# Cancer Prediction using Different Classification Algorithms

By

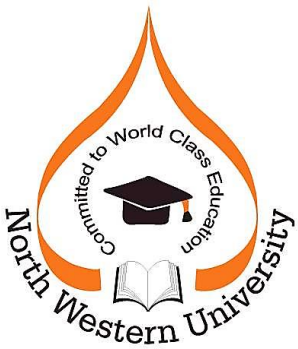
Md. Ariful Islam Bhuiyan
Student ID: 20152004010

Md. Taufik Akunjee
Student ID: 20152024010

Md. Hashikul Islam
Student ID: 20152019010

&

Rafiqul Islam
Student ID: 20152010010




SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

AT

NORTH WESTERN UNIVERSITY, KHULNA-9000

BANGLADESH

SEPTEMBER, 2019

# NORTH WESTERN UNIVERSITY, KHULNA

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

The undersigned hereby certify that have read and recommended for acceptance a thesis entitled "Cancer Prediction using Different Classification Algorithms" by Md. Ariful Islam Bhuiyan, Md. Taufik Akunjee, Md. Hashikul Islam and Rafiqul Islam in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

1. **Research Supervisor**

10/10/19

**Tajul Islam**
Senior Lecturer
Department of Computer Science and Engineering
North Western University, Khulna

2. **Second Examiner**

10/10/19

**Romana Rahman Ema**
Senior Lecturer
Department of Computer Science and Engineering
North Western University, Khulna

3. **Head of the Department**

10/10/19

**Tajul Islam**
Senior Lecturer
Department of Computer Science and Engineering
North Western University, Khulna

# NORTH WESTERN UNIVERSITY, KHULNA

Date: September, 2019

| | | |
|---|---|---|
| **Authors** | : | Md. Ariful Islam Bhuiyan, Md. Taufik Akunjee, Md. Hashikul Islam, Rafiqul Islam |
| **Title** | : | Cancer Prediction using Different Classification Algorithms |
| **Department** | : | Computer Science and Engineering |
| **Degree** | : | Bachelor of Science in Computer Science and Engineering |

Permission is herewith granted to North Western University to circulate and to have copied for-commercial purpose, at its discretion, the above title upon the request of individuals or institutions.

Md. Ariful Islam Bhuiyan

Md. Taufik Akunjee

Md. Hashikul Islam

Rafiqul Islam

# Abstract

At present in Bangladesh, there are 152 million people lives here, unneeded to announce that is the ninth largely populous country in the world. There are 15 to 16 lakh cancer suffered people in Bangladesh, approximately three lakh suffers recently gathered with cancer each distinct. There are over 100 types of cancer. Predicting cancer takes advantage of a crucial position in medical science. Naive Bayes, J48, K-star and K-Nearest neighbor algorithm have used this paper for predicting cancer disease. Naive Bayes runs easily and can easily handle large dataset. It works very fast. Counting missing values is a vital feature of J48 algorithm. In Weka tool, C4.5 algorithm is often implemented using through J48. K-star operates always on-the-fly, that means it's no need to be explicitly available and stored in main memory. K-Nearest neighbor is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. Weka tool is used to measure the accuracy of the cancer disease dataset including 20 types of cancer. In Naive Bayes the accuracy is 98.8%, J48 accuracy is 98.6%, K-star accuracy is 98.9% and K-Nearest neighbor accuracy is 98.8%. In K-Star accuracy is the best performance among all those four algorithms.

**Keywords:** Different Cancers, Classification, Naive Bayes, J48, K-star, K-Nearest neighbor, Prediction.

# Acknowledgments

At first, we are grateful to almighty Allah for giving us strength, patience and intelligence to perform our Thesis properly.

We would like to express our sincere gratitude to our supervisor Tajul Islam, Senior Lecturer, Department of Computer Science and Engineering, North Western University, Khulna for his valuable suggestions and proper guidance. Actually his wonderful contribution has motivated us to reach the goal.

We are grateful to all of our respective teachers for their suggestions, researches and weka tool for providing us necessary materials.

Our parents also give us support and encouragement to fulfill our graduation. Last but not the least, we are grateful to the Dr. Mrinal Kranti Sarkar and stuffs of Khulna Medical for helping us to collect data.

# Dedication

This study is wholeheartedly dedicated to our beloved parents, who have been our source of inspiration and gave us strength when we thought of giving up, who continually provide their moral, spiritual, emotional and financial support.

To our brothers, sisters, relatives, mentor, friends and classmates who shared their words of advice and encouragement to finish this study.

And lastly, we dedicated this book to the Almighty Allah, thank you for the guidance, strength, power of mind, protection and skills and for giving us a healthy life. All of these, we offer to you.

# Table of Contents

# Table Name

# Figure Name

# Chapter 1

## Introduction

### 1.1 Background

Cancer is the name given to a collection of related diseases. In all types of cancer, some of the body's cells begin to divide without stopping and spread into surrounding tissues. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Normally, human cells grow and divide to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place.

When cancer develops, however, this orderly process breaks down. As cells become more and more abnormal, old or damaged cells survive when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths called tumors.

Many cancers form solid tumors, which are masses of tissue. Cancers of the blood, such as leukemias, generally do not form solid tumors. Cancerous tumors are malignant, which means they can spread into, or invade, nearby tissues. In addition, as these tumors grow, some cancer cells can break off and travel to distant places in the body through the blood or the lymph system and form new tumors far from the original tumor.

Cancer cells differ from normal cells in many ways that allow them to grow out of control and become invasive. One important difference is that cancer cells are less specialized than normal cells. That is, whereas normal cells mature into very distinct cell types with specific functions, cancer cells do not. This is one reason that, unlike normal cells, cancer cells continue to divide without stopping.

Cancer cells are also often able to evade the immune system, a network of organs, tissues and specialized cells that protects the body from infections and other conditions. Although the immune system normally removes damaged or abnormal cells from the body, some cancer cells are able to "hide" from the immune system. Tumors can also use the immune system to stay alive and grow. For example, with the help of certain immune system cells that normally prevent a runaway immune response, cancer cells can actually keep the immune system from killing cancer cells.

Cancer is a genetic disease, that is, it is caused by changes to genes that control the way our cells function, especially how they grow and divide. Genetic changes that cause cancer can be inherited from our parents. They can also arise during a person's lifetime as a result of errors that occur as cells divide or because of damage to DNA caused by certain environmental exposures. Cancer causing environmental exposures include substances, such as the chemicals in tobacco smoke, and radiation, such as ultraviolet rays from the sun [1].

Cancer arises from the transformation of normal cells into tumor cells in a multistage process that generally process that generally progresses from a pre-cancerous lesion to a malignant tumour. These changes are the result of the interaction between a person's genetic factors.

WHO, through its cancer research agency, International Agency for Research on Cancer (IARC), maintains a classification of cancer-causing agents [2].

There are 13 to 15 lakh cancer patients in Bangladesh, with about two lakh patients newly diagnosed with cancer each year. Lung cancer and mouth-oropharynx cancer rank as the top two prevalent cancers in males. Other types of cancers are esophagus cancer and stomach cancer. In women, cancer cervix uteri and breast cancer are most prevalent. Other cancer types, which affect women, are mouth and oropharynx cancer, lung cancer, and esophagus cancer. There are around 150 qualified clinical oncologists and 16 pediatric oncologists working in the different parts of the country. Regular cancer treatment is available in 19 hospitals and 465 hospital beds are attached as indoor or day care facilities for chemotherapy in the oncology/radiotherapy departments. There are about 15 linear accelerators, 12 Co-60 teletherapy and 12 brachytherapy units currently available. Approximately, 56 cancer chemotherapeutic agents are obtainable in Bangladesh. Research facilities are available at tertiary care centers and a few multi country collaborative research activities are ongoing. Bangladesh has a unique National Cancer Control Strategy and Plan of Action 2009-2015 formulated with the assistance of WHO with an objective to develop and implement continuum of cancer care through a comprehensive cancer control program. Preventive measures taken to reduce the incidence of cancer include reduced tobacco smoking, change of dietary habit and reduced food adulteration, ensuring reproductive hygiene, increased physical activity, and reduced occupational hazard. Awareness buildup and media campaign are going on by organizing the general people, opinion leaders of the society, and boy and girls scout. Training of general physicians on cancer warning signs and setup of early cancer detection centers at each medical college and district levels are ongoing. Beside these, some other major cancer programs have taken place for early detection of breast, cervical and oral cancer by Bangladesh Government and NGOs such as ICDDR'B, BRAC, Ahsania Mission Cancer Hospital, BSMMU, Bangladesh Cancer Society, Ashic Foundation, AK Khan Healthcare Trust, CANSUP, Oncology club etc. Piloting of cervical cancer vaccination has recently been completed. Improving the cancer scenario overnight is not an easy task but policy makers may become interested and push this agenda forward, if the huge health impact and economic loss caused by cancer become evident to them. Besides, Bangladesh has accepted reduction of cancer morbidity and mortality targets set by United Nations and World Health Organization as a part of global non-communicable disease prevention agreement [5].

The WHO states that there are four key components to cancer control: cancer prevention, early detection, diagnosis and treatment and palliation. Developing countries face major challenges in each of these four areas. However, before proceeding to discuss these issues in more detail, it is important to first establish a definition for developing countries. Although the term 'developing country' is frequently used by leading international development organizations including UN agencies, the World Bank, World Trade Organization and WHO, there is no agreed single definition on what constitutes a developing and developed countries-some organizations determine the boundary between developed and developing country on the basis of economic indicators like gross domestic product (GDP) or gross national income (GNI), whilst the UN and other organizations refer to the human development index (HDI). The

terminology also does not recognize that there are often widely differing standards of living both between and within the countries defined as developing countries [3].

According to the World Health Organization, cancer is the second leading cause of death globally and a staggering 9.6 million people died of cancer just in 2018. To raise awareness of this fatal disease, and encourage its prevention, diagnosis and treatment, the world today is observing “World Cancer Day” with a range of activities. Different types of organizations and hospitals are also observing this day in Bangladesh with conferences, seminars and a range of awareness-raising campaigns. However, these activities probably carry little meaning for the thousands of Bangladeshi cancer patients and their family members who have suffered financially, physically, and mentally while continuing the treatment in Bangladesh.

The overwhelming treatment cost, wrong diagnosis, faulty treatment plan and shortage of trained doctors and treatment facilities have severely reduced Bangladesh's capacity to ensure proper treatment for its growing cancer patients. According to the 2018 report by the International Agency for Research on Cancer, every year an estimated 1.5 lakh people develop cancer in Bangladesh. However, there is only one functioning palliative care under government management at Bangabandhu Sheikh Mujib Medical University (BSMMU). There are only four specialized cancer hospitals in the country. Outside Dhaka, there is only one functional radiotherapy facility, at Chittagong Medical College Hospital, in operation for around three months now.

Due to absence of data, the International Agency for Research on Cancer published its report on Bangladesh's cancer scenario based on cancer registry data available in the neighbouring countries. Bangladesh also does not have a national protocol for treating cancer patients.

Actually, we don't even know the exact number of people suffering from cancer in the country as we don't have any population-based data on the prevalence of cancer. How many patients contract cancer each year? How many patients die of cancer every year? What are the most frequent cancers among Bangladeshis? How many people cannot access treatment? The NICRH, Bangladesh's apex cancer institute, or any of the country's medical institutes have no answers to these questions [4].

Cancer survival rates or survival statistics tell the percentage of people who survive a certain type of cancer for a specific amount of time. Cancer statistics often use an overall five-year survival rate.

Survival rates are usually given in percentages. For instance, the overall five-year survival rate for bladder cancer is 78 percent. That means that of all people who have bladder cancer, 78 of every 100 are living five years after diagnosis. Conversely, 22 out of every 100 are dead within five years of a bladder cancer diagnosis.

Cancer survival rates are based on research from information gathered on hundreds or thousands of people with a specific cancer. An overall survival rate includes people of all ages and health conditions who have been diagnosed with cancer, including those diagnosed very early and those diagnosed very late.

Doctor may be able to give more specific statistics based on stage of cancer. For instance, 56 percent, or a little more than half, of people diagnosed with early-stage lung cancer live for at least five years after diagnosis. The five-year survival rate for people diagnosed with late-stage lung cancer that has spread (metastasized) to other areas of the body is 5 percent [6].

Our body is made up of 100 million million cells. Cancer can start when just one of them begins to grow and multiply too much. The result is a growth called a tumour. Benign tumours are localized growths - they only cause problems if they put pressure on nearby tissues, such as the brain. Much more serious are malignant tumours, which invade the surrounding body tissues. Some malignant tumours also spread throughout the body via the bloodstream: a process called metastasis.

Cancer is abnormal cell growth, which is the result of damage - mutation - to certain crucial genes, the cell's instructions for making the proteins it needs to survive, grow and multiply. Many factors can affect the chances of the gene damage that may eventually lead to cancer, including cigarette smoke and other chemicals, a poor diet, ultraviolet radiation that causes sunburn, as well as some infections. Chemicals that cause cancer are called carcinogens.

Regular screening already detects some cancers in their early stages - for example smear tests for cervical cancer and mammograms for breast cancer. In the future, a simple blood or urine test could detect many other cancers very early on. Lifestyle changes or medicines could then help prevent serious problems later in life. Couples who have a history of inherited cancer are already screening potential embryos to ensure they have a child without the high-risk gene.

Currently, there are three major ways of treating cancer: surgery, radiotherapy and chemotherapy. Surgery, where the surgeon cuts out the diseased tissue, is the most effective treatment in the early stages of some types of cancer. Radiotherapy uses a beam of radiation to kill cancer cells. Doctors can do this with increasing accuracy as imaging techniques improve. Chemotherapy uses medicines that kill any cells that are multiplying rapidly. Cancer cells do this but so do other body cells like hair cells, which are also killed, resulting in hair loss during treatment. Scientists are developing more effective treatments for cancer that may become more widely available in the next few years. Some treatments block the action of damaged genes in cancer cells. Others aim to starve the tumour cells of the blood supply essential for their growth. Still others aim to direct the body's own defences to fight the disease - immune therapies. Researchers are also working on ways to stop cancer spreading and are even creating viruses that will kill only cancer cells [7].

Cancer can be a major cause of poverty. This may be due either to the costs of treating and managing the illness as well as its impact upon people's ability to work. This is a concern that particularly affects countries that lack comprehensive social health insurance systems and other types of social safety nets. The ACTION study is a longitudinal cohort study of 10,000 hospital patients with a first time diagnosis of cancer. It aims to assess the impact of cancer on the economic circumstances of patients and their households, patients' quality of life, costs of treatment and survival. Patients will be followed throughout the first year after their cancer diagnosis, with interviews conducted at baseline (after diagnosis), three and 12 months. A cross-section of public and private hospitals as well as cancer centers across eight member

countries of the Association of Southeast Asian Nations (ASEAN) will invite patients to participate. The primary outcome is incidence of financial catastrophe following treatment for cancer, defined as out-of-pocket health care expenditure at 12 months exceeding 30% of household income. Secondary outcomes include illness induced poverty, quality of life, psychological distress, economic hardship, survival and disease status. The findings can raise awareness of the extent of the cancer problem in South East Asia and its breadth in terms of its implications for households and the communities in which cancer patients live, identify priorities for further research and catalyze political action to put in place effective cancer control policies [8].

## 1.2 Motivation

Data mining technique has become a fundamental methodology for computing applications in medical informatics and is increasingly very rapidly in the medical field due to its success in the classification and prediction algorithms that helps doctors in Decision making. Data mining applications progressing and its implications are being manifested in the medical science area and others such as health-care organizations, epidemiology, patient caring and monitoring systems and different types of identification of unknown levels. Different types of algorithms are associated with data mining technique and having significantly helped to realize medical data more definitely, by determining pathological data from usual data, for assisting decision making as well as visualization and identification of hidden complex relation connecting typical characteristics of various patients groups. It goes without saying that it is quiet impossible to get a paper which is associated with cancer diseases prediction. The important thing of this paper is discussed about 20 types cancer prediction. Still now there are various types of cancers which is unknown to us. Such changes may be due to chance or to exposure to a cancer causing substance. The substances that cause cancer are called carcinogens. A carcinogen may be a chemical substance, such as certain molecules in tobacco smoke. The cause of cancer may be environmental agents, viral or genetic factors. Needless to say maximum people want to know the initial symptoms of cancer so that they can take proper initiate to prevent it. Actually the main reason of this paper is to make people conscious about cancer disease. Here Weka abbreviation have been used for Waikato Environment Knowledge for Analysis, this is open source data mining program, was newly developed at Waikato University New Zealand, and it is licensed under the GNU General Public License. There well known algorithm have been used here that is Naive Bayes, J48 algorithm, K-Star and K-Nearest neighbor. The dataset is verified by the doctors and it contains the data of undergoing cancers patients. 10-fold cross validation is used. Here we try to find out the accurate cancer what the patient had and hopefully this data set and the overall discussion helps the doctor to find out the actual cancer and if the patients want to know about the details of his cancer then the patient know this through this paper. We also use weka to find out the accuracy and other data attributes.

## 1.3 Contribution

In the history decade, data mining changes the discipline of information science, which inquires the properties of information and the process and techniques manipulated in the accession, calculation, dissemination and design of evidence. There is a vast territory of data mining system. It has been used with success in the information science. The World Health Organization (WHO) declared that 150,781 new cancers have been attacked in Bangladesh [9]. The prevalence is increasing at an alarming rate in a developing country like Bangladesh in current years (Ferly et al. 2010). Therefore, the early diagnosis of cancer is explicit but the diagnosis is expensive in the rising countries.

## 1.4 Thesis Organization

The remainder of the thesis is organized as the following:

### Chapter 2 (Overview):

It represents Overview of data mining, characteristics of data mining, Overview of machine learning and also its characteristics, missing value in data and dealing with it. And the last segment is data mining VS machine learning.

### Chapter 3 (Related Work):

It filled with various existing works on these procedures, how they implement, advantages and disadvantages.

### Chapter 4 (Methodology):

This phase discussed data collection, data organization, data features, data statement. It also discussed system design and machine learning algorithm.

### Chapter 5 (Implementation and Results):

In this chapter implementation and the result is discussed.

### Chapter 6 (Discussion):

The discussion part is the last chapter of this thesis. It concluded with the conclusion, limitations and future works.

# Chapter 2

## Overview

Cancer is a collection of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. These contrast with benign tumors, which do not spread. Possible signs and symptoms include a lump, abnormal bleeding, prolonged cough, unexplained weight loss, and a change in bowel movements. While these symptoms may indicate cancer, they can also have other causes. Over 100 types of cancers affect humans. But we have discussed 20 types of cancer here.

### 2.1 Overview of Data Mining:

Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data. The essential difference between the data mining and the traditional data analysis (such as query, reporting and on-line application of analysis) is that the data mining is to mine information and discover knowledge on the premise of no clear assumption [25].

### 2.1.1 Characteristics of Data Mining

Data mining management is a simple type of data gathering procedure wherein all the pertinent data experiences a type of recognizable proof procedure. What's more, in the long run toward the finish of this procedure, one can decide all the normal for the information mining process.

1. **Enlarged products of information:** While previous days, the learning mining structure can be determined with the support of their clients and clients, though in the existing data, one can obtain any number of data without the assistance of those customers. Moreover, following this sort of alteration in the opening framework, it likewise included one more issue and that is huge amounts of work. With the assistance of these data innovation, one can gain countless data with no additional weight or inconvenience.

2. **Presents insufficient learning:** The preponderance of the overall community present fragmented data regarding themselves in the destiny from the study started including the support of learning mining structures. Along these lines, people neglect the judgment of their data and that is the reason they give inadequate data about themselves in those reviews directed to support the mining frameworks.

3. **Entangled learning building**: Data mining is a structure wherein which all the data is assembled and fused with the support of data accumulation methods. Certain data gathering schemes are a greater amount of manual and rest are innovative. Consequently, the vast majority of the knowledge and assurance of these mining can be somewhat confounded than other structure of data innovation [26].

## 2.2 Overview of Machine Learning

Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available [27]. The procedures associated with AI are like that of information mining and prescient displaying. Both require scanning through information to search for examples and modifying project activities as needs be. Numerous individuals know about AI from shopping on the web and being served promotions identified with their buy. This happens in light of the fact that motors use AI to customize online advertisement conveyance in practically constant. Past customized promoting, other normal AI use cases incorporate extortion recognition, spam separating, and organize security danger location, prescient support and building news channels.

### 2.2.1 Characteristics of Machine Learning

Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.

**Supervised Learning:** Basically supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

**Unsupervised Learning:** Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data [28].

## 2.3 Missing value in Data

The concept of missing properties is important to understand so as to efficiently supervise knowledge. In the chance that the missing properties are not taken care of properly by the interpreter, at that point, he/she may finish up illustration an off base induction about the information. Because of imprudent dealing with, the consequence gotten by the scientist will contrast from ones where the missing qualities are available. Item non-reaction occurs when the respondent does not reply to particular inquiries because of stress, weariness or absence of information. The respondent may not react on the grounds that a few inquiries are delicate. This absence of answers would be viewed as missing qualities [29].

### 2.3.1 Dealing with missing Value

Here are some basic methods for managing missing information:

1. Encode NAs as - 1 or - 9999. This works sensibly well for numerical highlights that are prevalently positive in esteem and for tree-based models when all is said in done. This used to be an increasingly regular strategy in the past when the out-of-the container AI libraries and calculations were not skilled at working with missing information.
2. Case wise erasure of missing information. Here you essentially drop all cases/lines from the dataset that contain missing qualities. On account of an exceptionally huge dataset with not very many missing qualities, this methodology could conceivably work truly well. Notwithstanding, if the missing qualities are in cases that are likewise generally factually particular, this technique may truly slant the prescient model for which this information is utilized. Another serious issue with this methodology is that it will be unfit to process any future information that contains missing qualities. On the off chance that your prescient model is intended for generation, this could make difficult issues in organization.
3. Supplant missing qualities with the mean/middle estimation of the element in which they happen. This works for numerical highlights. The decision of middle/mean is frequently identified with the type of dispersion that the information has. For imbalanced information, the middle might be increasingly fitting, while for symmetrical and all the more ordinarily disseminated information, the mean could be a superior decision.
4. Name encodes NAs as another dimension of an unmitigated variable. This works with tree-based models and different models if the element can be numerically changed (one-hot encoding, recurrence encoding, and so forth.). This strategy does not function admirably with calculated relapse.
5. Run prescient models that ascribe the missing information. This ought to be done related to some sort of cross-approval conspire so as to maintain a strategic distance from spillage. This can be compelling and can help with the last model.
6. Utilize the quantity of missing qualities in an offered column to make another built element. As referenced above, missing information can regularly have heaps of valuable sign in its very own right, and this is a decent method to encode that data [30].

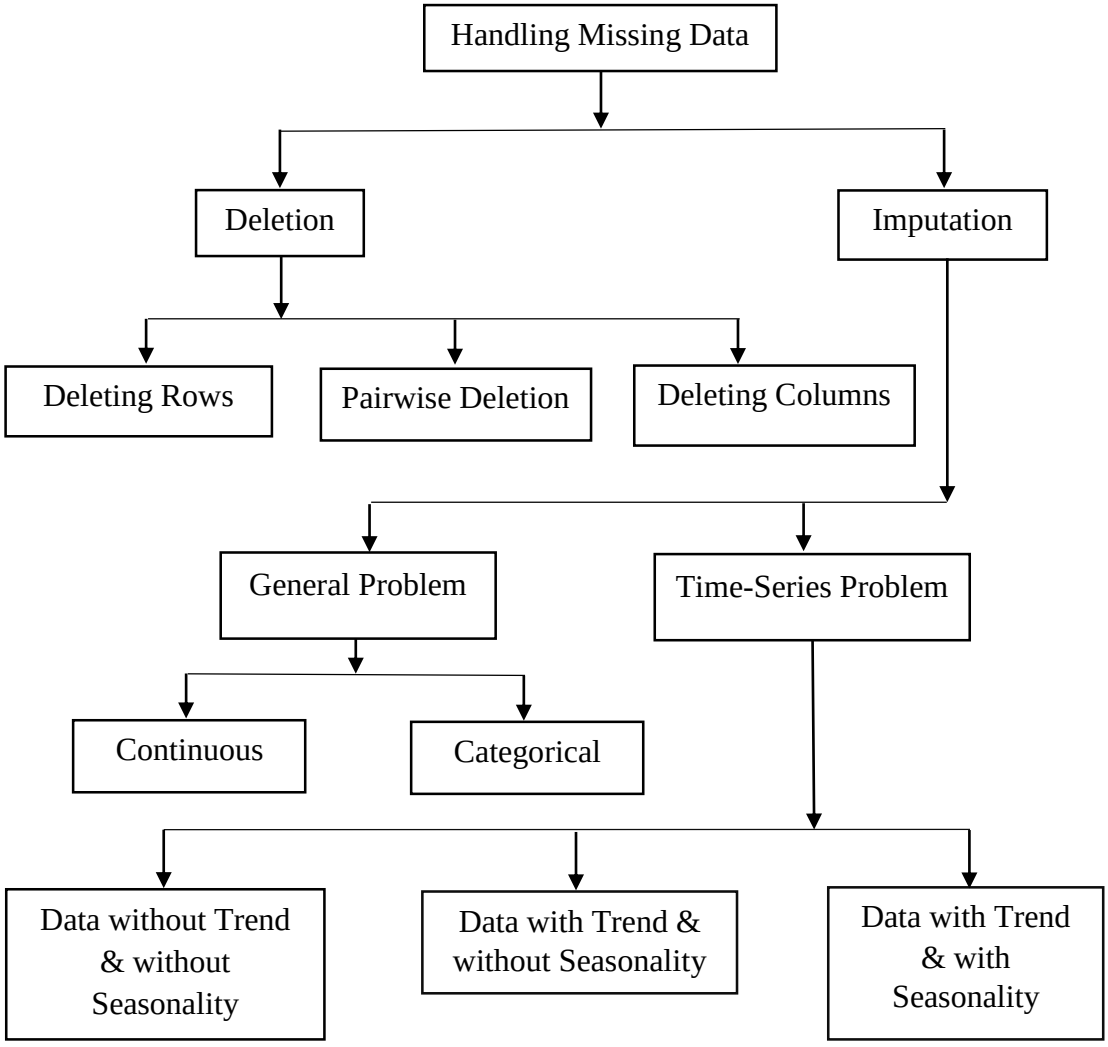



**Fig 2.1: Handling missing data**

## 2.4 Data Mining vs. Machine Learning

Data mining is considered the process of extracting useful information from a vast amount of data. It's used to discover new, accurate, and useful patterns in the data, looking for meaning and relevant information for the organization or individual who needs it. It's a tool used by humans. On the other hand, machine learning is the process of discovering algorithms that have improved courtesy of experience derived from data. It's the design, study, and development of algorithms that permit machines to learn without human intervention. It's a tool to make machines smarter, eliminating the human element. Data mining is designed to extract the rules from large quantities of data, while machine learning teaches a computer how to learn and comprehend the given parameters. Or to put it another way, data mining is simply a method of researching to determine a particular outcome based on the total of the gathered data. On the other side of the coin, we have machine learning, which trains a system to perform complex tasks and uses harvested data and experience to become smarter. Data mining relies on vast

stores of data (e.g., Big Data) which then, in turn, is used to make forecasts for businesses and other organizations. Machine learning, on the other hand, works with algorithms, not raw data [31].

# Chapter 3

# Related Works

## 3.1 Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability

Charles Edeki et al [10] suspects that none of the data mining and statistical learning algorithms used to breast cancer dataset outperformed the others in a certain process that it could be exposed the optimal algorithms and none of the algorithms accomplished poorly as to be dispelled from future prophecy pattern in breast cancer survivability tasks. Medical institutions looking to undertake a data mining approach to solve biological problems could be well-served by including statistical learning and data mining processes in their analytical and intervention efforts. Computer scientists, medical researchers and statisticians need to look at their own biological data availability for variables that might potentially link to prediction of cancer survivability. The selection of variables in this study was based on computational biology and bioinformatics literatures, breast cancer dataset available and domain knowledge of the researcher.

## 3.2 Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient

Ada et al [11] attempted to detect the lung tumors from the cancer images and the supportive tool is developed to check the normal and abnormal lungs and to predict survival rate and years of an abnormal patient so that cancer patient's lives can be saved. The early detection of lung cancer is a challenging problem, due to the structure of the cancer cells, where most of the cells are overlapped with each other. This paper presents the feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. After that they predict the survival rate of a patient by extracted features.

## 3.3 Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques

V. Krishnaiah et al [12] developed a prototype lung cancer disease prediction system using data mining classification techniques. The most effective model to predict patients with Lung cancer disease appears to be Naive Bayes followed by IF-THEN rule, Decision Trees and Neural Network. For Diagnosis of Lung Cancer Disease Naïve Bayes observes better results and fared better than Decision Trees. Lung cancer prediction system can be further enhanced and expanded. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining.

### 3.4 Predicting the Severity of Breast Masses with Data Mining Methods

Sahar A. Mokhtar et al [13] have analyzed three different classification models for the prediction of the severity of breast masses namely the decision tree, artificial neural network and support vector machine. The decision tree model is constructed using the Chi-squared automatic interaction detection method and the pruning method was used to find the optimal structure of the artificial neural network model and finally, the support vector machine has been built using polynomial kernel. The performances of the three models have been evaluated using statistical measures, gain and Roc charts. The support vector machine model outperformed the other two models on the prediction of the severity of breast masses.

### 3.5 A hybridized K-means clustering approach for high dimensional dataset

Rajashree Dash et al [14] a hybridized K-means algorithm has been proposed which combines the steps of dimensionality reduction through PCA, a novel initialization approach of cluster centers and the steps of assigning data points to appropriate clusters. Using the proposed algorithm a given data set was partitioned into k clusters. The experimental results show that the proposed algorithm provides better efficiency and accuracy comparison to an original k-means algorithm with reduced time. Limitations are the number of clusters (k) is required to be given as input. The method to find the initial centroids may not be reliable for a very large dataset.

### 3.6 Data clustering method for Discovering clusters in spatial cancer databases

Ritu Chauhan et al [15] concentrate on clustering algorithms such as HAC and KMeans in which, HAC is applied on K-means to determine the number of clusters. The quality of cluster is improved, if HAC is applied on K-means. The paper has referenced and discussed the issues on the specified algorithms for the data analysis. The analysis does not include missing records. The application can be used to demonstrate how data mining technique can be combined with medical data sets and can be effectively demonstrated in modifying the clinical research. This study clearly shows that data mining techniques are promising for clinical datasets.

### 3.7 Developing Prognostic Systems of Cancer Patients by Ensemble Clustering

Dechang Chen et al [16] introduced an ensemble clustering based approach to establish prognostic systems that can be used to predict an outcome or a survival rate of cancer patients. An application of the approach to lung cancer patients has been given. Generalizing or refining the work presented in this paper can be done in many ways. The algorithm EACCD actually is a two-step clustering method. In the first step, a dissimilarity measure is learnt by using PAM, and in the second step, the learnt dissimilarity is used with a hierarchical clustering algorithm to obtain clusters of patients. These clusters of patients form a basis of a prognostic system.

### 3.8 A study of digital mammograms by using clustering algorithms

S M Halawani et al [17] suggest that probabilistic clustering algorithms performed well than hierarchical clustering algorithms in which almost all data points were clustered into one cluster, maybe due to inappropriate choice of distance measure. This study presents clustering algorithms to study digital mammograms. Probabilistic clustering algorithms performed better than hierarchical clustering algorithm. Clustering results are competitive with classification results, indicating that clustering algorithms can be used as an important tool to study digital mammograms. Probabilistic clustering algorithms can also be used by radiologists to improve their prediction accuracy.

### 3.9 Improves Treatment Programs of Lung Cancer using Data Mining Techniques

Zakaria Suliman zubi et al [18] used some data mining techniques such as neural networks for detection and classification of lung cancers in X-ray chest films to classify problems aiming at identifying the characteristics that indicate the group to which each case belongs. According to the above concepts purposed and developed an automatic system for early detection of lung cancer by analyzing chest X-ray images using several steps. In generating the system, MATLAB has been the most important tools in implementing his paper. Medical image classification is an important thing in medicine. It allows biological structures to be isolated non-inva- sively. It is used for diagnostic purposes or practically applied in image guided surgeries; image classification has many forms and uses. Unfortunately, currently there is no classification strategy that can accommodate all its applications.

### 3.10 Detection of Brain Tumor using Modified K-Means Algorithm and SVM

Labeed K Abdul Gafoor et al [19] reported a methodology of segmentation of MRI images using wavelet and modified K-means algorithm. Wavelet transform made the algorithm noise free because wavelets provide frequency information as well as time-space localization. In addition, their multi-resolution character enables us to visualize image at various scales and orientations. Resolution reduction using wavelet depends on the amount of noise as well as the area of the target. Then k-means was applied to segment the MRI. K-means provides a very simple and efficient method of segmentation. Thereafter a parameter detection method has been employed to detect the tumor region. It proved that result is better by comparing with other two methods. On the other hand, this paper has shown that advanced technique of image processing and micro calcification detection which is useful in computer aided diagnosis. The intelligent systems development combined with health specialists' knowledge improve diagnostics associated to different pathologies. This method can be easily extended for brain tumor segmentation.

### 3.11 Application of data mining techniques to model breast cancer data

Shajahan et al [20] operated on the exercise of data mining techniques to pattern breast cancer data using decision trees to forebode the existence of cancer. Data collected held 699 evidence (patient records) with 10 characteristics and the output class as either effective or maleficent. The input used held sample code number, clump thickness, cell size and shape uniformity, cell growth and other results anatomical examination. The outcome of the supervised learning algorithm used displayed that the random tree algorithm had the maximum accuracy of 100% and the error rate of 0, while CART had the lowest accuracy with a rate of 92.99% but Naive Bayes had the accuracy of 97.42% with an error rate of 0.0258.

### 3.12 The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade

Britta Weigelt et al [21] declared that the improvement of gene exposure microarray technologies one decade ago has a passionate influence on the scientific society. The capacity to explore the evolution of thousands of genes in single research has discovered a new way for initial and translational experiments in breast cancer and bestowed the probability of responding questions that previously could not even be asserted. The use of ‘quantitative assessment' of genes more than histopathology worldly measurement of tumor properties would offer a more accurate imposition of the incessant tumor biology that recognizes clinical effect in breast cancer patients.

## 3.13 Cancer diagnosis using data mining technology

Dr. Syed Athar Masood [22] described that Cancer is a flock of malady in which the number of cells in the body augment unusually. Then these cells destroy other surrounding cells and their normal functions. It can breadth throughout the human body. Since it is an extremely treacherous disease, which diagnosis is very important. In some forms, it spreads within days. So the diagnosis of cancer at the early stages is very important. The challenge is to first diagnose the main type and then its subtypes. Hereby configuring of data mining classification tools to build a determination support method to recognize various types of cancer on the Genes dataset. Data mining technology aid in categorize cancer patients and this system aid to verify individual cancer patients by simply explore the data.

### 3.14 A hybridized K-means clustering approach for high dimensional dataset

Rajashree Dash et al [23] a hybridized K-means algorithm has been proposed which combines the steps of dimensionality reduction through PCA, a novel initialization approach of cluster centers and the steps of assigning data points to appropriate clusters. Using the proposed algorithm a given data set was partitioned in to k clusters in such a way that the sum of the total clustering errors for all clusters was reduced as much as possible while inter distances between clusters are maintained to be as large as possible. The experimental results show that the proposed algorithm provides better efficiency and accuracy comparison to original k-means algorithm with reduced time. Though the proposed method gave better quality results in all

cases, over random initialization methods, still there is a limitation associated with this, i.e. the number of clusters (k) is required to be given as input. Again the method to find the initial centroids may not be reliable for vary large dataset. Evolving some statistical methods to compute the value of k, depending on the data distribution is suggested for future research. Methods for refining the computation of initial centroids are worth investigating.

## 3.15 Early Detection of Cancer Using Data Mining

Neelam Singh et al [24] described that cancer is the leading cause of death worldwide. Therefore, the identification of genetic as well as environmental factors is very important in developing novel methods of cancer prevention. However, this is a multi-layered problem. Therefore a cancer risk prediction system is here proposed which is easy, cost-effective and time-saving. Large numbers of people in world have cancer. Most of them do not even know they have it. There is no remedy for cancer after completely affected. Death is inevitable. So the ability to predict cancer plays an important role in the diagnosis process. In this paper they have proposed an effective cancer prediction system based on data mining. They have provided an efficient approach for the extraction of significant pattern from data warehouse for efficient prediction of cancer. The proposed method is implemented using java. The proposed method can efficiently and successfully predict the risk of cancer.

# Chapter 4

# Methodology

## 4.1 Data Collection

In the case of predicting cancer, we collect the data of cancer patients. Creating dataset, we numerous data have been collect. Doctors help us for creating a experiment. Cancer patients also helps us to collect data. There are 1180 data and it contains 86 attributes and 1 class attributes. 86 attributes include symptoms and some tests part of cancers and 1 class attribute which is represented types of cancer. The cancer disease dataset format is .csv format.

### 4.1.1 Data organization

As cancer is a genetic disease, it has no specific age or time to attack. In my personal experience, we have seen some cancer patients those who were attacked at the age of 45 or more. One of my friends Roni Sardar, lives in chalna, Khulna and his father died 4years ago suffering cancer. Also, some patients, having met them in Khulna Medical College Hospital such as Kamrul Hasan is 53, Sajida Aktar (49) who is suffering in breast cancer and others like blood, lung and brain cancer etc.

#### 4.1.1.1 Data verified

All data of this paper have been verified by Cancer Specialist Dr. Mrinal Kanti Sarkar, Khulna Medical College Hospital, Khulna. He has written some books about cancer disease like brain cancer, prostate cancer, breast cancer and lung cancer. In those books, he has expressed that maximum data about cancer disease have been collected from Internet.

##### 4.1.1.1.1 Features

**Table 4.1 Data Features**

| Variables | Data Distribution | Types |
|---|---|---|
| Age | Minimum = 20 years<br>Maximum = 73 years<br>Mean = 46.714 | Numeric |
| Gender | Male = 607<br>Female = 573 | Nominal |
| Abdominal Pain | Positive Value = 373<br>Negative Value = 807 | Nominal |
| Abdominal Bloating | Positive Value = 177<br>Negative Value = 1003 | Nominal |
| Urinary Symptoms | Positive Value = 143<br>Negative Value = 1037 | Nominal |

| | | |
|---|---|---|
| Fatigue | Positive Value = 486<br>Negative Value = 664 | Nominal |
| Indigestion | Positive Value = 206<br>Negative Value = 974 | Nominal |
| Back Pain | Positive Value = 105<br>Negative Value = 1075 | Nominal |
| Constipation | Positive Value = 89<br>Negative Value = 1091 | Nominal |
| Swelling in Leg | Positive Value = 95<br>Negative Value = 1085 | Nominal |
| Frequently Urine | Positive Value = 76<br>Negative Value = 1104 | Nominal |
| Rash | Positive Value = 68<br>Negative Value = 1112 | Nominal |
| Fever | Positive Value = 270<br>Negative Value = 910 | Nominal |
| Weight Loss | Positive Value = 520<br>Negative Value = 660 | Nominal |
| Vomiting/Nausea | Positive Value = 229<br>Negative Value = 951 | Nominal |
| Diabetes | Positive Value = 69<br>Negative Value = 1111 | Nominal |
| Jaundice | Positive Value = 141<br>Negative Value = 1039 | Nominal |
| Swelling Patch | Positive Value = 79<br>Negative Value = 1101 | Nominal |
| Lump | Positive Value = 125<br>Negative Value = 1055 | Nominal |
| Sore Throat | Positive Value = 151<br>Negative Value = 1029 | Nominal |
| Hoarseness | Positive Value = 141<br>Negative Value = 1039 | Nominal |
| Nasal Obstruction | Positive Value = 80<br>Negative Value = 1100 | Nominal |
| Nose Bleeding | Positive Value = 112<br>Negative Value = 1068 | Nominal |
| Double Vision | Positive Value = 145<br>Negative Value = 1035 | Nominal |
| Numbness | Positive Value = 107<br>Negative Value = 1073 | Nominal |
| Jaw Pain | Positive Value = 86<br>Negative Value = 1094 | Nominal |

| | | |
|---|---|---|
| Difficulty Swallowing | Positive Value = 176<br>Negative Value = 1004 | Nominal |
| Ear Pain | Positive Value = 96<br>Negative Value = 1084 | Nominal |
| Loosening of Teeth | Positive Value = 64<br>Negative Value = 1116 | Nominal |
| Cough | Positive Value = 224<br>Negative Value = 956 | Nominal |
| Trouble Breathing | Positive Value = 126<br>Negative Value = 1054 | Nominal |
| Chest Pain | Positive Value = 263<br>Negative Value = 917 | Nominal |
| Intestinal Pain | Positive Value = 63<br>Negative Value = 1117 | Nominal |
| Diarrhea | Positive Value = 148<br>Negative Value = 1032 | Nominal |
| Blood Urine | Positive Value = 102<br>Negative Value = 1078 | Nominal |
| Loss of Appetite | Positive Value = 188<br>Negative Value = 992 | Nominal |
| Side Pain | Positive Value = 81<br>Negative Value = 1099 | Nominal |
| Anemia | Positive Value = 82<br>Negative Value = 1098 | Nominal |
| Weakness | Positive Value = 210<br>Negative Value = 970 | Nominal |
| Infection | Positive Value = 103<br>Negative Value = 1077 | Nominal |
| Swollen Lymph Nodes | Positive Value = 133<br>Negative Value = 1047 | Nominal |
| Night Sweats | Positive Value = 199<br>Negative Value = 981 | Nominal |
| Bone Pain | Positive Value = 184<br>Negative Value = 996 | Nominal |
| Tenderness | Positive Value = 57<br>Negative Value = 1123 | Nominal |
| Swelling in the Abdomen | Positive Value = 113<br>Negative Value = 1067 | Nominal |
| Chalky Stool | Positive Value = 73<br>Negative Value = 1107 | Nominal |
| Coughing up Blood | Positive Value = 102<br>Negative Value = 1078 | Nominal |

| | | |
|---|---|---|
| Shortness of Breath | Positive Value = 160<br>Negative Value = 1020 | Nominal |
| Headache | Positive Value = 147<br>Negative Value = 1033 | Nominal |
| Seizures | Positive Value = 84<br>Negative Value = 1096 | Nominal |
| Severe Itching | Positive Value = 74<br>Negative Value = 1106 | Nominal |
| Pain Lymph Nodes | Positive Value = 67<br>Negative Value = 1113 | Nominal |
| Groin Pain | Positive Value = 105<br>Negative Value = 1075 | Nominal |
| Muscle Weakness | Positive Value = 80<br>Negative Value = 1100 | Nominal |
| Respiratory Complications | Positive Value = 87<br>Negative Value = 1093 | Nominal |
| Changes in Hearing | Positive Value = 57<br>Negative Value = 1122 | Nominal |
| Unsteady | Positive Value = 60<br>Negative Value = 1120 | Nominal |
| Ultrasound | Positive Value = 305<br>Negative Value = 875 | Nominal |
| DRE | Positive Value = 50<br>Negative Value = 1130 | Nominal |
| PSA | Positive Value = 65<br>Negative Value = 1115 | Nominal |
| CT Scan | Positive Value = 601<br>Negative Value = 579 | Nominal |
| MRI | Positive Value = 425<br>Negative Value = 755 | Nominal |
| Cholangiography | Positive Value = 54<br>Negative Value = 1126 | Nominal |
| Angiography | Positive Value = 88<br>Negative Value = 1092 | Nominal |
| Laparoscopy | Positive Value = 139<br>Negative Value = 1041 | Nominal |
| Blood Test | Positive Value = 369<br>Negative Value = 811 | Nominal |
| Urine Test | Positive Value = 134<br>Negative Value = 1046 | Nominal |
| Endoscopy | Positive Value = 164<br>Negative Value = 1016 | Nominal |

| | | |
|---|---|---|
| Endoscopic Ultrasound | Positive Value = 63<br>Negative Value = 1117 | Nominal |
| X-Ray | Positive Value = 319<br>Negative Value = 861 | Nominal |
| Panoramic Radio Graph | Positive Value = 81<br>Negative Value = 1099 | Nominal |
| Bone Scan | Positive Value = 56<br>Negative Value = 1124 | Nominal |
| PET/PET-CT Scan | Positive Value = 296<br>Negative Value = 884 | Nominal |
| Physical Exam | Positive Value = 187<br>Negative Value = 993 | Nominal |
| Barium Swallow | Positive Value = 76<br>Negative Value = 1104 | Nominal |
| Esophagoscopy | Positive Value = 67<br>Negative Value = 1113 | Nominal |
| Bronchoscopy | Positive Value = 108<br>Negative Value = 1072 | Nominal |
| Photography | Positive Value = 90<br>Negative Value = 1090 | Nominal |
| IVP | Positive Value = 68<br>Negative Value = 1112 | Nominal |
| Bone Marrow Test | Positive Value = 158<br>Negative Value = 1022 | Nominal |
| Sputum Cytology | Positive Value = 92<br>Negative Value = 1088 | Nominal |
| Lymph Node Test | Positive Value = 93<br>Negative Value = 1087 | Nominal |
| Thoracentesis | Positive Value = 57<br>Negative Value = 1123 | Nominal |
| Thoracoscopy | Positive Value = 85<br>Negative Value = 1095 | Nominal |
| Mediastinoscopy | Positive Value = 46<br>Negative Value = 1134 | Nominal |
| Biopsy | Positive Value = 757<br>Negative Value = 423 | Nominal |
| Class | Lung Cancer = 49<br>Fallopian Tube Cancer = 39<br>Gallbladder Cancer = 29<br>Head and Neck Cancer = 34<br>Hypopharynx Cancer = 38<br>Kaposi Sarcoma Cancer = 47<br>Kidney Cancer = 38 | Nominal |

| | | |
|---|---|---|
| | Leukemia Cancer = 23<br>Liver Cancer = 41<br>Esophageal Cancer = 31<br>Lymphoma-Hodgkin Cancer = 37<br>Lymphoma-Non-Hodgkin Cancer = 39<br>Mesothelioma Cancer = 45<br>Chordoma Cancer = 32<br>Prostate Cancer = 46<br>Stomach Cancer = 35<br>Blood Cancer = 46<br>Brain Cancer = 40<br>Colorectal Cancer = 42<br>Pancreatic Cancer = 26<br>No Cancer = 423 | |

## 4.2 Data Statement

Cancer is when abnormal cells divide in an uncontrolled way. Some cancers may eventually spread into other tissues. Here 20 sorts of cancer works in represented which we well known. Generally we analysis the symptoms of the cancer then we analysis the tests which are must for cancer prediction. The symptoms are abdominal pain, fever, weight loss, cough, blood urine and muscle weakness. At first stage patient do not realize that they have cancer. It takes a long time to realize that they have cancer.

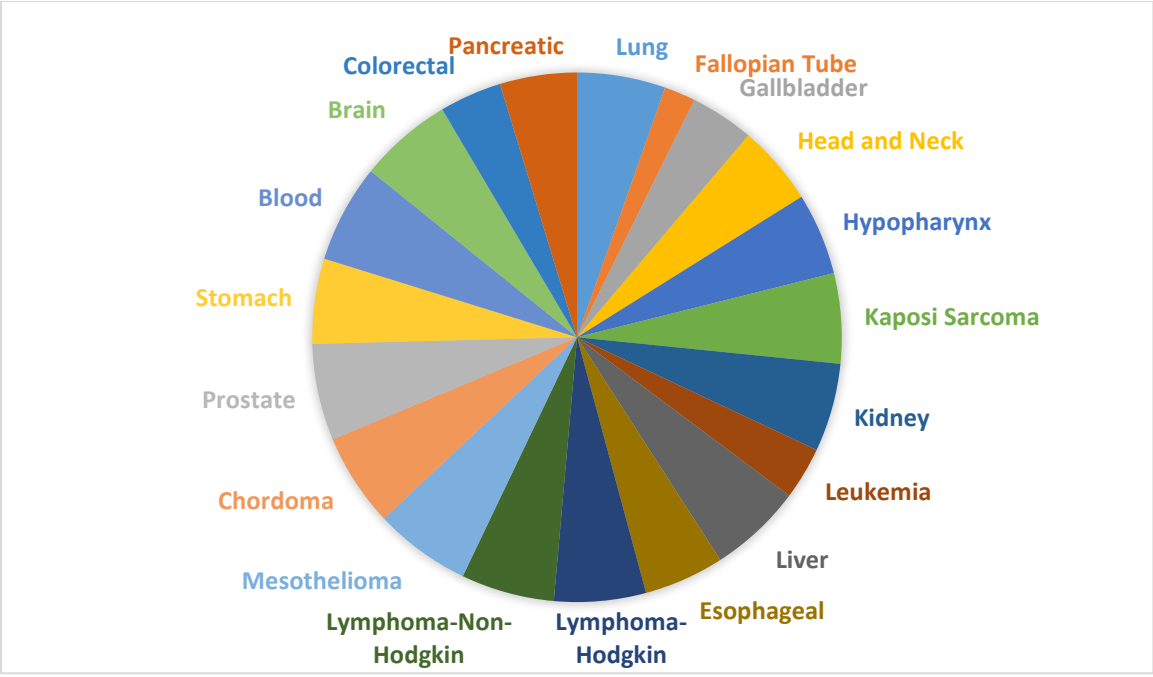



Fig 4.1: Cancer Types

The Dataset contains Fig 4.1 sorts of cancer which use in a dataset. Lung Cancer symptoms are very common for another disease. CT scan and X-Ray the major part of identifying the abnormalities lung. After that biopsy is the only medium which helps the doctor to identify cancer disease. For identifying fallopian tube cancer CT scan and MRI are must. Also CT scan, MRI, and Ultrasound are helped in identifying gallbladder cancer. Ultrasound, CT scan, MRI, Endoscopy are also identifying head and neck cancer and the biopsy helped must too sure about it. PET and Ultrasound are helped in identifying hypopharynx cancer. Kaposi sarcoma cancer identify when the CT scan, Bronchoscopy are positive to their result. Ultrasound sure to identifying kidney cancer. Blood test and Bone marrow test is positive then it confirm that is leukemia cancer. Abdominal pain, Swelling in the abdomen, Vomiting are the most common symptoms is liver cancer. CT scan, Ultrasound, MRI the major part of identifying the liver cancer. Barium Swallow and Endoscopy are identifying the esophageal cancer. Lymphoma-Hodgkin and Lymphoma-Non-Hodgkin are also identify when CT scan, MRI, Lymph node test are positive. Thoracentesis, CT scan also identifying the mesothelioma cancer. CT scan and MRI must be sure that identifying chordoma cancer. When Ultrasound, DRE, PSA tests are positive then it identifying Prostate cancer. Also CT scan and Ultrasound are identifying stomach cancer. For identifying blood cancer blood test is a must. Bone Marrow also helps in that purpose but the biopsy helped must too sure about it. CT scan and MRI the major part of identifying the abnormalities brain. Blood test gives an appropriate result for identifying colorectal cancer. CT scan, MRI and Ultrasound are also helped in identifying pancreatic cancer.

## 4.3 Architecture

System architecture for classification is given below:

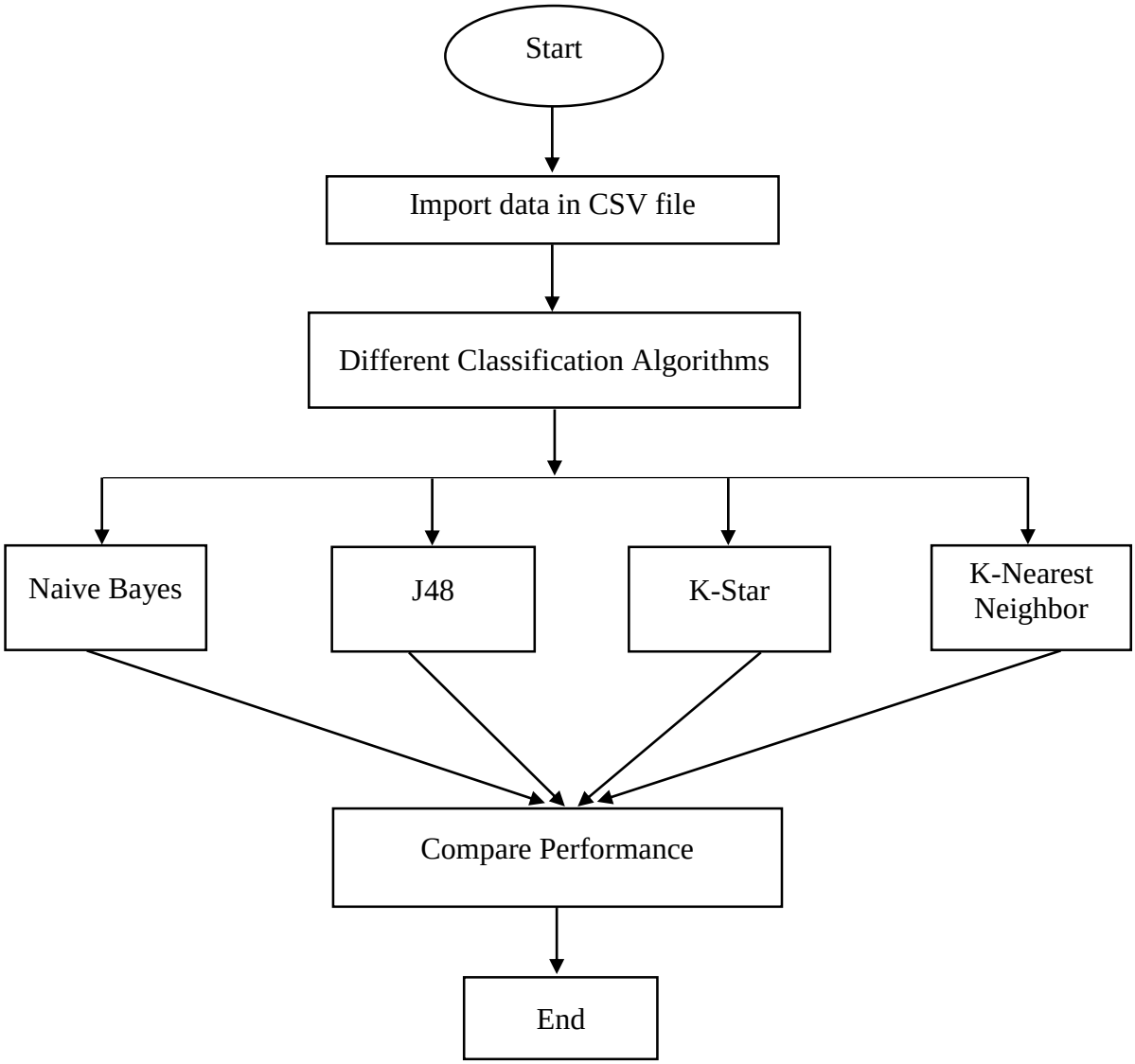



Fig 4.2: System architecture for classification

In Fig 4.2, we worked on Windows 10 operating system and Weka 3.8.3 version. First open Weka software then the experiment will be started. Imported the dataset as .csv file. Then applying different classification algorithm on the dataset like Naive Bayes, J48, K-star and K-Nearest neighbor. We got various data then compare it and the process is stopped.

System architecture for predicting cancer is given below:

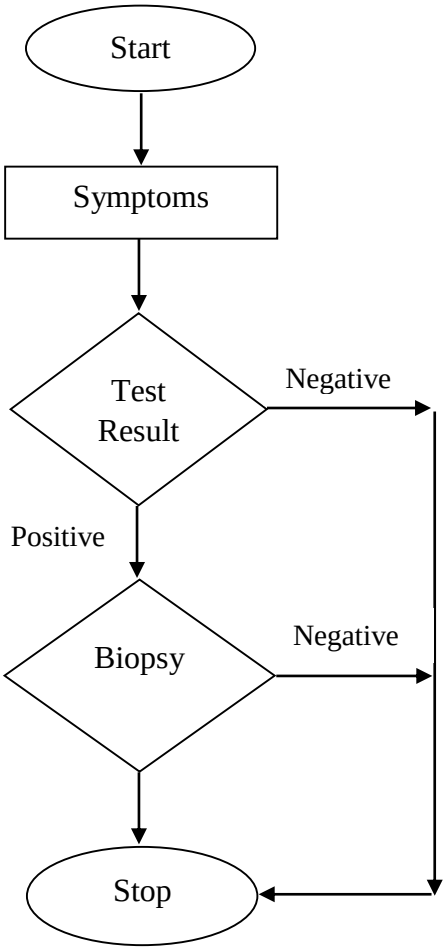



Fig 4.3: System architecture for predicting cancer

In Fig 4.3 the process is get started with the start. Symptoms are indicated as the training part. If any symptoms are true then it showed the result. If the test part indicates true then the biopsy is performed, otherwise the process is stopped. The test part and the biopsy both show negative then the process is stopped.

## 4.4 Machine Learning Algorithm

### 4.4.1 Naive Bayes

Naive Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. In short, it is a probabilistic classifier.
The Naive Bayes algorithm is called “naive” because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features.
For instance, if anyone trying to identify a fruit based on its color, shape, and taste, then an orange colored, spherical, and tangy fruit would most likely be an orange. Even if these features depend on each other or on the presence of the other features, all of these properties individually contribute to the probability that this fruit is an orange and that is why it is known as “naive.”

As for the "Bayes" part, it refers to the statistician and philosopher, Thomas Bayes and the theorem named after him, Bayes' theorem, which is the base for Naive Bayes Algorithm [32].
More formally, Bayes' Theorem is stated as the following equation:

$$P(A|B) = P(B|A)P(A)P(B)$$

Let us understand the statement first and then we will look at the proof of the statement. The components of the above statement are:

- $P(A|B)$: Probability (conditional probability) of occurrence of event A given the event B is true
- $P(A)$ and $P(B)$: Probabilities of the occurrence of event A and B respectively
- $P(B|A)$: Probability of the occurrence of event B given the event A is true.

The pseudo code for the naive bayes algorithm:
Input: Training Dataset T,
P = (p1, p2, p3,…, pn) // value of the predictor variable in testing dataset.
Output: class of testing dataset
Step:

1. Read the training dataset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat
   Calculate probability of all p using the gauss density equation in each class;
   Until the probability of all predictor variables (p1, p2, p3,….., pn) has been calculated
4. Calculate the likelihood for each class
5. Get the greatest likelihood; [39].

### 4.4.2 J48

J48 algorithm is called as optimized implementation of the C4.5 or improved version of the C4.5. The output given by J48 is the Decision tree. A Decision tree is same as that of the tree structure having different nodes, such as root node, intermediate nodes and leaf node. Each node in the tree contains a decision and that decision leads to our result as name is decision tree. Decision tree divide the input space of a data set into mutually exclusive areas, where each area having a label, a value or an action to describe or elaborate its data points. Splitting criterion is used in decision tree to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node [33].
The pseudo code of J48 algorithm:

1. Create a root node L;
2. IF (P belongs to same category C)
   {leaf node = L;
   Mark L as class C;
   Return L;}
3. For i =1 to n
   {Calculate Information_gain (Ig);}
4. ta = testing attribute;
5. L.ta = attribute having highest information_gain;

6. if (L.ta == continuous)
   {find threshold;}
7. For (Each P in splitting of P)
8. if (P is empty)
   {child of L is a leaf node;}
   else
   {child of L = dtree P}
9. calculate classification error rate of node L;
10. return L; [40]

### 4.4.3 K-Star

In classification problems, "each new instance is compared with existing ones using a distance metric, and the closest existing instance is used to assign the class to the new one" [34]. The principal difference of K* against other IB algorithms is the use of the entropy concept for defining its distance metric, which is calculated by mean of the complexity of transforming an instance into another; so it is taken into account the probability of this transformation occurs in a "random walk away" manner. The classification with K* is made by summing the probabilities from the new instance to all of the members of a category. This must be done with the rest of the categories, to finally select that with the highest probability [35].

To treat the missing values in datasets, Cleary & Trigg [36] assumed that the probability of transforming to that kind of values, is the mean of the probability of transforming to each of the specified attributes in the dataset. So, it is considered the expected distance to a random instance of that attribute.

Many authors of this area have used K* for different classification problems [37] and the results have been good.

However, there is not much information about how K* faces attribute and class noisy, and with mixed values of the attributes in the datasets. In the following section, we present an experimental study for comparing K* with some of the most influential DM algorithms, according to C4.5, Support Vector Machine (SVM), k-NN and Naive Bayes.

The pseudo code of K-star algorithm:

1. function divide (a,b);
   Input: Two n bit integers a and b, where b≥1
   Output: The quotient and remain of a divided by b
2. if a = 0 then
3. return (s,t) = (0,0)
   else
4. set (s,t) = divide ([a/b], b);
5. $s = 2 \times s$, $t = 2 \times t$;
6. if a is odd then
7. t = t+1
   end
8. if t≥b then
9. t = t – b, s = s+1

end
10. return (s,t)
end [41]

### 4.4.4 KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique [38].

The pseudo code of KNN algorithm:

1. Training algorithm
   For each training example <x, class (x)>, add the example to the list Traing
2. Classification algorithm (Rn–V)
   –Let V = {v1, …, vl} be a set of classes
   –Given a query instance Xq to be classified
   - Let X = {x1, …, xK} denote the K instances from Training that are nearest to Xq
   - Ai: 1…l, vote = {x € X | class (x) = Vi}
   - Return Vi such that |vote| is largest, [42]

# Chapter 5

## Implementation and Results

### 5.1 Implementation

Naïve Bayes, J48, K-star, K-Nearest neighbor are used to predict cancer disease. Weka tool is used for the purpose of measuring the accuracy of the cancer disease dataset including 20 types of cancer. 10-fold cross-validation is used for predicting cancer disease. Accuracy, Error rate, Sensitivity, Specificity, Precision and F-score.

**Table 5.1 Method of Data Formula**

| Measure | Formula |
|---|---|
| Accuracy | $\frac{TP + TN}{P + N}$ |
| Error rate | $\frac{FP + FN}{P + N}$ |
| Sensitivity | $\frac{TP}{P}$ |
| Specificity | $\frac{TN}{N}$ |
| Precision | $\frac{TP}{TP + FP}$ |
| F-score | $\frac{2 \times precision \times recall}{precision + recall}$ |

In Table 5.1 Here,

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative
- $P = TP + FN$
- $N = FP + TN$

### 5.2 Results

We tried to predict cancer disease implementing four types of algorithm and find the best accuracy. We use the Windows 10 and Weka 3.8.3 version. So our main work is to find the accuracy of the four classifier algorithm. We analyze 20 types of cancers accuracy, error rate, recall, specifity, precision and f-force. Using 10 folds cross validation and four classification algorithm, Weka gives us a confusion matrix. Confusion matrix gives us TP, FP, TN, FN values.

**Table 5.2 Comparative Performance of Various Algorithm on Dataset**

| Class | Accuracy | Error Rate | Recall | Specificity | Precision | F-Score |
|---|---|---|---|---|---|---|
| Naive Bayes | 98.8% | 1.2% | 95.2% | 99.8% | 95.7% | 95.1% |
| J48 | 98.6% | 1.4% | 95.6% | 99.6% | 95.6% | 95.6% |
| K-Star | 98.9% | 1.1% | 95.3% | 99.9% | 95.9% | 95.3% |
| K-Nearest neighbor | 98.8% | 1.2% | 95.5% | 99.8% | 96% | 95.6% |

Table 5.2 represents a comparison of four algorithm. From the table Naive Bayes accuracy is 98.8%, J48 accuracy is 98.6% K-Star accuracy is 98.9% and K-Nearest neighbor accuracy is 98.8%. So, here K-Star accuracy greater than the other three algorithms accuracy.

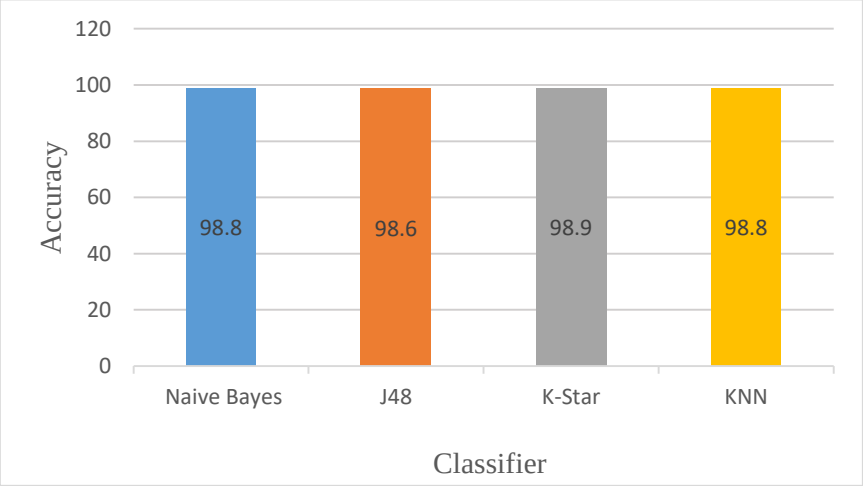



Fig 5.1: Graph of Accuracy

J48<Naive Bayes, KNN<K-Star

Accuracy refers to the ability of classifier. The number of greater the accuracy rate the classifier algorithm will be more accurately classified. After inserting dataset in Weka, we got a confusion matrix. Fig 5.1 represents the average accuracy graph of 4 algorithms named Naive Bayes, J48, K-Star and KNN classifier algorithm. K-Star accuracy is higher than others algorithm accuracy.

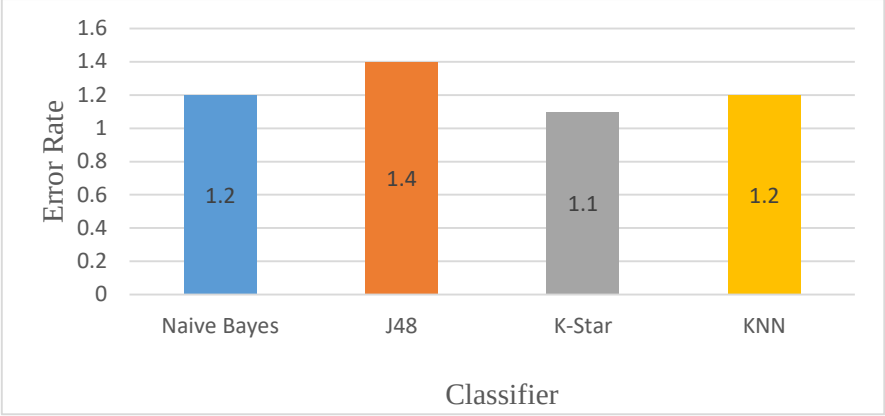



Fig 5.2: Graph of Error Rate

K-Star<Naive Bayes, KNN<J48

Error rate identifying the incorrect classifiers number. After inserting dataset in Weka, we got a confusion matrix. We put those values in the proper equation. Fig 5.2 represents the error rate graph of four algorithms named Naive Bayes, J48, K-Star and KNN. Error rate graph represents the opposite direction of accuracy graph. K-Star error rate is lower than others.

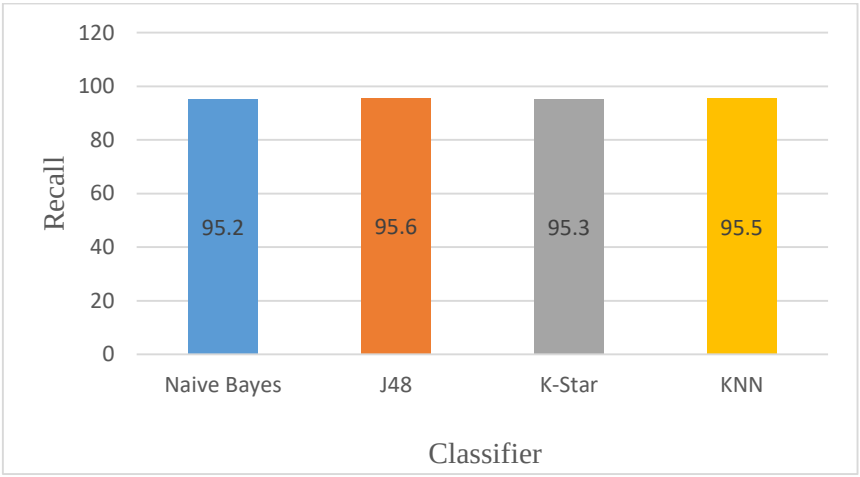



Fig 5.3: Graph of Recall

Naive Bayes<K-Star<KNN<J48

Recall is the ratio of true positive and positive value. After inserting dataset in Weka, we got a confusion matrix. We put those values in the proper equation. Fig 5.3 represents the average Recall graph of four algorithms named Naive Bayes, J48, K-Star and KNN. Here, J48 algorithm recall is high.

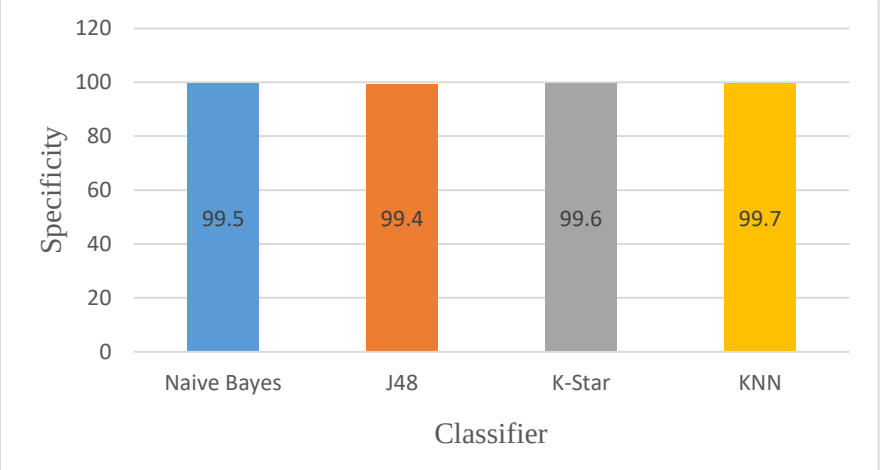



Fig 5.4: Graph of Specificity

J48<Naive Bayes<K-Star<KNN

Specificity is the ratio of true negative and negative. After inserting dataset in Weka, we got a confusion matrix. We put those values in the proper equation. Fig 5.4 represents the error rate graph of four algorithms named Naive Bayes, J48, K-Star and KNN. KNN Specificity is high.

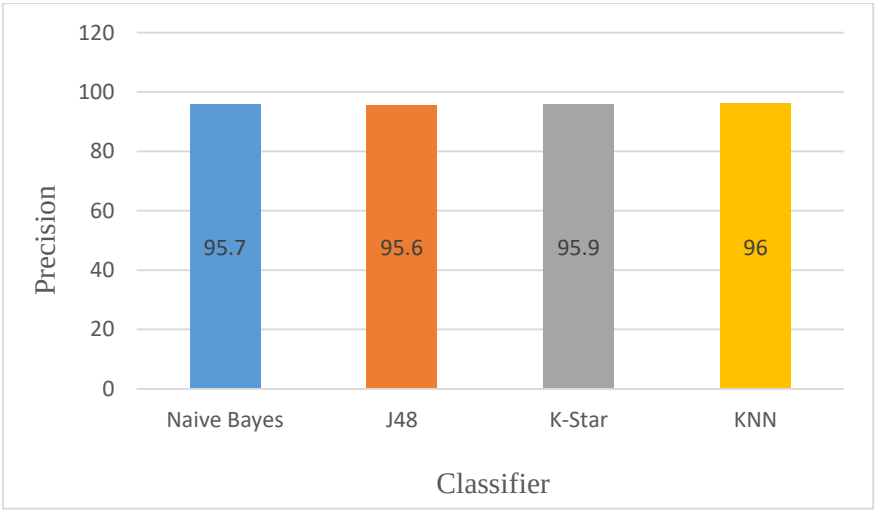



Fig 5.5: Graph of Precision

J48<Naive Bayes<K-Star<KNN

After inserting dataset in Weka, we got a confusion matrix. We put those values in the proper equation. Fig 5.5 represents the Precision graph of four algorithms named Naive Bayes, J48, K-Star and KNN. Again J48 Precision is low and KNN Precision is high.

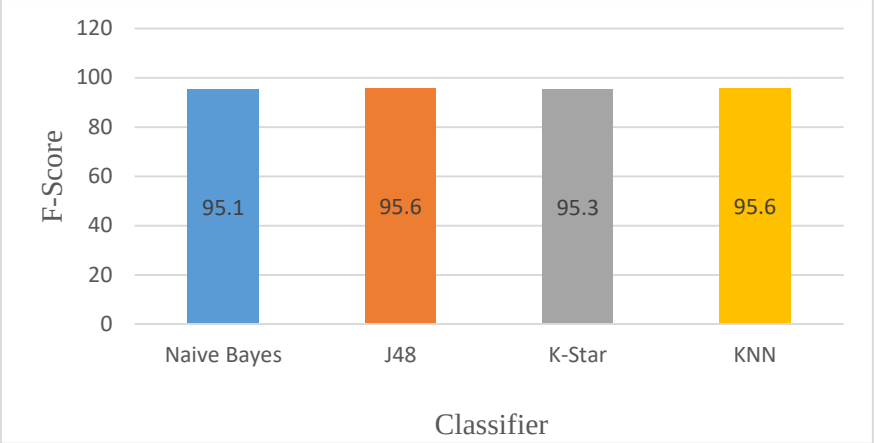



Fig 5.6: Graph of F-Score

Nave Bayes<K-Star<J48, KNN

F-Score is weighted average of the true recall and precision. After inserting dataset in Weka, we got a confusion matrix. We put those values in the proper equation. Fig 5.6 represents the F-Score graph of four algorithms named Naive Bayes, J48, K-Star and KNN. Here also KNN and J48 is high position and Naive bayes is low position.

**Table 5.2 Existing Work**

| Author | Year | Method | Accuracy |
|---|---|---|---|
| Charles Edeki and Shardul Pandya | 2012 | Logistics Regression, J48, ANN | Logistic Regression 71%, J48 70.17%, ANN 72.94% |
| Ritu Chauhan | 2010 | K-Means, HAC | It shows a graph. |
| Shar A. Moktar and Alaa M. Elsayad | 2013 | CAID, ANN, SVM | CAID 81.43%, ANN 81.13%, SVM 83.66% |

**Table 5.3 Our Work**

| Author | Year | Method | Accuracy |
|---|---|---|---|
| Md. Ariful Islam Bhuiyan, Md Taufik Akunjee, Md. Hashikul Islam, Rafiqul Islam | 2019 | Naive Bayes, J48, K-Star, KNN | Naive Bayes 98.8%, J48 98.6%, K-Star 98.9% and KNN 98.8% |

# Chapter 6

# Discussion

## 6.1 Conclusion

The paper is based on research medical dataset which is able to predict cancer disease. From the analysis we can say that most of the cancer mainly happen for smoking, drinking alcohol. We use four algorithm for creating the dataset. In this purpose we use Weka tool. Four algorithm use for identify confusion matrix. Confusion matrix gives the result of the classification algorithm. It contains information about actual and predicted classification. We have twenty classes lung cancer, fallopian tube cancer, gallbladder cancer, head and neck cancer, Hypopharynx cancer, Kaposi sarcoma cancer, kidney cancer, leukemia cancer, Liver cancer etc. Its help to find accuracy, error rate, recall, specificity, precision and F-score. Tables are created for clear perception. It gives a comparison between the classification algorithms named Naive bayes, J48 algorithm, K-star and KNN. The latest advances in cancer treatment have created a whole new outlook on how to treat cancer. These advances have developed from a deeper understanding of the molecular basis of cancer. Some of the earlier treatments are still valuable however they have some drawbacks. For example, surgery and radiation are effective but they only treat one local area of the cancer. Chemotherapy can treat cancer cells that are spread all over the body but they have extremely toxic side effects. All of these treatments are still in use today and will probably be in use for a while although they will not be the only kind of treatments. In this paper a doctor can find the actual cancer easily and its helps a doctor to find it. And the patient also understand which stage cancer he or she had. Here we discuss the whole causes and main reason and how to find the cancer. And we try to find the most accuracy that's why we ensure that which cancer he had.

## 6.2 Limitations

The number of data's are limited. We only use the classifier algorithm. Most of the cancer patients are over 35 years aged. Only focused on the patients have cancer or not and types of cancer. Our attribute length is also huge. We only find out 20 types of cancer but at present in this world almost 100 types of cancer present.

## 6.3 Future Work

In future we will try to add more innovation to a large improvement. We will try to extend dataset. We will try to develop new models prediction and survivability. We will try to find out more types of cancer. In future we add more data in our data set and we add more feature like specific reason of cancer and find out more accurate accuracy. In future we try another algorithm to find out the accuracy and we try to find out the all types of cancer and there causes and there actual reason and what types of treatment they need we also try to find out that.

# References


1. National cancer institute “What is Cancer” Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer [Accessed: 29.08.2019]
2. World Health Organization “Cancer” Available: https://www.who.int/news-room/fact-sheet/details/cancer [Accessed: 15.09.2019]
3. Challenges of Cancer Control in Developing Countries “Medscape” Available: https://www.medscape.com/viewarticle/752627_2 [Accessed: 15.09.2019]
4. The Daily Star “Cancer treatment in Bangladesh: Still a long way to go” Available: https://www.thedailystar.net/opinion/perspective/news/cancer-treatment-bangladesh-still-long-way-go-1696912 [Accessed: 15.09.2019]
5. South Asian Journal of Cancer “Comprehensive update on cancer scenario of Bangladesh” Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3889062/ [Accessed: 25.08.2019]
6. What it means for your prognosis “Cancer survival rate” Available: https://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer/art-20044517 [Accessed: 25.08.2019]
7. Science Museum “How does cancer begin” Available: http://whoami.sciencemuseum.org.uk/whoami/findoutmore/yourbody/whatiscancer/howdoescancerbegin [Accessed: 23.09.2019]
8. The ACTION study protocol “Socioeconomic impact of cancer in member countries of the Association of Southeast Asian Nations (ASEAN)” Available: https://www.ncbi.nlm.nih.gov/pubmed/22524800 [Accessed: 23.09.2019]
9. World Cancer Day 2019: “Cancer treatment still a dream” Available: https://www.dhakatribune.com/bangladesh/event/2019/02/04/world-cancer-day-2019-cancer-treatment-still-a-dream [Accessed: 20-08-2019]
10. Charles Edeki “Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability” Mediterranean Journal of Social Sciences Vol 3 (14) November 2012, ISSN: 2039-9340
11. Ada and Rajneet Kaur “Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient” International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, Issue. 4, April 2013, pg.1 – 6, ISSN 2320–088X
12. V.Krishnaiah “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques” International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45 www.ijcsit.Com ISSN: 0975-9646
13. A. Sahar “Predicting the Severity of Breast Masses with Data Mining Methods” International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org
14. Rajashree Dash “A hybridized K-means clustering approach for high dimensional dataset” International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66
15. Ritu Chauhan “Data clustering method for Discovering clusters in spatial cancer databases” International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010
16. Dechang Chen “Developing Prognostic Systems of Cancer Patients by Ensemble Clustering” Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume 2009, Article Id 632786
17. S M Halawani “A study of digital mammograms by using clustering algorithms” Journal of Scientific & Industrial Research Vol. 71, September 2012, pp. 594-600

18. Zakaria Suliman zubi "Improves Treatment Programs of Lung Cancer using Data Mining Techniques" Journal of Software Engineering and Applications, February 2014, 7, 69-77
19. Labeed K Abdulgafoor "Detection of Brain Tumor using Modified K-Means Algorithm and SVM" International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Applications NCRTCA 2013
20. Shajahaan, S.S., Shanthi, S. & ManoChitra, V. (2013). "Application of data mining techniques to model breast cancer data." International Journal of Emerging Technology and Advanced, 3(11), 3622-369
21. Britta Weigelt, Frederick L Baehner and Jorge S Reis-Filho "The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade," Journal of Pathology, J Pathol 2010; 220: 263–280, Published online 19 November 2009 in Wiley InterScience (www.interscience.wiley.com), DOI: 10.1002/path.2648
22. Muhammad Shahbaz, Shoaib Faruq, Muhammad Shaheen, Syed Ather Masood "Cancer diagnosis using data mining technology," Life Science Journal. 2012;9(1):308-313] (ISSN:1097-8135)
23. Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66
24. Neelam Singh and Santosh Kumar Singh Bhadauria "Early Detection of Cancer Using Data Mining" International Journal of Applied Mathematical Sciences, ISSN 0973-0176 Volume 9, Number 1 (2016), pp. 47-52, © Research India Publications, www.ripublication.com
25. An Overview on Data Mining "Semantic Scholar" Available: https://pdfs.semanticscholar.org/f44d/2c02e22ae27364e0bcfbfcb5bed74b0aa2e1.pdf [Accessed: 22.09.2019]
26. Wisestep "Data Mining" Available : https://content.wisestep.com/data-mining-purpose-characteristics-benefits-limitations[Accessed: 14.05.2019]
27. Special Report "Artificial Intelligence apps come of age" Available: https://searchenterpriseai.techtarget.com/definition/machine-learning-ML [Accessed: 22.09.2019]
28. Greeks for Greeks "Supervised and Unsupervised learning" Available: https://www.geeksforgeeks.org/supervised-unsupervised-learning/ [Accessed: 22.09.2019]
29. Statistics Solutions "Missing Values in Data" Available: https://www.statisticssolutions.com/missing-values-in-data [Accessed: 16.05.2019]
30. Zestfinance "Methods for Dealing with Missing Data" Available: https://www.zestfinance.com/blog/6-methods-for-dealing-with-missing-data [Accessed: 16.05.2019]
31. EDUCBA "Difference Between Data mining and Machine learning" Available : https://www.educba.com/data-mining-vs-machine-learning/ [Accessed : 14.05.2019]
32. Blog Developers "Naive Bayes" Available: https://www.hackerearth.com/blog/developers/introduction-naive-bayes-algorithm-codes-python-r/#targetText=Naive%20Bayes%20is%20a%20machine,high%20dimensional%20training%20data%20sets.&targetText=Naive%20Bayes%20is%20the%20first,for%20solving%20text%20classification%20problem [Accessed: 22.09.2019]
33. Vaithiyanathan, V., K. Rajeswari, Kapil Tajane, and Rahul Pitale. "Comparison of Different Classification Techniques Using Different Datasets." Vol.6, no. 2 (2013)

34. Witten, I.H., E. Frank, and M.A. Hall, Data Mining. Practical Machine Learning tools and techniques, ElSevier, Editor. 2011 [Accessed: 22.09.2019]
35. Cleary, J. and L. Trigg, K*: An Instance-based Learner Using an Entropic Distance Measure, in 12th International Conference on Machine Learning. 1995. p. 108-114 [Accessed: 22.09.2019]
36. Uzun, Y. And G. Tezel, Rule Learning With Machine Learning Algorithms And Artificial Neural Networks. Journal of Seljuk University Natural and Applied Science, 2012. 1(2) [Accessed: 22.09.2019]
37. Er, E., Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100. International Journal of Machine Learning and Computing, 2012. 2(4): p. 279 [Accessed: 23.09.2019]
38. Saedsayad "K-Nearest neighbor" Available: https://www.saedsayad.com/k_nearest_neighbors.htm [Accessed: 23.09.2019]
39. Researchgate "Pseudocode of Naive Bayes algorithm" Available: https://www.researchgate.net/figure/Pseudocode-of-naive-bayes-algorithm_fig2_325937073 [Accessed: 20.08.2019]
40. Researchgate "The pseudo code for the C45-J45 algorithm" Available: https://www.researchgate.net/figure/The-pseudo-code-for-the-C45-J48-algorithm_fig1_319716463 [Accessed: 29.08.2019]
41. Google "pseudocode for k-star algorithm & tbm" Available: https://www.google.com/search?q=pseudocode+for+k-star+algorithm&tbm=isch&sourch=univ&sa=X&ved=2ahUKEwjSsdzbl6rkAhXEq48KHa35BUwQ7AI6BAgGECQ&biw=1366&bih=657#imgrc=M9sn5GoDltZd6M: [Accessed: 30.08.2019]
42. Google "Pseudocode for KNN algorithm & sxrf" available: https://www.google.com/search?q=Pseudocode+For+KNN+algorithm&sxrf=ACYBGNRZGxZtfvMEG0GjvcufnndDUaCczg:1569951393294&tbm=isch&source=iu&ictx=1&fir=mY80rSNauDM3SM%253A%252CbUAobailtZpZEM%252C_&vet=1&usg+AI4_-kSWkcF_-tnlQj9OmLk0j8scrW1bNQ&sa=X&ved=2ahUkEwjgpfLYzPvkAhULwI8kHbf3C3MQ9QEwB3oECAQQBg#imgrc=-uk7xJh1iD3QZM:&vet=1 [Accessed: 30.08.2019]