CANCER PREDICTION APPLICATION USING DATA MINING

BY

Susmoy Mojumder

ID: 161–15–7563

Tapash Chandra Das

ID: 161–15–7586

Nayem Uddin Prince

ID: 161–15–6829

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

Supervised By

Mr. Narayan Ranjan Chakraborty

Assistant Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

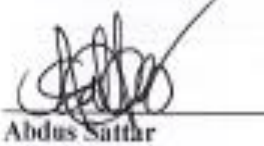
DECEMBER2019

APPROVAL

This Project/internship titled **"CANCER PREDICTION APPLICATION USING DATA MINING"**, submitted by Susmoy Mojumder ID:161-15-7563 Nayem Uddin Prince ID: 161-15-6829 Tapash Chandra Das ID: 161-15-7586 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 07-12-2019.

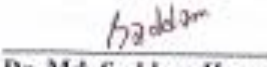
BOARD OF EXAMINERS

Dr. Syed AkhterHossain — **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

Shaon Bhatta Shuvo — **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 project has been done by us under the supervision of**Mr. Narayan RanjanChakraborty, Assistant Professor of CSE Department,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

14.12.19

Mr. Narayan RanjanChakraborty
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

SusmoyMojumder
ID: 161-15-7563
Department of CSE
Daffodil International University

Tapash Chandra Das
ID: 161-15-7586
Department of CSE
Daffodil International University

Nayem Uddin Prince
ID No: 161-15-6829
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. Narayan Ranjan Chakraborty, Assistant Professor of CSE Department, Daffodil International University.** Deep Knowledge & keen interest of our supervisor in the field of "*Cloud Computing, Big Data and IoT, Social Media Mining*" 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 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 patients of our parents.

ABSTRACT

Cancer is one of the mighty diseases in our country. Many people suffer from cancer and many people died because they don't get treatment at the proper time. If they get a chance to detect cancer in proper time might be they could alive. In our health care center have vast amounts of cancer patient's data. This data plays a vital role in decision making. The test is the manner by which to separate important learning from this information and follow up on it in a convenient way. To transform into information, proficient registering and data mining apparatuses must be utilized. This data can help to develop an expert system for decision assistance that can assist physicians in diagnosing and prediction cancer. We are using some algorithm to detect cancer. This research approach for forming the missing features based on K-NN and the distance function. This function gave the best accuracy. To seeing this accuracy how much probability has to suffer in cancer.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Malignant growth! Disease! At the point when we heard the name is heard, it hits the chest. I have been got notification from so often that this malady, no treatment. It is a long way from attempting to forestall the sickness before it occurs. Most malignant growths were not known to cause any. Cigarette smoking and lung disease were not known until the last thirty or forty years. Yet, over the most recent forty years, these things have changed.

We presently realize that radiation, cigarettes and numerous other ecological poisons cause malignant growth in the body. Relevantly, as radiation causes malignancy, controlled radiation is a significant and significant apparatus in disease treatment. (A typical reason for conventional individuals' feelings of trepidation and uncertainty about atomic power is the unavailability of uncontrolled radiation because of a significant mishap in the reactor. Be that as it may, to lessen the probability of such a mishap to a satisfactory degree, it isn't sensible to restrict it simply because of radiation. In nature, we discover radiation to some degree. The control of the 'level' of the radiation that has been impaired from the reactor how much certainty can be put on the framework.) We realize that specific hereditary ailments, stomach and bosom disease build the hazard. What's more, we know precisely how certain infections can infiltrate sound body cells and transform them into malignant growth cells.

In our country, we have 142 million people, which is the 9th most populous country in the world. There are 13 to 15 lakhs cancer patients in Bangladesh, with about 2 lakhs patients newly diagnosed with cancer each year. But most of the time people don't identify what the disease is! Even when they identify, it was too late. So we need to detect firstly with a Detection app & we are going to make a Cancer detection app that helps people to detect cancer in the First Stage and they can take proper treatment.

1.2 Motivation

A developing country like Bangladesh, we face a lot of problems with this mortality disease Cancer. Most of the Bangladeshis live below the poverty line. It is impossible for them to afford the cost of cancer.

Cancer therapy is not developed in our country. Many of the people we depend on overseas, which is very rare. In our country, many misunderstand whether or not they have cancer, a lot of time is wasted, and on the other hand, cancer becomes more terrible. At last, they were put to death. So we are looking to build a cancer detection application, that people can identify their cancer at home and take proper step easily.

Malignant growth is a weakening malady and has throughout the years given the therapeutic experts' restless evenings attempting to discover compelling, exact and dependable methods for diagnosing it. To this end consequently, it has gotten troublesome in giving a brief reaction to disease patients when it recently rises. This investigation attempted to recognize critical analytic highlights that best depicted malignant growth information utilizing.

1.3 Objective

Data Mining strategies are actualized together to make a novel strategy to analyze the presence of malignant growth for a specific patient. When starting to take a shot at a data mining issue, it is first important to bring all the data together into a lot of examples. The data must be gathered, coordinated, and tidied up.

To the created framework can be utilized by doctors and patients the same to effectively know an individual's disease status and seriousness without screening them for the testing disease.

- To reduce paperwork & testing.
- To saving money.
- To know results anytime anywhere.
- The user opinions have to be developed by using data mining which builds on gender.
- To developed classification accuracy and missing values.

1.4 Expected Outcome

Our application purpose has to build on machine learning and data mining technique which can help people to detect cancer cells easily and effectively. We put a lot of data here. When people fill up their data from it will show how much probability they have to cancer. In our medical industry are slow to given report, on the other hand, they took a lot of money and time. It will reduce our data mining technique. People don't need medical resources. In this thesis people information has been used as data sources for developing automatic data mining techniques and machine learning so it produces useful patterns and decisions to our people. That's how people could avoid this disease.

1.5 Research Questions

The main goal of our study is to answer the following research questions:

- What kind of data mining algorithm used currently in the healthcare sector for Cancer detection?
- How can the peculiar property that best describe data for the purpose of differentiating the cruel form of cancer prediction using data mining?
- How do missing component esteems improve expectations in deciding the exhibition accomplished by data mining calculations?
- How can be developed cancer prediction data mining model accuracy and missing values?

CHAPTER 2
BACKGROUND

2.1 Introduction

This part shows the ideas, which address the explanation behind the reason for proposal work. At first, we discussed human resource organization, how Knowledge Discovery in Databases (KDD) supports making decisions by staff evaluating. After that point, we talked about the means of KDD. In the work, we use KNN (K- Nearest Neighbor) algorithm.

Data mining is the way toward social affair the information by filtering through huge information, utilizing acknowledgment innovations. Information mining testing calculations that can help the procedure of arrangement, forecast, and example. For this reason, we utilized PC models got from Hospital information. That we gathered with our own. In this thought, we attempt to discover exactness and decrease by pre-customized strategies. As a result, a strong and solid working data mining calculations can be created to recognize malignant growth.

- Knowledge Discovery in Database (KDD)

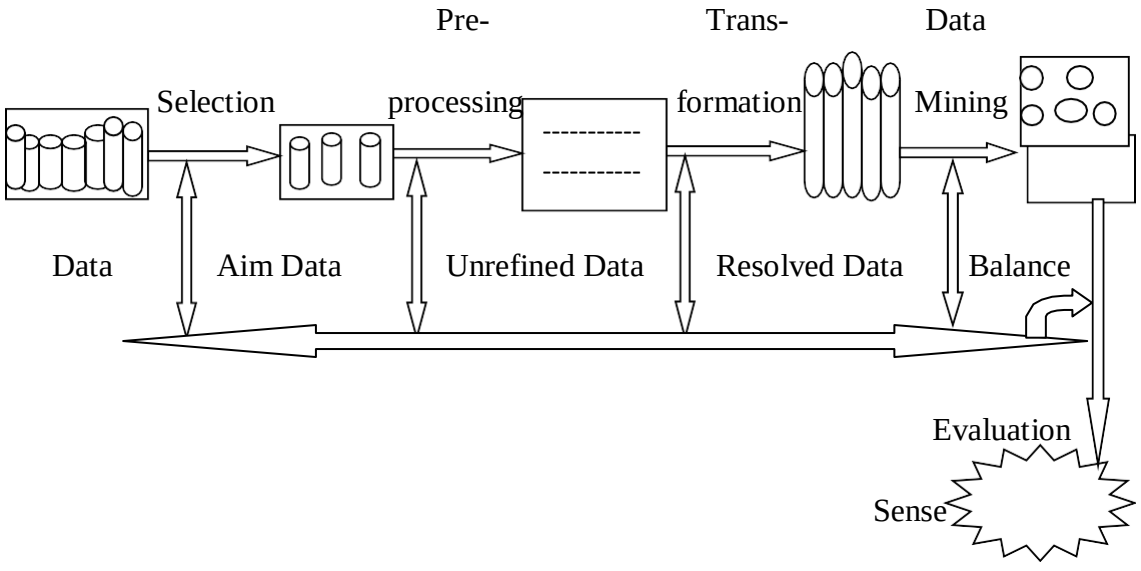


Fig 2.1: The KDD process steps

KDD is the general procedure of finding the most significant learning from information, we collected that why we use KDD. It's an integration of multiple techniques of data management such as data warehouse. As we can see (fig 2.1).

- Data all data includes topical for knowledge and the goals of the application.
- Creating a target dataset: including choosing a dataset or concentrating on a subset of factors or data samples.
- Preprocessed Data: include fundamental tasks, for example, expelling clamor or anomalies if proper, gathering the essential data to models or record for the commotion.
- The function of data mining: We include deciding the purpose of the model derived from the data mining algorithm.
- Balancing evaluation: Only place the main balancing of sense-based data.
- Knowledge presentation: Idea and sense present design we used to present the mined sense to the user

Data mining is the utilization of explicit calculations for concentrate designs from information. One the other hand, KDD is the process of data preparation, selection, cleaning, importing, transformed. In the proper implementation, we got some real data which going to put on an algorithm.

However, some person tried to shorten the term of knowledge Discover in Databases to Data mining, as it is the synonym of each other, not a shorter version.

2.2 Related Works

2.2.1 Model for cancer detecting

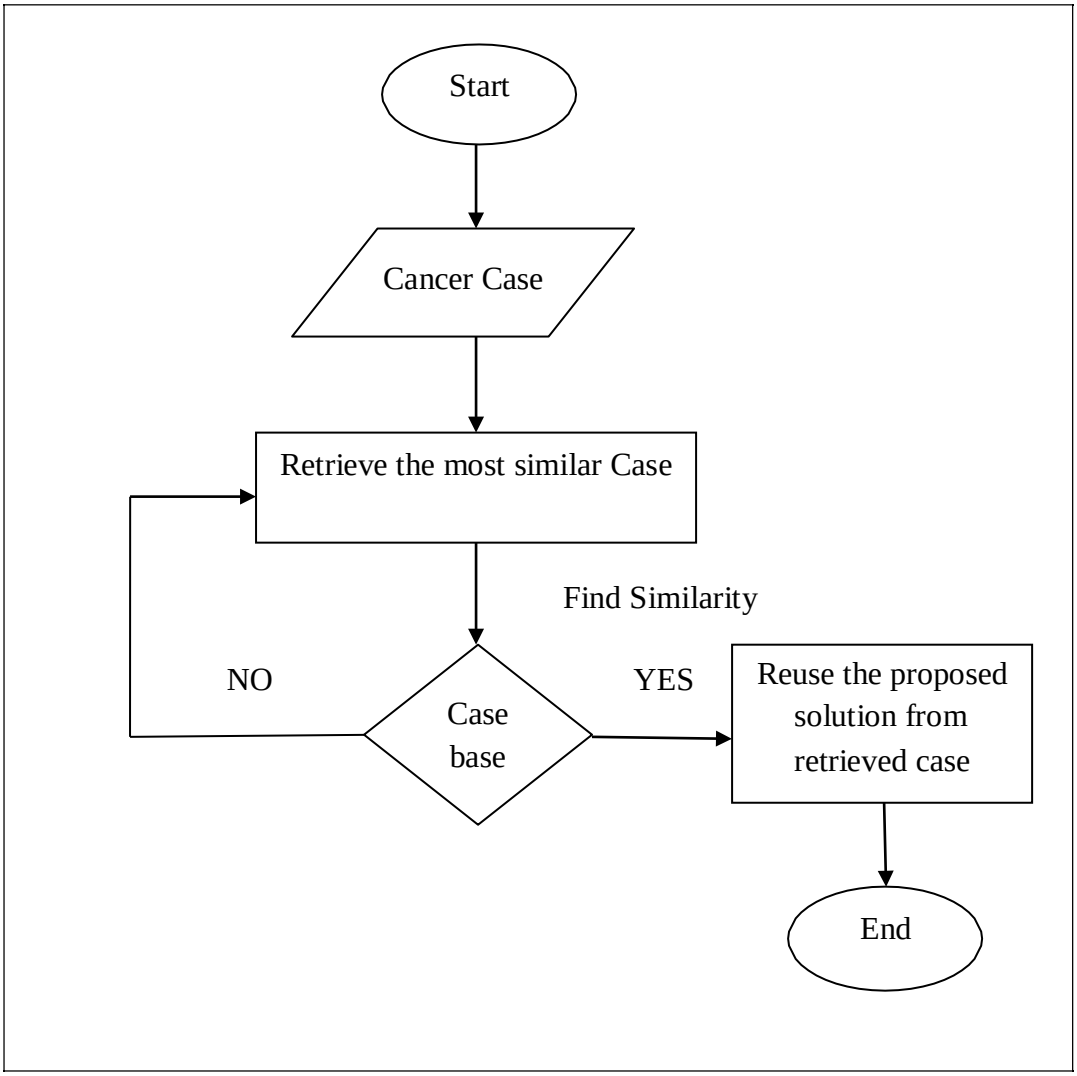


Fig 2.2: Model for cancer detecting [8]

Prerequisite investigation is moreover called as need planning which is the route toward choosing customer want for another system. These features, called essentials, must be

quantifiable, noteworthy and low down. Prerequisite examination incorporates visiting correspondence with structure customers to choose express segment wants.

Objectives of conflict or obscurity in necessities as mentioned by the various customers or social events of customers, evading of feature creep and documentation of all pieces of the endeavor improvement process from start to finish (fig 2.2).

2.2.2 Association Rules

Association rules play a vital role in the process of data mining. It's important class discipline within data that have been extensively studied by data mining. Association rule finds interesting associations and relationships with the data set. This rule show how it works. Association Rule – An implication expression of form A -> B, where A and B are any 2 item sets.

2.2.3 K-Nearest Neighbors algorithm

The k-closest neighbor's algorithmic program is one of the least complex AI calculations. It has just upheld the idea that articles that are close to each option can also have comparative qualities. So in the event that it can perceive the trademark choices of one of the items, it will be also anticipated for its closest neighbor. K-NN is partner act of spontaneity over the closest neighbor procedure. It depends for the most part on the arrangement that any new occasion will be grouped by the dominant part vote of its 'k' neighbors; - any place k is a positive number, now and then a little assortment.

K-N N is one of the chief simple and straightforward information preparing methods. It is referred to as Memory-Based Classification as the instructing models must be in the memory at run-time. When taking care of persistent characteristics the qualification between the traits is determined Euclidean separation. A genuine disadvantage once managing the Euclidean separation recipe is that the recurrence of the enormous worth bogs the little ones. When KNN is utilized for characterization, the yield is determined in light of the classification with the absolute best recurrence from the K-most comparable occurrences. Each occasion basically votes in favor of their group and in this manner the classes with the chief votes are taken for the forecast. We can see it graphically below (fig 2.3).

Class probabilities are determined on the grounds that the standardized recurrences of tests that have a place with each class inside the arrangement of K most comparable occurrences for another information example. For example, during a double characterization issue (class is zero or 1):

(Class=0) = count (class=0) / (count (class=0) +count (class=1))

In the event that utilizing K and having a much number of classes (for example 2) it is a smart thought to pick a K esteem with an odd number to stay away from a tie. Furthermore, the reverse, utilize a significantly number for K when having an odd number of classes.

Pseudo code of K-Neighbor

- Burden the preparation and test data
- Pick the estimation of K
- For each point in test information: - locate the Euclidean separation to all preparation information focuses
    - store the Euclidean separations in a rundown and sort it
    - pick the primary k focuses
    - appoint a class to the test point dependent on most of the classes present in the picked focuses

- End

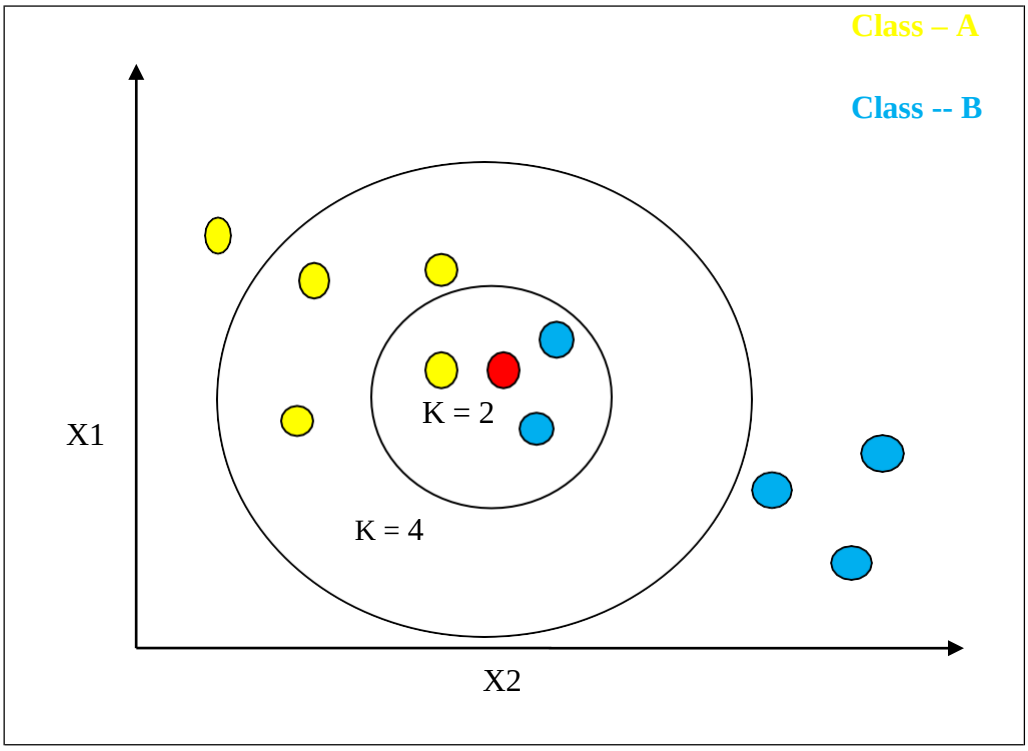


Fig 2.3: K-Neighbor

2.2.4 Advantage of K–NN

No Training Period: KNN is called Lazy Learner. It doesn't get the hang of anything in the preparation time frame. It doesn't get any discriminative capacity from the preparation information. As it were, there is no preparation period for it. It stores the preparation dataset and gains from it just at the hour of making continuous expectations. This makes the KNN calculation a lot quicker than different calculations that require preparing, for example, Linear Regression, and so on. K-NN can apply data from any kind of distribution. Data don't have to separable with a linear boundary-NN is pretty simple with better accuracy.

Since the KNN calculation requires no preparation before making expectations, new information can be included consistently which won't affect the precision of the calculation.

KNN is exceptionally simple to execute. There are just two parameters required to execute KNN the estimation of K and the separation work.

2.2.5 FP-Growth Algorithm

FP is the improved version of the prior method. FP is affected without the need for candidate generation. Its represent the database in the form FP tree. Here are 2 steps:

This takes place in 2 steps:-

- It reduces the data database showing repeat thing set into FP-tree. FP-tree is collecting using 2 passes on the dataset.
- The are peacock tree into a lot of the restrictive dataset and mines them specifically. Along these lines remove the recurrence thing set from FP-tree. Info: developed FP-tree.

OUTPUT: the total arrangement of continuous examples.

Strategy: Call FP-development (FP-tree, invalid).

System FP-improvement (Tree, α) {

- in the occasion that Tree contains a singular route P, by then
- For each blend do deliver structure β U α with help = least help of center points in β.
- Else for each header AI in the header of Tree do {
- Generate plan β = AIUα with assistance = AI.support;
- Construct β.s prohibitive model base and after that β.s prohibitive FP Tree β
- If Tree β = invalid
- Then call FP-improvement (Tree β, β)} } [4]

2.2.6 The execution process of FP method

Advantages of FP growth algorithm:

- FP growth algorithm is faster than any other algorithm, only 2passes over data set.
- It's more useful when someone works with a lot of data; FP compressed the dataset using FP tree.
- In the FP growth algorithm no candidate generates.
- The data is very self-supported, each node needs the root.

Disadvantages of FP growth algorithm:

- FP tree is not fit in memory.
- FP growth algorithm bit of expensive.

2.2.7 Decision Tree

The choice tree calculation goes under administered learning. We can utilize take care of both relapse and grouping issues. It relies upon the hub and hub relies upon the root. The choice tree contains hubs, branches, and a leaf. From the start hub on the tree or the top hub called root. Root is likewise a hub however its fundamental hub that why it's called root. Every single hub associated with a solitary or plural hub utilizing branches and end hubs in the tree that contains no cordial, it's called a leaf hub. The leaf hub demonstrates the result esteems. The choice breaks a major informational index in littler subsets that is the reason it's anything but difficult to work with. It works the information with two extra subsets. All the isolated focuses to see the best split among all the split focuses utilized. The data split recursively in their group and make an example which is a sommelier. In this preparing, exceptions don't affect their exhibition. They are not supporting Euclidian separation. Additionally, highlight scaling; we can locate inappropriate suspicions.

The tree will deal with every single clear cut and numerical factors as information. During this model element variable and the objective, variable work stunningly, which makes the work consummately.

The choice tree indicating the danger of malignancy, the aiding of information we gathered. The informational index contains three traits, Age, Gender, Intensity of indications, one objective property. The age estimation demonstrates the malignant growth likelihood Low, Medium, High. At the point when age is roar 25 likelihood will show Low, then again, its medium and high. We can see it graphically below (fig 2.4 :(a)).

❖ Decision tree Built from data:

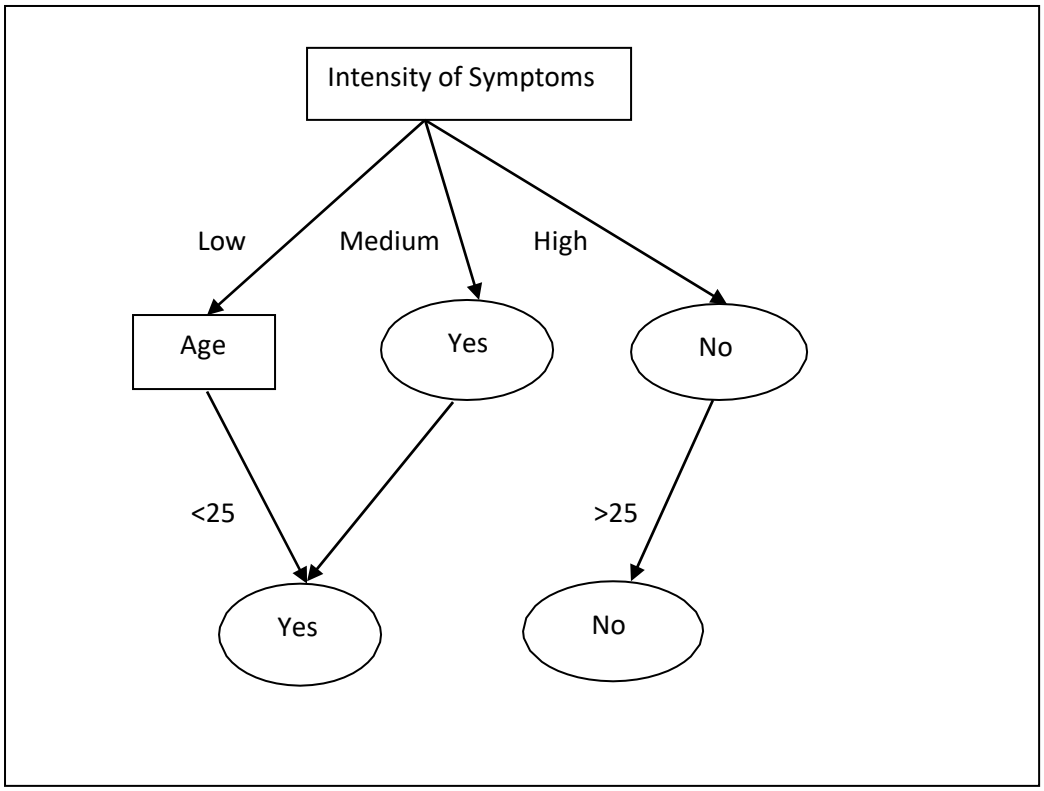


Fig 2.4 :(a) Decision tree[10]

2.2.8 The Cancer prediction Probability

This table (2.1) contains the cancer prediction probability. When we use to give the value on an algorithm, they show the result in positive and negative. We can see when the value is below 1 the result was positive on the other hand value was above 1 the result was negative. If someone gets the value below 1 his probability is high to suffer cancer. Our syndrome works with numerically. As a result below 1 showing the high probability to attract cancer.

Table 2.1: The Cancer prediction Probability

INDEX	A	B	C	D	E
1	4.6	3.2	1.6	0.4	positive
2	6	4	1.6	1.6	positive
3	6	3.4	1.5	0.4	positive
4	5.2	3.4	1.4	0.4	positive
5	4.7	3.5	1.6	0.6	positive
6	5	3.2	1.6	0.2	positive
7	5.6	3.4	1.5	0.2	positive
8	7	3.2	4.7	0.4	positive
9	6.4	3.2	4.7	1.6	negative
10	6.9	3.1	4.9	1.5	negative
11	5.6	2.3	4	1.5	negative
12	6.5	2.8	4.6	1.3	negative
13	5.7	2.8	4.5	1.2	negative
15	6.3	3.3	4.7	1.8	negative
16	4.8	2.4	3.3	1	negative

2.2.9 Decision tree for above dataset

The choice tree can be utilized for obscure informational collection. Thought to push the down tree (fig 2.4 :(b)) following arrives at leaf hub as model (Age=32, Gender=female, Intensity of indications = medium, Goal =?), where "?"It does will be "yes".

In case, it will push down a comparable branch and arrive at the age hub. At that point the tree tests the age an incentive from information. On the off chance that answers 30 again pushed down through a comparable branch.

Now the example arrives at the leaf hub, where it is named yes.

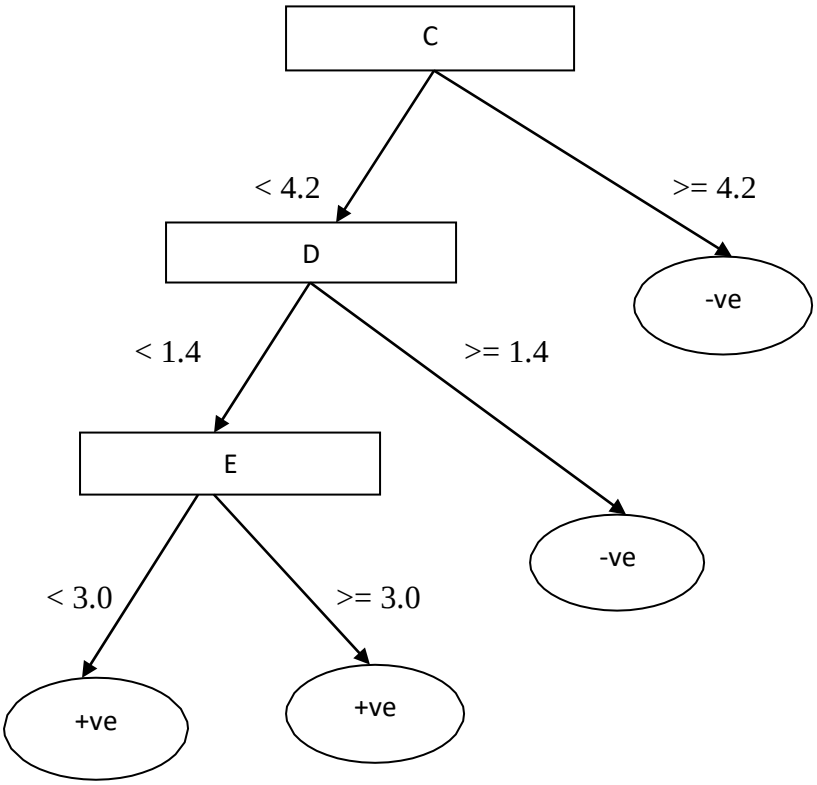


Fig 2.4 :(b)Decision tree[13]

This table contains the data set of our collection data. It's showing the syndrome level. Which age to suffer cancer possibility is higher its shows this. We can see age above 45 is a high possibility of suffering cancer also for males and females. On the other hand, early age, possibility goes down at low. At the result sometimes age matters to attack cancer. But if it catches at the right time it can be possible to avoid.

Table 2.2: Data Set

1) Age	Data Set		
	2) Gender	3) Intensity of Symptoms	4) Disease(goal)
30	Male	Medium	Yes
25	Male	Low	No
32	Female	Medium	Yes
44	Female	High	Yes
21	Female	Low	No
18	Male	Low	No
34	Female	Medium	Yes
60	Male	High	Yes
45	Male	High	Yes

2.3 Research Summary

In our research, we picked up many data from cancer patients. We talked to them. We also talked with the doctor. The doctor gives some vital information about cancer that we don't know before. We saw cancer patient's prescriptions. We learned some points from them. In data collecting, we go to some hospitals all after work we got some data. After that, we pre-processed the data with the algorithm. In that process, we need to erase some unwanted data. Then all the data ready for applied in the algorithm. We applied FP-Growth & K-NN algorithm. We are also using a decision tree.

Then we draw a flow chart for our working process. We followed the instruction of the flow chart. This model we applied to our project. We are making an application using that data set. This application should help us to detect cancer. We have all done our research implementation on Prediction.

2.4 Scope of the Problems

Our task is generally to predict the probability of cancer. If people are suffering cancer our app should aware of them before. In this working process, we need to follow some systematic way. For this grading purpose, we took different data that are overlapped with each other. We could not find our real probability to predict cancer. In our framework, people can easily satisfy if the probability was low but it's not good for their future. So the data implementation is very important for this work, it can show a high percentage result. People must be aware when they put data on the app if they put some wrong data results will not come purely. In that case, they can be misguided. Android mobiles can assume an essential job to advance our project and public awareness, also to make sure we have reliable data storage access and future expansion.

2.5 Challenges

We have to stand up to distinctive sorts of difficulties in our research. It is so difficult to gather all data from cancer patients and doctors. Some patients are not hospitable with us; they don't like to share information about their disease. So it's really challenging to convince them. We need to care full of data. Our topics were medical sites we don't have much knowledge about it. We need to learn this. That the biggest challenge we faced. Preparing the model of perfect requirement with appropriate data is also a challenge. Is our application work properly or not, find bugs. Also we should restore the data for future implementation is our main challenge.

2.5.1 Unsupervised Vs. Supervised Learning

There are two types of learning:

- supervised
- unsupervised
- Unsupervised learning: Kind of technique that operates by trying to find hidden structures in unlabeled data and also tries to understand data. It's not looking for something specific. It does not require labeled data. Unsupervised method of learning is less complex.
- Supervised learning: In supervised learning, the input data used is labeled. The supervised method of learning is complex. Calcification: KNN, Tress, Naive Bays. It's used for creating a decision function so that new data is classified based on the training set.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

The dataset used for this paper is publicly available and was created from public data. When some suffer from cancer, which kind of syndrome shows on their body and how they feel at that time all are included here. Here we discussed our working process, how we faltering the data. The program uses a curve-fitting algorithm to computer features from each one of the cells in the sample and calculates the value. The dataset contains a lot of cases. In this process its works fluently.

3.2 Research Subject and Instrumentation

The first objective was to attain a dataset of numerical values of various instances. We focus on the outcome, not outputs and treat out data as a critical asset and make decisions based on evidence. We disagree without being disagreeable, we always think about delivery end to end. We did our work ethically, honestly and deliver things that act in our user's best interests. After finalizing our dataset, we split the train test ration 80:30 in order to train an algorithm: Logistic Regression, Linear Discriminate Analysis, Gaussians, Gradient boosting Classifier, Ad Booster classifier, Decision Tree, K-Neighbor.

Here the exact accuracy results of different types of algorithm which we use in our project. In here we basically see the accuracy. Graphical view shows that accuracy point of itch algorithm (fig 3.1).

When we don’t want to see the code that time we can understand this project by viewing this graph how accrete it is this the important part of our project.

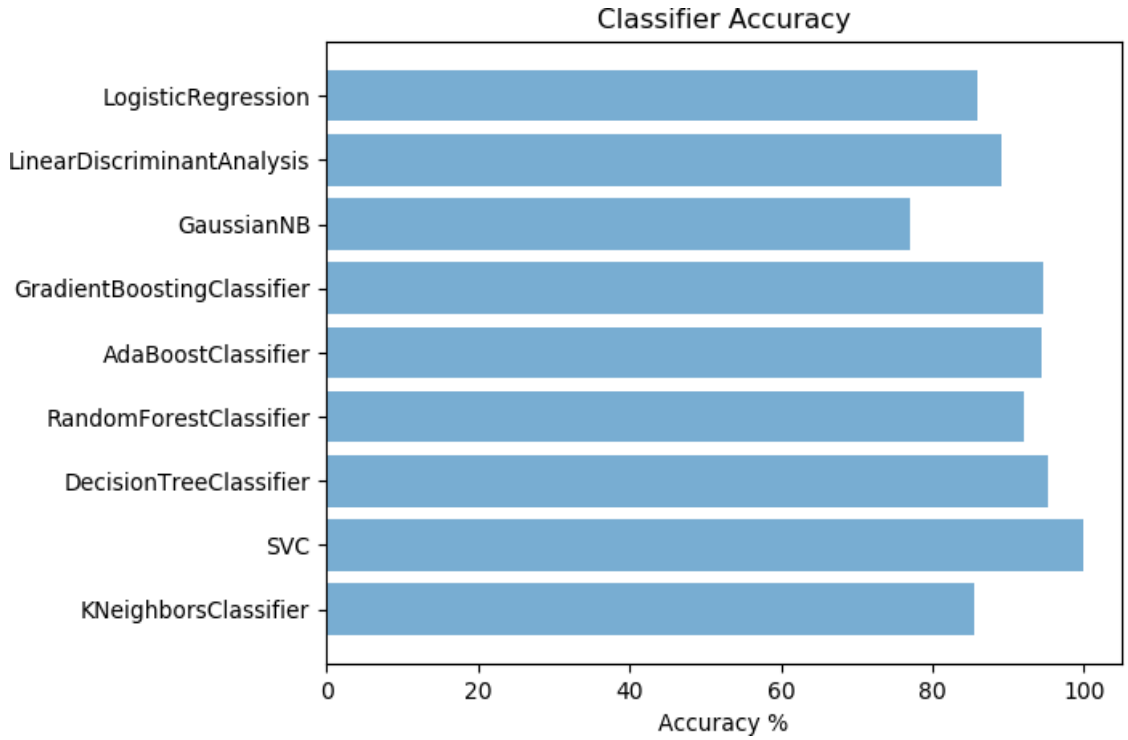


Fig 3.1: Classifier accuracy

3.3 Data Collection Procedure

Excellent information was required for acknowledging the best outcomes; it was significant in this way that its securing is profoundly dependent on the nature of the information accumulation process. We collect data from the hospital. Data in these databases were collected from the clinical environment, and have undergone proper organizational ethics approval processes. At first, we collect cancer-related data from a cancer hospital. Then we have select data that we work with. Apple and approach all data. After that, we evaluated data and visualization of our results.

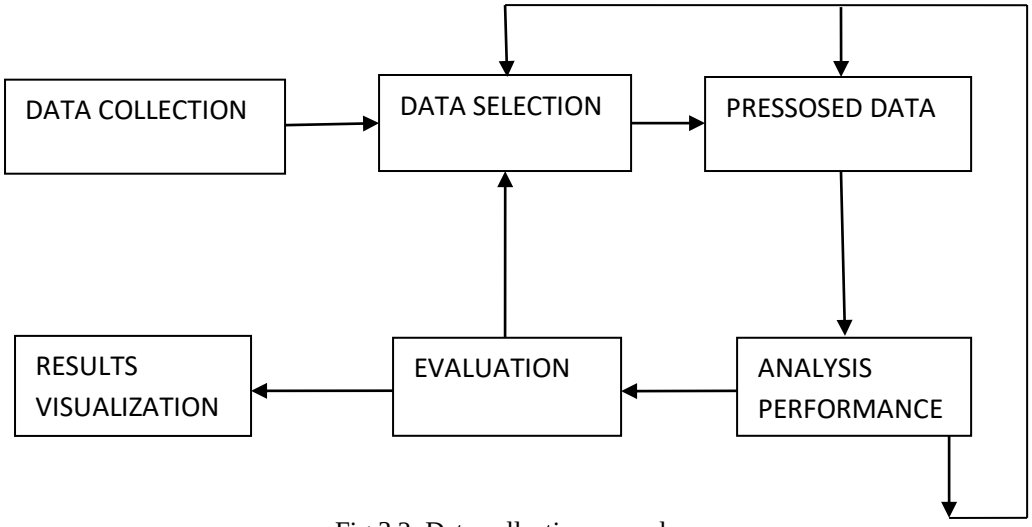


Fig 3.2: Data collection procedure

Here we describe the data flowchart (fig 3.2). How we collect data from various sources. That selects data which we take or which one we delete. Than we processed data which data we take than allays it finally visualize it. This is the process of data collection.

	Name	Gender	age	caugh with blood	chest pain	caugh thick	sudden growing and pain	shape change	left pain	burning	nipple rotten	infection	burning.1	be smooth and parly	rash	heavy sweating ,vomiting	ittchy skin	lose of nusia	extrime rash
0	Mamun	1	34	2	2	2	2	2	5	2	2	5	5	5	5	5	5	5	5
1	Rima	2	29	5	5	5	2	2	2	2	2	2	5	5	5	5	5	5	5
2	Alash	2	25	5	5	5	5	2	5	5	2	5	2	5	2	5	2	5	5
3	Tarif	1	56	5	5	5	5	5	5	5	5	5	5	5	5	2	2	2	2
4	Maruf	1	57	5	2	5	5	5	5	5	5	5	5	5	2	2	2	2	2

Fig 3.3: Collect data

3.4 Requirement analysis (Proposed)

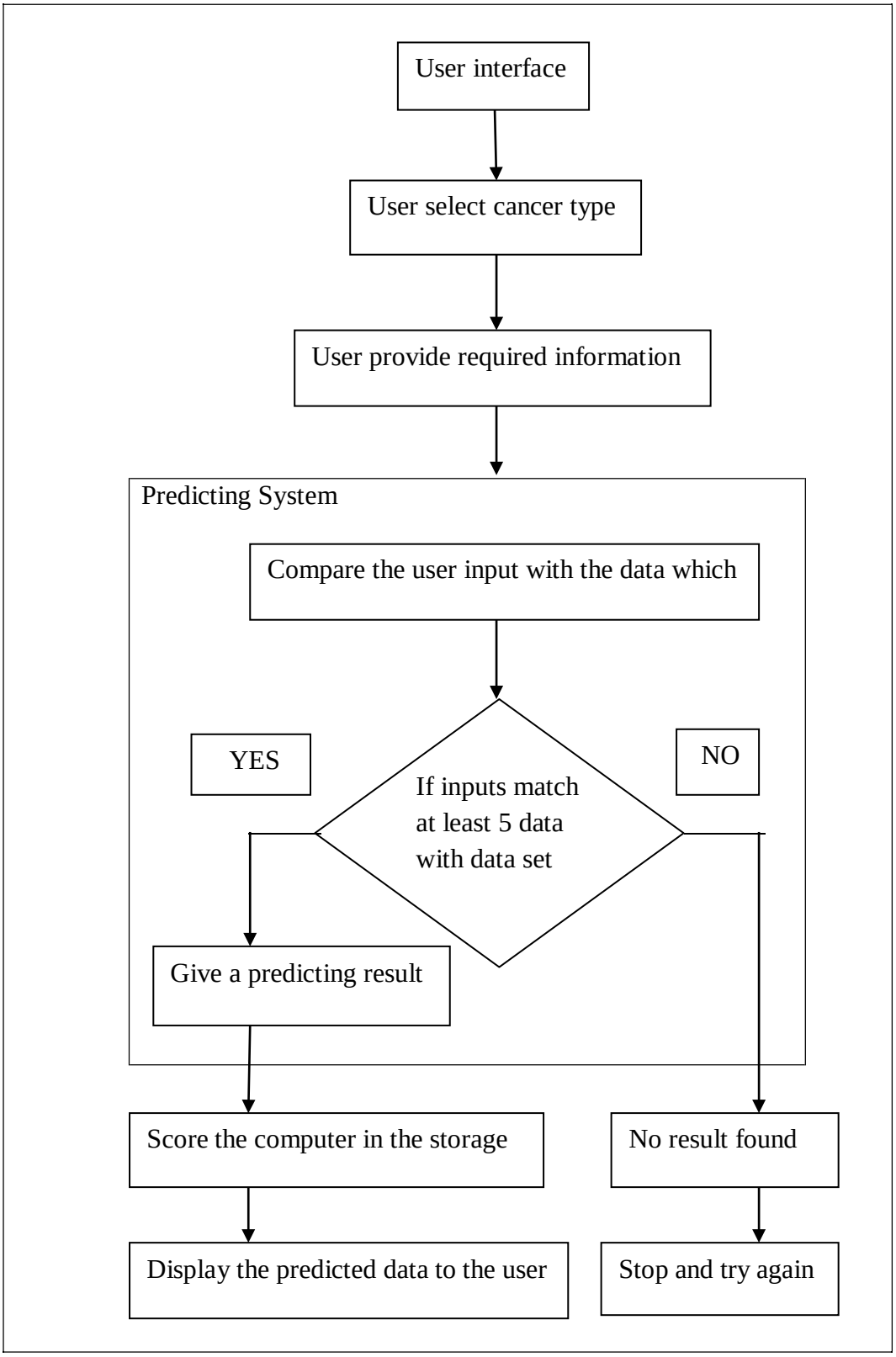


Fig 3.4: Requirement analysis

First of all user see the (fig 3.4) interface of our project than then user entered it. They see many type of cancer. User chose one type cancer and come into it .they give is symptom which requirement of our application. User must be give us proper information .than the given data match with our database than which data are similar application collect it. When at list five data are similar with our database and user symptom than time it gives a prediction result which in percentage. If minimum five data are not similar that time it will be showed that no results found and take back to user interface. This is how project is work.

3.5 Implementation Requirements

After an informational analysis of all features doing data collecting and implement, there is a requirement to complete our work. Our requirement helps to generate some algorithms to recognize cancer data properly.

Hardware/Software Requirements:

- Operating System (Windows 10)
- Language-PYTHON
- Laptop:
  - HP(8GB ram)
  - Daffodil PC (4GB ram)
  - Lenovo (8GB ram)
- GPU:
  - Nvidia
  - Amd
- Developing Tools:
  - JupiterNotebook
  - Spider
  - Anaconda

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

Cancer is a deadly disease and has over the years given the medical practitioners' sleepless nights trying to find effective, accurate and reliable ways of diagnosing it. In this chapter, we mainly target on how we use and methodology for Proposed Intelligent Model which involves the following: first, data preprocessing second classification engine third model testing. That's the procedure we follow. To include the ultimate result, to begin with, the crude data were collected from the hospital. We investigated some cancer patients also. We focus on the implementation data and the validity of classification.

4.2 Experimental Results

We trained our data using some classification. Those working processes are different and we got a different type of validity. We also implement some algorithm.

4.2.1 Fitting Dataset for classifier

To build a model using the pre-train model we divide our data into cancer type. We use a hundred people cancer data. We collect all types of cancer information. It is basically two types:

- For the training set, we use 80 percent data.
- For test data set we use 20 percent data

Here is our scattered data which we collect from various types of cancer patient. It's not well recognized and decorated for our work it also an untrained data. We collect it from various cancer patients from various hospitals. Here we collect many types of cancer symptom like Lange, Brest, skin and many more. We took it randomly.

❖ Raw data set

NAME	MALE	FEMALE	CANCER_ TYpe	AGE	CANCER_SYMPTOM	TEST	cancer_tes
tahir	m		laungh	28	ccaughwith blood		cbc
hashmot	m		laungh	25	chest pain		cbc
oli	m		laungh	55	medium chest pain		cbc
mahmud	m		laungh	67	caugh thick		cbc
tahira		f	breast	76	left side pain		cbc
asha		f	breast	43	suddeb growing andpain		cbc
riba		f	breast	32	lleft pain		cbc
ankue		f	breast	34	burning		cbc
usuf	m		blood	45	urine with blood		cbc
atik	m		blood	32	urine with blood		cbc
bonna		f	ovarian	34	burning and blood		cbc
bimola		f	breast	56	pain		cbc
bokitiar	m		skin	27	infection		cbc
bivah		f	ovarian	28	infection		cbc
rima		f	blood	21	blood clothing		cbc
ridita		f	breast	24	abnormal growth		cbc
roy	m		skin	18	infection		cbc
rokib	m		laungh	23	caugh thick		cbc
rakibv	m		blood	32	blood in nose		cbc
nabil	m		skin	43	indection		cbc
pabna	m		laungh	34	breathing problem		cbc
pranto	m		laungh	51	breathing problem		cbc
sharif	m		laungh	55	caugh thick		cc
shorif	m		laungh	56	chest pain		bc
shuborna		f	ovarian	12	blood flow suddenly		cbc
dulal	m		skin	13	infection		cbc
rahim	m		skin	74	infection		cbc
karim	m		stomac	43	pain		cbc
zayed	m		laungh	32	heavy caugh		cbc
ononto	m		laungh	33	caugh with blood		cbc
kausar	m		blood	15	blood flow by year		cbc
kaysar	m		blood	15	blood flow by nose		cbc
koli		f	breast	16	shape change and dull		cbc
kumli		f	breast	32	colour change		cbc
kobir	m		stomac	29	heavy pain in balley		cbc
koran kiro	m		laungh	20	burning		cbc
koili		f	skin	26	burning		cbc
kebla		f	breast	39	nipple discharged		cbc
kochi		f	ovarian	30	discomfort in pelvis		cbc

Fig 4.1.1: Raw data

urmi		f	blood		39	sortness of breath		cbc
karbi	m		skin		43	be smooth and parly		cbc
korobi		f	stomac		45	fell full stomac while taking foo		cbc
koboz	m		skin		42	felli ng icchi and bleed sometim		cbc
azan	m		laungh		15	hoeseness		cbc
amit	m		stomac		19	pain		cbc
riadf	m		skin		17	itchi		cbc
ridoy	m		stomac		34	vomiting		cbc
hridoy	m		throught		42	pain		cbc
shamoli		f	blood		42	vomiting		cbc
jhorna		f	breast		25	pain in nipples		cbc
jolsha		f	blood		26	heavy sweating		cbc
jhotika		f	breast		28	find a lump		cbc
jibon jiona	m		skin		65	burning		cbc
jotil	m		laungh		54	weight lose in appitite		cbc
terjo arzu	m		laungh		43	worst chest pain		cbc
arafat	m		laungh		27	worst chest pain		cbc
abir	m		laungh		1	worst chest pain		cbc
salauddin	m		throught		74	find infection		cbc
shail;a		f	throught		65	find infection		cbc
shayla\		f	throught		54	find infection		cbc
sifat	m		skin		43	infection		cbc
arif	m		stomac		34	vomiting		cbc
jojo	m		stomac		24	pain		cbc
ripon	m		throught		43	blood flow		cbc
rima		f	ovarian		60	infection		cbc
rumpa		f	breast		64	hight fever		cbc
ridita		f	breast		61	nipple rotten		cbc
shoma		f	blood		54	ittchy skin		cbc
shabik		f	blood		23	rash		cbc
sadik	m		blood		16	extrime rash		cbc
sadek	m		throught		17	pain		cbc
showmen	m		skin		14	develop a crust of scab		cbc
siam	m		throught		16	change in your voice		cbc
soikot	m		throught		35	sore throught		cbc
shoikotr	m		throught		34	pain		cbc
afza	m		throught		32	infection		cbc
bibizan		f	throught		16	wheezing		cbc
nilima		f	breast		53	nipple infection		cbc
nitol	m		stomac		34	fatig		cbc
taimur	m		stomac		26	felling bloated		cbc
tonmouy	m		throught		32	voice change		cbc
faisal	m		throught		23	pain		cbc
shumon	m		laungh		64	persistent heartburn		cbc

Fig 4.1.2: Raw data

4.2.2 Trained Dataset

We trained our data set using two models and also different ways. Those models are based on a high learning process. We ready our implemented data. We are putting attributes on working data. People should write their name and gender this is compulsory for all. Then they should infinity their age. We also input some syndrome related to cancer. If anyone affected with this syndrome, then probability will high for cancer.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	Gender	age	caugh with blood	chest pain	caugh thick	sudden growing and pain	shape change	left pain	burning	nipple rotten	infection	burning
2	Mamun	1	34	1	1	1	1	1	0	1	1	0	0
3	Rima	2	29	0	0	0	1	1	1	1	1	1	0
4	Alash	2	25	0	0	0	0	1	0	0	1	0	1
5	Tarif	1	56	0	0	0	0	0	0	0	0	0	0
6	Maruf	1	57	0	1	0	0	0	0	0	0	0	0
7	Apon	1	30	1	1	1	0	0	0	1	0	0	0
8	Asma	2	42	0	0	0	1	0	1	1	1	0	0
9	Meem	2	38	0	0	0	1	0	1	1	1	0	0
10	Walid	1	23	0	0	0	0	0	0	1	0	1	0
11	Renzina	2	45	0	0	0	0	0	0	1	0	1	0
12	Farid	1	39	1	0	0	0	0	0	0	0	0	0
13	Rashid	1	30	1	0	0	0	0	0	0	0	0	0
14	Zakir	1	28	1	0	0	0	0	0	0	0	0	0
15	Alamgir	1	37	1	0	0	0	0	0	0	0	0	0
16	Samia	2	26	0	0	0	1	1	1	1	1	0	0
17	Nasrin	2	21	0	0	0	0	0	0	1	0	1	0
18	Uzzal	1	52	1	1	1	0	0	0	1	0	0	0
19	Shihab	1	34	1	1	1	0	0	0	1	0	0	0
20	Afruza	2	21	1	1	1	0	0	0	1	0	0	0
21	Sumi	2	47	0	0	0	0	0	0	0	0	1	1
22	Shuvo	1	26	0	0	0	0	0	0	0	0	1	1
23	Priyanka	2	29	1	0	0	0	0	0	0	0	0	0
24	Zarin	2	56	1	0	0	0	0	0	0	0	0	0
25	Rony	1	30	1	1	1	0	0	1	1	0	0	0
26	Mahmudul	1	35	1	1	1	0	0	1	1	0	0	0
27	Rabeya	2	37	0	0	0	0	0	0	1	0	1	0
28	Belayet	1	40	0	0	0	0	0	0	1	0	1	0
29	Jesmin	2	27	0	0	0	1	1	1	1	1	0	0
30	Saiful	1	36	0	0	0	1	1	1	1	1	0	0
31	Sharif	1	29	0	0	0	1	1	1	1	1	0	0
32	Iffat	1	27	0	0	0	0	0	0	0	0	1	1
33	Sabrina	2	38	0	0	0	0	0	0	0	0	1	1
34	Tumpa	2	40	0	0	0	0	0	0	0	0	0	0
35	Hasnat	1	51	0	0	0	0	0	0	1	0	1	0
36	Safia	2	39	0	0	0	0	0	0	1	0	1	0
37	Ismat	1	70	0	0	0	1	1	1	1	1	0	0
38	Tasnin	2	53	0	0	0	1	1	1	1	1	0	0
39	Jarin	1	59	1	1	1	0	0	1	1	0	0	0
40	Ariful	1	39	1	1	1	0	0	1	1	0	0	0
41	Tofayel	1	45	0	0	0	0	0	0	0	0	1	1
42	Suma	2	21	0	0	0	0	0	0	0	0	1	1
43	Habibul	1	37	0	0	0	0	0	0	1	0	1	0

Fig 4.2: Training data

When we collect data from various patients, our data set was scattered than we process our data. Than we stared train our data set and we find it this type. Fig 4.2 is our trained data from scattered data. It's very important part of our project.

In this section we will the show classifier how used algorithm work and also accuracy. In here we used KNneighbor'sclassifier. It will take data from our implemented data. Then print results with accuracy. Here also we use Gaussian Boost classifier many of other for proper accuracy.

```
classifiers = [
    KNeighborsClassifier(5),
    SVC(C=.1, degree=1, kernel='poly', probability=True),
   # NuSVC(nu=.1, degree=1, kernel='poly', probability=True),
    DecisionTreeClassifier(),
    RandomForestClassifier(),
    AdaBoostClassifier(),
    GradientBoostingClassifier(),
    GaussianNB(),
    LinearDiscriminantAnalysis(),
    LogisticRegression()]

# Logging for Visual Comparison
log_cols=["Classifier", "Accuracy", 'Log Loss']
log = pd.DataFrame(columns=log_cols)
#clf = MultinomialNB()
for clf in classifiers:
    clf.fit(X_train, y_train)
    name = clf.__class__.__name__

    print("="*30)
    print(name)
    try:
        print('****Results****')
        train_predictions = clf.predict(X_test)
        acc = accuracy_score(y_test, train_predictions)
        print("Accuracy: {:.4%}".format(acc))
#         print("accuracy_score: \n",accuracy_score(y_test, train_predictions))
        print("precision_score: \n",precision_score(y_test, train_predictions))
        print("f1_score: \n",f1_scorey_test, train_predictions)
#         print("classification_report: \n",classification_report(y_test, train_predictions))
        print("confusion_matrix: \n",confusion_matrix(y_test, train_predictions))
        print("log_loss: \n",log_loss(y_test, train_predictions))
        log_entry = pd.DataFrame([[name, acc*100, log_loss(y_test, train_predictions)]], columns=log_cols)
        log = log.append(log_entry)
    except Exception as e:
        print (e)

print("="*30)
```

Fig 4.3: KNN classifier

Here our code how it work we can see their graphically (fig 4.3). Basically here we can see our main targeted algorithm k nearest neighbor how works. Here also have full code of k-nn. This is the main algorithm of our works that is why we give here our code and its working procedure.

This is the accuracy result for our work. Here we have given classifier accuracy and how algorithms are works. This highest result. Here we use many algorithm and classifier we see that individually how much effective our data set and how much accurate. When we test our project we use many algorithm and many classifier for knowing its result. Like we use svc algorithm, k nearest neighbor algorithm, decision tree classifier, random forest classifier many of them. Actually we want to know the repot of percentage that is why we use many types of algorithm here the result of our project (fig 4.4).

```
==============================
KNeighborsClassifier
****Results****
Accuracy: 85.7143%
Target is multiclass but average='binary'. Please choose another average setting.
==============================
SVC
****Results****
Accuracy: 100.0000%
Target is multiclass but average='binary'. Please choose another average setting.
==============================
DecisionTreeClassifier
****Results****
Accuracy: 95.2381%
Target is multiclass but average='binary'. Please choose another average setting.
==============================
RandomForestClassifier
****Results****
Accuracy: 100.0000%
Target is multiclass but average='binary'. Please choose another average setting.
==============================
AdaBoostClassifier
****Results****
Accuracy: 100.0000%
Target is multiclass but average='binary'. Please choose another average setting.
C:\Users\rafi\Anaconda3\lib\site-packages\sklearn\svm\base.py:196: FutureWarning: The default value of gamma will change from
'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoid t
his warning.
  "avoid this warning.", FutureWarning)
C:\Users\rafi\Anaconda3\lib\site-packages\sklearn\ensemble\forest.py:246: FutureWarning: The default value of n_estimators will
change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
==============================
GradientBoostingClassifier
****Results****
Accuracy: 100.0000%
Target is multiclass but average='binary'. Please choose another average setting.
==============================
GaussianNB
****Results****
Accuracy: 100.0000%
Target is multiclass but average='binary'. Please choose another average setting.
```

Fig 4.4: Accuracy of result

This portion is a possibilities report graph. Here we see that a curb line goes to down from up .it means those different types of cancer depends on shape and color. The line shows us possibilities of cancer depend on data.

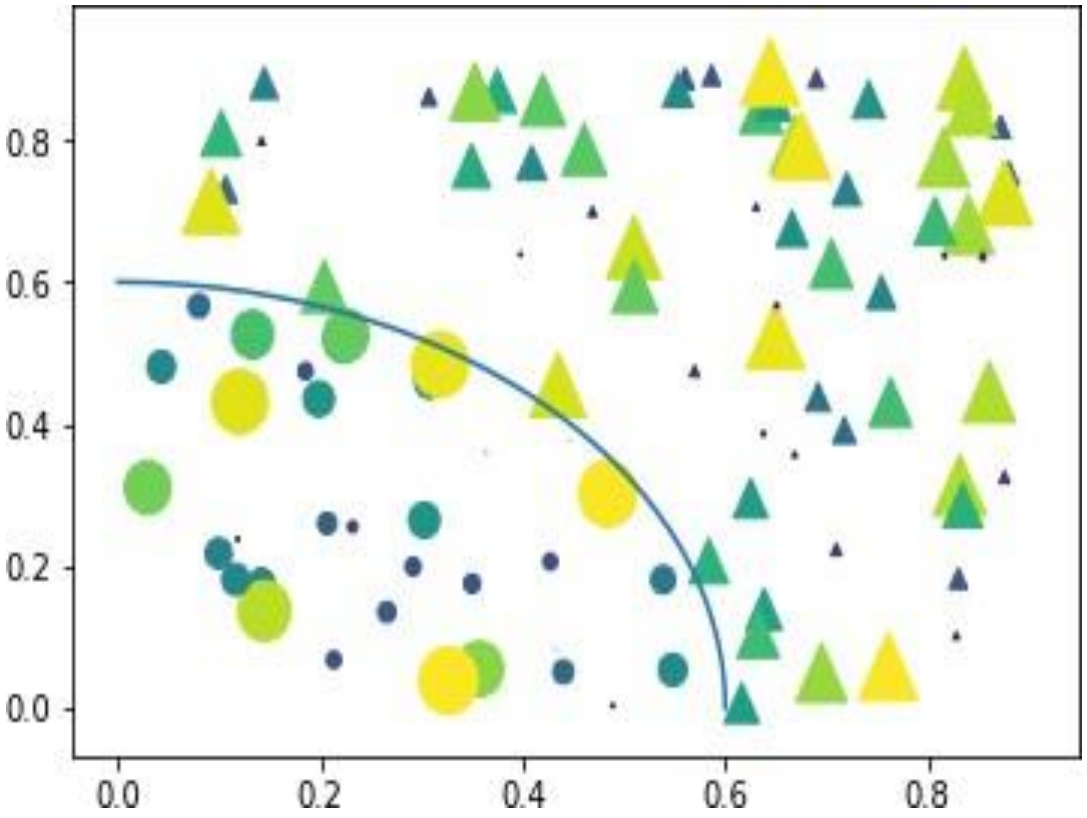


Fig 4.5: Possibility of cancer

When a user gives his data as a symptom he wants to know that how much possibility to grow cancer of his body. The graphical view shows this in graphical way. Here we saw that different color. Its each color signifies the different types of percentage. That's the reason of this graph; it's very important and crucial part of our project (fig 4.5).

Here we see that the correlation (fig 4.6) of different event means when symptom of a patient and reserved data are same that time it show dark blue color. Here three different types of color dark blue, medium blue, light blue, where what color will be showed it depends on input symptoms and data which are reserve.

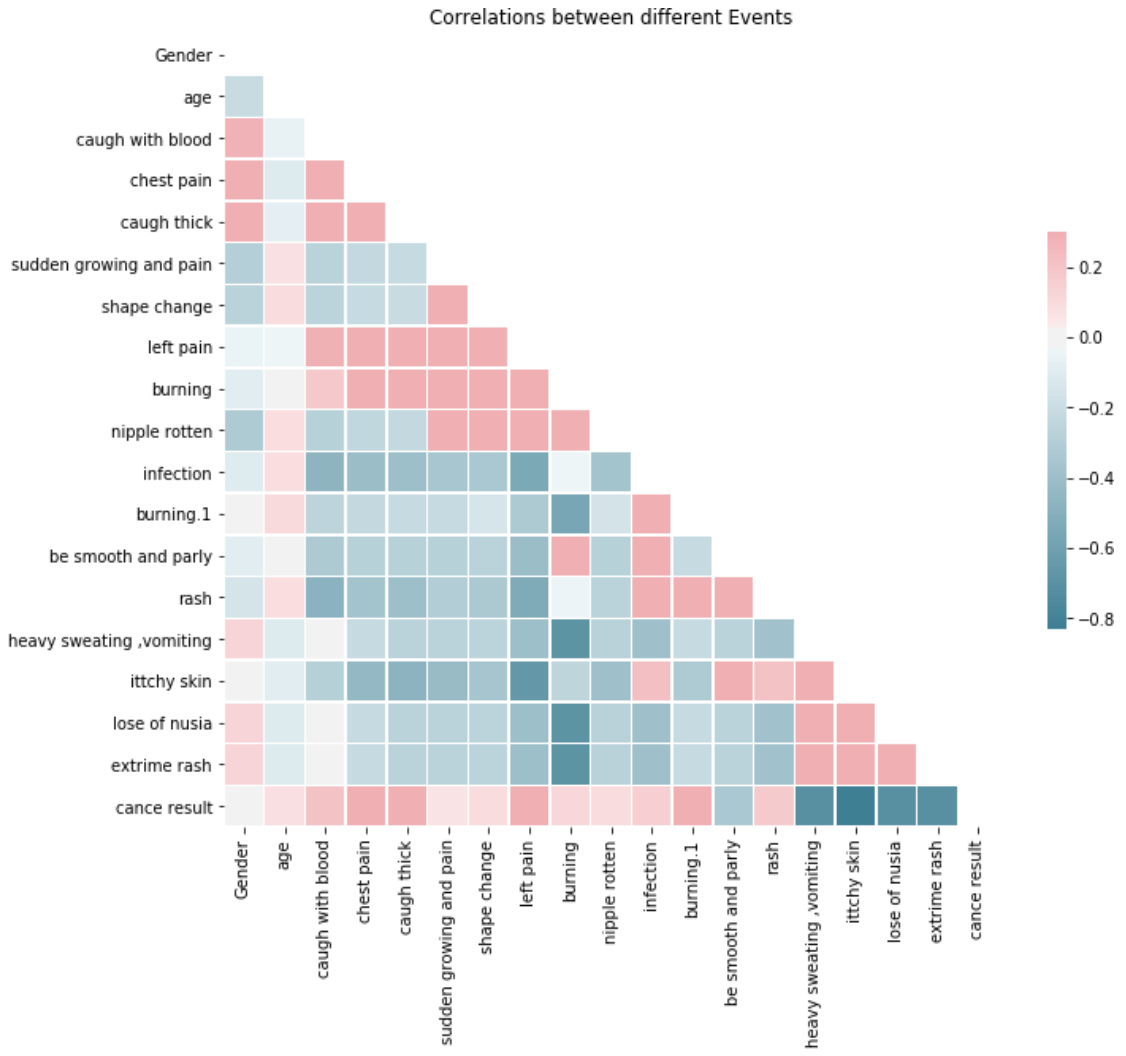


Fig 4.6: Accuracy of cancer

4.3 Descriptive Analysis

After complete, our work we see that all models are giving nearly the equivalent validity.SVC and Random cultivate set aside an excessive amount of effort for preparing and need high arrangement pc or workstation and other successive model needn't bother with high design PC and give better precision.

We utilize three model for preparing information at that point test this where we got practically 100% precision (fig 4.7) since we working with little information. In our examination, we can find that on the off chance that we train our information additional time every single model give better exactness and better yield. On the off chance that we see the table we can see, quicker Random Forest, Adam Boost Classifier, Gradient Boosting, Gaussian NB have higher exactness. K-Neighbors Classifier has lower precision. Direct Discriminate Analysis has medium lower precision. Be that as it may, at last, we can say all models are given a similar precision.

ALGORITHM	ACCURACY
K-Neighbors Classifier	85.7143%
SVC	100.0000%
Decision Tree	95.2381%
Random Forest	100.0000%
AdaBoostClassifier	100.0000%
Gradient Boosting	100.0000%
GaussianNB	100.0000%
LinearDiscriminateAnalysis	95.2381%
LogisticRegression	100.0000%

Fig 4.7: Algorithm and accuracy

4.4 Summary

Subsequent to getting legitimacy Faster K-NN and choice tree model is most solid for our exploration work. However, we can recommend that if need disease recognizes progressively or distinguish more than one ought to be utilize Faster K-NN. If we want to identify cancer with the small data set we should use Random Forest. But this may not give better output. One the other hand Linear Discriminate Analysis can work faster and give proper accuracy. Logistic Regression, Ada Boost Classifier this model also people can use for cancer detection. But in the end, we recommend using K-NN this provided the best accuracy.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the Study

In this chapter, we talk about our future work, some limitation and future implication of our work. We gave a diagram of the procedures of KDD. Additionally, we presented information mining

As an equivalent word of KDD and talked about in detail three calculations in arrangements zone, SVC, KNN and KNN and affiliation rules with FP-GROWTH. We used Cancer detection to help people. We use many models some give us higher accuracy, on the other hand, some give us low but they work properly. Some models can work with a small amount of data and some models need a big data set. We are showing the accuracy rate in the previous chapter. K-NN is faster to use cancer-detecting. Our work should be converted with mobile application.

5.2 Conclusion

Our project can predict cancer diseases. This framework processing data IV the most important and hardest part. Our main objective is to create an autonomic process that can detect cancer before it harms hardly. This framework is built with some important features added that could help people a various way. Our future work should concentrate on the expense of calculation. This is supposing that a computational approach is less expensive and can be able to create the best results.

5.3 Recommendation

- Vast data set will give better outturn on this prediction.
- The syndrome should give properly.

- Should use powerful device for training.
- Trained steps will give best validity.
- Future work should concentrate on the expense of calculation.

5.4 Implication for further Study

- To get better accuracy should train more time and collect more data, big train data set to give better outcomes.
- It can be converted in a mobile application, people can use easily.
- Medicine's name can be an add-in future.
- Doctor advise also can be added, in future, we add doctor number address, that's people can reach them easily.
- The six algorithms we used on this model, we wish to confirm the results we obtained are not biased thanks to the scale of our dataset. We would to like find out an even largest dataset and perform a similar analysis and see if the results are identical.

5.4.1 Limitation

- Our model can predict cancer only; it cannot suggest medicine for people.
- It's also cannot tell how many times suffer from cancer.
- The case study does not take a stance either on other types of cancer or on treatment.
- It does not detect the stage of cancer.
- It's fully dependent on the syndrome if syndrome was wrong it will give the wrong result.
- The case study does not take a stance either on other types of cancer or on treatment.
- The core subject is the medical data play a vital role here; it cannot detect a new type of cancer.

REFERENCES

[1] Breast cancer facts and figures 2003-2004 (2003). American Cancer Society.

[2] Stages | Mesothelioma | Cancer Research UK Breast cancer survival statistics September 26, 2017

[3] Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17: 223-232.

[4] deepsense.ai What is reinforcement learning? The complete guide July 05, 2018

[5] Hacker Noon Absolute Fundamentals of Machine Learning – Hacker Noon January 15, 2018

[6] Furundzic, D.; Djordjevic, M.; Bekic, A.J. Neural networks approach to early breast cancer detection. J. Syst. Archit. 1998, 44, 617–633. [CrossRef]

[7] Floyd, C.E.; Lo, J.Y.; Yun, A.J.; Sullivan, D.C.; Kornguth, P.J. Prediction of breast cancer malignancy using an artificial neural network. Cancer 1994, 74, 2944–2948. [CrossRef]

[8] Fogel, D.B.; Wasson, E.C.; Boughton, E.M. Evolving neural networks for detecting breast cancer. Cancer Lett. (1995), 96, 49–53. [CrossRef]

[9] Fogel, D.B.; Wasson, E.C.; Boughton, E.M.; Porto, V.W.; Angeline, P.J. Linear and neural models for classifying breast masses. IEEE Trans. Med. Imaging (1998), 17, 485–488. [CrossRef] [PubMed]

[10] Setiono, R. Extracting rules from pruned neural networks for breast cancer diagnosis. Ar-tif. Intell. Med. (1996), 8, 37–51. [CrossRef]

[11] Wilding, P.; Morgan, M.A.; Grygotis, A.E.; Shoffner, M.A.; Rosato, E.F. Application of backpropagation neural networks to diagnosis of breast and ovarian cancer. Cancer Lett. (1994), 77, 145–153. [CrossRef]

[12] Wu, Y.; Giger, M.L.; Doi, K.; Vyborny, C.J.; Schmidt, R.A.; Metz, C.E. Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. Radiology (1993), 187, 81–87. [CrossRef] [PubMed]

[13] H. A. Abbass, "An evolutionary artificial neural networks approach for breast cancer diagnosis." Artif.Intell. Med., vol. 25, no. 3, pp. 265–81, Jul. 2002.

[14] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.," Nat. Med., vol. 7, no. 6, pp. 673–9, Jun. 2001.

[15] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and ap-plications," Neurocomputing, (Dec. 2006.) vol. 70, no. 1–3, pp. 489–50

Test-2

ORIGINALITY REPORT

17%	3%	1%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	7%
2	Submitted to Aliah University Student Paper	3%
3	Submitted to Turun yliopisto Student Paper	1%
4	www.geeksforgeeks.org Internet Source	1%
5	Submitted to Griffith College Dublin Student Paper	1%
6	machinelearningmastery.com Internet Source	1%
7	Submitted to Birla Institute of Technology and Science Pilani Student Paper	1%
8	Submitted to University Politehnica of Bucharest Student Paper	1%
9	Submitted to Deakin University	

	Student Paper	<1%
10	Submitted to Middlesex University Student Paper	<1%
11	Mohammed H. Tafish, Alaa M. El-Halees. "Breast Cancer Severity Degree Predication Using Data Mining Techniques in the Gaza Strip", 2018 International Conference on Promising Electronic Technologies (ICPET), 2018 Publication	<1%
12	eprints.rclis.org Internet Source	<1%
13	Submitted to Botswana Accountancy College Student Paper	<1%
14	Submitted to Bridgepoint Education Student Paper	<1%
15	Rini Anggrainingsih, Nach Rowi Khoirudin, Haryono Setiadi. "Discovering drugs combination pattern using FP-growth algorithm", 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017 Publication	<1%
16	Submitted to Universiti Malaysia Pahang Student Paper	<1%

17	ijirt.org Internet Source	<1%
18	Submitted to Oxford Brookes University Student Paper	<1%
19	Submitted to Caledonian College of Engineering Student Paper	<1%
20	scholar.uwindsor.ca Internet Source	<1%
21	documents.mx Internet Source	<1%
22	Xiao-Feng Gu, Xiao-Juan Hou, Chen-Xi Ma, Ao-Guang Wang, Hui-Ben Zhang, Xiao-Hua Wu, Xiao-Ming Wang. "Comparison and improvement of association rule mining algorithm", 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2015 Publication	<1%
23	link.springer.com Internet Source	<1%
24	acims.asu.edu Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		