CANCER INCIDENCE RATE PREDICTION USING MACHINE LEARNING ALGORITHMS

A THESIS SUBMITTED TO THE GRADUATE SCHOOL OF APPLIED SCIENCES OF NEAR EAST UNIVERSITY

By
KÜBRA TUNCAL

In Partial Fulfilment of the Requirements for The Degree of Master of Science in Information Systems Engineering

NICOSIA, 2019

Kübra Tuncal: CANCER INCIDENCE RATE PREDICTION USING MACHINE LEARNING ALGORITHMS

Approval of Director of Graduate School of Applied Sciences

Prof.Dr. Nadire ÇAVUŞ

We certify this thesis is satisfactory for the award of the degree of Masters of Science in Information Systems Engineering

Examining Committee in Charge:

Assoc.Prof.Dr. Kamil Dimililer	Department of Automotive Engineering, NEU
Assoc.Prof. Dr. Yöney Kırsal Ever	Department of Software Engineering, NEU
Assist.Prof. Dr. Boran Şekeroğlu	Supervisor, Department of Information Systems Engineering, NEU

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as require by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name:

Signature:

Date:

To my family...

ACKNOWLEDGMENTS

My deepest gratitude is to my advisor, supervisor and Chairman Assist. Prof. Dr. Boran Şekeroğlu, for his encouregement, guidance, patience and support with his knowledge. His guidance helped me during the preparation of this thesis and this would not be possible without him.

Then, I would like to thank my mother, brother and sister for their support and ideas. Without them, everything would have been difficult for me.

Finally, I would like to express my lovely thoughts to Çağrı Özkan for his priceless patience.

ABSTRACT

Everyday, the frequency of incidence and mortality of cancer disease is rising. It is the most fatal disease in the world with several types and there is a few reliable data about incidence and mortality rates of cancer and its types. Thus, the prediction of the rates is challenging task for human beings. For this reason, several machine learning algorithms have been implemented to provide effective and rapid prediction of uncertain raw data with minimized error. In this thesis, Support Vector Regression, Backpropagation Neural Network, Radial Basis Function Neural Network, Decision Tree and Long-Short Term Memory Network is used to perform lung cancer incidence prediction for European continent those records have been started from 1993. All cancer types, Lung cancer, Prostate Cancer, Breast Cancer and Colorectum Cancer is considered in these predictions. Results show that the prediction of incidence rates is possible with high scores with all algorithms however, Support Vector Regression performed superior results than other considered algorithms.

Keywords: Machine learning models; cancer predictions; european cancer rates; mortality rates.

ÖZET

Kanser hastalığının görülme ve ölüm oranı hergün artmaktadır. Dünyadaki en ölümcül hastalık olan kanserin bir çok çeşidi vardır ve bu çeşitleriyle görülme ve ölüm oranlarını içeren çok az sayıda veri mevcuttur. Bu da, bu oranlarının tahminini insanlar tarafından yapılmasını oldukça zorlaştırmaktadır. Bu nedenle, bir çok makine öğrenme algoritması bu az ve ham veriler üzerinde etkili ve hızlı tahmin yürütme için uygulanmıştır. Bu tezde, Destek Vektör Tahmini, Radyal Basis Fonksiyon YSA, Geriyayılmalı YSA, Karar Verme Ağaçları ve Uzun-Kısa Dönem Hafıza Ağı uygulanarak Avrupa kıtasındaki 1993 yılından itibaren kanser görülme oranlarının tahmini yapılmıştır. Tüm kanser çeşitlerinin toplamı, Akciğer kanseri, Göğüs kanseri, Postat Kanseri ve Kolorektum kanseri bu tahminlerde dikkate alınmıştır. Sonuçlar göstermiştir ki, kanser görülme oranlarının yüksek bir başarı ile tahmini tüm algoritmalarca mümkündür fakat, Destek Vektör Tahmini en iyi sonuçları üretmiştir.

Anahtar Kelimeler: Makine öğrenme modelleri; kanser tahminleri; avrupa kanser oranları; ölüm oranları.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BPNN Backpropagation Neural Network

LSTM Long-Short Term Memory Neural Network

RBFNN Radial-Basis Function Neural Network

DT Decision Tree

SVR Support Vector Regression

EV Explained Variance

MSE Mean Squared Error

SVR Support Vector Regression

DNA Deoxyribonucleic Acid

WHO World Health Organization

CHAPTER 1

INTRODUCTION

1.1 Introduction

One of the most common health problems worldwide is cancer (Kachroo, Melek, and Kurian (2013)). Therefore, early diagnosis and early treatment is important in cancer. Although early diagnosis plays an important role in cancer, it is sometimes not possible to prevent rapidly spreading cancers and result in death (Bosetti, Malvezzi, Rosso, and et al. (2012)). In all types of cancer occurs in the cells that are the cornerstone of the body. In order to better understand the cancer, how cancer occurs is cancer; it is called the bad products that occur when cells are irregularly divided on the tissue or organ. It does not cause any problems in our body since it is conscious that healthy cells in our body multiply by multiplying and consciously multiply and how much they should die. However, there is no unconsciousness and proliferation in cancer cells. As a result of this unrestricted division and reproduction, they produce their masses as a size or a tumor.

Tumors are classified as benign and malignant. Although benign tumors are not cancer, they are usually taken and also known as tumors with non-recurrent structures. Malignant tumors are known as cancer. Cancer diseases may change and become deadly compared to tumors.

According to the 2018 data determined by the World Health Organization (Organization (2018)); the total population was determined as 7,632,819,272 of the the world and 18,078,957 were recorded as new cases. The mortality rates cover 9.555.027 of the total population.In addition, the number of cases in the last 5 years is 43.841.302.

According to the data of the World Health Organization in 2018, the number of new cases of men and women of all ages in the world is as follows: Lung, Breast, Colorectum, Prostate, Stomatch and Other Cancers explained this way. Again according to the latest data from the world health organization new cases 2018 [ref]; Lung 2.093.876 (11.6%), Breast 2.088.849 (11.6%), Colorectum 1.849.518 (10.2%), Prostate 1.276.106 (7.1%), Stomach 1.033.701 (5.7%) and Other Cancers 9.736.907 (53.9%) this information is included. Total new cases

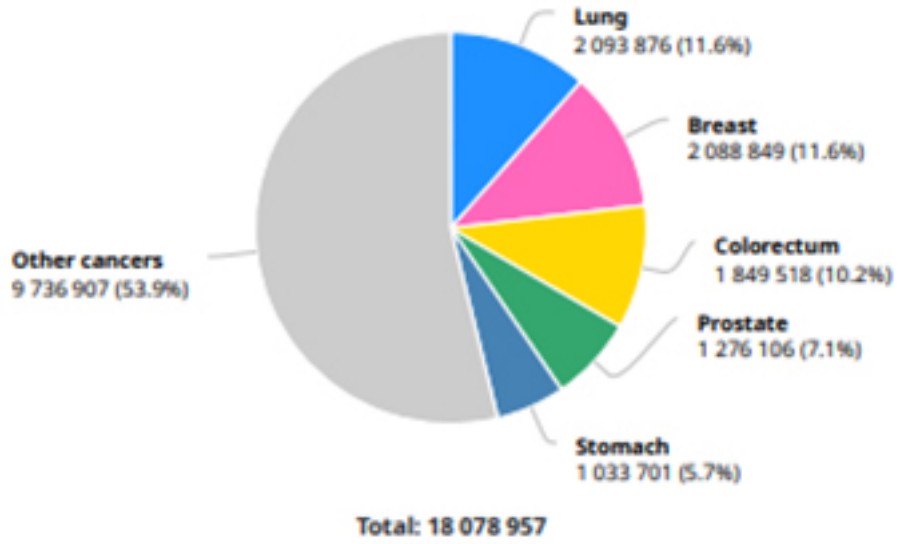


Figure 1.1: Number of new cases in 2018, both sexes, all ages

are indicated as 18.078.957. Both sexes have reached this information.The information in Figure 1.1 below is shown.

In more detail, if we examine the sexes separately for both males and females, new cases of cancer varieties for men of all age groups; Lung, Prostate, Colorectum, Stomach, Liver and Other cancers.Again, the statistical rates of male cancer varieties; Lung 1.368.524 (14.5In more detail, if we examine the sexes separately for both males and females, new cases of cancer varieties for men of all age groups; Lung, Prostate, Colorectum, Stomach, Liver and Other cancers.Again, the statistical rates of male cancer varieties; Lung 1.368.524 (14.5%), Prostate 1.276.106 (13.5%), Colorectum 1.026.215 (10.9%), Stomach 683.754 (7.2%), Liver 596.574 (6.3%) and Other cancers 4.505.245 (47.6%) as indicated.The types of cancer that belong to every age group for women are as follows; Breast, Colorectum, Lung, Cervix uteri, Thyroid and Other cancers.Again, the statistical rates of cancer varieties for women; Breast 2.088.849 (24.2%), Colorectum 823.303 (9.5%), Lung 725.352 (8.4%), Cervix uteri 569.847 (6.6%), Thyroid 436.344 (5.1%), Other cancers 3.978.844 (46.1%) and Total in 8.622 .539.All information in Figure 1.2 below is shown.

According to the information given by the World Health Organization in 2018, the number of people who died of cancer in the world is stated as 9.555.027, as mentioned above.The

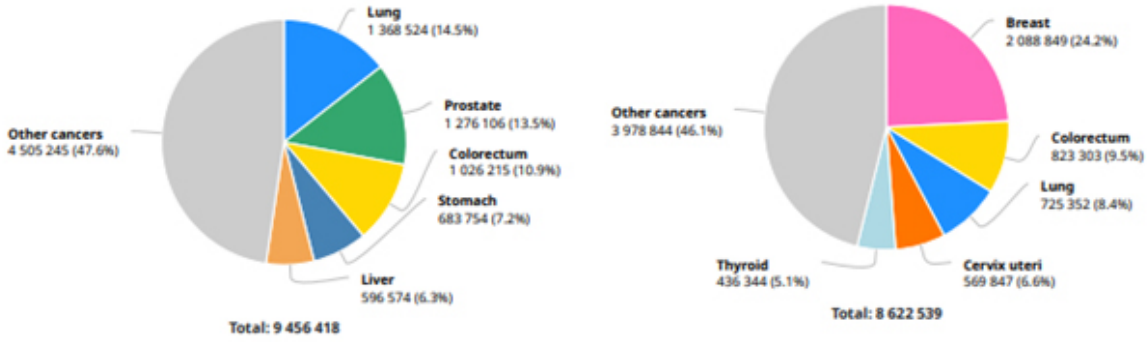


Figure 1.2: Number of new cases in 2018, males and females, all ages

types of cancers of all age groups, including two genders, in which the highest number of deaths occur worldwide; Lung, Colorectum, Stomach, Liver, Breast, Oesophagus, Pancreas, Prostate and Other cancers.The rates of these cancers are; Lung 1.761.007 (18.4%), Colorectum 880.792 (9.2%), Stomach 782.685 (8.2%), Liver 781.631 (8.2%), Breast 626.679 (6.6%), Oesophagus 508.585 (5.3%), Pancreas 432.242 (4.5%) , Prostate 358.989 (3.8%) and Other cancers as 3.422.417 (35.8%).

The highest number of New Cases cancer types worldwide, including two genders, belongs to all age groups; Lung, Breast, Colorectum, Prostate, Stomach, Liver, Oesophagus, Cervix Uteri and Other Cancers.The rates of cancer varieties; Lung 2.093.876 (11.6%), Breast 2.088.849 (11.6%), Colorectum 1.849.518 (10.2%), Prostate 1.276.106 (7.1%), Stomach 1.033.701 (5.7%), Liver 841.080 (4.7%) , Oesophagus 572.034 (3.2%), Cervix Uteri 569.847 (3.2%), Other Cancers 7.753.946 (42.9%) and Total 18.078.957. The information in Figure 1.3 below is shown.

Machine Learning started to be used efficiently in health sciences and especially in forecasting cancer research (Senturk and Senturk (2016)). Senturk (Senturk and Senturk (2016)) conducted a study on the database obtained from the UCI Machine Learning Repository using the Backpropagation Neural Network (BPNN) to achieve a 77% success in breast cancer classification.These investigations address faster data analysis and efficient estimation of large data; It also aims to obtain the best results by using machine learning techniques by excluding the disadvantages of human factors.

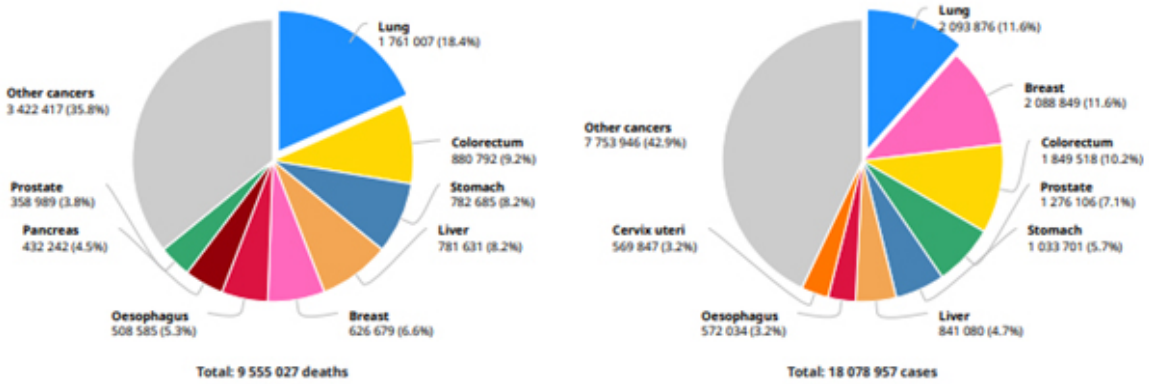


Figure 1.3: Number of deaths and new cases in 2018, both sexes, all ages

Kourou et al. Kourou, Exarchos, and Exarchos (2014) investigated several models in order to determine the efficiency of machine learning techniques in cancer prognosis and prediction. They concluded that the researches are focused on supervised models for the development of predictive algorithms.

Mohammadzadeh et al. (Mohammadzadeh, Noorkojuri, Pourhoseingholi, and et al. (2014)) used decision trees to predict the mortality rate of gastric cancer patients. They used the data of 216 patients and 74% of accuracy was achieved.

O'Lorcain et al. (O'Lorcain, Deady, and Comber (2006)) implemented Log and log-linear Poisson regression model for colorectal cancer prediction. They fit the model using the data of World Health Organization from 1950 to 2002 to predict Ireland mortality rates.

Malvezzi et al. (Malvezzi, P.Bertuccio, Levi, and et al. (2014)) used linear regression to predict cancer mortality rates of European Union and 6 other European countries.

Ribes et al. (Ribes, Esteban, Cléries, and et al. (2013)) used Bayesian models to predict both incidence and mortality rates of Catalonia up to 2020. They obtained the data from cancer registries in Spain and Catalonia.

Alhaj and Maghari (Alhaj and Maghari (2017)) considered Random Forest and Rule Induction Algorithms to predict the cancer survivability in Gaza strip. They concluded that Random Forest achieved more accurate result than rule induction algorithm by 74.6%.

Recently, Jung et al. (Jung, Won, Kong, and Lee (2017)) implement Jointpoint regression model to predict cancer incidence and mortality in Korea for 2019. They used Korea National Cancer Incidence Database in their research.

Malvezzi et al. (Malvezzi, Bosetti, Rosso, and et al. (2013)) made a comprehensive research about prediction studies and Jointpoint regression was implemented in order to predict lung cancer rates in Europe.

1.2 The Aim of the Thesis

The aim of this thesis is to implement several machine learning algorithms in order to predict cancer incidence rates of European countries with latest dataset and to analyse the prediction efficiency of obtained results for considered algorithms.

1.3 Thesis Overview

Main parts of the thesis are as shown below:

- Chapter 1 presents the introduction to the thesis and gives information about the Cancer disease and the machine learning applications in cancer researches.
- Chapter 2 explains detailed information about Cancer disease and the literature review related to this field.
- Chapter 3 gives information about considered machine learning algorithms.
- Chapter 4 introductes experimental design and data preparation phase.
- Chapter 5 presents obtained results and discussions.
- Chapter 6 concludes the work done in this thesis and suggests future works and improvements.

CHAPTER 2

CANCER DISEASE AND LITERATURE REVIEW

2.1 Cancer Cell

Cancer; Deoxyribonucleic Acid (DNA) damage in cells in our body is formed by collecting and at the same time begins to increase irregularly.The disease caused by the formation of DNA damage in these cells is called cancer.The fact that these events occur in the cell, which is the building block of our body, means that it goes out of its normal functioning in our body.The fact that the cell in our body goes outside the normal means that it increases irregularly, which causes the tumor.There are differences in appearance between a normal cell and a cancerous cell.The figure is shown in Figure 2.1 below.

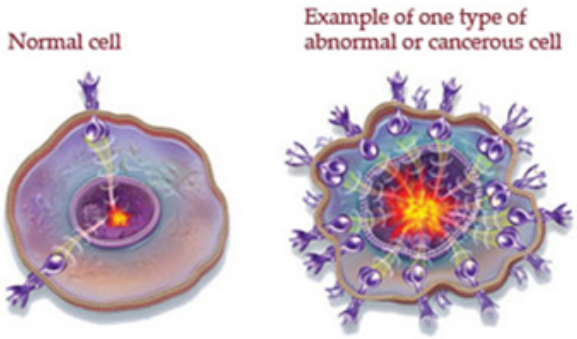


Figure 2.1: Normal cell versus cancer cell

This is the normal functioning of our body out of the way we need to elaborate a little more; It has approximately 100 Trillion cells of 200 various types with specific functions specialized in an adult human body (*Bilim ve Teknik* (2002)).Cells in our bodies form tissues, organs, organ systems and organisms. Since all cancers begin in the cell, we need to know exactly what is in the cell. There are nuclei, chromosomes and DNA in the cell.All the vital activities of the cell inside the cell nucleus are checked.There are also chromosomes in the cell nucleus.Chromosomes are in the human body as 23 pairs (46 in total).One of these chromosomes is used for sex determination.The remaining 22 pairs of chromosomes were composed of DNAs. Figure 2.2 shows the structure of the cell.

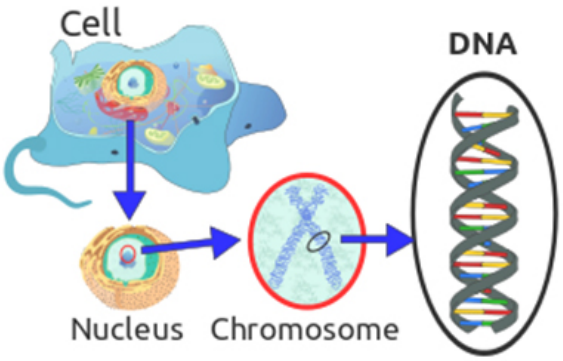


Figure 2.2: Cell structure

DNA is the molecule in which the vital activities in the cell are managed.The parts of the DNA are called Gene.DNA is a structure consisting of genes.According to the human genome project, the total number of genes in the range of 29.000-36.000 is found in the human body.In addition, it was found that an average of 3,000 nucleotides in a gene were obtained.The number of genes found in the chromosome of an organism is called Genome.The number of human genomes consists of 3,164,700,000 nucleotides. (Human Genome Project reference) DNA is a molecule structure that is like a long, staircase shape and forms a double helix.Each strand of the helix is called a nucleotide. There are four types of nucleotides in DNA.Each DNA nucleotide has one of four nitrogen bases (A = Adenine, G = Guanine, S = Cytosine, and T = Timine).These four bases together with various combinations form the genetic code. (Klug and Cummings, 2011). The DNA Structure in Figure 2.3 is shown below.

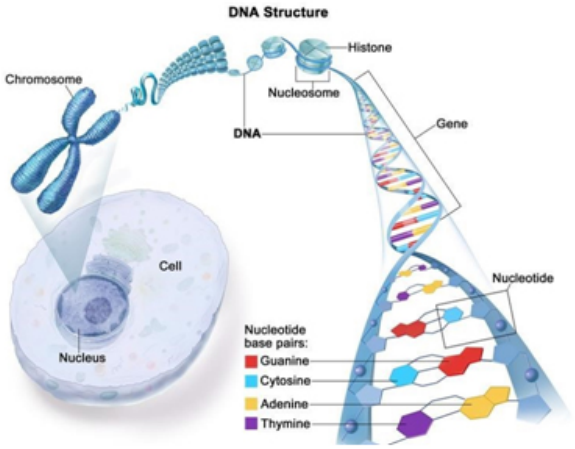


Figure 2.3: DNA structure

2.2 Statistical Data

According to the information provided by the World Health Organization in 2018, there are 35 cancer types in total.These are respectively; Lung, Breast, Prostate, Colon, Stomach, Liver, Rectum, Oesophagus, Cervixuteri, Thyroid, Bladder, Non-Hodgkinlymphoma, Pancreas, Leukaemia, Kidney, Corpusuteri, Lip, oral cavity, Brain, nervoussystem, Ovary, Melanoma of skin, Gallbladder , Larynx, Multiplemyeloma, Nasopharynx, Oropharynx, Hypopharynx, Hodgkinlymphoma, Testis, Salivaryglands, Anus, Vulva, Kaposisarcoma, Penis, Mesothelioma and Vagina. These types of cancer; Newcases, deaths and 5 prevalence rates are statistically different. Figure 2.4 below shows the details in detail.

	New cases				Deaths				5-year prevalence (all ages)	
Cancer	Number	Rank	(%)	Cum.risk	Number	Rank	(%)	Cum.risk	Number	Prop.
Lung	2 093 876	1	13.01	2.75	1 761 007	1	19.92	2.22	2 129 964	27.91
Breast	2 088 849	2	12.97	5.03	626 679	4	7.09	1.41	6 875 099	181.78
Prostate	1 276 106	3	7.93	3.73	358 989	8	4.06	0.60	3 724 658	96.73
Colon	1 096 601	4	6.81	1.31	551 269	5	6.24	0.54	2 785 583	36.49
Stomach	1 033 701	5	6.42	1.31	782 685	2	8.86	0.95	1 589 752	20.83
Liver	841 080	6	5.22	1.08	781 631	3	8.84	0.98	675 210	8.85
Rectum	704 376	7	4.38	0.91	310 394	10	3.51	0.35	1 876 453	24.58
Oesophagus	572 034	8	3.55	0.78	508 585	6	5.75	0.67	547 104	7.17
Cervix uteri	569 847	9	3.54	1.36	311 365	9	3.52	0.77	1 474 265	38.98
Thyroid	567 233	10	3.52	0.68	41 071	25	0.46	0.05	1 997 846	26.17
Bladder	549 393	11	3.41	0.65	199 922	14	2.26	0.18	1 648 482	21.60
Non-Hodgkin lymphoma	509 590	12	3.17	0.61	248 724	12	2.81	0.27	1 353 273	17.73
Pancreas	458 918	13	2.85	0.55	432 242	7	4.89	0.50	282 574	3.70
Leukaemia	437 033	14	2.71	0.48	309 006	11	3.50	0.33	1 174 433	15.39
Kidney	403 262	15	2.50	0.52	175 098	17	1.98	0.20	1 025 730	13.44
Corpus uteri	382 069	16	2.37	1.01	89 929	21	1.02	0.21	1 283 348	33.93
Lip, oral cavity	354 864	17	2.20	0.46	177 384	16	2.01	0.23	913 514	11.97
Brain, nervous system	296 851	18	1.84	0.36	241 037	13	2.73	0.30	771 110	10.10
Ovary	295 414	19	1.83	0.72	184 799	15	2.09	0.45	762 663	20.17
Melanoma of skin	287 723	20	1.79	0.35	60 712	23	0.69	0.07	965 623	12.65
Gallbladder	219 420	21	1.36	0.25	165 087	18	1.87	0.18	233 820	3.06
Larynx	177 422	22	1.10	0.25	94 771	20	1.07	0.13	488 900	6.41
Multiple myeloma	159 985	23	0.99	0.20	106 105	19	1.20	0.12	376 005	4.93
Nasopharynx	129 079	24	0.80	0.16	72 987	22	0.83	0.10	362 219	4.75
Oropharynx	92 887	25	0.58	0.13	51 005	24	0.58	0.07	280 508	3.68
Hypopharynx	80 608	26	0.50	0.11	34 984	26	0.40	0.05	119 130	1.56
Hodgkin lymphoma	79 990	27	0.50	0.08	26 167	27	0.30	0.03	275 947	3.62
Testis	71 105	28	0.44	0.14	9 507	34	0.11	0.02	284 073	7.38
Salivary glands	52 799	29	0.33	0.06	22 176	29	0.25	0.03	123 460	1.62
Anus	48 541	30	0.30	0.06	19 129	31	0.22	0.02	127 599	1.67
Vulva	44 235	31	0.27	0.09	15 222	32	0.17	0.03	132 269	3.50
Kaposi sarcoma	41 799	32	0.26	0.04	19 902	30	0.23	0.02	88 379	1.16
Penis	34 475	33	0.21	0.09	15 138	33	0.17	0.04	93 850	2.44
Mesothelioma	30 443	34	0.19	0.04	25 576	28	0.29	0.03	31 250	0.41
Vagina	17 600	35	0.11	0.04	8 062	35	0.09	0.02	43 877	1.16
All cancer sites	18 078 957	-	-	20.20	9 555 027	-	-	10.63	43 841 302	574.38

Figure 2.4: Cancer statistics for different types

World statistics are explained in more detail in 6 different continents including Africa, Latin America and the Caribbean, North America, Asia, Europe and Oceania.According to the data of World Health Organization 2018, rates varying according to continents and the incidence of cancer, the number of deaths from cancer according to population rates and men and women of all ages are mentioned separately.

2.2.1 Africa

According to World Health Organization 2018 data, population of Africa continent is 1.287.920.608, number of new cancer cases is 1.055.172, number of deaths is 693.487 and number of prevalent cases (5 years) is specified as 1.930.912. In Africa continent, two cases of cancer and new cases of cancer belonging to each age group respectively; Breast, Cervix Uteri, Prostate, Liver, Colorectum and Other Cancers. Statistical ratios of these cancers are; Breast 168.690 (16%), Cervix Uteri 119.284 (11.3%), Prostate 80.971 (7.7%), Liver 64.779 (6.1%), Colorectum 61.846 (5.9%) and Other Cancers 559.602 (53%). As it is mentioned above, the total number of new cases is 1.055.172. Figure 2.5 shows the statistical data of Africa continent in details.

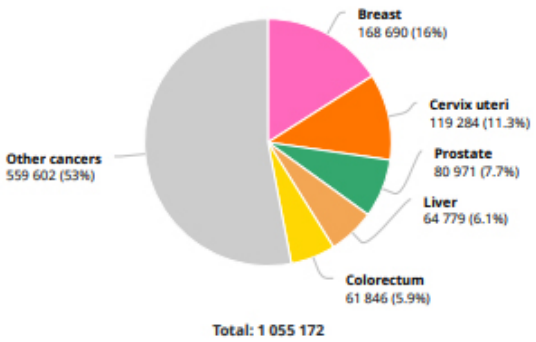


Figure 2.5: Africa, Number of new cases in 2018, both sexes, all ages

2.2.2 Latin America and the Caribbean

In 2018 data, World Health Organization announced the population of Latin America and the Caribbean continent 652.011.967, number of new cases 1.412.732, number of deaths 672.758 and number of prevalent cases (5 year) 3.336.468. New cases of cancer disease of two sexes and all age groups in the Latin America and the Caribbean continent are; Breast, Prostate, Colorectum, Lung, Stomach and Other Cancers.

The rates of cancer types are as Breast 199.734 (14.1%), Prostate 190.385 (13.5%), Colorectum 128.006 (9.1%), Lung 89.772 (6.4%), Stomach 67.058 (4.7%) and Other Cancers 737.777 (52.2%). As it is stated above, new cases in total is 1.412.732. Figure 2.6 shows the statistical data for Latin America and the Caribbean continent.

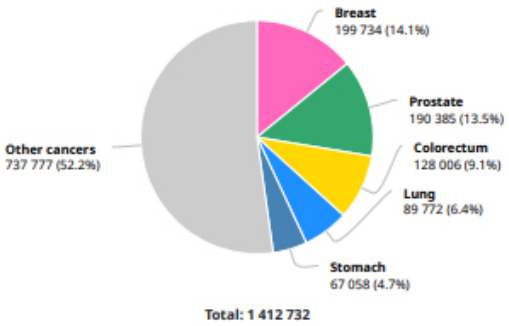


Figure 2.6: Latin America and the Caribbean, Number of new cases in 2018, both sexes, all ages

2.2.3 North America

World Health Organization announnced that the total population in North America is 363.844.506. It is also noticed that the number of new cases is 2.378.785, number of deaths 698.266 and number of prevalent cases (5-year) 8.132.437. North America continent for both sex and all age groups belonging to the new cases cancer types are Breast, Lung, Prostate, Colorectum, Bladder and Other Cancers. The rates of cancer types are given as Breast 262.347 (11%), Lung 252 746 (10.6%), Prostate 234.278 (9.8%), Colorectum 179.771 (7.6%), Bladder 91.689 (3.9%) and Other cancers 1.357.954 (57.1%). Figure 2.7 shows graphical statistics for North America continent.

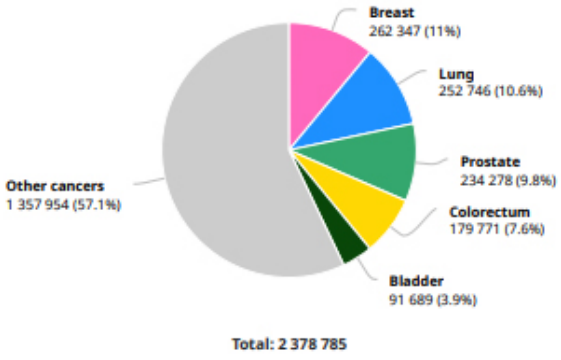


Figure 2.7: North America, Number of new cases in 2018, both sexes, all ages

2.2.4 Oceania

In same data sheet, World Health Organization declared that the total population of Oceania is 41.261.185. Number of new cases, number of deaths and number of prevalent cases are

mentioned as 251.674, 69.974 and 921.628 respectively.

New Cases cancer types belonging to all sexes in Oceania and in all age groups respectively; Breast, Prostate, Colorectum, Melanoma skin, Lung and Other Cancers are given as cancer types. The numerical values of cancer types are indicated as Breast 24.551 (9.8%), Prostate 23.496 (9.3%), Colorectum 22.332 (8.9%), Melanoma of skin 17.246 (6.9%), Lung 16.937 (6.7%), Other Cancers as 147.112 (58.5%) and Total 251.674. Graphical representation of statistical values can be seen in Figure 2.8.

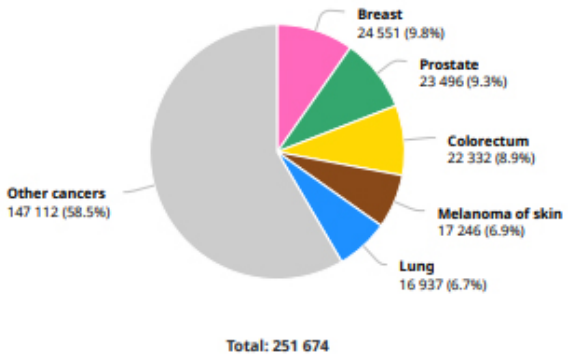


Figure 2.8: Oceania, Number of new cases in 2018, both sexes, all ages

2.2.5 Asia

According to World Health Organization 2018 data in Asia; Total Population 4.543.943.980, Number of New Cases 8.750.932, Number of Deaths 5.477.064 and Number of Prevalent Cases (5-year) 17.387.570.In Asia, New Cases cancer types for both sexes and for all age groups, respectively; Cancer types are given as Lung, Colorectum, Breast, Stomach, Liver and Other Cancers.Statistical data of cancer types; Lung 1.225.029 (14%), Colorectum 957.896 (10.9%), Breast 911.014 (10.4%), Stomach 769.728 (8.8%), Liver 609.596 (7%), Other Cancers 4.277.669 (48.9%) and as mentioned above Total 8.750.932 new cases as provided. The information in Figure 2.9 is shown below

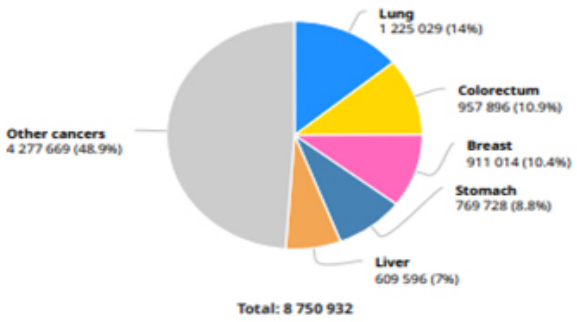


Figure 2.9: Europe, Number of new cases in 2018, both sexes, all ages

2.2.6 Europe

According to World Health Organization, total population of Europe in 2018 is 743.837.100. It is also declared that the number of new cases is 4.229.662, number of deaths is 1.943.478 and number of prevalent cases (5-year) is 12.132.287. New cases cancer types belonging to all sexes in Europe continent and all age groups are Breast, Colorectum, Lung, Prostate, Bladder and Other Cancers respectively. Incidence rates are as Breast 522.513 (12.4%), Colorectum 499.667 (11.8%), Lung 470.039 (11.1%), Prostate 449.761 (10.6%), Bladder 197.105 (4.7%), Other Cancers 2.090.577 (49.4%) and Total and 4.229.662. Figure 2.10 shows the statistical information in graphical representation.

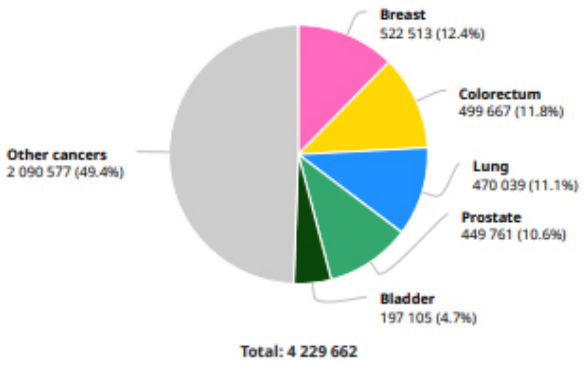


Figure 2.10: North America, Number of new cases in 2018, both sexes, all ages

Cancer types and rates in men and women of cancer in Europe continent is announced as new types of cancer for men of all age groups, Prostate, Lung, Colorectum, Bladder, Kidney and Other Cancers. The numerical values of the types of cancers are given as Prostate 449.761 (20%), Lung 311.843 (13.9%), Colorectum 271.600 (12.1%), Bladder 153.849 (6.8%), Kid-

ney 84.928 (3.8%), Other Cancers 975.537 (43.4%) and Total 2.247.518.

New Cases for women of all age groups in Europe continent are Breast, Colorectum, Lung, Corpus Uteri, Melanoma of skin and Other Cancers. The rates of cancer types are given as Breast 522.513 (26.4%), Colorectum 228.067 (11.5%), Lung 158.196 (8%), Corpus Uteri 121.578 (6.1%), Melanoma of skin 73.041 (3.7%), Other Cancers 878.749 (44.3%) and Total 1.982.144. Figure 2.11 shows all information for all genders in Europe.

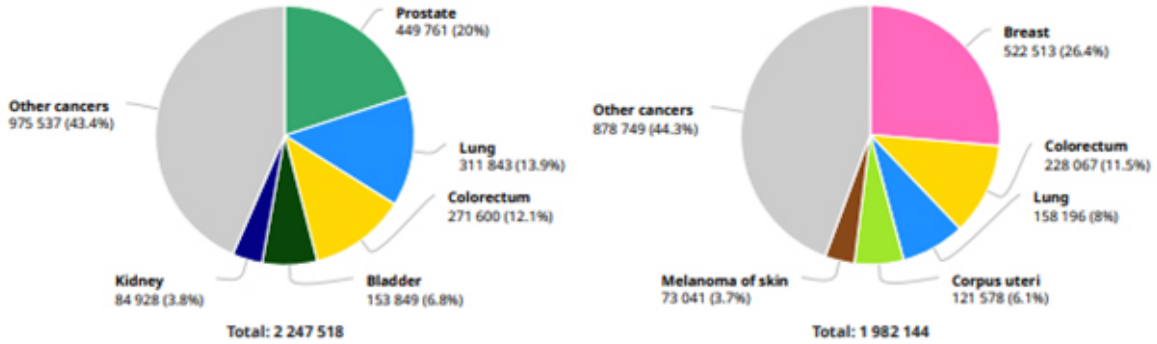


Figure 2.11: Europe, Number of new cases in 2018, both sexes, all ages

2.3 Types of Cancer

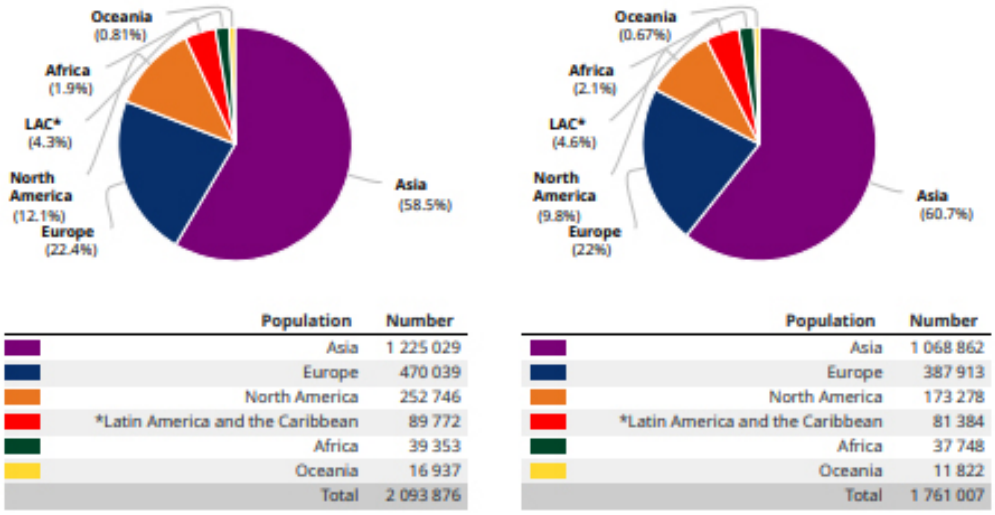
In this section, four cancer types, Lung, Breast, Prostate and Colorectum which are considered to be analysed in this thesis, will be explained. Also statistical data about each type will be presented.

2.3.1 Lung Cancer

Recently, lung cancer is the most common cause of cancer in males and second in females after breast cancer. It is known that 80-90% of lung cancer is caused by smoking. Figure 2.12 shows data for lung cancer incidence and mortality rates by continents and regions, and Figure 2.13 shows these rates by genders.

	Incidence						Mortality					
	Both sexes		Males		Females		Both sexes		Males		Females	
	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)
Eastern Africa	5 891	0.32	3 296	0.40	2 595	0.26	5 733	0.32	3 230	0.40	2 503	0.26
Middle Africa	2 260	0.33	1 285	0.41	975	0.26	2 151	0.32	1 240	0.40	911	0.26
Northern Africa	19 537	1.23	16 008	2.10	3 529	0.40	18 838	1.17	15 655	2.02	3 183	0.36
Southern Africa	8 416	1.89	5 634	3.04	2 782	1.03	7 945	1.80	5 348	2.92	2 597	0.96
Western Africa	3 249	0.19	2 087	0.26	1 162	0.13	3 081	0.19	2 058	0.26	1 023	0.12
Caribbean	11 006	2.29	6 540	2.89	4 466	1.75	9 475	1.90	5 866	2.49	3 609	1.36
Central America	10 262	0.68	5 962	0.86	4 300	0.53	8 987	0.58	5 368	0.74	3 619	0.43
South America	68 504	1.57	39 255	1.99	29 249	1.22	62 922	1.44	36 824	1.86	26 098	1.08
North America	252 746	4.27	133 950	4.76	118 796	3.82	173 278	2.64	91 957	3.02	81 321	2.29
Eastern Asia	950 015	4.10	633 284	5.61	316 731	2.60	815 635	3.36	557 985	4.74	257 650	1.99
South-Eastern Asia	113 182	2.03	78 453	3.07	34 729	1.11	100 731	1.84	70 504	2.82	30 227	0.98
South-Central Asia	111 042	0.77	80 415	1.14	30 627	0.40	103 862	0.72	74 991	1.06	28 871	0.38
Western Asia	50 790	2.77	41 309	4.72	9 481	0.91	48 634	2.69	39 773	4.64	8 861	0.85
Central and Eastern Europe	149 083	3.54	109 928	6.29	39 155	1.52	131 359	3.08	99 266	5.67	32 093	1.20
Western Europe	145 656	4.25	90 239	5.35	55 417	3.22	114 236	3.04	74 003	4.04	40 233	2.11
Southern Europe	100 929	3.55	72 411	5.33	28 518	1.92	86 633	2.80	63 933	4.36	22 700	1.39
Northern Europe	74 371	3.67	39 265	4.03	35 106	3.34	55 685	2.52	30 114	2.88	25 571	2.18
Australia and New Zealand	15 584	3.11	8 345	3.27	7 239	2.96	10 573	2.01	6 026	2.30	4 547	1.73
Melanesia	886	1.57	549	2.16	337	1.05	848	1.51	532	2.09	316	0.99
Polynesia	262	5.01	173	6.84	89	3.24	208	4.03	144	5.57	64	2.54
Micronesia	205	4.49	136	6.22	69	2.78	193	4.18	130	5.98	63	2.40
Low HDI	15 006	0.34	9 024	0.45	5 982	0.25	14 526	0.34	8 854	0.45	5 672	0.25
Medium HDI	218 954	1.10	154 539	1.62	64 415	0.60	201 673	1.01	142 855	1.50	58 818	0.54
High HDI	974 629	3.45	659 344	4.86	315 285	2.11	872 386	3.01	601 774	4.35	270 612	1.74
Very high HDI	884 313	3.73	545 020	4.92	339 293	2.66	671 628	2.60	430 958	3.62	240 670	1.68
World	2 093 876	2.75	1 368 524	3.80	725 352	1.77	1 761 007	2.22	1 184 947	3.19	576 060	1.32

Figure 2.12: Lung Cancer incidence and mortality statistics worlwide and by region



Population	Number
Asia	1 225 029
Europe	470 039
North America	252 746
*Latin America and the Caribbean	89 772
Africa	39 353
Oceania	16 937
Total	2 093 876

Population	Number
Asia	1 068 862
Europe	387 913
North America	173 278
*Latin America and the Caribbean	81 384
Africa	37 748
Oceania	11 822
Total	1 761 007

Figure 2.13: Lung Cancer Incidence and Mortality, both sexes

2.3.2 Breast Cancer

Breast cancer is a structure that usually occurs in milk channels in the breast. It is a type of cancer which is thought to occur due to the secretion of estrogen hormone in the body in the long term. In addition, genetically similar to every type of cancer, the fact that someone had previously had cancer in the family may also be one of the factors that increase this risk factor.

According to the 2018 data of the WHO, breast cancer is prevalence in the world, while it is ranked fifth in the death rate. Figure 2.14 and 2.15 shows incidence and mortality rates by regions and by genders respectively.

	Incidence						Mortality					
	Both sexes		Males		Females		Both sexes		Males		Females	
	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)
Eastern Africa	40 310	3.15	-	-	40 310	3.15	20 165	1.62	-	-	20 165	1.62
Middle Africa	14 486	2.89	-	-	14 486	2.89	7 864	1.64	-	-	7 864	1.64
Northern Africa	53 917	5.06	-	-	53 917	5.06	20 058	1.96	-	-	20 058	1.96
Southern Africa	14 820	4.93	-	-	14 820	4.93	5 002	1.60	-	-	5 002	1.60
Western Africa	45 157	3.92	-	-	45 157	3.92	20 983	1.92	-	-	20 983	1.92
Caribbean	14 097	5.50	-	-	14 097	5.50	5 496	1.95	-	-	5 496	1.95
Central America	35 349	4.17	-	-	35 349	4.17	9 341	1.14	-	-	9 341	1.14
South America	150 288	6.19	-	-	150 288	6.19	37 721	1.45	-	-	37 721	1.45
North America	262 347	9.32	-	-	262 347	9.32	46 963	1.38	-	-	46 963	1.38
Eastern Asia	476 509	4.15	-	-	476 509	4.15	119 678	0.93	-	-	119 678	0.93
South-Eastern Asia	137 514	4.17	-	-	137 514	4.17	50 935	1.61	-	-	50 935	1.61
South-Central Asia	241 077	2.81	-	-	241 077	2.81	123 060	1.53	-	-	123 060	1.53
Western Asia	55 914	4.81	-	-	55 914	4.81	16 904	1.45	-	-	16 904	1.45
Central and Eastern Europe	149 024	6.10	-	-	149 024	6.10	49 951	1.80	-	-	49 951	1.80
Western Europe	169 640	9.90	-	-	169 640	9.90	41 629	1.65	-	-	41 629	1.65
Southern Europe	119 577	8.51	-	-	119 577	8.51	28 064	1.41	-	-	28 064	1.41
Northern Europe	84 272	9.63	-	-	84 272	9.63	18 063	1.46	-	-	18 063	1.46
Australia and New Zealand	22 062	10.16	-	-	22 062	10.16	3 631	1.37	-	-	3 631	1.37
Melanesia	2 116	5.30	-	-	2 116	5.30	1 046	2.73	-	-	1 046	2.73
Polynesia	252	7.46	-	-	252	7.46	78	2.46	-	-	78	2.46
Micronesia	121	4.44	-	-	121	4.44	47	1.71	-	-	47	1.71
Low HDI	105 620	3.40	-	-	105 620	3.40	52 846	1.78	-	-	52 846	1.78
Medium HDI	402 800	3.34	-	-	402 800	3.34	183 827	1.61	-	-	183 827	1.61
High HDI	666 731	4.29	-	-	666 731	4.29	184 014	1.12	-	-	184 014	1.12
Very high HDI	912 469	8.16	-	-	912 469	8.16	205 616	1.44	-	-	205 616	1.44
World	2 088 849	5.03	-	-	2 088 849	5.03	626 679	1.41	-	-	626 679	1.41

Figure 2.14: Breast cancer incidence and mortality statistics worlwide and by region

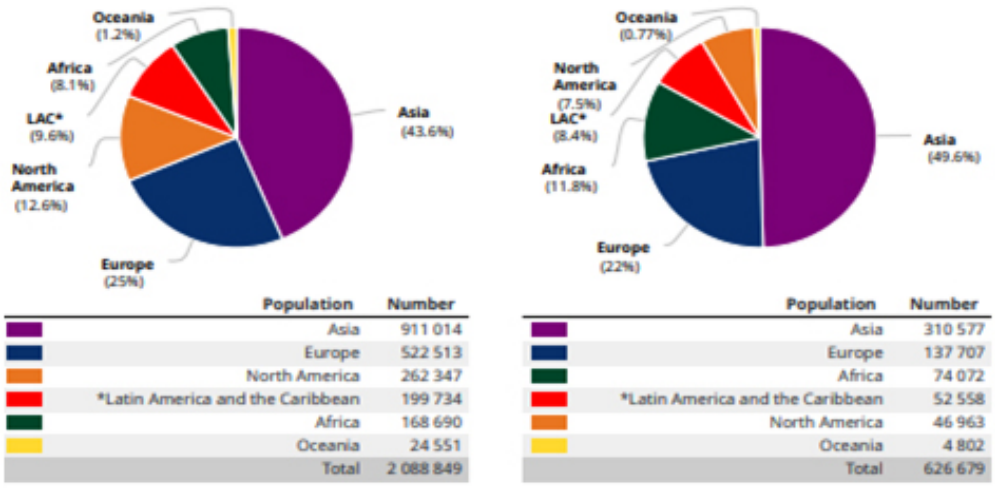


Figure 2.15: Breast Cancer Incidence and Mortality, both sexes

2.3.3 Colorectal Cancer

Colorectal cancer According to the data of the World Health Organization in 2018, it ranks second in women after breast cancer and third in men. Colorectal cancer, which is the third type of cancer for both sexes, is the second most common cause of death in the world.

Bowel cancer is the last part of our digestive system, which is also known as the type of cancer that occurs in the large intestine. Colorectal cancer is considered to be one of the risk factors, eating habits in the high amount of nutrients are preferred, while high amounts of fiber food is not preferred.

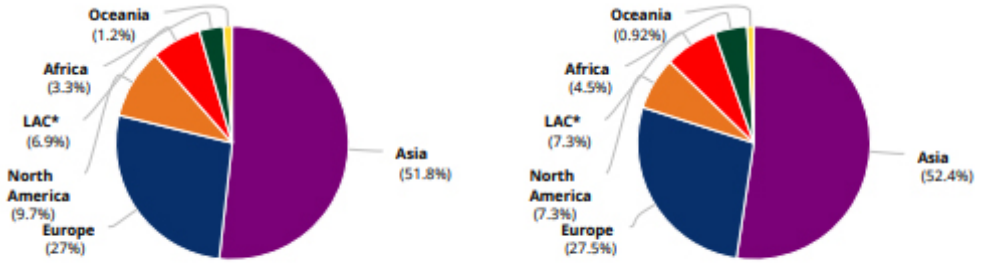
The prevalence of colorectal cancer worldwide is shown in Figure 2.16 and Figure 2.17 shows the rates of colorectal cancer according to continents as incidence and mortality by genders.

	Incidence						Mortality					
	Both sexes		Males		Females		Both sexes		Males		Females	
	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)
Eastern Africa	17 125	0.89	7 933	0.90	9 192	0.89	12 201	0.65	5 802	0.68	6 399	0.62
Middle Africa	6 010	0.84	2 895	0.83	3 115	0.85	4 562	0.64	2 232	0.64	2 330	0.64
Northern Africa	18 810	1.05	9 696	1.14	9 114	0.97	10 902	0.57	5 801	0.64	5 101	0.50
Southern Africa	7 167	1.52	3 637	1.92	3 530	1.24	3 801	0.74	1 979	0.98	1 822	0.57
Western Africa	12 734	0.74	6 489	0.79	6 245	0.70	8 568	0.52	4 440	0.56	4 128	0.49
Caribbean	10 886	2.03	5 016	2.06	5 870	2.01	6 259	0.99	2 898	1.04	3 361	0.95
Central America	19 520	1.26	9 959	1.41	9 561	1.13	9 614	0.58	4 910	0.66	4 704	0.52
South America	97 600	2.13	48 061	2.39	49 539	1.90	48 793	0.95	24 563	1.11	24 230	0.81
North America	179 771	2.96	93 898	3.38	85 873	2.58	64 121	0.87	33 752	1.05	30 369	0.71
Eastern Asia	736 573	3.06	426 342	3.73	310 231	2.40	325 128	1.11	183 346	1.38	141 782	0.84
South-Eastern Asia	95 223	1.66	53 542	2.06	41 681	1.31	52 475	0.86	29 384	1.08	23 091	0.67
South-Central Asia	88 033	0.57	53 534	0.71	34 499	0.43	63 401	0.41	39 852	0.53	23 549	0.29
Western Asia	38 067	1.88	21 490	2.24	16 577	1.55	20 418	0.94	11 240	1.12	9 178	0.77
Central and Eastern Europe	164 998	3.59	84 951	4.68	80 047	2.83	94 545	1.80	48 025	2.43	46 520	1.37
Western Europe	138 820	3.30	75 948	4.02	62 872	2.63	61 304	1.04	33 323	1.33	27 981	0.76
Southern Europe	119 949	3.76	69 446	4.84	50 503	2.78	53 975	1.21	30 991	1.62	22 984	0.84
Northern Europe	75 900	3.68	41 255	4.32	34 645	3.08	32 659	1.10	17 367	1.32	15 292	0.89
Australia and New Zealand	21 217	4.10	11 444	4.71	9 773	3.52	7 424	1.10	3 893	1.30	3 531	0.90
Melanesia	906	1.51	557	2.05	349	1.05	561	0.93	372	1.34	189	0.58
Polynesia	113	2.09	67	2.59	46	1.62	30	0.55	23	0.86	7	0.26
Micronesia	96	2.10	55	2.59	41	1.59	51	1.17	31	1.57	20	0.77
Low HDI	38 047	0.81	18 491	0.83	19 556	0.80	27 363	0.60	13 565	0.62	13 798	0.58
Medium HDI	174 011	0.82	101 633	1.01	72 378	0.64	111 853	0.51	66 429	0.64	45 424	0.38
High HDI	737 862	2.50	414 653	2.96	323 209	2.06	360 627	1.06	201 076	1.29	159 551	0.84
Very high HDI	898 751	3.56	490 997	4.38	407 754	2.82	380 563	1.16	202 947	1.47	177 616	0.89
World	1 849 518	2.27	1 026 215	2.75	823 303	1.83	880 792	0.92	484 224	1.14	396 568	0.72

Figure 2.16: Colorectal Cancer incidence and mortality statistics worlwide and by region

2.3.4 Prostate Cancer

As stated WHO in 2018 data, prostate cancer which is one of the most common types of cancer, ranks fourth. It is in the eighth rank in the death rate.



Population	Number
Asia	957 896
Europe	499 667
North America	179 771
*Latin America and the Caribbean	128 006
Africa	61 846
Oceania	22 332
Total	1 849 518

Population	Number
Asia	461 422
Europe	242 483
North America	64 121
*Latin America and the Caribbean	64 666
Africa	40 034
Oceania	8 066
Total	880 792

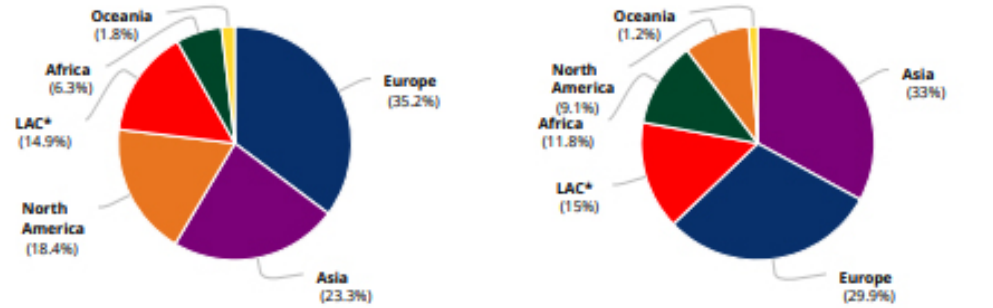
Figure 2.17: Colorectal Cancer Incidence and Mortality, both sexes

The prostate plays an important role in the male reproductive system. Prostate is a secretory gland for the formation and maintenance of viable and healthy sperm. Cancer occurs in the prostate gland. Therefore, it is a type of cancer seen only in men. It is located in the lower part of the bladder and it is a diaper which is involved in the prevention of urinary incontinence except that the sperm is alive and healthy. It is usually associated with aging as a risk factor. In addition, the risk factor is increased if a family has already had such a cancer. Because prostate cancer is a hormone-related structure, it is used in hormone therapy in cancer treatment methods.

All data of the World Health Organization in 2018 are given in Figure 2.18 and Figure 2.19 shows rates according to the continents by genders.

	Incidence						Mortality					
	Both sexes		Males		Females		Both sexes		Males		Females	
	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	New cases	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)	Deaths	Cum. risk 0-74 (%)
Eastern Africa	20 816	2.74	20 816	2.74	-	-	12 790	1.37	12 790	1.37	-	-
Middle Africa	11 666	4.07	11 666	4.07	-	-	7 133	2.13	7 133	2.13	-	-
Northern Africa	11 770	1.45	11 770	1.45	-	-	5 148	0.23	5 148	0.23	-	-
Southern Africa	12 950	7.04	12 950	7.04	-	-	4 699	2.07	4 699	2.07	-	-
Western Africa	23 769	3.72	23 769	3.72	-	-	12 528	1.77	12 528	1.77	-	-
Caribbean	17 563	7.93	17 563	7.93	-	-	8 605	2.02	8 605	2.02	-	-
Central America	33 711	5.21	33 711	5.21	-	-	9 921	0.87	9 921	0.87	-	-
South America	139 111	7.49	139 111	7.49	-	-	35 272	1.09	35 272	1.09	-	-
North America	234 278	9.50	234 278	9.50	-	-	32 686	0.64	32 686	0.64	-	-
Eastern Asia	193 638	1.66	193 638	1.66	-	-	68 472	0.32	68 472	0.32	-	-
South-Eastern Asia	35 386	1.51	35 386	1.51	-	-	14 914	0.41	14 914	0.41	-	-
South-Central Asia	41 145	0.59	41 145	0.59	-	-	27 015	0.34	27 015	0.34	-	-
Western Asia	27 046	3.33	27 046	3.33	-	-	8 026	0.59	8 026	0.59	-	-
Central and Eastern Europe	98 138	5.61	98 138	5.61	-	-	33 684	1.44	33 684	1.44	-	-
Western Europe	160 684	9.76	160 684	9.76	-	-	32 014	0.75	32 014	0.75	-	-
Southern Europe	99 548	8.15	99 548	8.15	-	-	20 522	0.58	20 522	0.58	-	-
Northern Europe	91 391	10.77	91 391	10.77	-	-	21 095	0.94	21 095	0.94	-	-
Australia and New Zealand	22 096	10.89	22 096	10.89	-	-	3 961	0.70	3 961	0.70	-	-
Melanesia	1 078	4.20	1 078	4.20	-	-	401	1.03	401	1.03	-	-
Polynesia	217	8.07	217	8.07	-	-	67	1.79	67	1.79	-	-
Micronesia	105	4.52	105	4.52	-	-	36	1.20	36	1.20	-	-
Low HDI	53 890	3.01	53 890	3.01	-	-	31 129	1.46	31 129	1.46	-	-
Medium HDI	94 077	1.02	94 077	1.02	-	-	48 954	0.40	48 954	0.40	-	-
High HDI	324 685	2.41	324 685	2.41	-	-	120 204	0.53	120 204	0.53	-	-
Very high HDI	802 294	7.92	802 294	7.92	-	-	158 335	0.72	158 335	0.72	-	-
World	1 276 106	3.73	1 276 106	3.73	-	-	358 989	0.60	358 989	0.60	-	-

Figure 2.18: Prostate Cancer incidence and mortality statistics worlwide and by region



Population	Number
Europe	449 761
Asia	297 215
North America	234 278
*Latin America and the Caribbean	190 385
Africa	80 971
Oceania	23 496
Total	1 276 106

Population	Number
Asia	118 427
Europe	107 315
*Latin America and the Caribbean	53 798
Africa	42 298
North America	32 686
Oceania	4 465
Total	358 989

Figure 2.19: Prostate Cancer Incidence and Mortality, both sexes

CHAPTER 3

MACHINE LEARNING TECHNIQUES

3.1 Overview

In this chapter, basic definitions of Machine Learninig will be presented and then, five Machine Learning algorithms which are considered in this work; Backpropagation neural networks (BPNN), Radial-Basis Function Neural Networks (RBFNN), Support Vector Regression (SVR) and Decision Trees (DT) and Long-Short Term Memory neural network (LSTM) will be introduced.

3.2 Machine Learning

Machine Learning is a subclass of computer science and aim is to teach data to get proper response from machine according to the model charachteristics to predict, classify or cluster the data.

Learning occurs in four different way as supervised, unsupervised, semi-supervised and reinforcement (Burkov (2019)).

3.2.1 Supervised Learning

In supervised learning, the dataset is the collection of labeled examples:

$$\{(x_i, y_i)\}_{i=1}^{N} \tag{3.1}$$

Each element x_i among N is called a feature vector (Burkov (2019)).

The aim of supervised learning is to use this feature vector as input and outputs information to label for this feature vector in order to make classifications and predictions.

3.2.2 Unsupervised Learning

In supervised learning, the dataset is the collection of unlabeled examples not like as in supervised learning:

$$\{(x_i)\}_{i=1}^{N} \tag{3.2}$$

Again x_i among N is called a feature vector but there is not any corresponding labels for these feature vectors (Burkov (2019)).

The aim is to create a model that takes a feature vector x as input and either transforms it into another vector or into a value that can be used to solve especially a clustering problems.

3.2.3 Semi-supervised Learning

In that type of learning, dataset contains both labeled and unlabeled data. Usually, the quantity of unlabeled examples is much higher than the number of labeled examples and the goal is same as in supervised learning.

3.2.4 Reinforcement Learning

In reinforcement learning, machine acts in an environment. It percieves the states as a vector of features.

It executes actions in each state and different actions bring different rewards.

The goal is to learn a policy and a policy is a function f that takes inputs of a state and outputs an optimal action to execute in that state (Burkov (2019)).

3.3 Backpropagation Learning Algorithm

Backpropagation is a learning algorithm for multi-layer perceptron that updates weights of each neuron using gradient descent algorithm. Initial weights are generally randomly assigned and it starts by feeding inputs to the net and calculating total potential of following

hidden layer by corresponding weights as shown in Equation 3.3.

$$neth_1 = w_1 * i_1 + w_2 * i_2 + b_1 \tag{3.3}$$

where w and i are weights and corresponding inputs of neuron respectively.

Activation function produces the output of each neuron and same calculations are repeated until output layer. Outputs of corresponding input neurons are calculated using Sigmoid Activation function as shown below:

$$outh_1 = \frac{1}{1 + e^{-neth_1}} \tag{3.4}$$

At that layer, actual outputs are compared by targets and error is calculated. According to these error values, weights are updated until the convergence of neural network using Gradient-Descent Algorithm.

Backpropagation learning algorithm was used in several real-life applications in classification, prediction and optimization problems (Adali and Sekereoglu (2012), Senturk and Senturk (2016)).

Figure 3.1 shows the general architecture of BPNN.

3.4 Support Vector Regression

Support Vector Regression is a kind of Support Vector Machines with a few changes to accept real value outputs instead of binary numbers. It is effectively used in prediction of data while minimizing error by maximizing the margin of hyperplane. SVR was used successfully in prediciton problems recently (Sekeroglu, Dimililer, and Tuncal (2019)).

Figure 3.2 presents the general architecture of SVR.

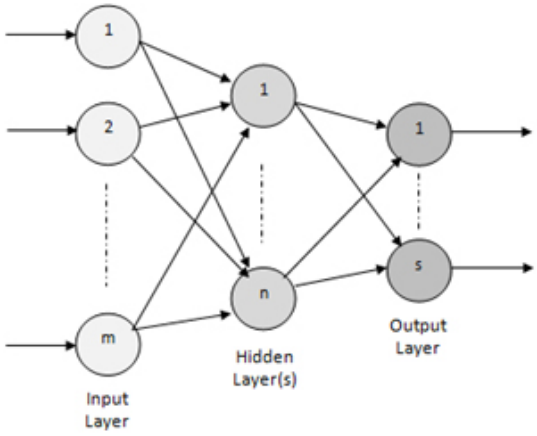


Figure 3.1: Architecture of backpropagation neural network

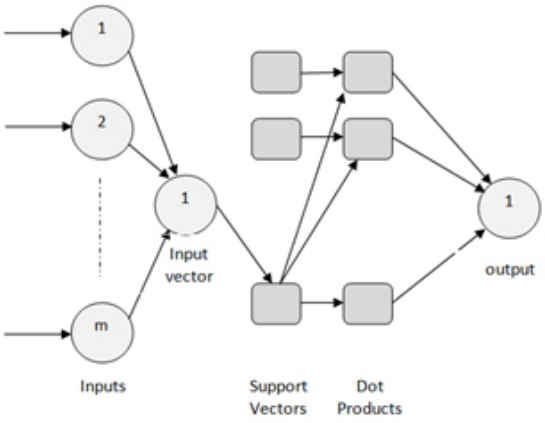


Figure 3.2: Architecture of support vector regression

3.5 Long-Short Term Memory Neural Network

LSTM is an effective special version of recurrent network and generally used for classification and prediction problems (Chen, Liu, and Liu (2017)). Four major components are formed its architecture: cell, input gate, output gate and forget gate. It uses gradients to update weights however, it remembers previous errors and this improves the error minimization of network in a short time.

Figure 3.3 demonstrates the general architecture of LSTM neural network.

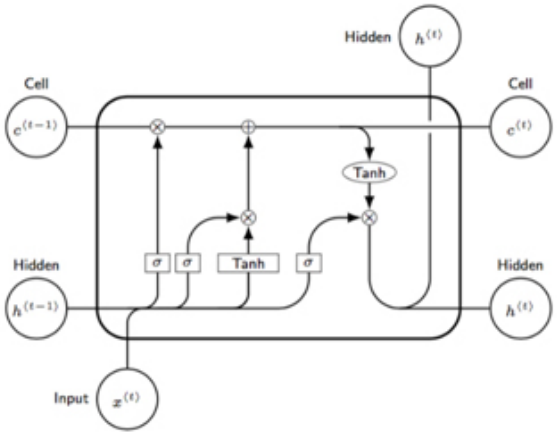


Figure 3.3: Architecture of LSTM (Image courtesy of stackexchange.com)

3.6 Radial-Basis Function Neural Network

RBFNN consists input, hidden, and output layer. It is limited to have exactly single hidden layer. It increases dimension of feature vector. Inputs of hidden neurons are calculated as same as BPNN which was given in Equation 3.3 and output of hidden neurons are calculated by using Radial Basis Functions which is shown below:

$$h(x) = e^{-\frac{(x-c)^2}{r^2}} \tag{3.5}$$

It can be used both for classification and prediction problems (Chang, Liang, and Chen (2001)). General architecture of RBFNN can be seen in Figure 3.4.

3.7 Decision Trees

Decision Trees were proposed for the classification problems. Then, they are modified to be used in regression models and, their simplicity and efficiency with large number of variables and cases make them popular for prediction problems (Geofrey Dougherty (2013)).

They are using divide-and-conquer strategy from root leaf to final leaf. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label or prediction value.

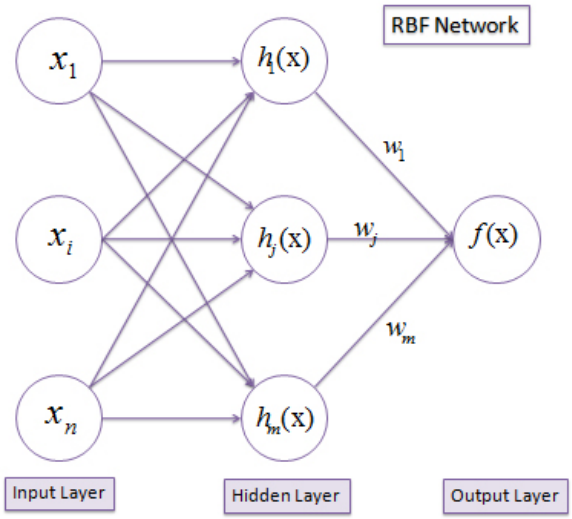


Figure 3.4: Architecture of RBF neural network (Image courtesy of towardscience.com)

Attribute selection can be performed by Information Gain or Gini Index to minimize the probable trees and to optimize the accuracy of the created tree. Information Gain is based on entropy which is given Equation 3.6 and Gini index is a metric to measure how often a randomly chosen element would be incorrectly identified.

$$I(x) = -\sum_{x \in X} p(x) log p(x) \tag{3.6}$$

General architecture of Decision Trees is shown in Figure 3.5.

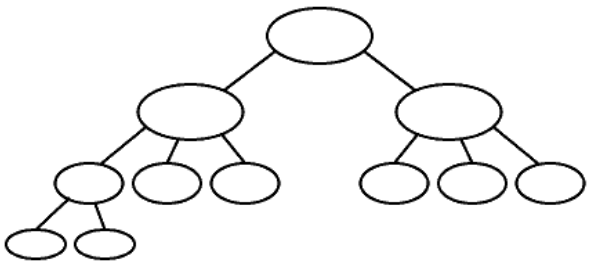


Figure 3.5: Architecture of decision trees

CHAPTER 4

EXPERIMENTAL DESIGN

4.1 Overview

In this section, design of experiments, characteristics of considered dataset, used data imputation techniques and evaluation strategies will be introduced.

4.2 Dataset

World Health Organization (WHO) published a data report that contains incidence and mortality rates of 2012 (Organization (2012)) for each cancer type according to the continents and genders belongs to these continents. In this thesis, European continent is considered to be used in prediction of cancer incidence rates for male and female separately.

2018 dataset is under preparation by WHO and it has been shared partially. Thus, it is decided to use 2012 dataset.

In 2012 European Dataset, incidence rates of 29 cancer types for 22 countries as Austria (with 3 regions), Bulgaria, Belarus, Croatia, Cyprus, Denmark, Estonia, France (with 9 regions), Germany (with 2 regions), Iceland, Ireland, Italy (with 8 regions), Lithuania, Malta, Netherlands, Norway, Poland, Slovakia, Slovenia, Spain (with 9 regions), Switzerland (with 6 regions) and UK (with 11 regions) are declared both for male and female group. Some records starts from 1953 however some of them starts from 1998 to 2012. Only records of two countries were ended in 2010. These countries are Italy and Slovakia.

For this reason, it is decided to consider the years between 1993-2012 in this thesis which the most records are occured in the dataset and causes minimum data imputation technique that affects the learning of models.

Four cancer types with the highest incidence rates for male and female are considered in this thesis. These are Lung Cancer, Breast Cancer, Prostate Cancer, Colorectum Cancer. In

addition to this, experiments are performed for total of all 22 cancer types in order to test the stability and efficiency of Machine Learning techniques.

4.3 Region Selection

As it is mentioned above, Austria, France, Germany, Italy, Spain, Switzerland and UK have different number of regions that consists different number of incidence rates. Instead of taking average of these regions, the region that has the maximum incidence rate was selected in order to represent whole country.

4.4 Data Imputation

Data imputation is the replacing missing values in the data with some new value. In this thesis, it is decided to use nearest neighbor value to fill the missing values.

Missing years of Italy and Slovakia which were 2011 and 2012, were replaced by the value of 2010.

4.5 Data Normalization

After replacing all missing values, data is normalized between 0 and 1 for each attribute by using the following equation:

$$(\bar{X})^{(j)} = \frac{x^{(j)} - min^{(j)}}{max^{(j)} - min^{(j)}} \tag{4.1}$$

4.6 Evaluation Strategies

Evaluation of obtained results is performed according to 3 criteria, Mean Square Error (MSE), R^2 Score and Explained Variance (EV) Score which are the main indicators of the success of predicted results (Sekeroglu et al. (2019)).

Mean Square Error calculates the squares of error of estimator and it is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y})^2 \tag{4.2}$$

where n is the total number of samples and Y_i and $\hat{Y}_i$ are the predicted and expected outputs of estimator respectively.

Explained Variance Score is another evaluation criteria of an estimator and also known as the regression sum of squares. It is defined as:

$$EV = \sum_{i=1}(f_i - \hat{y})^2 \tag{4.3}$$

where f_i is the predicted values and $\hat{y}$ is real sample.

R^2 Score is variance of predictable sample from the independent sample. It is defined as:

$$R^2 = \frac{EV_s}{UV_s} \tag{4.4}$$

where UV is unexplained variations of samples.

4.7 Design of Experiments

For each considered cancer type and for total incidence rates, five Machine Learning models which were described in Chapter 3, are trained by using 60%, 70% and 80% of total data. During these training, experiments are divided into two groups as Male and Female group. Then, each obtained data was analysed according the evaluation criteria explained above separately for each group.

During the analysis of obtained results, all obtained results for each training ration are com-

pared with each other in order to determine optimal model for this prediction problem and also it is also tried to observe the effect of training ratio on the efficiency of prediction.

4.8 Selection of the Parameters of Machine Learning Models

Each machine learning model has its own uniqe hyperparameter in order to increase learning ability and prediciton performance. In this section, used parameters and reasons will be explained briefly.

4.8.1 Parameters for Decision Tree Regressor

As it is mentioned in Chapter 3, decision trees have 2 attribute selection techniques. However, these techniques can be used for classification problems. Thus, in this thesis, attribute selection criterion is used as mean squared error which is used for prediction.

4.8.2 Parameters for Support Vector Regressor

In Support Vector Regression, most frequently used kernel function, radial basis function is decided to be used. After several experiments γ and ε values are determined to be used as 0.005 and 0.01 respectively.

4.8.3 Parameters for Backpropagation Neural Network

From the characteristics of dataset, 19 inputs fed to backpropagation directly. After performing several experiments, 2 hidden layer was decided to be used with Sigmoid Activation Function for each. Optimum results were obtained by 500 hidden units in each hidden layer. Maximum iterations were limited to 3000 in order to avoid over-fitting.

4.8.4 Parameters for Radial Basis Function Neural Network

In radial basis function neural network, learning rate was determined as 0.09 and maximum iterations were limited to 4000 in order to avoid over-fitting. Radial-basis functions were used in hidden layer as it is expected.

4.8.5 Parameters for Long-Short Term Memory Neural Network

In LSTM, 3 hidden layers were added to the architecture to increase the prediction ability of the model. In output layer, Sigmoid Activation Function was used and maximum iterations were limited to 200.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Overview

In this section, performed experiments, obtained results, analyses on these results and discussions will be presented in details.

5.2 Experimental Results

5.2.1 Male Group Results

As it is mentioned in Chapter 4, five ML models were trained by considering three different training ratio for each group.

For lung cancer results of male group, obtained results showed that SVR produced more accurate results than other models in all training ratios when R^2 and EV Scores are considered. When MSE is considered, again SVR achieved highest results except 80% of training ratio which Decision Tree produced mininum error in this ratio either its R^2 and EV Scores are lower than SVR. Table 5.1 shows obtained results for Lung Cancer.

Table 5.1: Results for lung cancer of male group with different training ratios

60% Training					
Result	**DT**	**SVR**	**BP**	**RBFNN**	**LSTM**
MSE	0.0020	**0.0009**	0.0156	0.0146	0.0339
R^2	0.797	**0.988**	0.811	0.832	0.616
EV	0.821	**0.988**	0.819	0.842	0.695
70% Training					
MSE	0.0014	**0.0012**	0.024	0.0032	0.0659
R^2	0.778	**0.986**	0.737	0.964	0.311
EV	0.780	**0.987**	0.746	0.972	0.432
80% Training					
MSE	**0.0004**	0.0013	0.0325	0.0067	0.0124
R^2	0.843	**0.988**	0.724	0.923	0.891
EV	0.896	**0.989**	0.749	0.925	0.899

Figure 5.1 - 5.5 shows prediction graphs of DT, SVR, BP, RBFNN and LSTM for 70% of training ratio respectively.

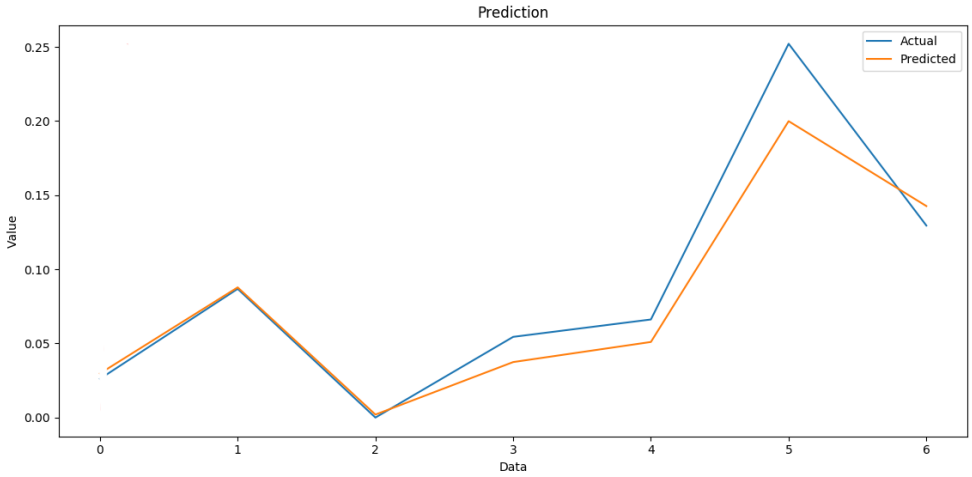


Figure 5.1: Prediction graph of decision tree for lung cancer with 70% of training ratio

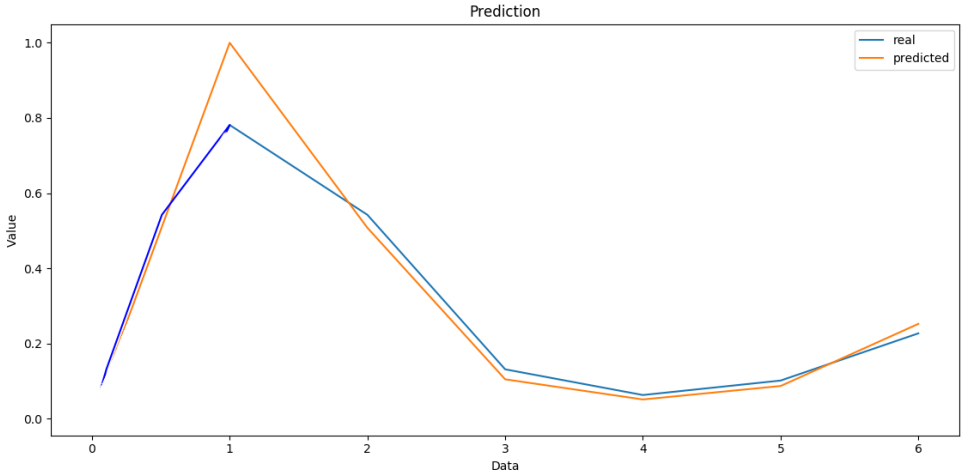


Figure 5.2: Prediction graph of support vector regressor for lung cancer with 70% of training ratio

For prostate cancer results of male group, obtained results showed that SVR produced more accurate results than other models similar to the results obtained in lung cancer predictions. When MSE is considered, again SVR achieved highest results except 70% of training ratio which Decision Tree produced mininum error in this ratio but again highest performance was

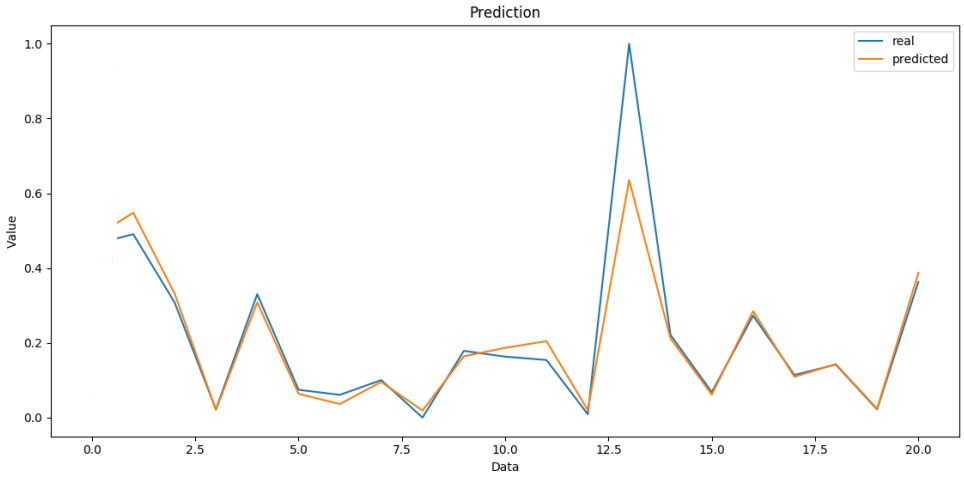


Figure 5.3: Prediction graph of backpropagation for lung cancer with 70% of training ratio

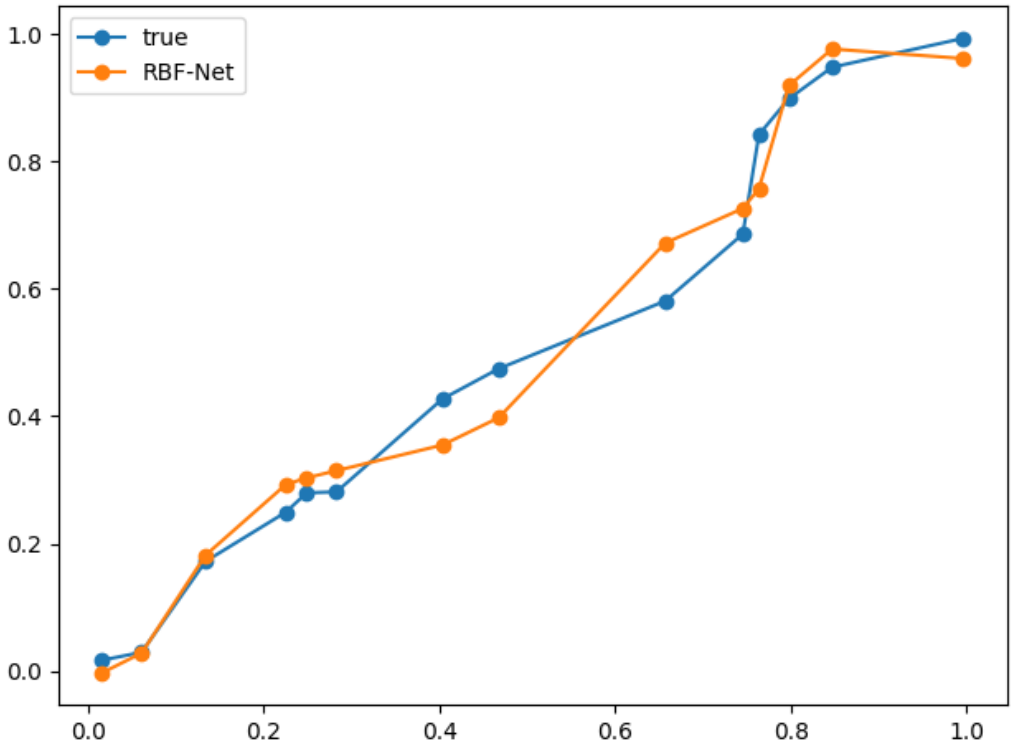


Figure 5.4: Prediction graph of radial basis function nn for lung cancer with 70% of training ratio

achieved by SVR in R^2 and EV Scores. LSTM was not able to produce any prediction result for this data. Table 5.2 shows obtained results for Prostate Cancer.

Example prediction graphs of DT, SVR and RBFNN for prostate cancer prediction with 60%

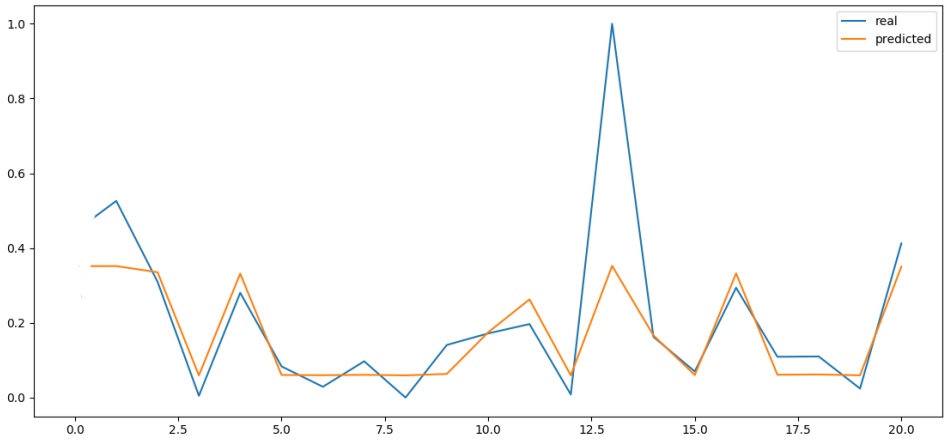


Figure 5.5: Prediction graph of LSTM for lung cancer with 70% of training ratio

Table 5.2: Results for prostate cancer of male group with different training ratios

60% Training					
Result	**DT**	**SVR**	**BP**	**RBFNN**	**LSTM**
MSE	0.0061	**0.0013**	0.0389	0.0110	NA
R^2	0.694	**0.984**	0.577	0.842	NA
EV	0.790	**0.989**	0.626	0.843	NA
70% Training					
MSE	**0.0013**	0.0014	0.0508	0.0073	NA
R^2	0.925	**0.984**	0.452	0.895	NA
EV	0.928	**0.986**	0.543	0.955	NA
80% Training					
MSE	0.0043	**0.0012**	0.0575	0.0047	NA
R^2	0.780	**0.989**	0.511	0.952	NA
EV	0.784	**0.990**	0.639	0.954	NA

of training ratio is shown in Figure 5.6, 5.7 and 5.8 respectively.

For colorectum cancer results of male group, obtained results showed that similar to other cancer types, SVR produced more accurate results for all evaluation criteria except MSE for 80% of training ratio. In that ratio, DT produced suprisingly lower error value however EV and R^2 scores are not successful enough that means overfitting occurred during the training. LSTM could not produce any prediction result for this data also. Table 5.3 shows obtained results for Colorectum Cancer.

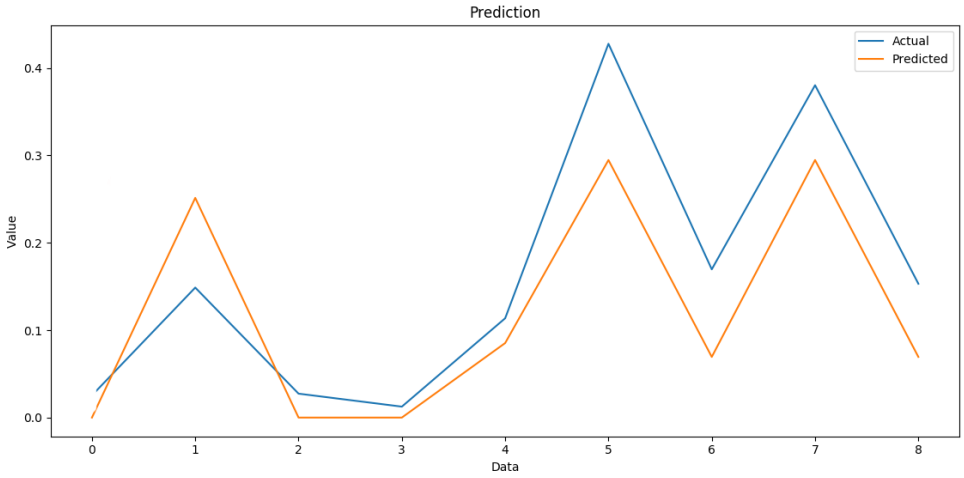


Figure 5.6: Prediction graph of DT for prostate cancer with 60% of training ratio

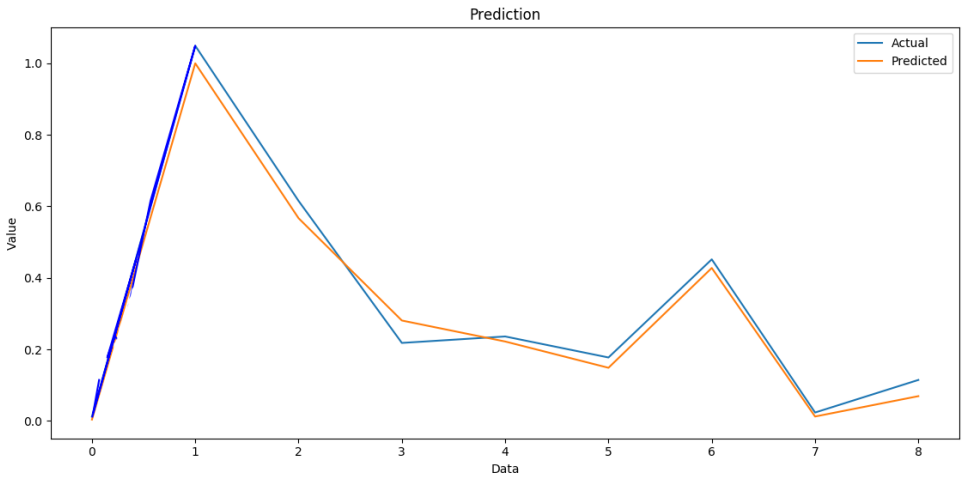


Figure 5.7: Prediction graph of SVR for prostate cancer with 60% of training ratio

Prediction graphs of SVR and RBFNN for colorectum cancer is shown in Figure 5.9 and 5.10 respectively.

When all types of cancers considered, similar results obtained by SVR but closer results are obtained by backpropagation which were not obtained in other experiments. LSTM was able to produce some prediction results for this dataset however the prediction results are not superior when they are compared to the results produced by SVR and backpropagation neural network.

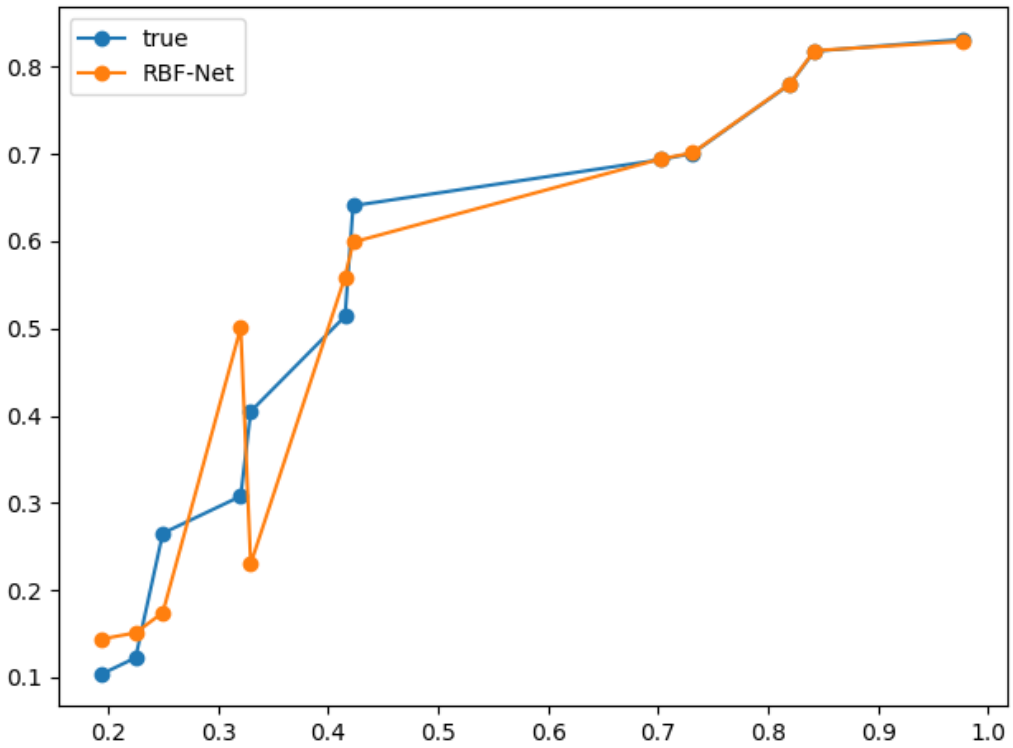


Figure 5.8: Prediction graph of RBFNN for prostate cancer with 60% of training ratio

Table 5.3: Results for colorectum cancer of male group with different training ratios

60% Training					
Result	**DT**	**SVR**	**BP**	**RBFNN**	**LSTM**
MSE	0.0039	**0.0022**	0.0380	0.0141	NA
R^2	0.760	**0.973**	0.552	0.816	NA
EV	0.802	**0.983**	0.592	0.922	NA
70% Training					
MSE	0.0046	**0.0025**	0.0541	0.0076	NA
R^2	0.6919	**0.977**	0.406	0.912	NA
EV	0.706	**0.982**	0.530	0.977	NA
80% Training					
MSE	**0.0008**	0.0012	0.0697	0.0197	NA
R^2	0.6921	**0.989**	0.396	0.792	NA
EV	0.697	**0.990**	0.592	0.843	NA

Table 5.4 shows obtained results for All Cancer for Men with different training ratios.

Prediction graphs of DT, SVR and BP for all cancer types with 70% of training ratio is shown in Figure 5.11, 5.12 and 5.13 respectively.

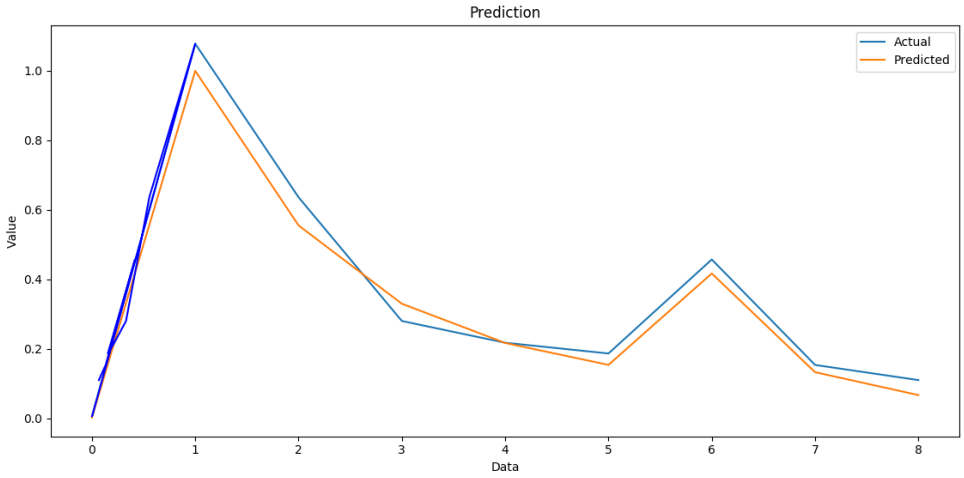


Figure 5.9: Prediction graph of SVR for colorectum cancer with 60% of training ratio

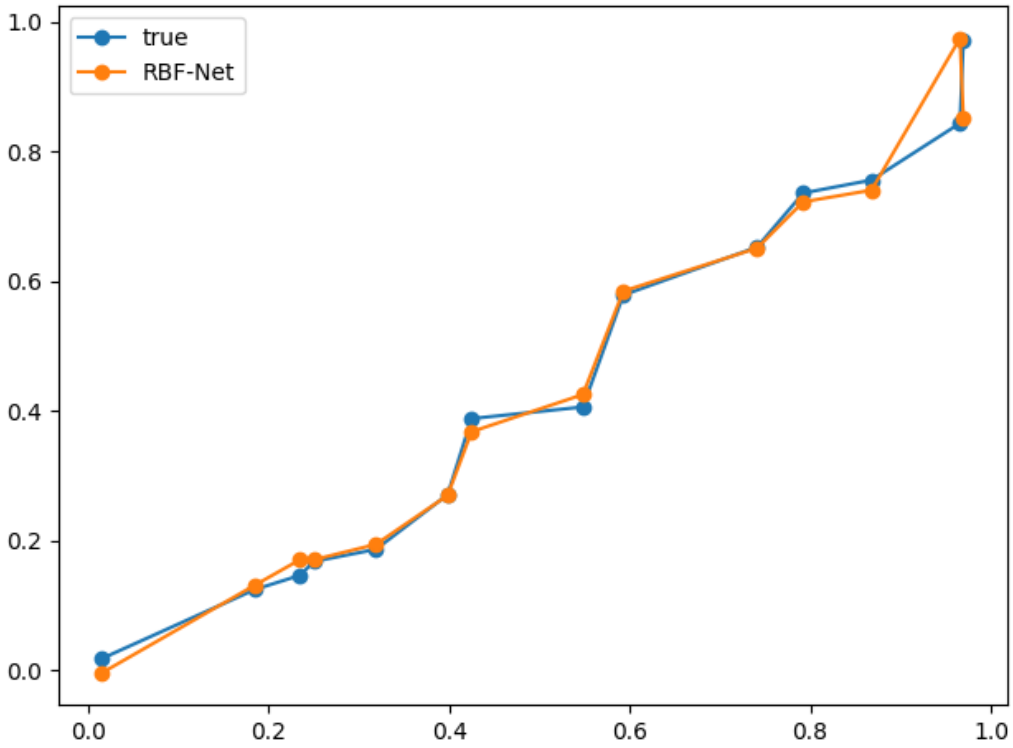


Figure 5.10: Prediction graph of RBF for prostate cancer with 70% of training ratio

5.2.2 Discussions on Male Group Results

As tables show above, Support Vector Regression achieved more accurate results in all experiments of male group.

Table 5.4: Results for all types of cancers of male group with different training ratios

60% Training					
Result	**DT**	**SVR**	**BP**	**RBFNN**	**LSTM**
MSE	0.0012	**0.0001**	0.0002	0.0048	0.0027
R^2	0.870	**0.990**	0.985	0.918	0.860
EV	0.886	**0.990**	0.985	0.918	0.879
70% Training					
MSE	0.0015	**0.0001**	0.0003	0.0195	0.0038
R^2	0.863	**0.991**	0.979	0.792	0.845
EV	0.876	**0.992**	0.981	0.821	0.894
80% Training					
MSE	0.0013	**0.0002**	0.0004	0.0116	0.0021
R^2	0.865	**0.991**	0.980	0.816	0.956
EV	0.865	**0.992**	0.983	0.869	0.936

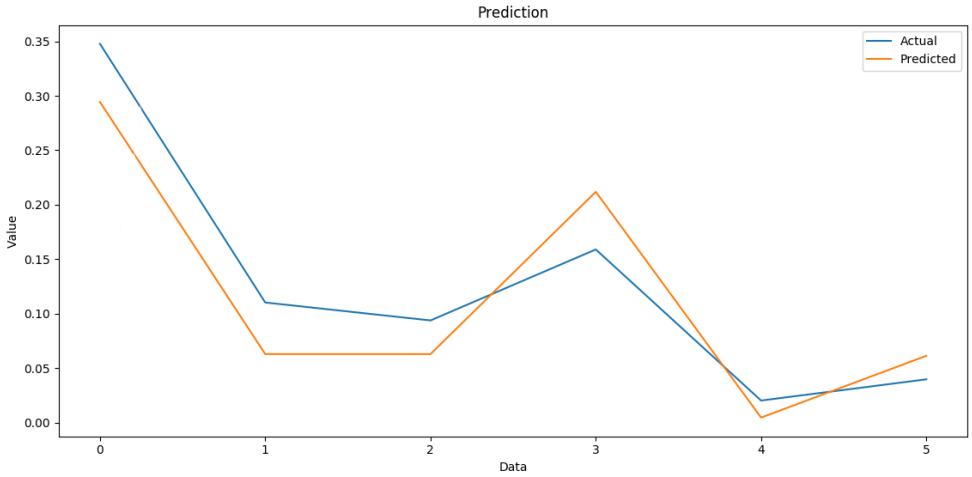


Figure 5.11: Prediction graph of DT for colorectum cancer with 70% of training ratio

Considering R^2 and EV Scores directly indicates that SVR outperforms other models for prediction problems and it is followed by RBFNN except few examples of experiments which is BP in all cancer types and DT in prostate cancer using 70% of training.

Generally MSE and other considered criterias have linear relationship however in some situations, DT produces outstanding lowest MSE values but not sufficient EV and R^2 scores that shows us the effect of overfitting.

Another comparison is performed using training ratios within the experiments. Obtained

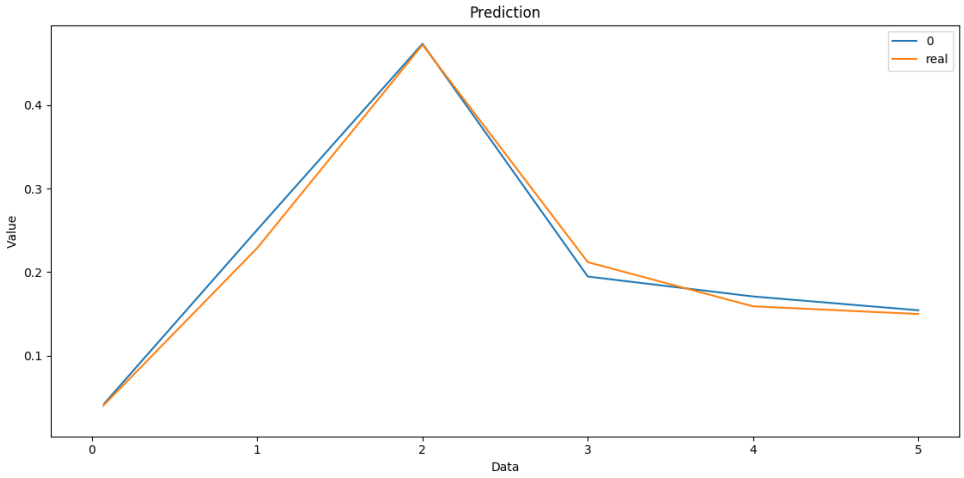


Figure 5.12: Prediction graph of SVR for all cancer for male group with 70% of training ratio

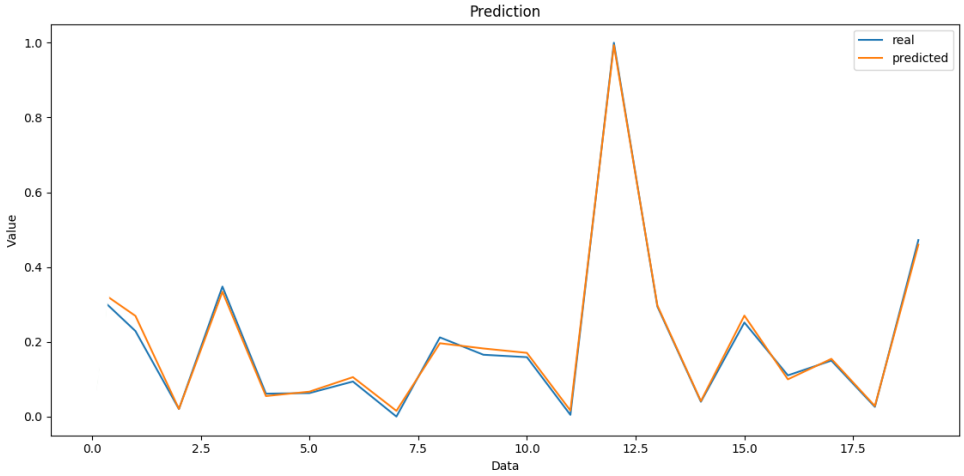


Figure 5.13: Prediction graph of BP for all cancer for male group with 70% of training ratio

results show that there is not linear relationship between training ratios and prediction results. Using higher ratios does not mean that the prediction results will raise.

5.2.3 Female Group Results

For lung cancer results of female group, obtained results showed that RBFNN produced more accurate results than other models in 60% and 70% of training ratios when R^2 and EV Scores are considered. Increment of training ratio caused RBFNN not converge efficiently the data and in 80% of training ratio, SVR produced more accurate results. When MSE is

considered, it is observed that the increment of training ratio causes DT to minimze MSE but not produce optimal results in R^2 and EV Scores. Table 5.5 shows obtained results for Lung Cancer.

Table 5.5: Results for lung cancer of female group with different training ratios

60% Training					
Result	**DT**	**SVR**	**BP**	**RBFNN**	**LSTM**
MSE	**0.0015**	0.0058	0.0337	0.0040	0.0236
R^2	0.905	0.938	0.645	**0.951**	0.829
EV	0.906	0.943	0.692	**0.971**	0.858
70% Training					
MSE	**0.0004**	0.0072	0.0454	0.0037	0.0243
R^2	0.917	0.933	0.585	**0.960**	0.833
EV	0.932	0.938	0.648	**0.964**	0.891
80% Training					
MSE	**0.0002**	0.0056	0.0602	0.0086	0.1033
R^2	0.756	**0.960**	0.570	0.891	0.432
EV	0.822	**0.965**	0.679	0.920	0.542

Prediction graphs of RBFNN with 60% and 70%, and SVR with 70% of trainin ratio for lung cancer of femal group is shown in Figure 5.14, 5.15 and 5.16 respectively.

For breast cancer results of female group, obtained results showed that SVR produced more accurate results than other models in all training ratios when R^2 and EV Scores are considered.

When MSE is considered, it is clear that DT is superior and also increment of training ratios causes decrement of MSE suddenly. However, this does not help DT in other evaluation metrics to produce outstanding results as in MSE criteria. Table 5.6 shows obtained results for brast cancer of female group.

Prediction graphs of SVR and RBFNN with 60% and BP 80% of training ratios for breast cancer of female group is shown in Figure 5.17, 5.18 and 5.19 respectively.

For colorectum cancer results of female group, obtained results showed that similar to other

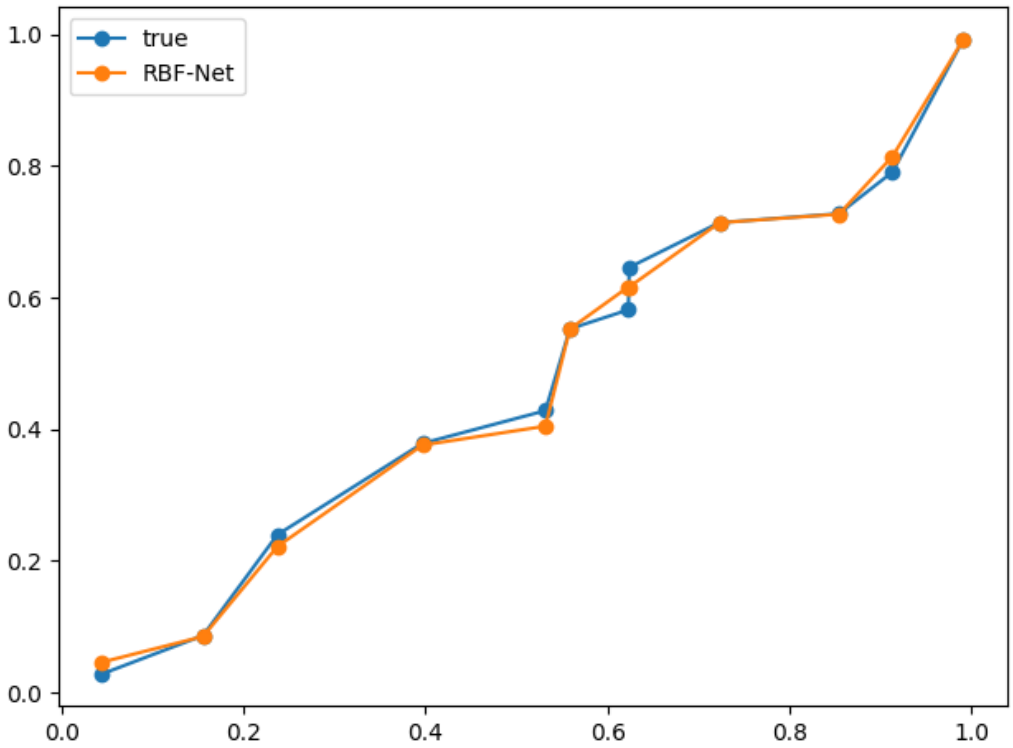


Figure 5.14: Prediction graph of RBFNN for lung cancer of female group with 60% of training ratio

Table 5.6: Results for breast cancer of female group with different training ratios

60% Training					
Result	**DT**	**SVR**	**BP**	**RBFNN**	**LSTM**
MSE	**0.0005**	0.0028	0.0501	0.0149	NA
R^2	0.927	**0.967**	0.421	0.708	NA
EV	0.928	**0.971**	0.514	0.868	NA
70% Training					
MSE	**0.0002**	0.0035	0.0633	0.0109	NA
R^2	0.932	**0.963**	0.359	0.839	NA
EV	0.934	**0.969**	0.477	0.949	NA
80% Training					
MSE	**0.0001**	0.0050	0.0823	0.0110	NA
R^2	0.926	**0.960**	0.354	0.901	NA
EV	0.967	**0.970**	0.532	0.911	NA

cancer types, SVR produced more accurate results for all evaluation criteria except for 80% of training ratio. In that ratio, DT produced suprisingly lower error value however in EV and R^2 scores RBFNN produced higher rates which are more accurate than other results. Table 5.7 shows obtained results for Colorectum Cancer.

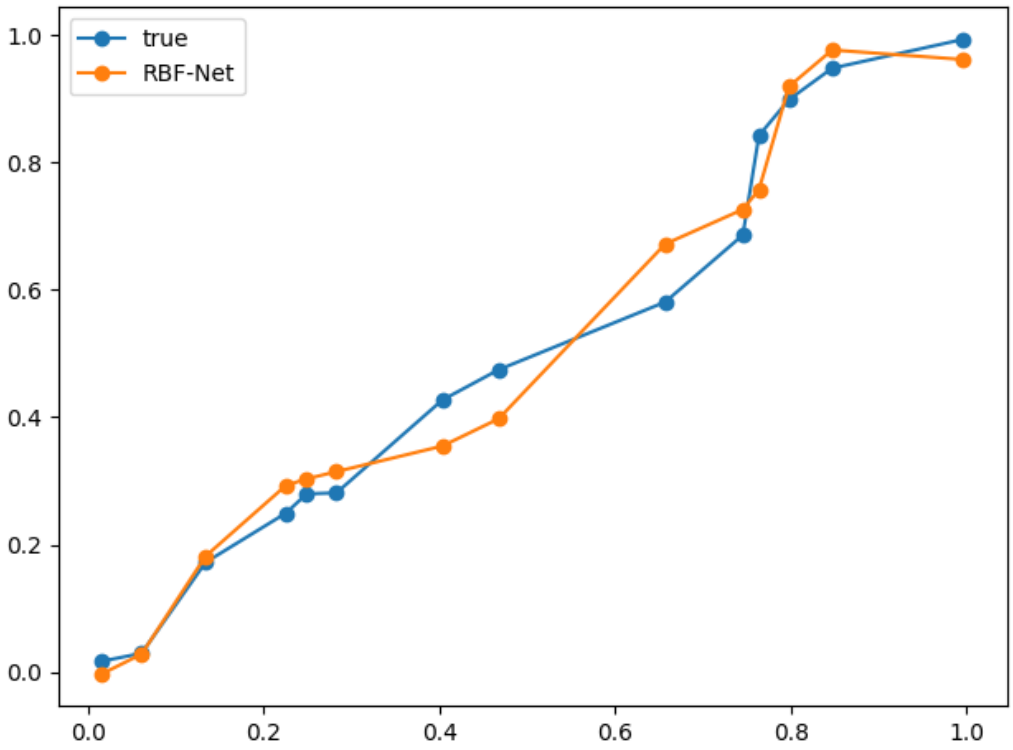


Figure 5.15: Prediction graph of RBFNN for lung cancer of female group with 70% of training ratio

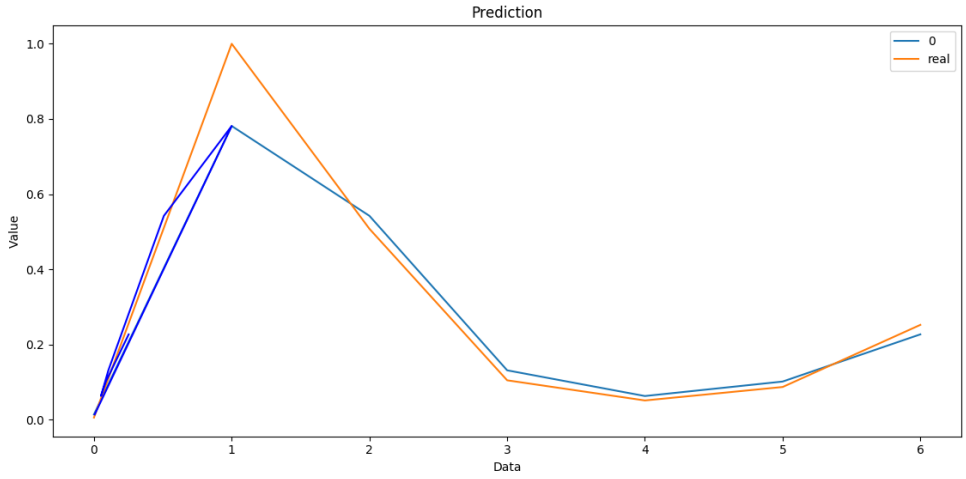


Figure 5.16: Prediction graph of SVR for lung cancer of female group with 70% of training ratio

Prediction graphs of DT, SVR, RBFNN and LSTM can be seen in Figure 5.20-5.23 respectively.

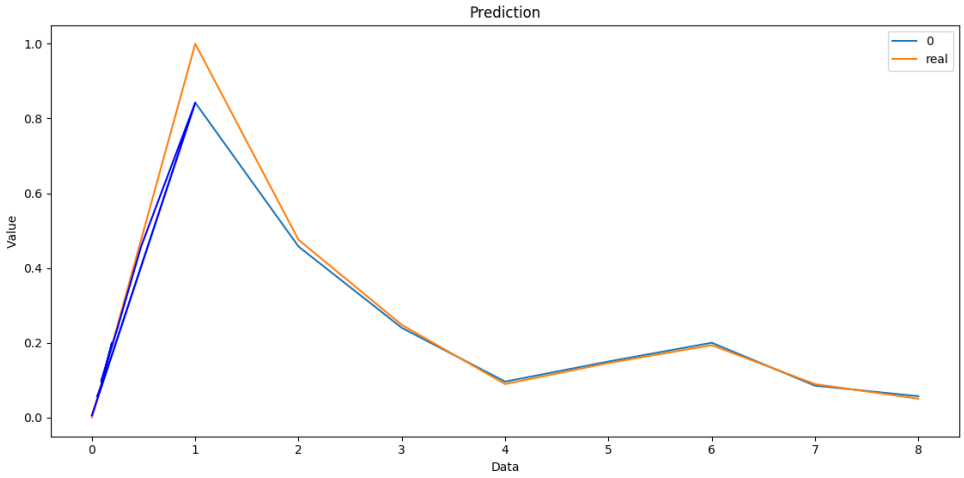


Figure 5.17: Prediction graph of SVR for breast cancer of female group with 60% of training ratio

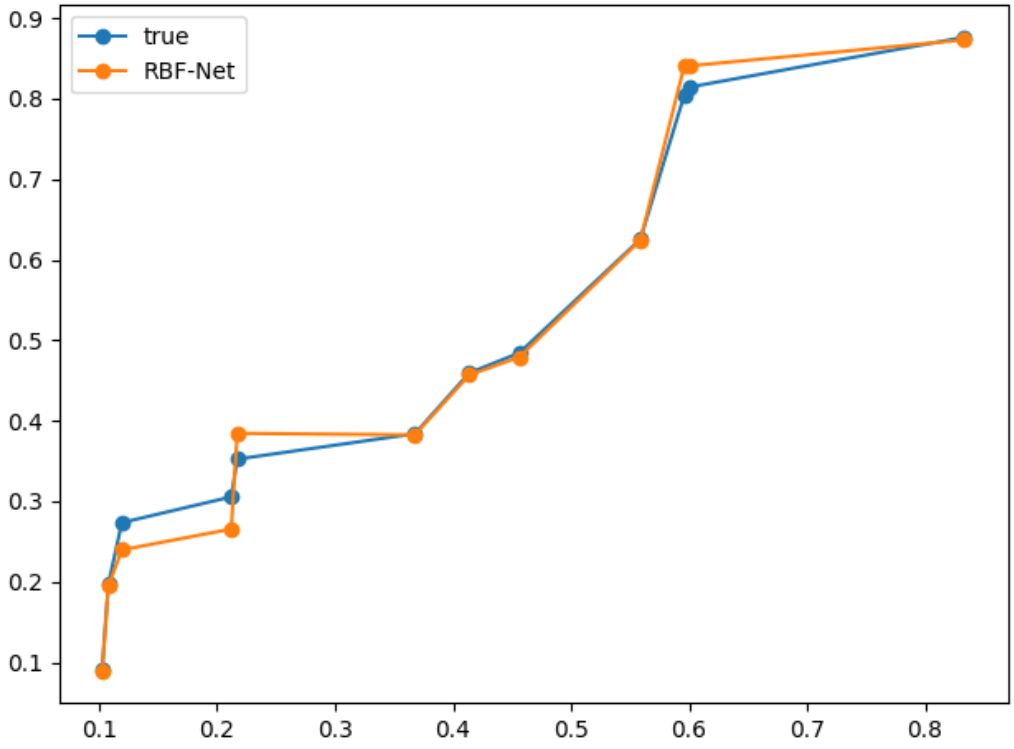


Figure 5.18: Prediction graph of RBF for breast cancer of female group with 60% of training ratio

When all types of cancers considered, SVR achieved highest results in all training ratios but in 80% of training ratio, BP and RBFNN produced exactly same EV score as SVR. However when R^2 Score and MSE is considered, they are close but not equal to SVR.

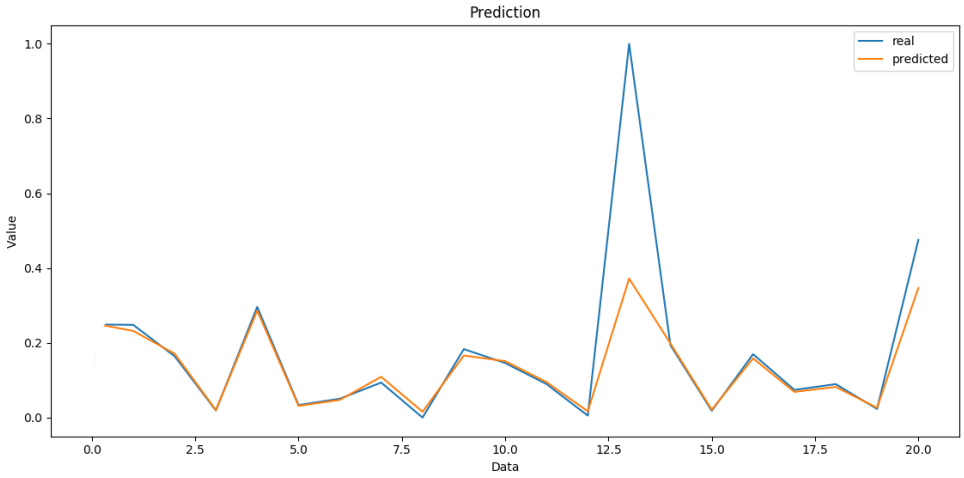


Figure 5.19: Prediction graph of SVR for breast cancer of female group with 80% of training ratio

Table 5.7: Results for colorectum cancer of male group with different training ratios

60% Training					
Result	**DT**	**SVR**	**BP**	**RBFNN**	**LSTM**
MSE	0.0027	**0.0024**	0.0026	0.0150	0.0263
R^2	0.900	**0.980**	0.788	0.873	0.793
EV	0.922	**0.984**	0.836	0.906	0.793
70% Training					
MSE	0.0045	**0.0023**	0.0374	0.0107	0.0340
R^2	0.819	**0.981**	0.705	0.897	0.646
EV	0.821	**0.985**	0.785	0.914	0.745
80% Training					
MSE	**0.0005**	0.0007	0.0446	0.0097	0.0422
R^2	0.848	0.793	0.686	**0.889**	0.613
EV	0.873	0.809	0.832	**0.971**	0.740

Table 5.8 shows obtained results for All Cancer for female with different training ratios.

Prediction graphs of SVR, BP and RBFNN with 80% of training ratio can be seen in Figure 5.24-5.26 respectively.

5.2.4 Discussions on Female Group Results

Interestingly, unstable results are obtained in Female Group results. Even SVR produced more accurate results in EV and R^2 Scores, in some experiments, BP and RBFNN produced

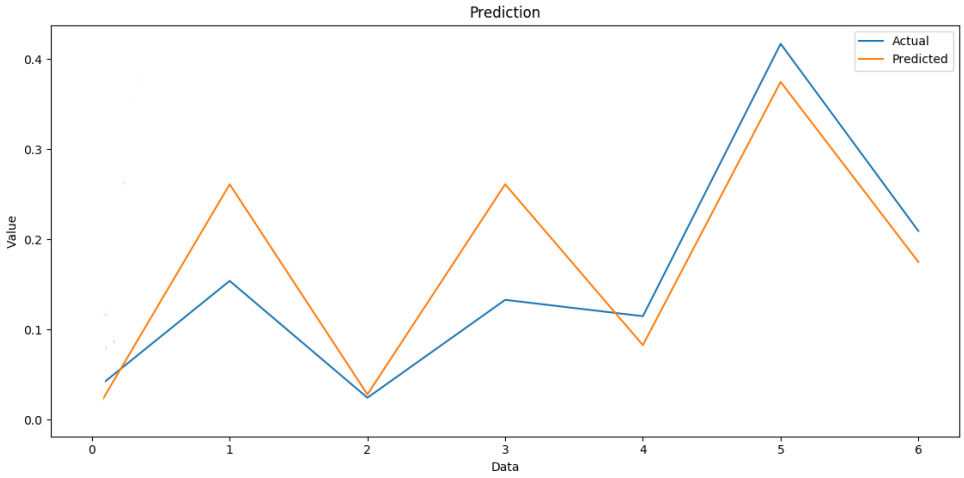


Figure 5.20: Prediction graph of DT for colorectum cancer of female group with 70% of training ratio

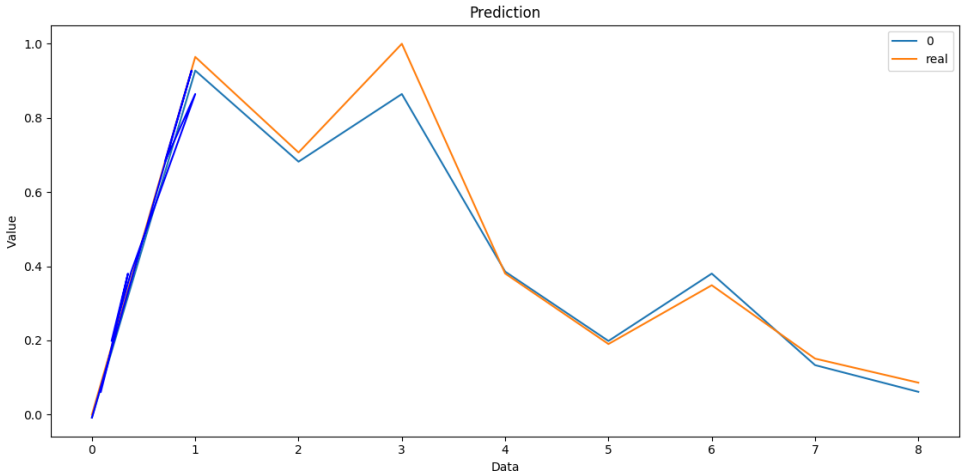


Figure 5.21: Prediction graph of SVR for colorectum cancer of female group with 60% of training ratio

similar, better or equal results as produced by SVR.

Similar to male group, Decision Tree minimizes error while increasing training samples, but this does not help model to increase the prediciton success.

Even the datasets are similar, the behaviour of models changed between Male and Femal groups. This may caused by small changes between the internal or general relationships of features within the dataset.

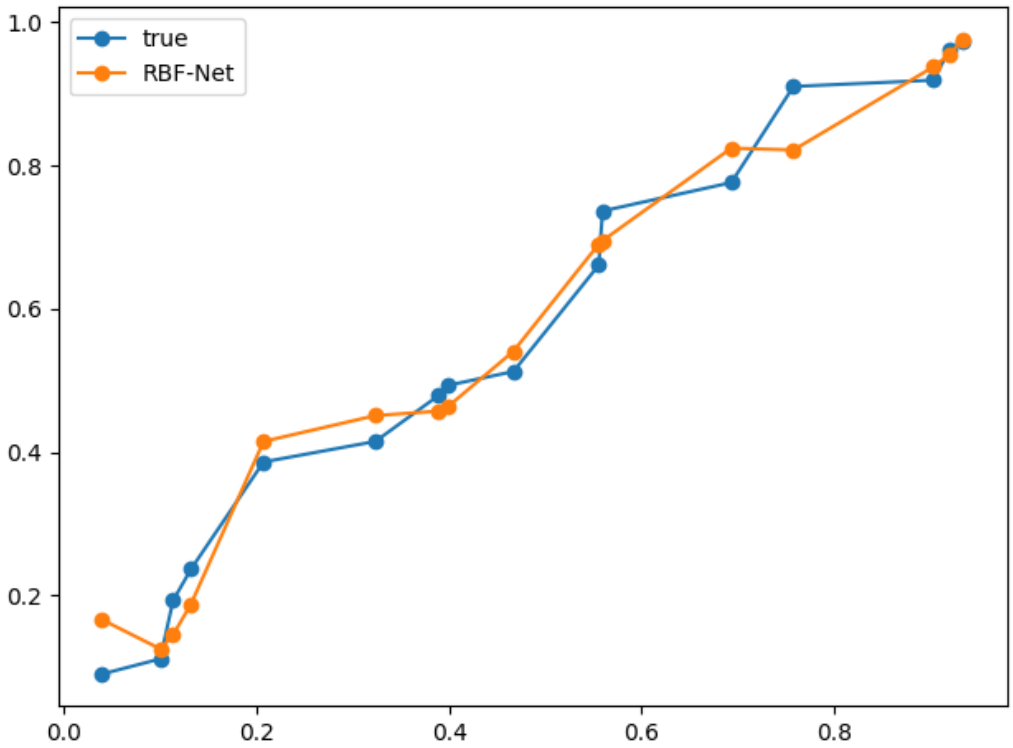


Figure 5.22: Prediction graph of RBf for colorectum cancer of female group with 80% of training ratio

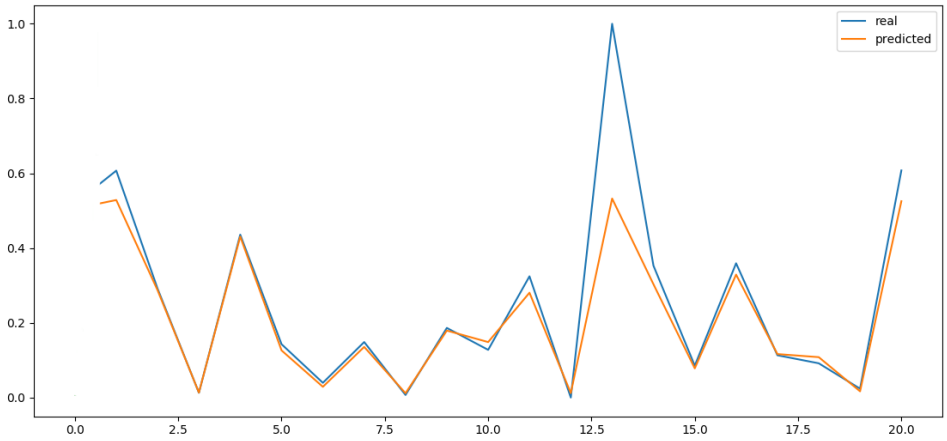


Figure 5.23: Prediction graph of LSTM for colorectum cancer of female group with 60% of training ratio

Considering the training ratios within the experiments, obtained results show similar relationship that there is not linear relationship between training ratios and prediction results. Using higher ratios does not mean that the prediction results will raise.

Table 5.8: Results for all types of cancers of female group with different training ratios

60% Training					
Result	**DT**	**SVR**	**BP**	**RBFNN**	**LSTM**
MSE	0.0012	**0.00009**	0.00017	0.0083	0.0066
R^2	0.872	**0.994**	0.990	0.921	0.770
EV	0.880	**0.995**	0.992	0.925	0.928
70% Training					
MSE	0.0002	**0.0001**	0.0015	0.0104	0.0019
R^2	0.977	**0.995**	0.931	0.813	0.946
EV	0.980	**0.995**	0.935	0.907	0.971
80% Training					
MSE	0.0009	**0.0001**	0.0002	0.0036	0.0049
R^2	0.911	**0.994**	0.992	0.992	0.898
EV	0.915	**0.995**	**0.995**	**0.995**	0.930

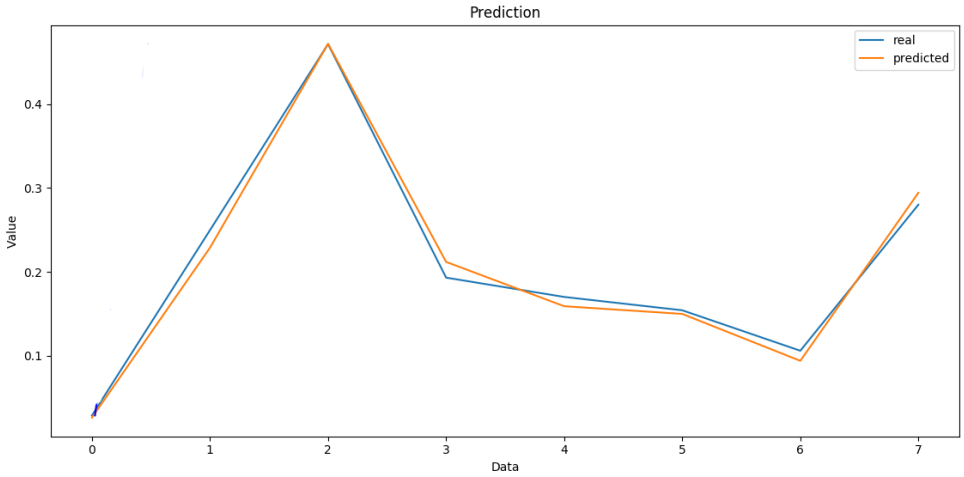


Figure 5.24: Prediction graph of SVR for all cancer types of female group with 80% of training ratio

Table 5.9 presents the generalized optimum results that were obtained within this thesis.

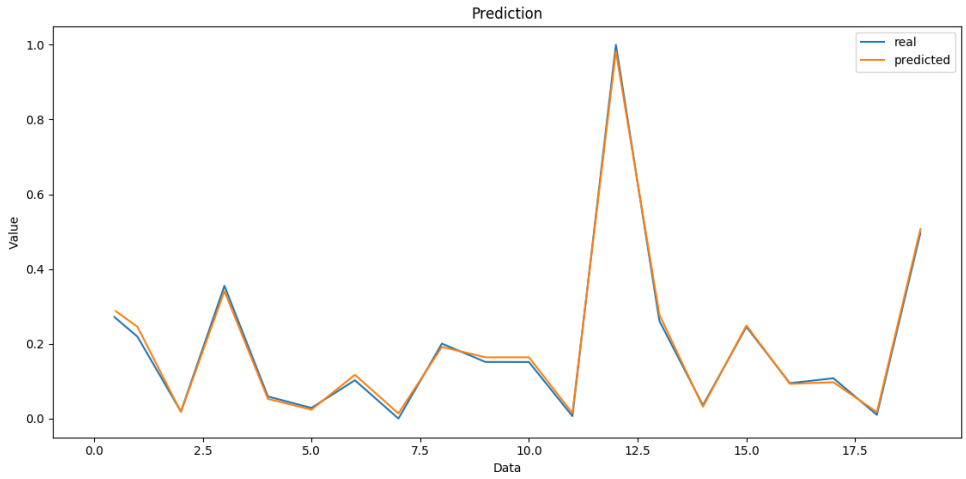


Figure 5.25: Prediction graph of BP for all cancer types of female group with 80% of training ratio

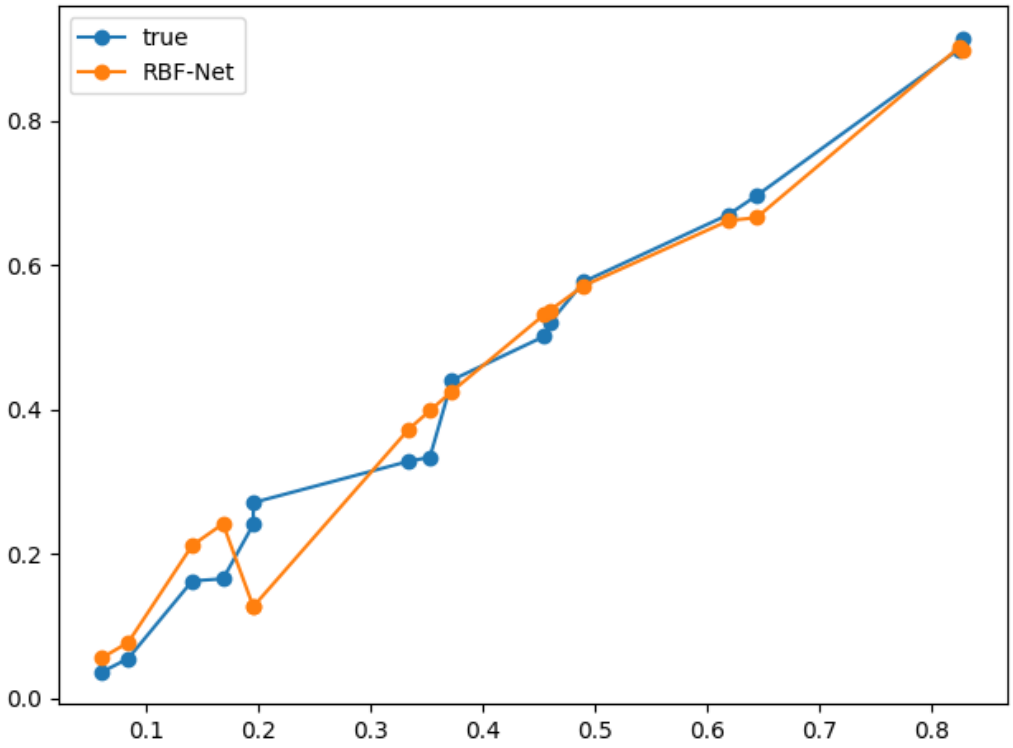


Figure 5.26: Prediction graph of RBF for all cancer types of female group with 80% of training ratio

Table 5.9: Most accurate results for MSE

Experiment	**MSE / R^2 / EV**	**Training Ratio**
Lung (M)	DT / SVR / SVR	80% / 80% / 80%
Lung (F)	DT / RBFNN / RBFNN	80% / 70% / 60%
Prostate (M)	SVR / SVR / SVR	80% / 80% / 80%
Breast (F)	DT / SVR / SVR	80% / 60% / 60%
Colorectum (M)	DT / SVR / SVR	80% / 80% / 80%
Colorectum (F)	DT / SVR / SVR	80% / 70% / 70%
All Types (M)	SVR / SVR / SVR	60-70% / 70-80% / 70-80%
All Types (F)	SVR / SVR / (SVR-BP-RBFNN)	60% / 70% / (ALL)

CHAPTER 6

CONCLUSIONS

6.1 Conclusions

Cancer is the main cause of death in the world and prediction researches for both incidence and mortality rates have increased popularity nowadays.

In this thesis, five machine learning models are implemented in order to predict cancer incidence rates according to the WHO 2012 report. Obtained results show that Support Vector Regression is superior to other models in the prediction of considered Lung, Prostate, Breast, Colorectal and all cancer types. It can also be concluded that increment of training data may help to increase prediction ability of models however, during this increment, decrement of Mean Squared Error does not offer successful prediction as it is in Decision Tree.

LSTM was not able to make predictions in some of these datasets and when it proceed, it never achieve considerable results.

Backpropagation neural network was one of the most unstable model in the experiments and never achieve optimum results.

Radial Basis Function Neural Network was also performed unstable predictions in some experiments but generally produces better results than other methods except Support Vector Regression.

Future work will include the implementation of more machine learning algorithms for the prediction of all 25 cancer types. Also, 2018 report of WHO will be considered when it will be distributed to the researchers.

REFERENCES

Adali, T., & Sekereoglu, B. (2012). *Analysis of micrornas by neural network for early detection of cancer.* Procedia Technology Elsevier, vol.1.pp.449-452, 2012.

Alhaj, M. A. M., & Maghari, A. Y. A. (2017). *Cancer survivability prediction using random forest and rule induction algorithms.* 2017 8th International Conference on Information Technology (ICIT).

Bilim ve teknik. (2002).

Bosetti, C., Malvezzi, M., Rosso, T., & et al., P. B. (2012). *Lung cancer mortality in european women: Trends and predictions.* Lung Cancer 78 (2012) 171-178, Elsevier.

Burkov, A. (2019). *The hundred-page machine learning book* (1st ed.). Author.

Chang, F., Liang, J., & Chen, Y. (2001). *Flood forecasting using radial basis function neural networks.* IEEE Trans, on Systems, Man, and Cybernetics, Part C: Applications and Reviews.

Chen, Z., Liu, Y., & Liu, S. (2017). *Mechanical state prediction based on lstm neural network.* 2017 36th Chinese Control Conference (CCC). pp.3876-3881,2017.

Geofrey Dougherty. (2013). *Pattern recognition and classification*. Springer.

Jung, K.-W., Won, Y.-J., Kong, H.-J., & Lee, E. S. (2017). *Prediction of cancer incidence and mortality in korea,2019.* Cancer Research and Treatment 2019;51(2):431-437.

Kachroo, S., Melek, W. W., & Kurian, C. (2013). *Evaluation of predictive learners for cancer incidence and mortality.* The 4th IEEE International Conference on E-Health and Bioengineering - EHB2013.

Kourou, K., Exarchos, T. P., & Exarchos, K. P. (2014). *Machine learning applications in cancer prognosis and prediction.* Computational and Structural Biotechnology Journal 13 (2015) 8-17, Elsevier.

Malvezzi, M., Bosetti, C., Rosso, T., & et al., P. B. (2013). *Lung cancer mortality in european men: Trends and predictions.* Lung Cancer 80 (2013) 138-145, Elsevier.

Malvezzi, M., P.Bertuccio, Levi, F., & et al., C. L. V. (2014). *European cancer mortality predictions for the year 2014.* Annals of Oncology 25:1650-1656,2014.

Mohammadzadeh, F., Noorkojuri, H., Pourhoseingholi, M., & et al., S. S. (2014). *Predicting the probability of mortality of gastric cancer patients using decision tree.* Irish Journal of Medical Science (2015) 184:277-284, Springer.

O'Lorcain, P., Deady, S., & Comber, H. (2006). *Mortality predictions for colon and anorectal cancer for ireland, 2003-17.* Colorectal Disease, 8, 393-401.

Organization, W. H. (2012). Global cancer observatory, 2012 cancer report [Computer software manual].

Organization, W. H. (2018). Global cancer observatory, 2018 cancer report [Computer software manual].

Ribes, J., Esteban, L., Cléries, R., & et al., J. G. (2013). *Cancer incidence and mortality projections up to 2020 in catalonia by means of bayesian models.* Clinical and Translational Oncology (2014) 16:714-724, Springer.

Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). *Student performance prediction and classification using machine learning algorithms.* 8th International Conference on Educational and Information Technology (ICEIT 2019).

Senturk, Z. K., & Senturk, A. (2016). *Yapay sinir ağları ile göğüs kanseri tahmini.* El-Cezeri Journal of Science and Engineering Vol: 3, No: 2, 2016 (345-350).