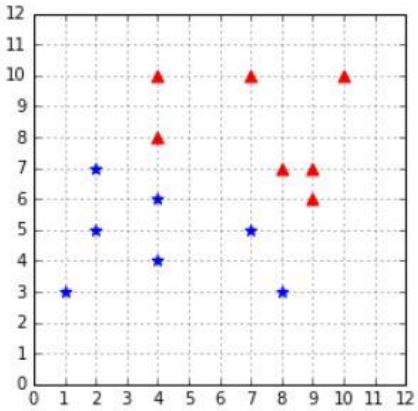

Figure 2.9. A linearly separated dataset.

This is a linearly separable data in (Figure 3.9), a hyperplane could be used to perform a binary classification, with the vector $W = (0.4,1.0)$ and $b = -9$ as shown in (Figure 3.10):

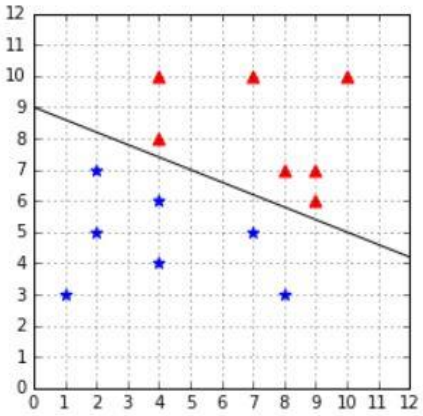

Figure 2.10. A hyperplane separating data.

Assume that each vector is X_i with a label y_i which can be +1 or -1(each value for a class), then a hypothesis function can be defined which is h

$$h(X_i) = \begin{cases} +1 \; if \; W \cdot X + b \geq 0 \\ -1 \; if \; W \cdot X + b < 0 \end{cases} \tag{3.19}$$

And this is equivalent to:

$$h(X_i) = sign(W \cdot X + b) \tag{3.20}$$

For example $X = (8,7)$ which is above the hyperplane its calculation will be as shown below:

$$W \cdot X + b = 0.4 \times 8 + 1 \times 7 - 9 = 1.2 \text{ So } h(X) = +1 \tag{3.21}$$

Another example is $X = (1,3)$ which stands below the hyperplane and $h(X)$ will be -1 because:

$$W \cdot X + b = 0.4 \times 1 + 1 \times 3 - 9 = -5.6 \tag{3.22}$$

With another trick it is possible to make h function simpler by removing the constant b, by adding a component $x_0 = 1$ to the vector $X_i = (x_1, x_2, \dots, x_n)$ so it will convert to: $\hat{X}_i = (x_0, x_1, \dots, x_n)$ also another component will be added to $w_0 = b$ to the vector $W = (w_1, w_2, \dots, w_n)$ then it will become $\widehat{W} = (w_0, w_1, \dots, w_n)$.

After this process the two crucial components that impact the shape of the hyperplane are included in W, because after the augmenting the vectors became like this: $X = (x_0, x_1, x_2)$ and $W = (b, a, -1)$ and it is known that b and a are the main components that define the look of the hyperplane. Due to this any change in the vector W will cause a change in the hyperplane as shown in the (Figure 3.11) below:

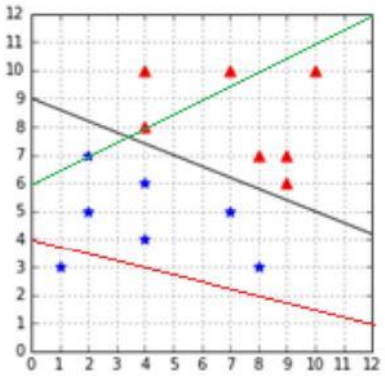

Figure 2.11. Different hyperplanes due to different values of w.

A variable β can be defined as $\beta = W \cdot X + b$ so the point that has the minimum distance between it and the hyperplane is needed, because of that

$$B = \min_{i=1\dots m} \beta_i \tag{3.23}$$

So if there were k hyperplanes the hyperplane with the biggest value will be selected by the equation ($\max_{i=1\dots k} B_i$).

But there will be a problem with the negative values because the β will decrease with the farther point, and the purpose is finding the nearest point to the hyperplane. Therefore, another trick will be added to the function:

$$B = \min_{i=1\dots m} |\beta_i| \tag{3.24}$$

Or there is another way by multiplying y which was the label of the points

$$f = y \times \beta \tag{3.25}$$

$$f = y(W \cdot X + b) \tag{3.26}$$

So, if f was positive the point is correctly classified, and if it was negative the point is classified incorrectly, and each point will be examined:

$$F = \min_{i=1\dots m} f_i \tag{3.27}$$

$$F = \min_{i=1\dots m} y_i(W \cdot X_i + b) \tag{3.28}$$

The hyperplane in this case can be rescaled by multiplying W and b by a number, for example 10, so if $W = (2,1)$ and $b = 5$ then they will be $W = (20,10)$ and $b = 50$. Due to that, the margin will be wider, and there will be an infinitive number of hyperplanes, to solve this problem W and b will be divided by $\|W\|$ the norm of W:

$$\gamma = y(\frac{W}{\|W\|} \cdot X + \frac{b}{\|W\|}) \tag{3.29}$$

As before, the margin will be computed by the way below:

$$M = \min_{i=1\dots m} \gamma_i \tag{3.30}$$

$$M = \min_{i=1\dots m} y_i(\frac{W}{\|W\|} \cdot X + \frac{b}{\|W\|}) \tag{3.31}$$

After these steps another important process is need to be done, which is optimization. The optimization problem aims to find the minimum or the maximum value of a function respecting a variable x. This thesis will not focus on the mathematical details of the optimization problem, but there are some resources that are useful for understanding this problem like Convex Optimization (Boyd & Vandenberghe, 2004). (Kowalczyk, 2017)

The previous explaining was about choosing the best hyperplane between two classes, which are linearly separable, but there are many problems that could not be separated it by a line.

3.5.5. Soft margin SVM

Soft-margin SVM is an improvement over the hard-margin SVM, this improvement make the classifier able to classify data accurately even if there was a noisy data. In other words, there is an issue with the hard margin SVM because not every dataset can be separated linearly, due to the outlier points. The outlier points have two cases; one of them is the case that a data point is closer to the other class points than its class as shown in (Figure 3.12), and this causes reducing in the margin, another case happens when a point is among the other class points as shown in (Figure 3.13), and this case breaks the linear separability.

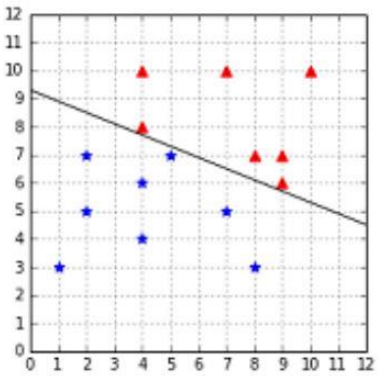

Figure 2.12. The dataset is linearly separable but with a narrow margin.

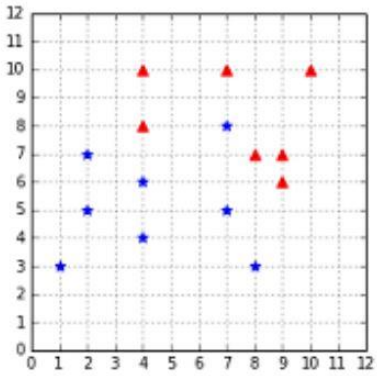

Figure 2.13. The point (7,8) broke the linear separability.

To solve this issue Vapnik and Cortes updated the SVM algorithm in 1995, this modification allows the classifier to make some mistakes, because the goal is not to avoid any mistake, but the aim is to classify in an accurate way, this modification was by adding another variable to the constraints of the optimization which is ζ (zeta). So the constraint:

$$y_i(W \cdot X_i + b) \geq 1 \tag{3.32}$$

Will become:

$$y_i(W \cdot X_i + b) \geq 1 - \zeta_i \tag{3.33}$$

There is another problem and that is every time a huge number could be chosen for ζ and the constraints will be satisfied, therefore to avoid this problem:

$$\text{minimize}_{W,b,\zeta} \quad \frac{1}{2}\|W\|^2 + \sum_{i=1}^{m} \zeta_i \tag{3.34}$$

$$\text{Subject to } y_i(W \cdot X_i + b) \geq 1 - \zeta_i \quad \text{for any } i = 1, \dots, m$$

Will become:

$$\text{minimize}_{W,b,\zeta} \qquad \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m} \zeta_i \tag{3.35}$$

$$\text{Subject to} \qquad y_i(W \cdot X_i + b) \geq 1 - \zeta_i$$
$$\zeta \geq 0 \quad \text{for any } i = 1, \dots, m$$

These mathematical details are shown in Comparison of L1 and L2 Support Vector Machines (Koshiba & Abe, 2003).

The differences in C values cause the differences in the hyperplane, but generally the small C gives wider margin with some incorrect classifications, and the huge C gives a hard margin with tolerance to zero violation of the constraints, so that value of C must be selected that reduce the influence of noisy data on the accuracy. There are some examples of the different values of C and its influence on the hyperplane in two problems shown in (Figure 3.14), (Figure 3.15) and (Figure 3.16):

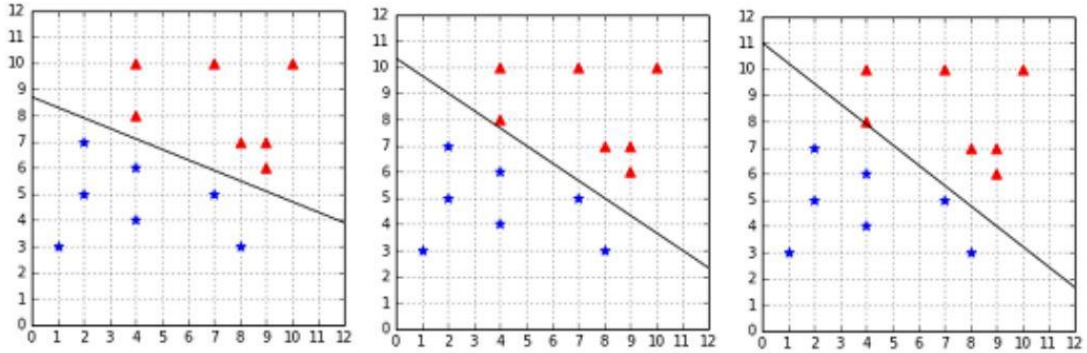

Figure 2.14. Effect of C=+infinity, C=1, and C=0.01 on a linearly separated dataset.

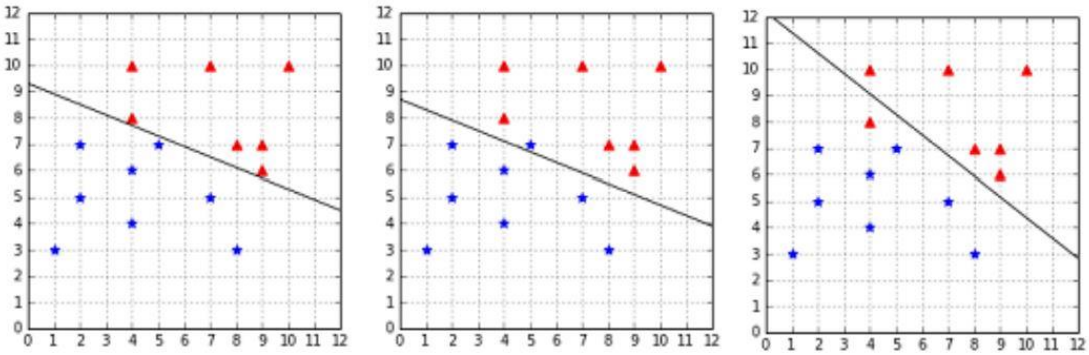

Figure 2.15 Effect of C=+infinity, C=1, and C=0.01 on a linearly separable dataset with an outlier.

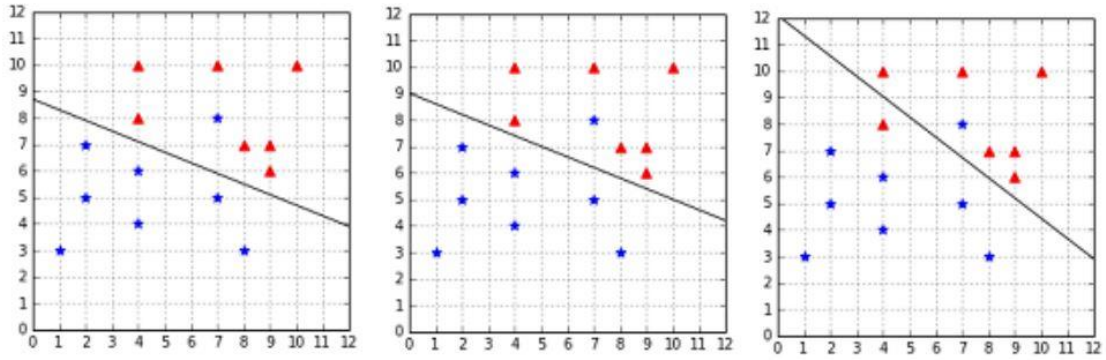

Figure 2.16. Effect of C=3, C=1, and C=0.01 on a non-separable dataset with an outlier.

So it is clear that there is no specific value of C that is satisfied with every problems, but each problem and dataset has a value of C that is the best value of it.

A recommended way to reach the best value of C is by using grid search and cross-validation (Kowalczyk, 2017).

Another formulation for the soft margin is 2-norm soft margin or L2 regularized soft margin, this formula is near to the 1-norm formula which is been explained before except adding ζ^2 instead of ζ as it shown below:

$$\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m}\zeta_i^2 \qquad (3.36)$$

3.5.6. Kernels

It is clear that there are many datasets that could not be separated linearly, like the (Figure 3.17) below:

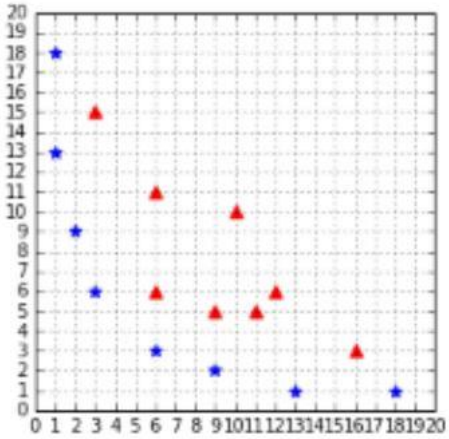

Figure 2.17. This data cannot be separated by a straight line.

This dataset is not able to be separated linearly in two dimensions, but what about converting it to three dimensions? It is possible to make these data points in three dimensions by applying polynomial mapping using the function $\emptyset: R^2 \rightarrow R^3$ defined by:

$$\emptyset(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{3.37}$$

It is clear that the data could be transacted not only to three dimensions, but it can be transacted to four, five, ten or hundred dimensions.

3.5.6.1. What is a kernel and its basic concept?

There is a drawback in the previous process, the problem is that every training data point must be transacted to three dimensions, and every testing data point also must be transacted, and this process need much time.

To solve this problem, the kernel can be used. Because kernel is a function that gives the result of a dot product like if it has been performed in another space.

For example: instead of transforming $X_1 = (3,6)$ and $X_2 = (10,10)$ to three dimensions and then performing dot product (which the result will be 8100), another way (a kernel) can be used, as below:

$$3^2 \times 10^2 + 2 \times 3 \times 10 \times 6 \times 10 + 6^2 \times 10^2 = 8100 \tag{3.38}$$

Without any doubt this kernel will save time.

3.5.6.2. Linear kernel

Linear kernel is the simplest type of kernels, if X and X' were vectors, the function of this kernel will be:

$$K(X, X') = X \cdot X' \tag{3.39}$$

3.5.6.3. Polynomial kernel

The function of this kernel is:

$$K(X, X') = (X \cdot X' + c)^d \tag{3.40}$$

There are two parameters in this function; c which is a constant, and d which is the degree of the kernel. If the degree d of a polynomial kernel was 1 and there was no constant the kernel will be the same as linear kernel, but when the degree increases the hyperplane becomes more complex and it could be influenced by individual data points. However, if the degree was so high, another drawback will happen which called "overfitting", in this case the hyperplane will be very close to the data points and this will be a problem for the testing data points. Examples of the previous explanation are shown below in (Figure 3.18), (Figure 3.19) and (Figure 3.20):

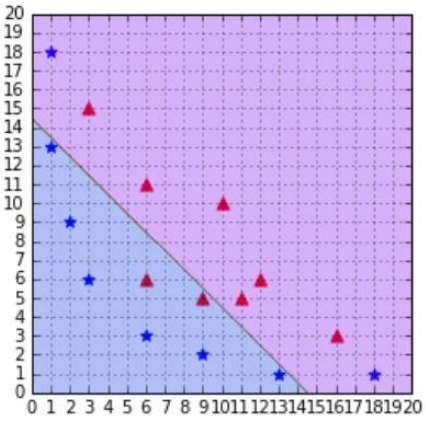

Figure 2.18. A Polynomial kernel degree=1.

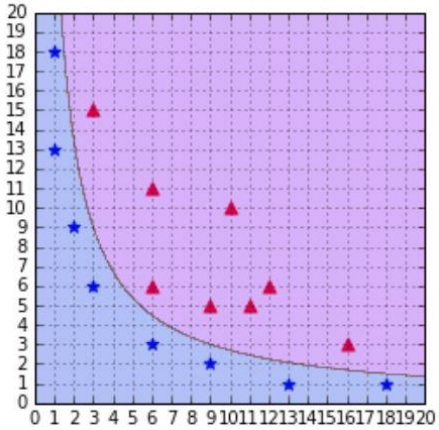

Figure 2.19. A Polynomial kernel degree=2.

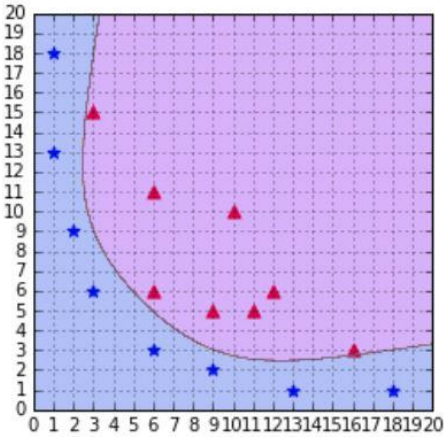

Figure 2.20. A polynomial kernel degree=6.

3.5.6.4. Radial Basis Function (RBF) or Gaussian kernel

There are some problems that even polynomial kernel could not solve it, because the dataset is very complex, like the problem in (Figure 3.21) below:

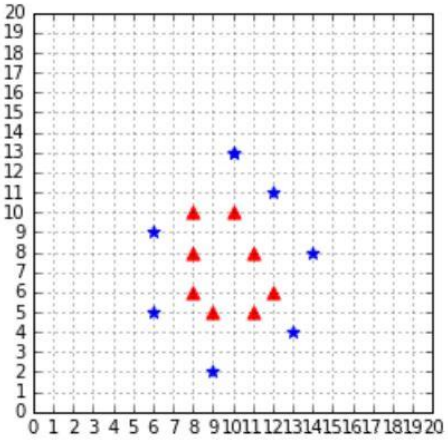

Figure 2.21. The data is more difficult to work with.

Such a problem the polynomial kernel cannot classify it, for example if the polynomial kernel been performed on this problem, the result will be as seen in (Figure 3.22):

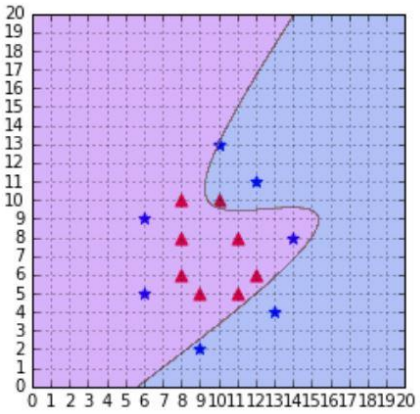

Figure 2.22. A polynomial kernel is not able to separate the data (degree=3, C=100).

So for this type of cases another complex kernel is needed which is RBF or Gaussian kernel, this type of kernels has a different function which is this:

$$K(X, X') = \exp(-\gamma \|X - X'\|^2) \tag{3.41}$$

The value of radial basis function depends on distance from a training point, and when the value of gamma γ increases the predicting area around the training data points decreases, like the graphs (Figure 3.23),(Figure 3.24) and (Figure 3.25):

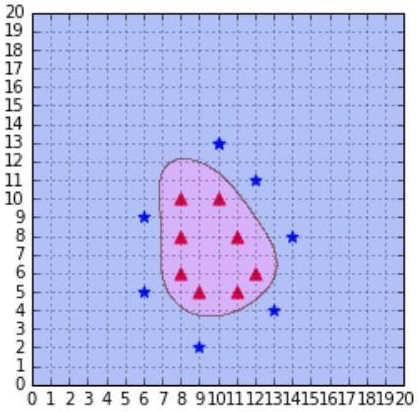

Figure 2.23. The RBF kernel classifies the data correctly with gamma = 0.1.

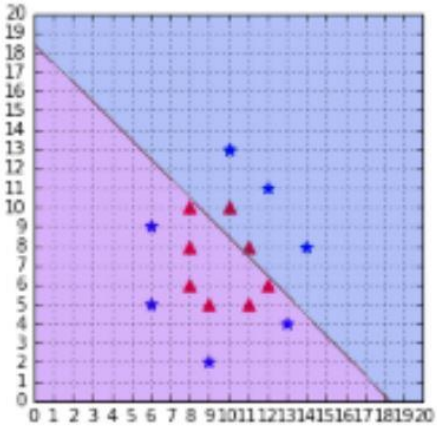

Figure 2.24. The RBF kernel classifies the data correctly with gamma = 1e-5.

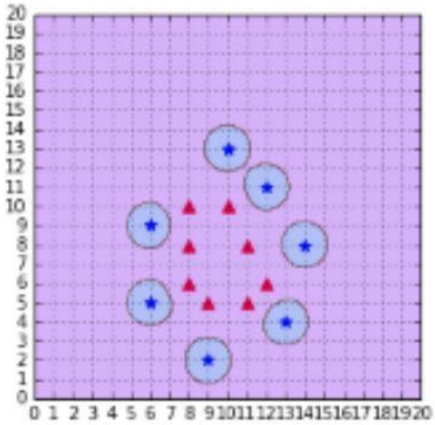

Figure 2.25. The RBF kernel classifies the data correctly with gamma = 2.

There are many other types of kernels, but these were the most famous among them. For more information about this subject (Kowalczyk, 2017), (Cristianini & Shawe-Taylor, 2000) and (Souza, 2010).

3.6. Dataset

The dataset that is used in this study is based on the data that had been collected by the U.S. Department of Health and Human Services/ National Institutes of Health/ National Cancer Institute, in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial in the U.S., the data were collected from nearly 155,000 participants between 1993 and 2001, and the data about cancer patients were collected in 2009 to know which patients had died, and the death cause, the time of the death and other information, so the average follow-up time of the patients' cases was 12.4 years. PLCO data tables include huge amount of information about the participants' cancer disease, their diet, history of other personal diseases, family diseases, participants' physical activity, occupation, marital status, geographic place and much more information.

3.7.1. Data preprocessing and the final dataset

Data tables that had been used were too complex and huge, and it contained a big number of columns, and this big number of variables was a problem for the classifiers and it contained many columns that are not important for the classification process, therefore the number of columns were reduced and just such columns remained which were related to the risk factors of colorectal cancer depending on information that were collected from an expert doctor (Mizory, 2019) and special cancer websites (cancer.org, cdc.gov and cancercenter.com). The data about the participants were related to these risk factors:

- Age: The probability of colorectal cancer diagnosis increases after the age 40, and it upwards more sharply after 50, due to that more than 90% of colorectal cancer cases occur after 50 (Haggar & Boushey, 2009). However it increased among

younger people also (O'Connel, et al., 2003) (O'Connell, Maggard, Livingston, & Yo, 2004).

- Personal history of polyps: A person with polyps history has a bigger risk of colorectal cancer, than who has no previous polyps history (De Jong, et al., 2005). Approximately 95% of colorectal cancer cases develop from these polyps (Society, 2019). Diagnosis and removal of the polyps before it transform malignantly may decrease colorectal cancer risk (Grande, et al., 2008).
- Personal history of inflammatory bowel disease (IBD): The term inflammatory bowel disease (IBD) used to describe ulcerative colitis and Crohn disease. This disease increases the overall risk of colorectal cancer (Grande, et al., 2008). Due to that, people who have of inflammatory bowel disease are need to be screened for colorectal cancer more frequently. In the person who has colon cancer comorbidity predicts his survival and it might effect on the treatment's benefit. However in rectal cancer cases comorbidity insignificantly impacted survival (Eaghen, Bakker, Bochove, & Loffeld, 2015).
- Family history of cancer: Colorectal cancer occurs mostly in people with no family colorectal cancer history. But still one in five developed colorectal cancer cases occurs in people who have another colorectal cancer patient in their families. So anyone has one colorectal cancer patient or more among his first-degree relatives has a bigger risk of this disease. There are stronger family history situations, such as someone who has a younger than age 60 colorectal cancer patient in his family, or someone who has two or more first-degree relative patients (Grande, et al., 2008).
- Diet: A diet with high rate of fat or red meats and processed meats raises the risk of colorectal cancer. Also the meats that are cooked in high temperature, these types of cooking make chemicals that might increase the risk of this disease. On the other hand, a diet that is high in vegetables and fruit can decrease the risk.
- Physical activity and obesity: It is proved that higher physical activity levels are related to the low colorectal cancer risk (Boyle & Langman, 2000). The obesity and being overweight considerably effect on the rates of estrogens and insulin, and

these hormones are believed to influence colorectal cancer risk (De Jong, et al., 2005).

- Smoking: Researchers got the result that 12% of colorectal cancer deaths occurs due to smoking (Zisman, Nickolov, Brand, Gorchow, & Roy, 2006). Tobacco includes carcinogens that increase the risk of colorectal cancer and the growth of this disease and larger polyps found in the colon and rectum were associated with long-term smoking (Grande, et al., 2008).
- Heavy alcohol use: Alcohol may increase the risk of developing colorectal cancer, and it leads to this disease at a younger age (Patel, 2017).

Also the dataset contained the participants' cancer situation, which was simplified into one column with only two values (0 for healthy participants and 1 for the patients in any stage of the disease), so the data table contained 22 columns finally.

3. RESULTS AND DISCUSSION

The programming language that had been used in this study is Python, Python is one of the best programming language in the world, because of its easiness, good performance, useful libraries, wide range of applications, and its wide use in the deep learning. Therefore, popular companies use this language like YouTube, Instagram, Pinterest, Survey Monkey, Quora, Mozilla, and Spotfy.

The process that this program works on consists of the data preparation for the classification, and implementing the classification. Firstly the program reads the data from a comma separated values (.csv) file. As explained before this file includes 22 columns and more than 60000 rows, each row includes the information about an individual participant, and approximately 700 from are patients of colorectal cancer, which means 1.1% of the whole participants, this shows that data is unbalanced, therefore only the first 5000 participants are chosen for the next steps, by decreasing the number of participants, the ratio of patients increases to near to 14%, and this is a good balance between the two classes.

After that, the data was separated into two subset; training subset, and testing subset. The purpose is that the program will train on the training subset to predict the patterns that new data points from testing subset will stand in. Splitting the dataset was done by the function train_test_split from sklearn library, this function has four major parameters; the first one carries the dataset without the labeling column, the second one carries the labels' columns, the third parameter is for the size of testing subset, and the last one is random_state. The size of testing subset is chosen to be 25% of the whole data, and this quarter is chosen randomly, the random_state saves the results of the randomly splitting, so during the implementation of the program more than one time the subsets will be the same if the random_state value did not change.

After this process, the classification algorithm which is SVM in this case is implemented on the training subsets, and then the program is ready to predict the data points in testing subset. This process is shown in the (Figure 4.1) below:

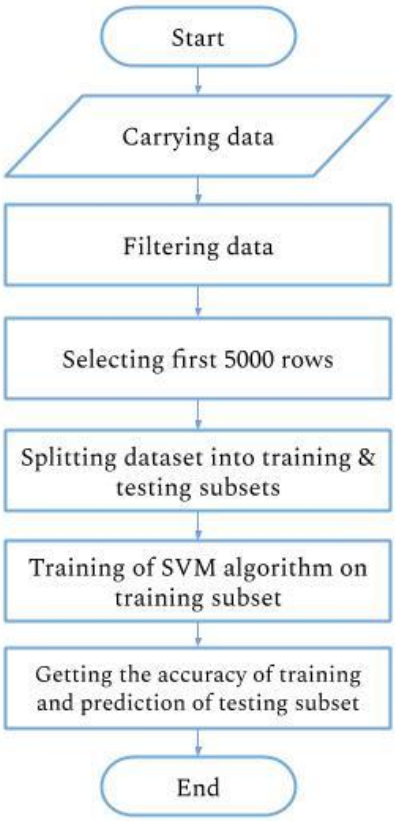

Figure 3.1. The simple classification process.

In this study three types of kernels had been used, which are; linear kernel, polynomial kernel, and Gaussian (RBF) kernel. As explained before, each kernel has its variables that can influence the result, so to get good results another development has been done on the process. Firstly, in the splitting step, when the data is been separated into two subset, the size of testing subset is 25% of the whole data, so from four rows of the first class participants the program chooses one row to put in the testing subset, these data points are chosen randomly. Also, the data points are chosen from the second class in the same way. Each time, which the dataset is split the result of the splitting or the new subsets are saved by a different random state (random_state) value, so the differences in the random state cause change in the accuracy of the training and prediction even if it was tiny.

To solve this problem, 100 different random state values had been used in this classification program, so by getting the results of all 100 classifications and getting the

average of the results this problem is solved and the performance of the classifier is known.

The linear kernel is influenced also by the value of constant c, so in this program for each random state the classification had implemented using 10 different c values, and in the end of the process the 100 classification results of each c value are averaged to know the best c value. As a result, the classification had been implemented 1000 times, and the average result of each c value had been gotten, in addition to the best training accuracy, and the best testing prediction. So the process has been done in a more complicated way as shown in the (Figure 4.2):

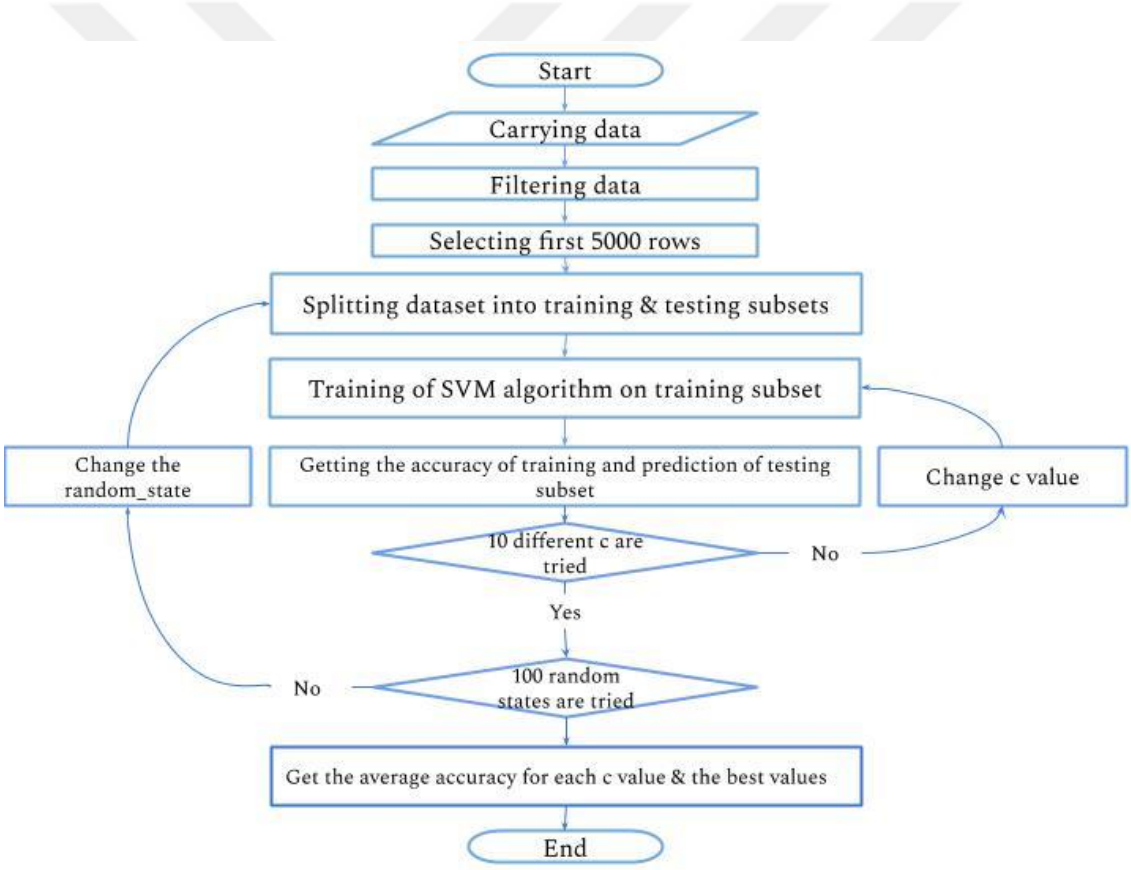

Figure 3.2. Study linear kernel SVM implementation.

Another kernel that is used in this study is the Gaussian or Radial Basis Function (RBF) kernel, in this type of kernels the value of gamma is important, so

instead of implementing the classification with different c values, 10 different values of gamma had been used, and in the same previous way the average accuracy results of each gamma value and the best results are been gotten. The process steps are shown in the (Figure 4.3):

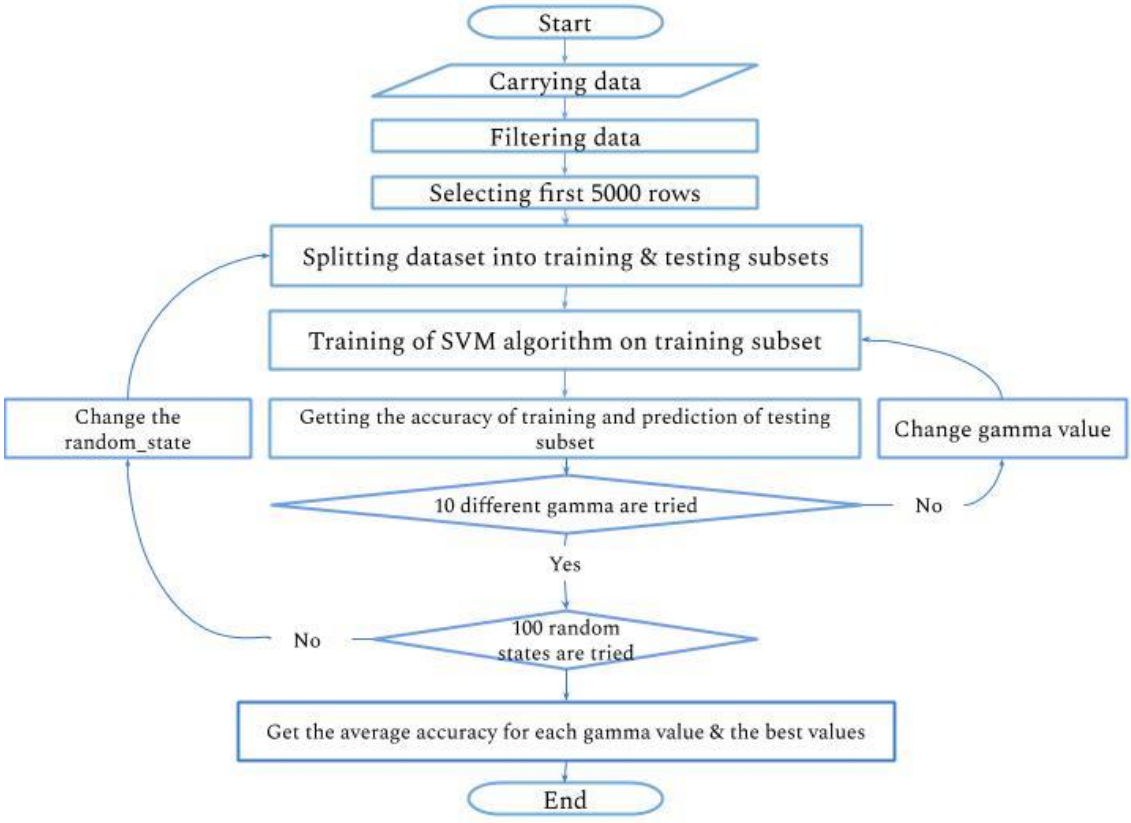

Figure 3.3. Study RBF kernel SVM implementation.

The last type of kernels that is used in this study is polynomial kernel, the results of this type of kernels are influenced by two parameters; the constant c and the number of dimensions d, so to get the best results for each random state 10 different c values are used, and for each c value the classification had been implemented using 10 different numbers of dimensions d, so the classification had been done 10000 times, and the average accuracy for each d value is gotten. Also, the best values among the 10000 implementations had been gotten.

4.1. Results

In the linear kernel the average accuracy that the program reached by each value of constant c are shown in the (Table 4.1):

Table 3.1. Linear kernel SVM accuracy averages depending on the value of C.

Constant c values	**Training accuracy average**	**Testing accuracy average**
$\boldsymbol{c = 0.125}$	98.10%	98.12%
$\boldsymbol{c = 0.25}$	98.13%	98.16%
$\boldsymbol{c = 0.5}$	98.15%	98.19%
$\boldsymbol{c = 1}$	98.17%	98.20%
$\boldsymbol{c = 2}$	98.17%	98.21%
$\boldsymbol{c = 4}$	98.18%	98.21%
$\boldsymbol{c = 8}$	98.18%	98.21%
$\boldsymbol{c = 16}$	98.19%	98.22%
$\boldsymbol{c = 32}$	98.20%	98.23%
$\boldsymbol{c = 64}$	98.21%	98.24%

And the best values are demonstrated in the (Table 4.2):

Table 3.2. Best accuracies of linear kernel SVM.

	Accuracy	**Kernel constant *c***
Best training accuracy	98.45%	$c = 32$
Best testing accuracy	98.88%	$c = 4$

The other kernel that is used is Gaussian (RBF), the (Table4.3) below shows the average accuracies among different gamma values:

Table 3.3. RBF kernel SVM accuracy averages depending on the value of gamma.

Gamma values	**Training accuracy average**	**Testing accuracy average**
$\boldsymbol{gamma = 0.00001}$	88.96%	88.97%
$\boldsymbol{gamma = 0.0001}$	95.77%	95.79%
$\boldsymbol{gamma = 0.001}$	96.49%	96.52%
$\boldsymbol{gamma = 0.01}$	97.10%	96.68%
$\boldsymbol{gamma = 0.1}$	98.80%	87.93%
$\boldsymbol{gamma = 1}$	100%	86.36%

Table 3.3. RBF kernel SVM accuracy averages depending on the value of gamma (continue).

Gamma values	**Training accuracy average**	**Testing accuracy average**
$\boldsymbol{gamma = 10}$	100%	86.36%
$\boldsymbol{gamma = 100}$	100%	86.36%
$\boldsymbol{gamma = 1000}$	100%	86.36%
$\boldsymbol{gamma = 10000}$	100%	86.36%

The best accuracies of the classification using RBF kernel of SVM algorithm are shown in the (Table 4.4):

Table 3.4. Best accuracies of RBF SVM.

	Accuracy	**Kernel gamma value**
Best training accuracy	100%	$gamma \geq 1$
Best testing accuracy	97.52%	$gamma = 0.01$

The third kernel that is used is polynomial kernel, as explained before, this kernel has two parameters that influence the accuracy result, which are; constant c and the number of dimensions or degree d, so the SVM algorithm had been implemented 10 times with 10 different d values with each of 10 values of c, as a result SVM algorithm had performed 100 times with each of 100 random states, the average training accuracies of each degree with every constant are shown in the (Table 4.5) below (each column is a different d value and each row is a different c value):

Table 3.5. Polynomial kernel SVM training accuracy averages depending on C and d.

	d=1	**d=2**	**d=3**	**d=4**	**d=5**	**d=6**	**d=7**	**d=8**	**d=9**	**d=10**
C=0.125	90.32 %	93.08 %	94.45 %	95.13 %	95.36 %	95.55 %	95.77 %	95.98 %	96.13 %	96.29 %
C=0.25	92.66 %	94.85 %	95.76 %	95.92 %	96.13 %	96.34 %	96.45 %	96.51 %	96.63 %	96.66 %
C=0.5	94.53 %	96.04 %	96.14 %	96.24 %	96.57 %	96.68 %	96.81 %	96.99 %	97%	97.06 %
C=1	96.04 %	96.34 %	96.48 %	96.77 %	96.96 %	97.07 %	97.08 %	97.19 %	97.24 %	97.33 %
C=2	96.25 %	96.47 %	96.80 %	97.07 %	97.16 %	97.29 %	97.42 %	97.46 %	97.51 %	97.59 %
C=4	96.45 %	96.78 %	97.12 %	97.31 %	97.45 %	97.51 %	97.61 %	97.69 %	97.75 %	97.74 %

Table 3.5. Polynomial kernel SVM training accuracy averages depending on C and d (continue).

	d=1	**d=2**	**d=3**	**d=4**	**d=5**	**d=6**	**d=7**	**d=8**	**d=9**	**d=10**
C=8	96.69 %	97.14 %	97.44 %	97.57 %	97.66 %	97.76 %	97.79 %	97.81 %	97.84 %	97.86 %
C=16	96.98 %	97.49 %	97.64 %	97.72 %	97.77 %	97.87 %	97.94 %	97.98 %	97.98 %	97.96 %
C=32	97.44 %	97.79 %	97.87 %	97.91 %	97.98 %	98.03 %	98.03 %	98.07 %	98.10 %	98.08 %
C=64	97.80 %	97.94 %	98.02 %	98.08 %	98.11 %	98.12 %	98.12 %	98.14 %	98.14 %	98.17 %

(Table 4.6) shows the average testing accuracies:

Table 3.6. Polynomial kernel SVM testing accuracy averages depending on the value of C and d.

	d=1	**d=2**	**d=3**	**d=4**	**d=5**	**d=6**	**d=7**	**d=8**	**d=9**	**d=10**
C=0.125	90.33 %	93.07 %	94.49 %	95.16 %	95.42 %	95.59 %	95.79 %	96.03 %	96.13 %	96.26 %
C=0.25	92.71 %	94.89 %	95.79 %	95.96 %	96.16 %	96.39 %	96.49 %	96.54 %	96.62 %	96.68 %
C=0.5	94.57 %	96.09 %	96.19 %	96.28 %	96.61 %	96.67 %	96.85 %	96.99 %	97.04 %	97.09 %
C=1	96.09 %	96.39 %	96.53 %	96.81 %	97%	97.13 %	97.13 %	97.21 %	97.25 %	97.33 %
C=2	96.32 %	96.51 %	96.84 %	97.12 %	97.22 %	97.32 %	97.45 %	97.48 %	97.49 %	97.51 %
C=4	96.50 %	96.83 %	97.19 %	97.38 %	97.48 %	97.51 %	97.60 %	97.68 %	97.77 %	97.63 %
C=8	96.72 %	97.20 %	97.49 %	97.60 %	97.67 %	97.77 %	97.77 %	97.77 %	97.74 %	97.71 %
C=16	97.03 %	97.52 %	97.65 %	97.73 %	97.80 %	97.86 %	97.90 %	97.88 %	97.81 %	97.72 %
C=32	97.49 %	97.79 %	97.85 %	97.91 %	97.97 %	97.97 %	97.93 %	97.90 %	97.84 %	97.77 %
C=64	97.80 %	97.93 %	98.01 %	98.06 %	98.05 %	98.01 %	97.94 %	97.89 %	97.83 %	97.78 %

The best accuracies of polynomial kernel based SVM are demonstrated in the (Table 4.7):

Table 3.7. Best accuracies of polynomial SVM.

	Accuracy	**Kernel c value**	**Number of dimensions(d)**
Best training accuracy	98.48%	$c = 64$	$d = 6$
Best testing accuracy	98.88%	$c = 64$	$d = 6$

4.2. Best Results in Detail

The details of the best results among all implementations of each kernel are shown below according to each kernel. In addition to the best results with implementation on other different numbers of participants for the comparison.

In the linear kernel the details of the best result are shown in the (Table 4.8):

Table 3.8. Best accuracies of linear kernel SVM

Participants' number	**Patients' percentage**	***c***	**Training accuracy**	**Testing accuracy**	**specificity**	**sensitivity**
5000	14%	4	97.92%	98.88%	100%	90.79%
1400	50%	0.125	97.71%	99.71%	100%	99.39%
2800	25%	0.125	97.71%	99.571%	100%	98.37%
7000	10%	32	97.94%	98.74%	100%	87.06%
10000	7%	0.125	97.99%	98.80%	100%	80.12%
60675	1.16%	0.125	98.87%	98.91%	100%	0.0%

The best result of the RBF kernel are in the (Table 4.9):

Table 3.9. Best accuracies of RBF kernel SVM

Participants' number	**Patients' percentage**	**gamma**	**Training accuracy**	**Testing accuracy**	**specificity**	**sensitivity**
5000	14%	0.01	96.93%	97.52%	99.91%	79.17%
1400	50%	0.01	96.38%	98%	99%	96.67%
2800	25%	0.01	96.38%	98%	100%	92.39%
7000	10%	0.001	97.11%	98%	100%	78.26%
10000	7%	0.001	97.83%	98.60%	100%	78.52%
60675	1.16%	0.001	98.87%	98.91%	100%	0.0%

In the polynomial kernel the best result details were these that are showen in (Table 4.10):

Table 3.10. Best accuracies of polynomial kernel SVM.

Participants' number	**Patients' percentage**	***d***	***c***	**Training accuracy**	**Testing accuracy**	**specificity**	**sensitivity**
5000	14%	6	64	97.78%	98.88%	100%	90.79%
1400	50%	4	32	97.52%	99.43%	100%	98.78%
2800	25%	5	64	97.71%	99.42%	100%	97.82%
7000	10%	6	8	97.70%	98.69%	100%	86.47%
10000	7%	2	32	97.96%	98.80%	100%	81.59%
60675	1.16%	1	0.125	98.87%	98.91%	100%	0.0%

4.3. Discussion

The results that are shown above show that linear kernel is better than other kernels working on this dataset, because it could not be said that a type of kernels is the best kernel for every problem, at the same time linear kernel is the simplest kernel mathematically. Also, the other kernels usually do well when its parameters are in a situation that make a similar hyperplane to linear kernel's hyperplane.

The best accuracy that obtained was accurate by 97.92% on training subset, and the accuracy of predicting testing subset data points was 98.88%, and that was a superb result.

Overall, these results are good and this is because of two major points; The efficiency of SVM algorithm in the classification, and the more effective is the small number of patients, because generally the algorithm does not make mistakes in the first class who are healthy, but the wrongly predictions are mostly from the second class which includes colorectal cancer patients, therefore the mistakes even if they were not acceptable in the second class they are few in comparison to the first class, for example if the classifier did not predict any the accuracy will not be less than 86% if all first class data points were predicted truly.

4. CONCLUSION

This study aimed to take step forward in the prediction and early detection of a dangerous disease that causes the death of thousands of people around the world every year, which is colorectal cancer. While this disease is curable, easy for treatment, costs less, and less dangerous in its early levels, so the importance of this study is related to the importance of saving numerous lives yearly, everywhere in our world.

In this thesis the difference had been shown between this study and the earlier studys and research that are near to its subject and purpose. In addition to that, cancer disease generally, and colorectal cancer specially had been explained briefly. Also, the technology that had been used in the study and helpful information had shown that cause the clearness of the main subject and objectives of this study.

The results of this study had been gotten after implementing the program for more than 12000 times, and it had been demonstrated and explained in detail in the previous chapter.

As told before, this study is a good step in this field, but this does not mean that this field does not need more efforts and research to reach its final goals, but the researchers in the medical side must do their best to collect such useful data to reach better and critical data patterns that makes the prediction easier and more effective, and the programmers and data scientists must work continuously to develop new methods for better classification of data, and predict these types of diseases that make the danger on our lives.

In future, a crucial work that should be focus on, is collecting data in our local areas, because a dataset that based on the people of the United States of America may not be useful to predict colorectal cancer or other diseases in the people of Middle East, and then programs should be developed based on the new local datasets.

If academicians and researchers helped each other -and they must do- a better and healthier future will be waiting for the new generations.

REFERENCES

Akay, M., 2009. support vector machines combined with feature selection for breast cancer diagnosis. ***Expert systems with appliations, 36***:3240-3247.

Alba, E., Garcia-Nieto, J., Jourdan, L., Talbi, E., 2007. Gene Selection in Cancer Classification using GPSO/SVM and GA/SVM Hybrid Algorithms. ***Evolutionary Computation (CEC), 2007 IEEE Congress***, 284-290.

Anand, A., Suganthan, P., 2009. Multiclass cancer classification with support vector machines class-wise optimized genes and probability estimates. *Journal of* ***Theoretical Biology, 259***:533-540.

Artificial intelligence, 2019. ***https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html***. SAS. Date accessed: 25.11.2019

Boyd, S., Vandenberghe, L., 2004. ***Convex Optimization***. Cambridge University Press.

Boyle, P., Langman, J., 2000. ABC of colorectal cancer: Epidemiology. ***BMJ, 321***:805-808.

Chang, L., 2018, January 27. Understanding Cancer -- the Basics. ***https://www.webmd.com/cancer/understanding-cancer-basics***.WebMD. Date accessed: 25.11.2019

Chen, X., 2003. Gene Selection for Cancer Classification Using Bootstrapped Genetic Algorithms and Support Vector Machines. ***Computational Systems Bioinformatics Conference, 2003. CSB 2003. Proceeding 2003 IEEE,*** 504.

Chu, F., Wang, L., 2005. Applications of Support Vector Machine to cancer classification with microarray data. ***International Journal of Neural Systems, 15***:475-484.

Colorectal Cancer Prevention (PDQ)–Patient Version, 2019, March 15. ***https://www.cancer.gov/types/colorectal/patient/colorectal-prevention-pdq*** National Cancer Institute. Date accessed: 25.11.2019

Copeland, B. (2019, May 9). Artificial Intelligence. ***https://www.britannica.com/technology/artificial-intelligence***. Encyclopedia Britannica. Date accessed: 25.11.2019

Cristianini, N., Shawe-Taylor, J., 2000. ***An Introduction to Support Vector Machines.*** Cambridge University Press.

De Jong, A., Morreau, H., Nagengast, F., Mathus-Vliegen, E., Kleibeuker, J., Griffioen, G., . . . Vasen, H., 2005. Prevalence of adenomas among young individuals at average risk for colorectal cancer. ***Am J Gastroenterol, 100***:139-143.

Del Rio, M., Molina, F., Bascoul-Mollevi, C., Copois, V., Bibeau, F., Chalbos, P., . . . Ychou, M., 2007. Gene Expression Signature in Advanced Colorectal Cancer Patients Select Drugs and Response for the Use of Leucovorin, Fluorouracil, and Irinotecan. ***Journal of Clinical Oncology, 25***:773.

Duan, K., Rajapakse, J., 2004. SVM-RFE peak selection for cancer classification with mass spectrometry. ***Proceeding of the 3rd Asia-Pacific Bioinformatics conference***, 191-200.

Duan, K., Rajapakse, J., Nguyen, M., 2007. One-Versus-One and One-Versus-All Multiclass SVM-RFE for Gene Selection in Cancer Classification. ***European***

Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, 45-56.

Duan, K., Rajapakse, J., Wang, H., Azyuaje, F., 2005. Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data. ***IEEE transactions on nanobioscience, 4***:228-234.

Eaghen, E., Bakker, S., Bochove, A., Loffeld, R., 2015. Impact of age and comorbidity on survival in colorectal cancer. ***Journal of Gastrointestinal Oncology, 6***:605-612.

Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., . . . Hartwell, A., 2003. The case for early detection. ***Nature Reviews Cancer, 3***:243-252.

Gandhi, R., 2018, June 7. Support Vector Machine — Introduction to Machine Learning Algorithms. ***https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47***. Medium. Date accessed: 25.11.2019

Grande, M., Milito, G., Attina, G., Caddedu, F., Muzi, M., Nigro, C., . . . Farinon, A., 2008. Evaluation of clinical, laboratory and morphologic prognostic factors in colon cancer. ***World J Surg Oncol., 6***:98.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification using Support Vector Machines. ***Machine Learning, 46:***389-422.

Haggar, F., Boushey, R., 2009. Colorectal Cancer Epidemology: Incidence, Mortality, Survival, and Risk Factors. ***Clinics in Colon and Rectal Surgery, 22:***191-197.

Hong, J., Cho, S., 2008. A probabilistic multi-class strategy of one-vs-rest support vector machines for cancer classification. ***Neurocomputing, 71***:3275-3281.

Jemal, A., Bray, F., Mellisa, M., Ferlay, J., Ward, E., Forman, D., 2011. Global Cancer Statistics. ***CA: A Cancer Journal for Clinicans, 61:***69-90.

Kominami, Y., Yoshida, S., Tanaka, S., Sanomora, Y., Hirakawa, T., Raytchev, B., . . . Chayama, K., 2016. Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. ***Gastrointestinal endoscopy, 83:***643-649.

Koshiba, Y., Abe, S., 2003. Comparison of L1 and L2 Support Vector Machines. ***Proceedings of the International Joint Conference on Neural Networks, 2003.*** Portland: IEEE.

Kowalczyk, A., 2017. ***Support Vector Machines Succinctly***. Morrisville: Syncfusion.

Kumar, C., 2018, September 1. Artificial Intelligence vs Machine Learning. ***https://medium.com/@chethankumargn/artificial-intelligence-vs-machine-learning-3c599637ecdd***. Medium. Date accessed: 25.11.2019

Kumar, C., 2018, August 31. Artificial Intelligence: Definition, Types, Examples, Technologies. ***https://medium.com/@chethankumargn/artificial-intelligence-definition-types-examples-technologies-962ea75c7b9b***. Medium. Date accessed: 25.11.2019

Lanza, G., Ferracin, M., Gafa, R., Veronese, A., Spizzo, R., Pichiorri, F., . . . Nergini, M., 2007. mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. ***Molecular cancer, 6:***54.

Li, S., Chen, G., Zhang, Y., Guo, Z., Liu, Z., Xu, J., . . . Lin, L., 2014. Identification and characterization of colorectal cancer using Raman spectroscopy and feature selection techniques. ***Optics express, 21:***25895-25908.

Li, X., Peng, S., Chen, J., Lü, B., Zhang, H., Lai, M., 2012. SVM-T-RFE: A novel gene selection algorithm for identifying metastasis-related genes in colorectal cancer using gene expression profiles. ***Biochemical and biophysical research communications, 419:***148-153.

Lin, H., Friederichs, J., Black, M., Mages, J., Rosenberg, R., Guilford, P., . . . Reeve, A., 2007. Multiple Gene Expression Classifiers from Different Array Platforms Predict Poor Prognosis of Colorectal Cancer. ***Clinical Cancer Research, 13:***498-507.

Liu, J., Cutler, G., Li, W., Pan, Z., Peng, S., Heoy, T., . . . Ling, X., 2005. Multiclass cancer classification and biomarker discovery using GA-based algorithms. ***Bioinformatics, 21:***2691-2697.

Liu, Y., 2004. Active Learning with Support Vector Machine Applied to Gene Expression Data for Cancer Classification. ***Journal of chemical information and computer sciences, 44:***1936-1941.

Machhale, K., Nandpuru, H., Kapur, V., Kosta, L., 2015. MRI Brain Cancer Classification Using Hybrid Classifier (SVM-KNN). ***Industrial Instrument and Control (ICIC) 2015 international Conference,*** (pp. 60-65).

Mao, Y., Zhou, X., Pi, D., Sun, Y., Wong, S., 2005. Multiclass Cancer Classification by Using Fuzzy Support Vector Machine and Binary Decision Tree with Gene Selection. ***Journal of BioMed Reasearch International 2005, 2:***160-171.

Maulik, U., Chakraborty, D., 2014. Fuzzy Preference Based Feature Selection and Semi supervised SVM for Cancer Classification. ***IEEE transaction on nanobioscience, 13:***152-160.

Mishra, D., Sahu, B., 2011. Feature Selection for Cancer Classification: A Signal-to-noise Ratio Approach. ***International Journal of Scientific Engineering Research, 2:***1-7.

Mizory, R., 2019, May 2. ***Interview with Dr. Ramadhan Mizory.*** (B. Mateen, Interviewer)

Nikumbh, S., Ghosh, S., Jayaraman, V., 2012. Biogeography-Based Informative Gene Selection and Cancer Classification Using SVM and Random Forests. ***Evolutionary Computation (CEC), 2012 IEEE Congress,*** 1-6.

O'Connel, J., Maggard, M., Liu, J., Etzioni, D., Livingston, E., Ko, C., 2003. Rates of colon and rectal cancers are increasing in young adults. ***Am Surg., 69***:866-872.

O'Connell, J., Maggard, M., Livingston, E., Yo, C., 2004. Colorectal cancer in the young. ***Am J Surg., 187:***343-348.

Patel, S., 2017, June 2. Machine Learning 101. ***https://medium.com/machine-learning-101/https-medium-com-savanpatel-chapter-6-adaboost-classifier-b945f330af06***. Medium. Date Accessed: 25.11.2019

Peng, S., Xub, Q., Lingc, X., Pengd, X., Dua, W., Chenb, L., 2003. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. ***FEBS letters, 555:***358-362.

Polat, K., Güneş, S., 2007. Breast cancer diagnosis using least square support vector mahine. ***Digital signal processing, 17:***694-701.

Rifkin, R., Mukherjee, S., Tamayo, P., Ramaswamy, S., Yeang, C., Angelo, M., . . . Mesirov, J. 2003. An Analytical Method for Multiclass Molecular Cancer Classification. ***Siam Review, 45:***706-723.

Rouse, M., 2018, May. machine learning (ML). TechTarget: ***https://searchenterpriseai.techtarget.com/definition/machine-learning-ML***. TechTarget. Date accessed:25.11.2019

Sewak, M., Vaidiya, P., Chan, C., Duan, Z., 2007. SVM Approach to Breast Cancer Classification. ***Computer and computational sciences 2007. IMSCCS 2007. Second International Multi-symposium,*** 32-37.

Shen, L., Tan, E., 2005. Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data. ***IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),*** 2:166-175.

Society, A. C., 2019. ***Cancer Facts & Figures 2019.*** Atlanta: American Cancer Society.

Souza, C. 2010, March 17. Kernel Functions for Machine Learning Applications. ***http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/***. Cesar Souza. Date accessed: 25.11.2019

Statnikov, A., Aliferis, C., 2008. Are Random Forests better than Support Vector Machines for microarray-based cancer classification. ***BMC Bioinformatics, 9:***319.

Statnikov, A., Aliferis, C., Tsamardinos, I., Hardin, D., Levy, S., 2005. A comprehensive evaluation of multi-category classification methods for microarray gene expression cancer diagnosis. ***Bioinformatics, 21:***631-643.

Subashini, T., Ramalingam, V., Palanivel, S., 2009. Breast mass classification based on cytological patterns using RBFNN and SVM. ***Expert systems with applications, 36:***5284-5290.

Tang, Y., Zhang, Y., Huang, Z., Hu, X., 2005. Granular SVM-RFE Gene Selection Algorithm for Reliable Prostate Cancer Classification on Microarray Expression Data. ***Bioinformatics and Bioengneering, 2005. BIBE 2005. Fifth IEEE Symposium,*** 290-293.

Wang, H., Huang, G., 2010. Application of support vector machine in cancer diagnosis. ***Medical oncology, 28:***613-618.

Wang, L., Chu, F., Xie, W., 2007. Accurate Cancer Classification Using Expressions of Very Few Genes. ***IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 4:***40-53.

Wang, Y., Makedon, F., 2004. Application of Relief-F Feature Filtering Algorithm to Selecting Informative Genes for Cancer Classification using Microarray Data. ***Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceeding 2004 IEEE,*** 497-498.

Wang, Y., Tetko, I., Hall, M., Frank, E., Facius, A., Mayer, K., Mewes, H., 2005. Gene selection from microarray data for cancer classification – a machine learning approach. ***Computational biology and chemistry, 29:***37-46.

What Is Cancer?, 2015, February 9. ***https://www.cancer.gov/about-cancer/understanding/what-is-cancer***. Cancer. Date accessed: 25.11.2019

What is Machine Learning? A definition., 2017, March 7. ***https://expertsystem.com/machine-learning-definition/***. Expert System. Date accessed: 25.11.2019

Yu, J., Chen, Y., Zheng, S., 2004. An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics. ***World Journal of Gastroenterology: WJG, 10:***3127.

Yu, S., Yu, J., Ge, W., Hu, H., Yuan, Y., Zheng, S., 2011. SPARCL1, Shp2, MSH2, E-cadherin, p53, ADCY-2 and MAPK are prognosis-related in colorectal cancer. ***World Journal of Gastroenterology, 17:***2028-2036.

Zhang, X., Lu, X., Shi, Q., Xu, X., Hon-chiu, E., Harris, L., . . . Wong, W., 2006. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. ***BMC Bioinformatics, 7:***197.

Zhou, X., Tuck, D., 2007. MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. ***Bioinformatics, 5:***427-443.

Zhu, J., Hastie, T., 2004. Classification of gene microarrays by penalized logistic regression. ***Biostatistics, 5:***427-443.

Zisman, A., Nickolov, A., Brand, R., Gorchow, A., Roy, H., 2006. Associations between the age at diagnosis and location of colorectal cancer and the use of alcohol and tobacco: implications for screening. ***Arch Intern Med., 166:***629-634.

APPENDICES

APPENDIX 1. GENİŞLETİLMİŞ TÜRKÇE ÖZETİ

ÖZET

MATEEN, Bawar
Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı
Tez Danışmanı : Doç. Dr. Rıdvan SARAÇOĞLU
Ocak 2020, 69 sayfa

Kolorektal kanser, kolon veya rektumda başlayan yaygın bir kanser hastalığı türüdür. Kadınlar arasında en sık görülen ikinci kanser türü, erkekler arasında ise üçüncü kanser türüdür. Kolorektal kanser her yıl dünya çapında binlerce insanın ölümüne neden olmaktadır. Kolorektal kanser lokalize, tedavi edilebilir olduğundan ve erken evrelerinde tedavi için daha az maliyete ihtiyaç duyduğundan, erken teşhis ve tahmini, kolorektal kanser tedavisinin başarı şansını arttırır. Bu çalışmada, destek vektör makineleri algoritması kullanılarak kolorektal kanser tahmin ve tespit edilmeye çalışılmıştır. Katılımcıların yaşam tarzına dayanan bir veri kümesinin kullanılmasıyla, bu tür veri kümeleriyle kolorektal kanseri tahmin edilmiştir. Bu veri seti her katılımcı hakkında 22 bilgi içermektedir. Sonuçlar kanserin, destek vektör makineleri algoritması kullanılarak bu tür bir veri kümesi ile yüksek bir doğrulukla öngörülebilir olduğunu göstermektedir. Eğer daha fazla kolorektal kanser hastasının bilgisi toplanırsa ve veriler bölgesel olarak toplanıp yine bölgesel modellemeler yapılırsa daha iyi performansla çalışıp daha doğru sonuçlar elde edilebilecektir.

Anahtar Kelimeler: Çekirdek Fonksiyonları, Destek Vektör Makineleri, Kolerktal Kanser, Yapay Zeka.

1. GİRİŞ

Günümüzde yeni teknolojiler hayatımızın önemli bir parçası olmuştur. Teknolojiyi kullanmanın etkinliği ve zahmetsiz oluşu nedeniyle insanlar teknolojiyi daha önce ulaşmadıkları çeşitli yeni alanlarda kullanmaya çalışmaktadır.

Öte yandan, sağlık insanların hakkında çok kafa yordukları olgulardan biridir. İnsanlar, hayatlarına zarar veren ve yaşam kalitelerini düşüren sebepleri araştırır ve sağlıkları için tehlikeli olan hastalıkları bilmek, tedavi etmek, teşhis etmek ve öngörmek için zaman, para ve çaba harcarlar.

Teknoloji hayatımızdaki yerini aldığından beri, tıbbi amaçlar için sıklıkla kullanılmaktadır ve teknolojinin önemli ölçüde yardımcı olduğu tıbbi alanlardan biri, hastalıkların tahminidir. Araştırmacılar, Yapay Zeka algoritmalar ile hastalıkları tahmin etmek için çok sayıda araştırma yapmışlardır. Çok tehlikeli ve genellikle de ölüme neden olan bu hastalıklardan en önemlisi de kanserdir.

Bu çalışma, sağlık ve teknoloji arasındaki kombinasyonun yeni bir parçası olabilecektir. Çünkü etkili bir algoritma olan Destek Vektör Makinesi (SVM) kolorektal kanseri tahmin etmek için kullanılmıştır. Önceki çalışmalarda sağlıklı veya kolorektal kanser hastası olan çok sayıda katılımcının bilgisine dayanan bir veri seti kullanılarak yapılmıştır ancak genellikle kanserli dokuların tanınmasına dayalıdır. Bu çalışmada ise, çeşitli veri setlerinde etkili bir algoritma olduğu kanıtlanan bir algoritma olan SVM algoritması, yaşam tarzı tabanlı veri setini kullanarak kolorektal kanserini öngörmek için kullanılmıştır ve bu açıda ilk çalışmalardan biridir.

1.1 Projenin amacı

Bu projenin temel amacı kolorektal kanser hastalığının SVM algoritması ile öngörüsüdür. Ayrıca, çalışmada, veri kümesi için ve çekirdek fonksiyonu parametrelerinin en iyi değerleri, sınıflandırma ve tahmindeki doğruluğu en iyi seviyeye çıkaracak şekilde belirlenmeye çalışılmıştır.

1.2 Projenin önemi

Kolorektal kanser tüm dünyadaki en yaygın kanser türlerinden biridir, kadınlarda ikinci, erkeklerde üçüncü kanser tipler sıradadır ve dünya genelinde 1.2 milyon yeni teşhisi mevcut olup, sadece 2008'de ise 608.700 kişinin ölümüne sebep olmuştur.

Kolorektal kanserin erken tespiti ve öngörülmesi, tedavinin başarı şansını arttırdığı için çok önemlidir. Genel olarak kanser ve özellikle de kolorektal kanser lokalizedir, tedavi edilebilir ve erken aşamalarında tedaviler için daha az maliyet gerektirir. Bu nedenle, kolorektal kanser gibi tehlikeli bir hastalığın öngörülmesi için bir sistem tasarımı çok önemli bir çalışmadır. Buna ek olarak, çalışma katılımcıların yaşam tarzı ve kolorektal kanserin risk faktörleri ile ilgili bir veri seti kullandığı için önemlidir.

2. KAYNAK BİLDİRİŞLERİ

Kanser yaygın bir hastalıktır, bu nedenle insanlar tanı ve tedavisi için mümkün olan her şeyi yapmaktadırlar. SVM sınıflandırma algoritması, etkinliği nedeniyle kanser hastalığının öngörülmesi ve sınıflandırılmasında yaygın olarak kullanılmıştır. Bu konuda yazılan makalelerin çoğu DNA Mikro-dizi veri seti kullanmıştır, ancak az bir kısmı kolorektal kanserle, çoğu ise diğer kanser türleriyle ilgilidir.

3. MATERYAL VE METOD

3.1. Kanser

Yaşamlarımız boyunca vücudumuzun hücreleri sağlıklı bir şekilde bölünür ve yeni hücreler diğer hücrelerin yerine geçer, bu işlem kontrollü bir şekilde gerçekleşir, ancak bu hücreler kontrolsüz bir şekilde bölünüp çoğaldıklarında, bu kanser olarak bilinen hastalığa neden olur.

Kanser tümörleri malignndir, bu vücudun diğer bölgelerini istila edebileceği anlamına gelir. Kanserin başka vücut parçalarına geçebileceği diğer bir yol, kanser hücreleri ayrılıp ilk kanser tümöründen uzak organlara giderken kan veya lenf sistemidir.

3.2. Kolorektal Kanser

Eğer kanser kolonda başlarsa, buna kolon kanseri denir ve rektumda başlarsa rektal kanser olarak adlandırılır. Bu iki kanser türü kolorektal kanser olarak bilinir.

Kolorektal kanser, tüm dünyada en yaygın kanser türlerinden biridir, kadınlarda ikinci en yaygın kanser türüdür ve erkeklerde üçüncü türdür. Kolorektal kanserlerin çoğu, rektumun veya kolonun polip olarak adlandırılan iç astarında bir büyüme olarak başlar. Zamanla, kanser polipleri birkaç katman içeren kolon veya rektumun iç duvarında büyüyebilir. Kolon veya rektum katmanlarında olan kanser hücreleri kan veya lenf damarlarına aktarılabilir, böylece vücutta her yere ulaşabilir.

3.3. Yapay Zeka (AI)

“Yapay zeka”, 1956'da yani günümüzden 63 yıl önce ortaya atılan bir terimdir. Yapay zeka için çok sayıda tanım vardır, bunlardan biri “Dijital bir bilgisayar veya bilgisayar kontrollü bir robotun, genellikle akıllı varlıklarla ilgili görevleri yerine getirme kabiliyeti” dir. Bir diğer tanım “Yapay zeka insan zekası süreçlerinin makinelerin, özellikle bilgisayar sistemlerindeki simülasyonudur, ve bu süreçler ise öğrenme, muhakeme ve kendi kendini düzeltme” dir.

3.4. Makine Öğrenmesi

Yapay zekanın başlıca bir türü Makine öğrenmesidir (ML), ve önceden bir programlamaya gerek kalmadan sonuçların tahmininde kullanılan ve uygulamaları daha doğru hale getirmeye çalışan bir algoritmalar bütünüdür. Makine öğrenmenin temel fikri, girdi verilerini alabilecek algoritmalar oluşturmak ve çıktıları tahmin etmek için istatistiksel analizleri kullanmaktır.

3.5. Sınıflandırma Algoritmaları

Sınıflandırma algoritmaları, verileri sınıflara kategorize eden yöntemlerdir, bu algoritmalar önceden karşılaşmadığı bir bilginin sınıfını tahmin etmek için bir kısım veriler ile (eğitim verisi) çalışmak zorundadır.

Destek Vektör Makineleri, Karar Ağacı ve Yapay Sinir Ağları, sınıflandırma algoritmalarının bazı örnekleridir.

3.6. Veri

Bu projede kullanılan veri seti ABD'de ABD Sağlık Bakanlığı ve İnsan Hizmetleri Dairesi / Ulusal Sağlık Enstitüleri / Ulusal Kanser Enstitüsü/ Prostat, Akciğer, Kolorektal ve Yumurtalık (PLCO) Kanser Tarama Denemesi'nde toplanan verilere dayanmaktadır.

1993 ve 2001 yılları arasında yaklaşık 155.000 katılımcıdan veriler toplanmış ve 2009 yılında, hangi hastaların öldüğünü ve ölüm nedenini, ölüm zamanını ve diğer bilgileri tespit için kanser hastalarıyla ilgili veriler toplanmıştır. Ayrıca bu veri setinde, hasta vakalarının ortalama takip süresi 12.4 yıldır.

Veri setinde yer alan veri tabloları çok karmaşık ve büyüktür yani çok sayıda sütun içermektedir. Bu büyük sayıdaki değişkenler sınıflandırıcılar için bir problem oluşturur ve sınıflandırma işlemi için önemli olmayan birçok sütunu içeriyordur. Bu veri azaltılmış ve sadece bir uzman doktordan ve özel kanser sitelerinden toplanan bilgilere bağlı olarak kolorektal kanserin risk faktörleriyle ilgili olan sütunlar geriye kalmıştır. Katılımcı verileri şu risk faktörleriyle ilgilidir: yaş, kişisel polip öyküsü, kişisel enflamatuar barsak hastalığı öyküsü (IBD), aile kanseri öyküsü, diyet, fiziksel aktivite ve obezite, sigara ve ağır alkol kullanımı.

4. BULGULAR

Bu projede kullanılan programlama dili Python'dur, Python dünyanın en yaygın kullanılan programlama dillerinden biridir, önemli özellikleri ise kolaylık, iyi performans, kullanışlı kütüphaneler, geniş uygulama yelpazesi ve derin öğrenmede geniş kullanımıdır. Bu nedenle, popüler şirketler; YouTube, Instagram, Pinterest, Survey Monkey, Quora, Mozilla ve Spotfy gibi bu dili kullanmaktadırlar.

Bu programın üzerinde çalıştığı süreç, sınıflandırma için veri hazırlama ve sınıflandırmanın uygulanmasından ibarettir.

İlk olarak, program verileri virgülle ayrılmış değerler (.csv) dosyasından okur. Bu dosya 22 sütun ve 60000'den fazla satır içermektedir ve her satır tek bir katılımcı hakkında bilgiler içerir. Yaklaşık 700 kişi kolorektal kanser hastasıdır, yani tüm katılımcıların % 1.1'idir. Bu veri setinin dengesiz olduğunu gösterir. Bu nedenle sonraki adımlar için farklı sayılarda katılımcı seçilerek farklı senaryolar oluşturulmuştur. Örneğin 5000 kayıt seçildiğinde hastaların tüm kayıtlara oranı% 14 olarak gerçekleşir.

Bundan sonra, veriler iki alt gruba ayrılır; eğitim alt kümesi ve test alt kümesi. Amaç, programın test altkümesi veri noktalarının sınıflarını tahmin etmek için eğitim altkümesi kullanılarak eğitilmesidir.

Bu işlemden sonra, SVM sınıflandırma algoritması eğitim alt gruplarına uygulanır ve daha sonra program test alt kümesindeki veri noktalarını tahmin etmeye hazırdır.

Bu çalışmada üç çeşit çekirdek fonksiyonu kullanılmıştır; Doğrusal çekirdek, polinom çekirdek ve Gaussian (RBF) çekirdeği. Her bir çekirdeğin sonucu etkileyebilecek değişkenleri vardır.

Sonuçlar, doğrusal çekirdeğin, bu veri setinde çalışan diğer çekirdeklerden daha iyi olduğunu göstermiştir. Bir çekirdeğin her sorun için en iyi çekirdek olduğu söylenemez. Aynı zamanda, doğrusal çekirdek matematiksel olarak en basit çekirdektir.

Elde edilen en iyi doğruluk, eğitim alt kümesinde 97.92% oranında doğru olup test alt kümesi veri noktalarının tahmin edilmesinin doğruluğu %98.88'dir.

Genel olarak, sonuçlar başarılıdır ve bu ise iki ana noktadan kaynaklanmaktadır; SVM algoritmasının sınıflandırmadaki etkinliği ve daha etkili olan hasta sayısının azlığıdır. Genellikle algoritma birinci sınıfta sağlıklı olan verilerde hata yapmamakta, ancak yanlış tahminler çoğunlukla kolorektal kanseri hastalar içeren ikinci sınıftan gelmektedir.

5. SONUÇLAR

Bu çalışmada, her yıl dünyada binlerce insanın ölümüne neden olan tehlikeli bir hastalık olan kolorektal kanserin öngörülmesinde ve erken teşhisinde bir adım atma hedeflenmiştir.

Bu hastalık tedavi edilebilir, erken evrelerinde teşhisi durumunda daha az tehlikeli, tedavisi daha kolay, maliyeti daha düşüktür ve bu çalışmanın önemi ise, dünyanın her yerinde her yıl sayısız hayat kurtarma çabasının önemi ile ilgilidir.

Daha önce de belirtildiği gibi, bu çalışma erken teşhis alanında iyi bir adımdır, ancak bu durum nihai hedefe ulaşmak için daha fazla çabaya ve araştırmaya ihtiyaç duyulmadığı anlamına gelmez. Hastalık tahminlerini daha kolay ve daha etkili hale getiren daha iyi ve kritik veri modellere ulaşmak için sağlıkçılar da erken teşhis çalışmaları için verileri toplamada ellerinden geleni yapmalıdır. Programcıların ve veri bilimcilerinin, verilerin daha iyi sınıflandırılması için yeni yöntemler geliştirmek ve sürekli olarak yaşamımıza zarar veren bu tür hastalıkları tahmin etmek için sürekli çalışması gerekir.

Gelecekte, üzerinde durulması gereken önemli bir çalışma bölgesel veri toplamaktır. Zira Amerika Birleşik Devletleri halkına dayanan bir veri seti kolorektal kanseri veya Orta Doğu insandaki diğer herhangi bir hastalığı tahmin etmede faydalı olmayabilir. Bunun yanı sıra yeni bölgesel veri kümelerine dayanarak programlar geliştirilmelidir. Akademisyenler, araştırmacılar ve sağlıkçılar birbirlerine yardım ederlerse yeni nesiller daha sağlıklı olabilir.

CURRICULUM VITAE

Bawar was burn in July 8, 1994, in Duhok city. He spend his first year of age in Kani Masi district, and he spend the rest of his in Duhok. He studied at University of Duhok, Faculty of Science, Computer Science Department.

He started his master degree study in Van Yüzüncü Yıl University in Turkey in 2017.

He is a Kurdish native speaker, and he speaks Arabic and English.

T.C
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
LİSANSÜSTÜ TEZ ORİJİNALLİK RAPORU

Tarih: 17./01./2020

Tez Başlığı / Konusu:
Prediction of Colorectal Cancer Using Support Vector Machines Algorithm

Yukarıda başlığı/konusu belirlenen tez çalışmamın Kapak sayfası, Giriş, Ana bölümler ve Sonuç bölümlerinden oluşan toplam 39 sayfalık kısmına ilişkin, 16./01./2020 tarihinde şahsım/tez danışmanım tarafından Turnitin intihal tespit programından aşağıda belirtilen filtreleme uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 0 (Sıfır) dır.

Uygulanan filtreler aşağıda verilmiştir:
- Kabul ve onay sayfası hariç,
- Teşekkür hariç,
- İçindekiler hariç,
- Simge ve kısaltmalar hariç,
- Gereç ve yöntemler hariç,
- Kaynakça hariç,
- Alıntılar hariç,
- Tezden çıkan yayınlar hariç,
- 7 kelimeden daha az örtüşme içeren metin kısımları hariç (Limit inatch size to 7 words)

Van Yüzüncü Yıl Üniversitesi Lisansüstü Tez Orijinallik Raporu Alınması ve Kullanılmasına İlişkin Yönergeyi inceledim ve bu yönergede belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini bilgilerinize arz ederim.

17.1.2020
Tarih ve İmza

Adı Soyadı: Bawar Safwat Hussein MATEEN
Öğrenci No: 179100010028
Anabilim Dalı: Elektrik Elektronik Müh. ABD
Programı: —

Statüsü: Y. Lisans ☑ Doktora ☐

DANIŞMAN ONAYI
UYGUNDUR
Doç. Dr. Ridvan Saraçoğlu
(Unvan, Ad Soyad, İmza)

ENSTİTÜ ONAYI
UYGUNDUR
(Unvan, Ad Soyad, İmza)