

**SYMPTOM WISE AGE PREDICTION OF CANCER PATIENT USING  
CLASSIFIER COMPARISON AND FEATURE SELECTION**

**BY**

**SUD MOHAMMAD RASHID  
ID: 161-15-6813**

**MD. NAYEM FERDOUS KHAN  
ID: 161-15-7088**

**AVIJIT BISWAS  
ID: 161-15-6715**

This Report Presented in Partial Fulfillment of the Requirements for the Degree  
of Bachelor of Science in Computer Science and Engineering

**Supervised By**

**Mr. Majidur Rahman**  
Lecturer  
Department of CSE  
Daffodil International University

**Co-Supervised By**

**Antara Mahmud**  
Lecturer  
Department of CSE  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

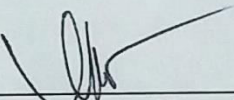
**DHAKA, BANGLADESH**

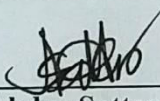
**DECEMBER 2019**

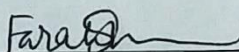
## APPROVAL

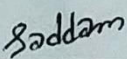
This Project/internship titled “**SYMPTOM WISE AGE PREDICTION OF CANCER PATIENT USING CLASSIFIER COMPARISON AND FEATURE SELECTION**”, submitted by Md. Nayem Ferdous Khan, ID No: 161-15-7088, Sud Mohammad Rashid, ID No: 161-15-6813 and Avijit Biswas, ID No: 161-15-6715 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 5-12-2019.

### BOARD OF EXAMINERS

  
\_\_\_\_\_  
**Dr. Syed Akhter Hossain** **Chairman**  
**Professor and Head**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

  
\_\_\_\_\_  
**Abdus Sattar** **Internal Examiner**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

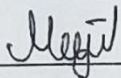
  
\_\_\_\_\_  
**Farah Sharmin** **Internal Examiner**  
**Senior Lecturer**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

  
\_\_\_\_\_  
**Dr. Md. Saddam Hossain** **External Examiner**  
**Assistant Professor**  
Department of Computer Science and Engineering  
United International University

## DECLARATION

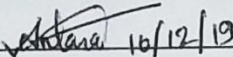
We hereby declare that, this thesis has been done by us under the supervision of **Mr. Majidur Rahman, Lecturer, Department of CSE** at Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

### Supervised by:



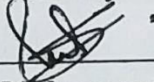
**Mr. Majidur Rahman**  
Lecturer  
Department of CSE  
Daffodil International University

### Co-Supervised by:

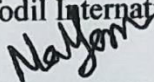


**Antara Mahmud**  
Lecturer  
Department of CSE  
Daffodil International University

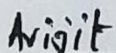
### Submitted by:



**Sud Mohammad Rashid**  
ID: 161-15-6813  
Department of CSE  
Daffodil International University



**Md. Nayem Ferdous Khan**  
ID: 161-15-7088  
Department of CSE  
Daffodil International University



**Avijit Biswas**  
ID: 161-15-6715  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Mr. Majidur Rahman, Lecturer, Department of CSE**, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Prof. Dr. Syed Akhter Hossain Professor and Head, Department of CSE** for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

## **ABSTRACT**

Cancer has become one of the most life threatening disease over the past few decades. Especially on Bangladesh the number of people being affected by cancer is increasing in an agitating rate. Again cancer, diagnosed after a certain stage, inevitably leads towards death. To abate this vicious upheaval of cancer, awareness has no other alternative. Our research primarily focuses on detection of certain age group, according to the corresponding cancer diagnosis and relevant factors. In order to do so, we have implemented logistic regression, support vector machine and convolutional neural network on the original dataset. Afterwards, two feature selection methods (Feature Importance Ranking Method and Recursive Feature Elimination) have been applied on the dataset to extract out the most significant features. The three classifier comparison has been implied on both the feature selection methods. It is found that the classifier accuracy on the extracted features is significantly better in case of Recursive Feature Elimination rather than Feature Importance Ranking Method.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGES</b>
Board of examiners	i
Declaration of the Student	ii
Acknowledgement	iii
Abstract	iv
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-3</b>
1.1 Introduction	1-2
1.2 Motivation	2
1.3 Research question	2
1.4 Research methodology	3
1.5 Research objective	3
1.6 Report layout	3
<b>CHAPTER 2: BACKGROUND</b>	<b>4-6</b>
2.1 Related work	4-5
2.2 Bangladesh perspective	5
2.3 Government Goals and Regulation	5-6
<b>CHAPTER 3: RESEARCH METHODOLOGIES</b>	<b>7-14</b>
3.1 Working procedure	7
3.1.1 Flow chart of working procedure	9
3.1.2 Preprocessing of Data	10
3.1.3 Feature Selection	10
3.1.3.1 Feature Importance Ranking Measure	10
3.1.3.2 Recursive Feature Elimination	10-11
3.1.4 Classification	11
3.1.4.1 Logistic Regression	11-12
3.1.4.2 Support Vector Machine	13
3.1.4.3 Convolutional Neural Network	13-14

<b>CHAPTER 4: RESULT AND OBSERVATIONS</b>	<b>15-19</b>
4.1 Experimental Analysis	15
4.2 Comparative Analysis	15-19
<b>CHAPTER 5: CONCLUSION AND FUTURE WORK</b>	<b>20</b>
5.1 Conclusion	20
5.2 Future work	20
<b>REFERENCE</b>	<b>21-22</b>

## LIST OF FIGURE

<b>FIGURE NAME</b>	<b>PAGES</b>
Figure-3.1: Stages of Working Procedure	9
Figure-3.2: Logistic Regression	12
Figure-3.3: Support Vector Machine	13
Figure-3.4: Convolutional Neural Network	14
Figure-4.1: Features Selected after Applying FIRM	17
Figure-4.2: Features Selected after Applying RFE	17



## LIST OF TABLE

<b>TABLE NO.</b>	<b>PAGES</b>
Table-3.1: Range Of Values Related To the Dataset	7-8
Table-4.1: Classification Report before Feature Selection	16
Table-4.2: Classification Report after FIRM	18
Table-4.3: Classification Report after RFE	19

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

For the past few decades, not only in Bangladesh but throughout the world, cancer has become a fatal disease. The likelihood of being affected by cancer has augmented for people of all ages. According to World Health Organization (WHO), cancer is the second leading cause of death all over the country which has resulted about 9.6 million people to die only in the year 2018 [1].

An unbridled division of aberrant cells in any part of the body causes cancer. Some types of cancer precipitate meteoric cell growth, while others result in cells to grow and divide at a slower rate. Most of the body's cells have specific functions and fixed lifespans. The natural phenomenon of cell death is called apoptosis [2] which is beneficial for health. Older cells need to die for promoting the generation of newer cells in the body. Cancerous cells lack the constituents that help them to destruct themselves. As a consequence, they disseminate throughout the body and absorb the nutrients and oxygen which would be used by other healthy cells. Moreover cancerous cells can be proven to be detrimental by producing tumors, impede the immune system and cause other harms the body to function regularly.

Numerous factors can cause cancer, among which some are preventable, whereas others are not. Among the preventable factors, smoking, excess body weight, consumption of alcohol, poor nutrition etc. obstinate are mentionable. A little bit of consciousness can help us check these factors. However the unpreventable factors of cancer can be proved as more deleterious since they affect one's body quietly. Genetic factors can highly contribute towards growing cancer inside the body. Again with age, genetic code changes which can contribute to causing cancer. That's why among the unpreventable factors, age is one of the most significant one.

Data mining has been proven to be a boon in medical field. It has provided numerous aspects that can be used to analyze and process information that can be manipulated for disease recognition, prediction and prevention measures. Again, the cancer situation of Bangladesh is veritably alarming where 0.2 million people get affected by

cancer every year and 0.15 million people die from cancer [3]. Though eradicating cancer completely from a country is not possible, however, application of data mining can lessen the severity to an appreciable extent. Our main goal of this research is to determine the age group of people in Bangladesh who are more vulnerable to cancer.

The rest of the paper is organized as follows. Section II illustrates the review of related works. Again, our working procedure and experimental results and observations are represented in Section III and Section IV respectively. Finally, we have provided our concluding remarks along with the scope and directions of our future research in Section V.

## **1.2 Motivation**

Bangladesh, at 142 million individuals, is that the ninth most thickly settled country within the world. There are thirteen to fifteen 100000 cancer patients in Bangladesh, with regarding two lakh patients new diagnosed with cancer every year. Thus it's a serious threat to our country because it is increasing day by day.

But technology oriented research about cancer in Bangladesh is not enough. So we got the motivation to work on it in “Machine Learning”. By this research we would like to help people through creating awareness about cancer. It will also help in our medical sector.

## **1.3 Research Question**

Our main research question is about patient’s age. Age is the most important single risk factor of cancer. Risk increases gradually after age 50, and half of all cancers occur at age 66 and above. The median age at diagnosis is 61 years for breast cancer, 66 years for prostate cancer and 68 years for colorectal cancer and 70 years for lung cancer

## **1.4 Research Methodology**

We used two types of feature selection and three types of Classifications.

Feature Selection are:

1. Feature Importance Ranking Measure
2. Recursive Feature Elimination

Classifications are:

1. Logistic Regression
2. Support Vector Machine
3. Convolutional Neural Network

## **1.5 Research Objective**

- Cancer prediction in Bangladesh.
- People can know the possibility of cancer through Cause, Age, and Gender.
- As our goal to build a model which can predict cancer so people can take precautionary steps to protect themselves.

## **1.4 Report Layout**

This segment follows the parts of each progression that we utilized in our report in short.

Chapter 1: Introduction

Chapter 2: Background

Chapter 3: Research Methodology

Chapter 4: Result & Observation

Chapter 5: Conclusion and Future work

## **CHAPTER 2**

### **BACKGROUND**

#### **2.1 Review Works**

In 2002 Djavan and Remzi worked regarding early detection of prostate cancer implementing 2 artificial neural networks. The risk factors of this cancer are age, familial history of cancer, and quality [4]. Again in 2004 Shieu-Ming and Tian-Shyug conferred a hybrid breast cancer diagnostic model by group action artificial neural networks and multivariate adaptive regression splines. The carcinoma diagnostic tasks are performed on one FNAC dataset [5].

After 5 years In 2009 Murat and Cevdet developed an automatic diagnosing system for detection of breast cancer based on association rules (AR) and neural network (NN).The planned AR + NN system performance is compared with NN model. They additionally used Apriori rule in preliminaries stage and got accuracy 95.6% [6]. Later in 2010 Joshi and rana designed and developed a Brain Cancer Detection and classification system using Artificial Neural Network in magnetic resonance imaging pictures of various patients with Astrocytoma sort of brain tumors [7]. Also a Neuro Fuzzy Classifier had been developed to acknowledge different types of brain cancers. The dataset of MRI pictures was collected from Radiology Department of Tata Memorial Hospital.

Afterwards, in 2015 an integrated method has been worked out by Wan-Ting and Wei-Fan, which helps in sorting out the deviations among the symptoms demonstrated in the precedent cases of death caused from oral cancer. This method primarily combines the clustering and classification features of data mining. Carcinoma historical cases are well dealt with the help of Decision Tree and Artificial Neural Network, the results of which were differentiated with that of the Logistic Regression [8]. In the year 2018 Dejun and lu bestowed an unsupervised feature learning framework by integrating a principal component analysis and auto encoder neural network to spot completely different characteristics from organic phenomenon

profiles. They implemented AdaBoost algorithm to predict clinical outcomes in breast cancer [9].

Simultaneously Sanjay and Nair have applied the Naive bayes (NB) classifier algorithm beside Univariate selection, recursive elimination and a hybrid feature selection methodology for a correct detection of breast cancer. They got the most effective result for testing set in hybrid methodology exactness 0.84 and sensitivity 0.79 [10]. Agein Vikas and Saurabh have also applied Naive bayes, Radial Basis function Network, J48 on same carcinoma. They used the dataset provided by UCIrvine Machine Learning repository located in breast cancer Wisconsin sub-directory. In their proposed system Naive bayes performed accuracy of 97.36% that was the state of the art in 2018 [11].

On the same year Nasser and Samy developed for detecting the absence or presence of carcinoma in human body using Artificial Neural Network. Their model showed 96.67% accuracy to predict the presence of lung cancer. Whereas, Bashir and Khan used Decision Tree, Logistic Regression, Logistic Regression SVM, Nave Bayes and Random Forest algorithms along with feature selection techniques to improve the accuracy of prediction of heart disease [12]. They used an open source UCI data set and got the accuracy for Logistic Regression 82.56%, Logistic Regression (SVM) 84.85%

## **2.2 Bangladesh Perspective**

Bangladesh is a densely populated country. There are 13 to 15 lakh cancer patients in Bangladesh, with about two lakh patients newly diagnosed with cancer each year. As enough research is not available in this field, cancer prediction can be a big deal in computer science field in Bangladesh perspective.

## **2.3 Government Goals and Regulation**

Government goal should be decreasing cancer growing rate gradually in every year. Through regulation, education, and support programmers, governments will create an atmosphere during which tobacco use is reduced and peoples maintain smart levels of physical activity, healthy bodyweight, and smart nutrition. Cancer interference and

therefore the creation of a culture of health is a vital mission of government, on the far side that of the normal health-focused departments like health ministries; it's within the domain of governmental agencies concerned in environmental protection, activity safety, and transportation. Cancer interference and health promotion are within the realm of the board, the board of education, and therefore the board of health.

## CHAPTER 3

### RESEARCH METHODOLOGY

#### 3.1 Working procedure

In our research, we try to estimate a particular age group affected by cancer due to various reasons. To carry out this task, we work with real dataset collected from Jalalabad Ragib-Rabeya Medical College. Our dataset contains 108 patients' data with their age, gender, types of carcinoma and relevant causes of carcinoma and other vital information. We have segregated our overall working procedure into three distinct stages. The stages consist of preprocessing of data, feature selection and data classification respectively. The overall details has been illustrated in the figure 1.

Table-3.1: Range Of Values Related To the Dataset

Data Variable	Observed Values
Gender	Male, Female
Area	Biriyani bazar, Jawwa, Akhalia, Borolekha etc
Presence of Antineoplastic Agents	Yes, No
Presence of Hodgkin Lymphoma	Yes, No
Presence of Multiple Myeloma	Yes, No
Presence of Benzene Exposure	Yes, No
Presence of Genetic Factors	Yes, No
Identification of 1st Degree Relative	Yes, No
Due to Smoking	Yes, No
Diet High in Saturated Fat	Yes, No
Due to GERD	Yes, No
Due to Alcohol Abuse	Yes, No
Due to obesity	Yes, No
Due to Presence in Ancestors	Yes, No
Presence of Human T Cell	Yes, No
Presence of Epstain Barr Virus	Yes, No
Presence of Helicobacter Pylorii	Yes, No
Due to Industrial Hazard	Yes, No
Presence of Cirrhosis	Yes, No



<b>Data Variable</b>	<b>Observed Values</b>
Presence of Diabetes	Yes, No
Presence of Iron Storage Disease	Yes, No
Due to HIV Infection	Yes, No
Due to Gastroesophageal Reflux Disease	Yes, No
Due to Radiation to Head	Yes, No
Presence of Arsenic	Yes, No
Due to Advancing Age	Yes, No
Due to Intestinal Metaplasia Due	Yes, No
Risk with Increased Exposure of Radiation	Yes, No
Due to Chromosomal Abnormalities	Yes, No
Due to Possible Evolution from Normal Plasma Cell	Yes, No
Due to Inherited Gene Mutation	Yes, No
Due to Inflammatory Bowel Disease	Yes, No
Due to Smoked Foods	Yes, No
Due to Familial Adenomatous Polyposis	Yes, No
Due to Retinoblastoma	Yes, No
Due to Lymphedema	Yes, No
Due to Estrogen Exposure	Yes, No
Presence of Human Papilloma Virus	Yes, No
Due to Multiple Sexual Partners	Yes, No
Use of OCP	Yes, No
Due to Actinic Keratosis	Yes, No
Due to Actinic Cheilitis	Yes, No
Presence of Leukoplakia	Yes, No
Presence of Bowen Disease	Yes, No
Due to Epithelial Tumors	Yes, No
Due to Stromal Tumors	Yes, No
Due to Germ Cell Tumors	Yes, No
Diagnosis	Acute lymphoblastic leukemia, Bone carcinoma, Carcinoma of gastro-esophageal junction, carcinoma of ascending colon etc.
Age	10 to 19, 20 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, Above 70

### 3.1.1 Flow chart of working procedure

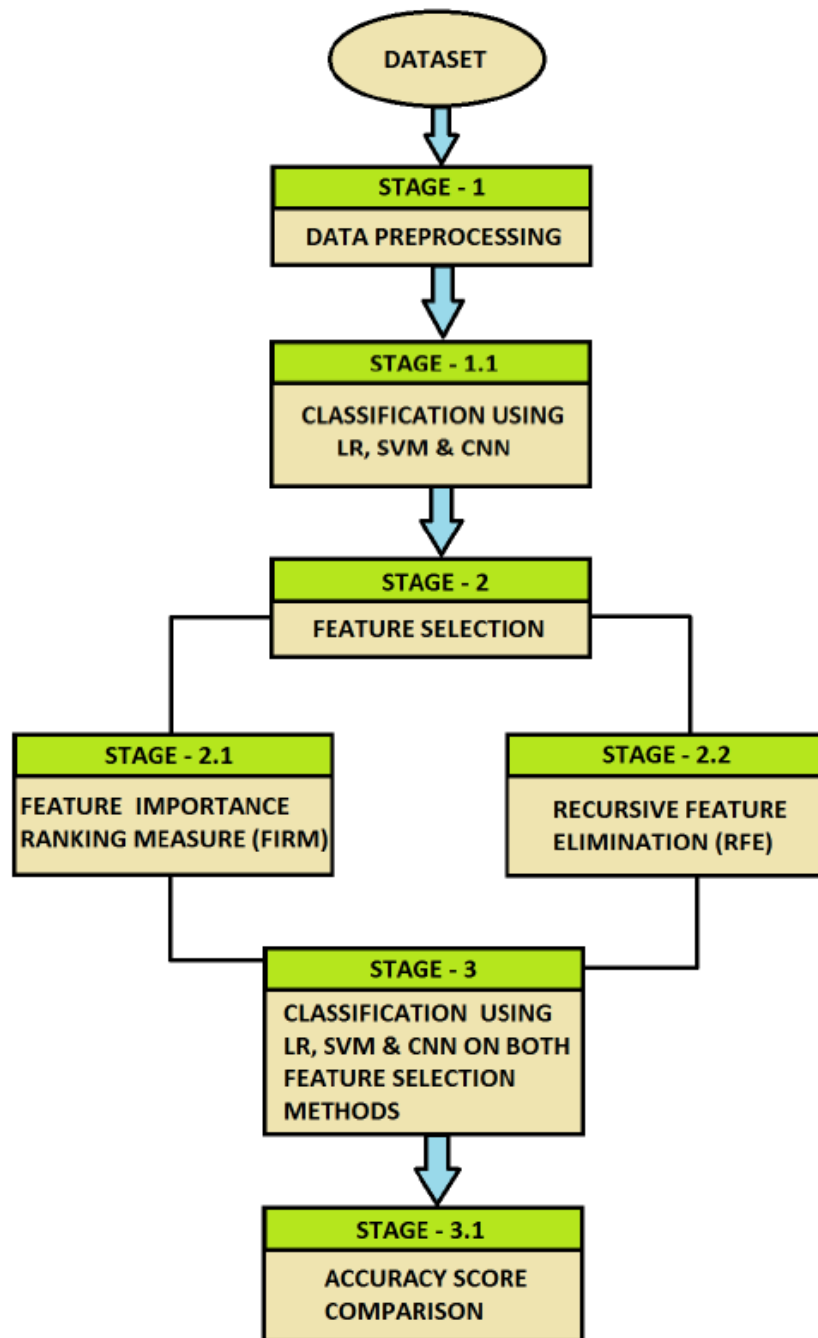


Figure-3.1: Stages of Working Procedure

### **3.1.2 Preprocessing of Data**

At the very outset of our working procedure, we arrange the dataset in such a way that it results in better classification result. To bolster this cause, we rearrange the distinct age into a particular age group, which is the output class of our dataset. The output class is categorized into seven distinct classes which is highlighted in table III.

### **3.1.3 Feature Selection**

In our research, we have employed two major feature selection approaches, to find out the major factors affecting the cause. These are: a) Feature Importance Ranking Measure (FIRM) and b) Recursive Feature Elimination (RFE).

#### **3.1.3.1 Feature Importance Ranking Measure**

Feature Importance Ranking Measure (FIRM) is a method, employed for feature selection by removing zero importance options. The importance of a feature is that the increase within the prediction error of the model once we have a tendency to permute the feature's values, which breaks the link between the feature and also the true outcome [13]. A feature is "important" if shuffling its values will increase the model error, as a result of during this case the model relied on the feature for the prediction. On the other hand, a feature is "unimportant" if shuffling its values leaves the model error unchanged, as a result of during this case the model neglected the feature for the prediction.

#### **3.1.3.2 Recursive Feature Elimination**

Recursive Feature Elimination (RFE) may be a feature choice technique that matches a model and removes the weakest feature (or options) until the desired range of features is reached. Features are graded by the model's attributes, and by recursively eliminating a little range of options per loop, RFE tries to eliminate the dependencies and collinearity which will exist within the model.

RFE needs such a variety of features to stay, but it's usually not proverbial prior to what percentage options square measure valid. To seek out the optimum variety of features, cross-validation is employed with RFE to attain a totally different feature

subsets and choose the simplest rating assortment of features. The RFECV visualizer plots the range of options within the model, in conjunction with their cross-validated check score and variability and visualizes the chosen number of features.

### **3.1.4 Classification**

In our working methodology, we have employed two baseline classifiers, Logistic Regression (LR) and Support Vector Machine (SVM), along with Convolutional Neural Network (CNN). We have applied these classifiers both on the pre-feature selection stage and post-feature selection stage.

#### **3.1.4.1 Logistic Regression**

Logistic regression is that the acceptable multivariate analysis to conduct once the dependent variable is divided (binary) [14]. Like all regression analyses, logistic regression may be a prognostic analysis. Logistic regression is employed to explain information and to elucidate the connection between one dependent binary variable and one or a lot of nominal, ordinal, interval or ratio-level freelance variables. The logistic regression model uses the logistical to perform to squeeze the output of a linear equation between zero and one [15]. The diagrammatic approach of logistic regression is illustrated in the figure 2. The logistic function is outlined as:

$$logstic(\eta) = \frac{1}{1 + \exp(-\eta)}$$

The step from linear regression to logistic regression is quite simple. Within the regression model, we've got modeled the link between outcome and features with a linear equation as depicted in equation 2.

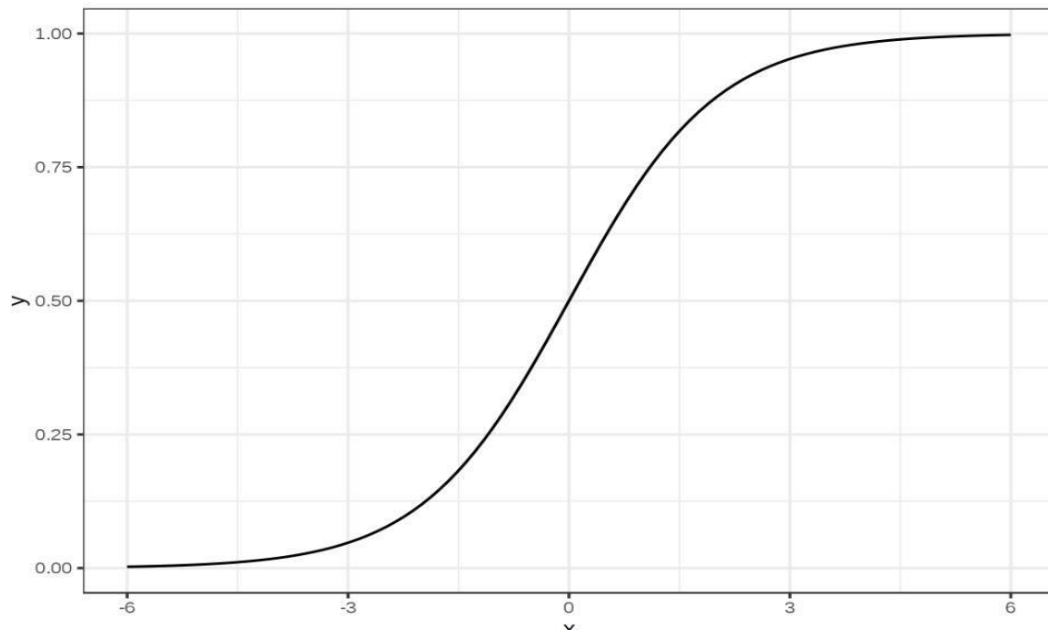


Figure-3.2: Logistic Regression.

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_p x_p^{(i)} \quad (2)$$

For classification, we tend to like chances between zero and one, thus we tend to wrap the correct aspect of the equation into the logistic function. This forces the output to assume solely values between zero and one.

$$P(y^{(i)} = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_p x_p^{(i)}))} \quad (3)$$

### 3.1.4.2 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally outlined by a separating hyperplane [16]. In different words, given labeled training information (supervised learning), the formula outputs the best hyperplane that sorts new examples. In two dimensional area, this hyperplane is a line dividing a plane into two components wherever in every category lay in either aspect. Hyperplanes are call boundaries that facilitate classify the information points. Data points will be attributed to totally different categories that falling on either aspect of the hyperplane. Support vectors are information points that are nearer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we have a tendency to maximize the margin of the classifier. Deleting the support vectors can modify the position of the hyperplane.

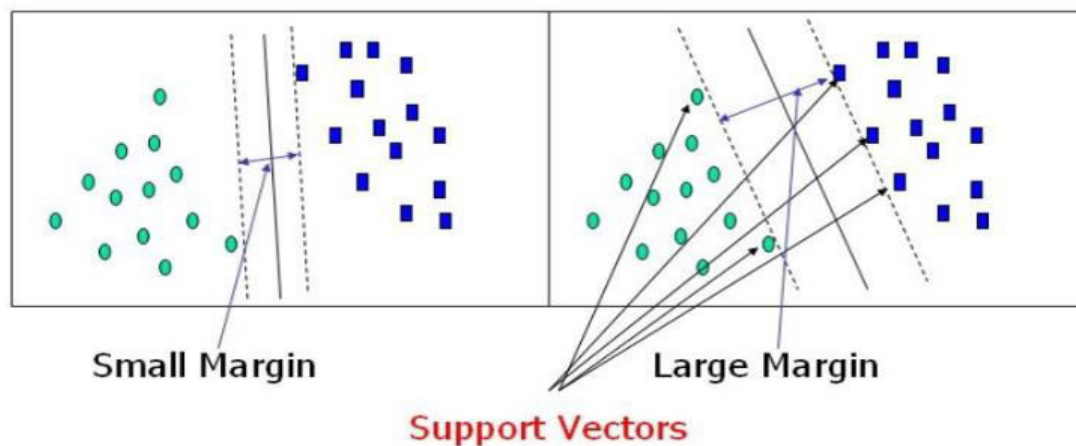


Figure-3.3: Support Vector Machine.

### 3.1.4.3 Convolutional Neural Network

A specific reasonably such a deep neural network is the convolutional network, which is often noted as CNN or ConvNet. It's a deep, feed forward artificial neural network. Keep in mind that feed-forward neural networks are also known as multi-layer perceptron's (MLPs) [17]. Convolutional neural networks are one of the foremost powerful innovations within the field of computer vision. They have performed a great deal higher than ancient computer vision and have made state-of-the-art results. The hidden layers of a CNN usually include convolutional layers, pooling layers, fully

connected layers, and normalization layers. Here it merely means rather than mistreatment the conventional activation functions outlined higher than, convolution and pooling functions are used as activation functions.

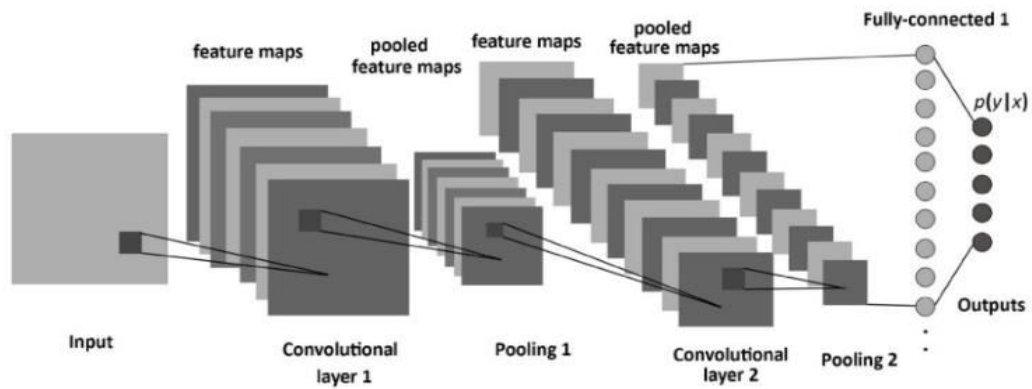


Figure-3.4: Convolutional Neural Network.

## CHAPTER 4

### RESULT AND OBSERVATIONS

#### 4.1 Experimental Analysis

We have investigated our experimental results with most preferably used metrics for performance evaluation of classifiers. These are: Precision, Recall, F1 Score, Support and Accuracy. The mathematical of all these metrics are stated as follows:

$$\textit{Precision} = \frac{\textit{truepositive}}{\textit{truepositive} + \textit{falsepositive}} \dots\dots\dots (4)$$

$$\textit{Recall} = \frac{\textit{truepositive}}{\textit{truepositive} + \textit{falsenegative}} \dots\dots\dots (5)$$

$$F = 2 \times \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}} \dots\dots\dots (6)$$

$$\textit{Accuracy} = \frac{\textit{truepositive} + \textit{truenegative}}{(\textit{truepositive} + \textit{truenegative} + \textit{falsepositive} + \textit{falsenegative})} \dots\dots\dots (7)$$

#### 4.2 Comparative Analysis

At first, we run Logistic Regression (LR), Support Vector Machine (SVM) and Convolutional Neural Networks (CNN) on our dataset before applying any feature selection technique. We observe that LR has performed better under the circumstances, rather than SVM and CNN. A detailed illustration of the result analysis, before feature selection stage, has been demonstrated in table II. After



applying FIRM as feature selection method, we get higher accuracy for CNN rather than LR and SVM.

Table-4.1: Classification Report before Feature Selection

Classifiers	Class	Precision	Recall	F1-score	Support	Accuracy
LR	0	0.00	0.00	0.00	0	59.09
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.50	0.50	0.50	2	
	4	0.64	0.88	0.74	8	
	5	0.50	0.57	0.53	7	
	6	1.00	0.50	0.67	2	
SVM	0	0.00	0.00	0.00	0	45.45
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.00	0.00	0.00	2	
	4	0.50	0.75	0.60	8	
	5	0.57	0.57	0.57	7	
	6	0.00	0.00	0.00	2	
CNN	0	0.00	0.00	0.00	0	50.00
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.00	0.00	0.00	2	
	4	0.58	0.88	0.70	8	
	5	0.57	0.57	0.57	7	
	6	0.00	0.00	0.00	2	

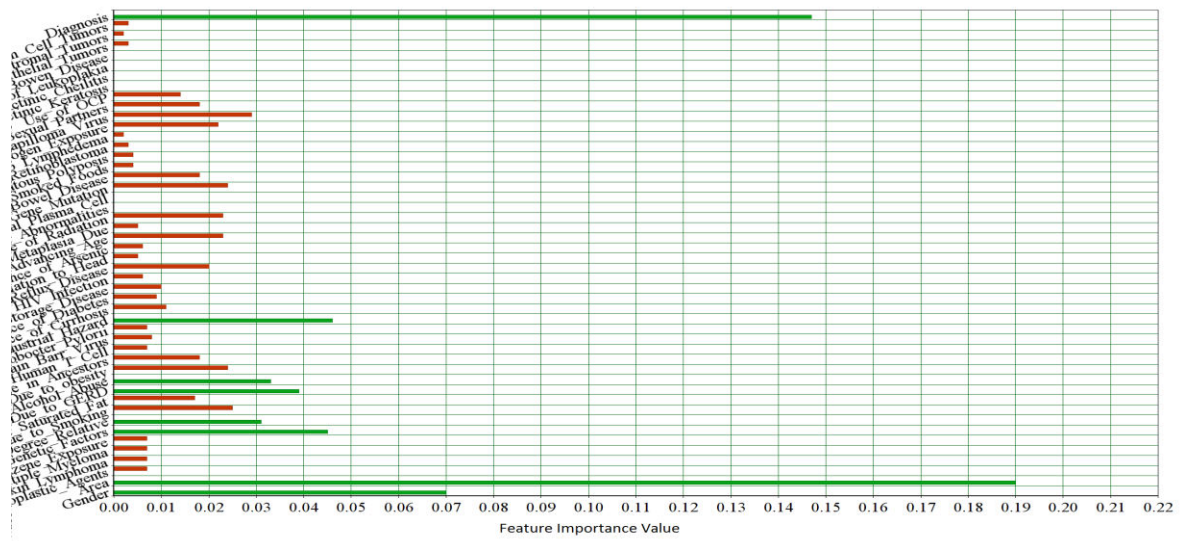


Figure-4.1: Features Selected after Applying FIRM.

Features extracted after FIRM application has been illustrated in the figure 5 and detailed illustration of the result is shown in the table III.

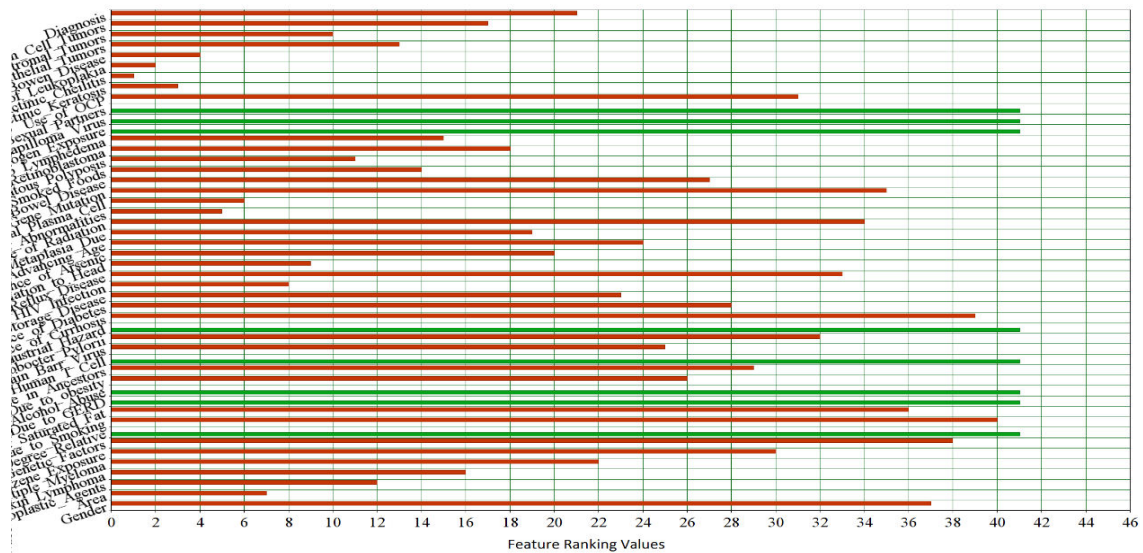


Figure-4.2: Features Selected after Applying RFE.

Finally after applying RFE technique, all the classifiers have performed better than that of the FIRM technique, as

Table-4.2: Classification Report after FIRM

Classifiers	Class	Precision	Recall	F1-score	Support	Accuracy
LR	0	0.00	0.00	0.00	0	45.45
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.00	0.00	0.00	2	
	4	0.41	0.88	0.56	8	
	5	0.60	0.43	0.50	7	
	6	0.00	0.00	0.00	2	
SVM	0	0.00	0.00	0.00	0	40.91
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.00	0.00	0.00	2	
	4	0.50	0.62	0.50	8	
	5	0.50	0.57	0.53	7	
	6	0.00	0.00	0.00	2	
CNN	0	0.00	0.00	0.00	0	50.00
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.00	0.00	0.00	2	
	4	0.44	0.88	0.58	8	
	5	0.80	0.57	0.67	7	
	6	0.00	0.00	0.00	2	

Table-4.3: Classification Report after RFE

Classifiers	Class	Precision	Recall	F1-score	Support	Accuracy
LR	0	0.00	0.00	0.00	0	54.55
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.50	0.50	0.50	2	
	4	0.64	0.88	0.74	8	
	5	0.60	0.43	0.50	7	
	6	0.25	0.50	0.33	2	
SVM	0	0.00	0.00	0.00	0	45.45
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.00	0.00	0.00	2	
	4	0.64	0.88	0.74	8	
	5	0.36	0.57	0.44	7	
	6	0.00	0.00	0.00	2	
CNN	0	0.00	0.00	0.00	0	59.09
	1	0.00	0.00	0.00	2	
	2	0.00	0.00	0.00	1	
	3	0.50	0.50	0.50	2	
	4	0.70	0.88	0.78	8	
	5	0.67	0.57	0.62	7	
	6	0.25	0.50	0.33	2	

Shown in table IV. As a result, the features, which have been extracted using RFE, are more Relevant to bolster the cause. Here, the extracted features are: "Identification of 1st Degree Relative", "Due to GERD", "Due to Alcohol Abuse", "Presence of Human T Cell", "Due to Industrial Hazard", "Due to Estrogen Exposure", "Presence of Human Papilloma Virus", "Due To Multiple Sexual Partners".

## CHAPTER 5

### CONCLUSION AND FUTURE WORK

#### 5.1 Conclusion

Bangladesh, having approximately 1.5 million cancer patients, is in a daunting situation where the number of affected people is increasing day by day. Cancer basically starts with genetic changes in a single cell. Yet if it is left untreated, then it expands and invades other parts of the body and ultimately may also cause death. If awareness is not made on this, then the affect would be more terrible in future. That is why it's essential to understand the symptoms of cancer at an early stage.

Moreover, death from cancer is inescapable if diagnosed after a certain stage. However, if people are made aware of indications of cancer for distinct ages, the number of people dying from cancer could be lessened.

Through this research, we have attempted to determine the symptoms indicating various types of cancer for different age groups in Bangladesh by using data mining techniques. To do so, we have applied two feature selection techniques: 1. Feature Importance and 2. Recursive Feature Elimination along with three types of classification algorithms: 1. Logistic Regression, 2. Support Vector Machine, 3. Convolutional Neural Network. Among these Convolutional Neural Network gave us better accuracy.

#### 5.2 Future Work

Our future plan is to explore the regional distinction effects of cancer situation and intensity. Moreover, it was difficult for us to manage real time dataset. So, in future we aim to evaluate the accuracy of the model on a bigger dataset.

## REFERENCES

- [1] Cancer Treatment in Bangladesh, available at <<<https://www.thedailystar.net/opinion/perspective/news/cancer-treatmentbangladesh-still-long-way-go-1696912>>>, last accessed at 2019-08-31 at 10.17 PM
- [2] Cancer Overview causes, available at <<<https://www.medicalnewstoday.com/articles/323648.php>>>, last accessed on 2019-08-31 at 10.51 PM.
- [3] Cancer Patients in Bangladesh, available at <<<https://www.thedailystar.net/country/news/over15-lakh-cancer-patients-bangladesh-who-1696903>>>, last accessed on 2019-08-31 at 11.42 PM.
- [4] B. Djavan, M. Remzi, A. Zlotta, C. Seitz, P. Snow, and M. Marberger, "Novel artificial neural network for early detection of prostate cancer," *Journal of Clinical Oncology*, vol. 20, no. 4, pp. 921–929, 2002.
- [5] S.-M. Chou, T.-S. Lee, Y. E. Shao, and I.-F. Chen, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines," *Expert systems with applications*, vol. 27, no. 1, pp. 133–142, 2004.
- [6] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert systems with Applications*, vol. 36, no. 2, pp. 3465–3469, 2009.
- [7] D. M. Joshi, N. Rana, and V. Misra, "Classification of brain cancer using artificial neural network," in *2010 2nd International Conference on Electronic Computer Technology*. IEEE, 2010, pp. 112–116.
- [8] W.-T. Tseng, W.-F. Chiang, S.-Y. Liu, J. Roan, and C.-N. Lin, "The application of data mining techniques to oral cancer prognosis," *Journal of medical systems*, vol. 39, no. 5, p. 59, 2015.
- [9] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer," *IEEE Access*, vol. 6, pp. 28 936–28 944, 2018.
- [10] A. Sanjay, H. V. Nair, S. Murali, and K. Krishnaveni, "A data mining model to predict breast cancer using improved feature selection method on real time data," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018, pp. 2437–2440.
- [11] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018.
- [12] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving heart disease prediction using feature selection approaches," in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. IEEE, 2019, pp. 619–623.
- [13] Feature Importance, available at <<<https://christophm.github.io/interpretablemlbook/featureimportance.html#theory-3>>>, last accessed on 2019-08-31 at 11.57 PM.
- [14] What is Logistic Regression, available at <<<https://www.statisticssolutions.com/whatis-logistic-regression/>>>, last accessed on 2019-09-1 at 12.45 AM.

[15] Logistic Regression, available at <<https://christophm.github.io/interpretable-mlbook/logistic.html>>>, last accessed on 2019-09-1 at 01.15 AM.

[16] Support Vector Machine, available at <<<https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms934a444fca47>>>, last accessed on 2019-09-1 at 01.36 AM.

[17] Data Camp Convolutional Neural Network, available at <<[https://www.datacamp.com/community/tutorials/convolutionalneuralnetworkspython?utm\\_source=adwordsppc&utm\\_campaignid=1455363063&utm\\_medium=6583631748&utm\\_device=c&utm\\_keyword=&utm\\_matchtype=b&utm\\_network=g&utm\\_adposition=1t1&utm\\_creative=332602034358&utm\\_targetid=aud299261629574dsa47406581915&utm\\_locationinterests=&utm\\_locationphysicalms=9074047&utm\\_glid=Cj0KCQwhdTqBRDNARIsABsOI94pzD4FL1nStJmYuRNawI4IgKGZDF8GooyN6cQSFzeMonpx6IXb8aAmUcEALw\\_wcB](https://www.datacamp.com/community/tutorials/convolutionalneuralnetworkspython?utm_source=adwordsppc&utm_campaignid=1455363063&utm_medium=6583631748&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adposition=1t1&utm_creative=332602034358&utm_targetid=aud299261629574dsa47406581915&utm_locationinterests=&utm_locationphysicalms=9074047&utm_glid=Cj0KCQwhdTqBRDNARIsABsOI94pzD4FL1nStJmYuRNawI4IgKGZDF8GooyN6cQSFzeMonpx6IXb8aAmUcEALw_wcB)>>, last accessed on 2019-09-1 at 01.45 AM.