PREDICTION THE RISK OF BREAST CANCER FROM BODY MASS INDEX (BMI), GLUCOSE AND INSULIN.

BY

Sadia Sultana

ID: 151-15-4689

Mst. Lotifa Akter Shirin

ID: 151-15-5098

AND

Md. Saif Anwar

ID: 151-15-5126

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

Supervised By

Mr. Aniruddha Rakshit
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Md. Rayhan Amin
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2018

APPROVAL

This Project titled **"**Prediction The Risk of Breast Cancer from Body Mass Index (BMI), Glucose and Insulin", submitted by *Sadia Sultana, ID No: 151-15-4689 *,*Mst. Lotifa Akter Shirin, ID No: 151-15-5098* and *Md. Saif Anwar, ID No: 151-15-5126* 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 (BSc) and approved as to its style and contents. The presentation has been held on 11th December, 2018.

BOARD OF EXAMINERS

_____

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

_____

Dr. Sheak Rashed Haider Noori **Internal Examiner**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Md. Zahid Hassan **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Mohammad Shorif Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Aniruddha Rakshit, Lecturer, Department of CSE** in 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:

Mr. Aniruddha Rakshit
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Md. Rayhan Amin
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Sadia Sultana
ID: 151-15-4689
Department of CSE
Daffodil International University

Mst. Lotifa Akter Shirin
ID: 151-15-5098
Department of CSE
Daffodil International University

Md. Saif Anwar
ID: 151-15-5126
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty Allah 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. Aniruddha Rakshit**, **Lecturer**, Department of CSE, Daffodil International University, Dhaka. His deep Knowledge & keen interest of our supervisor in the field of "*Breast Cancer*" 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 **Almighty Allah** and **Prof. Dr. Syed Akhter Hossain, Professor and Head**, Department of CSE, for his kind help to finish our project and also to other faculty member and 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

The purpose of our research is to predict the risk of breast cancer from body mass index (BMI).Body mass index stands for, proper ratio of weight according to height. Overweight and obesity causes various types of physical complexity. Cancer is also related with this. Still now cancer is an alarming word. We just get scared after hearing this word. This disease is related with lots of physical complexity. We have to go through lots of hassles to identify the probability of cancer. Sometimes when it is identified in human body then it's become too late to recover. So we wanted to predict the risk of cancer for them who are in risk. Thus they will be able to take precautionary steps and can avoid disease like cancer. It is so an easy task, still we tried to find out the relation between obesity and breast cancer. We tried our level best to generate a valid result.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Breast cancer is the most common form of cancer for worldwide women. Though both men and women are in the risk of breast cancer. In 2017 a random study shows that almost 252,710 new cases of fatal breast cancer occurs both in men and women. Men are in risk who are above 60 years old. Women are in the risk of after age of 18. So it is clear that women are in the higher risk than men.

In this paper, we tried to focus on the case of women only. And only those cancer which is associated with higher body mass index (BMI), higher range of Glucose, and lowest range of Insulin.

Body mass index (BMI) is supposed as an important indicator of breast cancer. Every 5kg/m² increases in BMI causes 2% increases in breast cancer.

Another one is Glucose. Higher range of glucose causes extra body fat. Fat body cell produces various types of unnecessary hormone. Which can increases physical complexity.

Third one is Insulin. Islets of Langerhans are the regions of the pancreas that contain its endocrine cells. Endocrine cell produces Insulin hormone. Insulin works as an enzyme to break down Glucose in our body. If there is a lack of insulin, then Glucose increases in our body and that cause obesity. So that is the relation among them.

There are two conditions. One is premenopausal and another one is post-menopausal. It is not mandatory that an obese women will have cancer, but after age of 18 lifetime weight gaining is a significant risk sign for breast cancer. If BMI >31.1 kg/m², then there is risk factor of cancer. The relationship between body mass index (BMI or weight/height²) and breast cancer risk is modified by menopausal status.

1.2 Motivation

In current world breast cancer is increasing day by day in. It is a most common cancer for women. Over weight is considered as a significant risk factor for breast cancer. Breast cancer risk depends on menopausal status. After menopause women are in risk of breast cancer that related to obesity. There is a hormone called estrogen which carries female characteristic and also bear signal for cell division. A longer exposure to estrogen increase the fat tissue that's why increase the obesity. So every woman should have the clear concept about the relation between breast cancer and body mass index (BMI). In our paper we tried to mean that unhealthy body mass index (BMI) is the risk factor for breast cancer

The old saying “prevention is better than cure”.
As our aim is to predict breast cancer from BMI. So, if women get to know about the risk of the cancer, they can be aware about their obesity. While it is likely that obese people who reduce weight will also reduce their cancer risk. Thus precautionary methods can also be taken to prevent or stay away from this diseases.

1.3 Relative Questions

Some question arise with the term of breast cancer which is associated with overweight and obesity. There is also some factors depends on it. Some general type questions is given below-

- Is obese people must have cancer?
- Is it depends on menopausal status?
- How BMI is associated with breast cancer?
- How much people die in every year due to breast cancer?
- What is the rate of recovery from breast cancer?
- How people can recover from cancer?
- Are men is free from the risk of breast cancer?

1.4 Research Area and Possible Causes of Breast Cancer

While doing research work about possibility for breast cancer, some constant factor are considered. Results are modified with this area of study. Factors which is considered is given below-

- Age, Height, Weight.
- Geographical region.
- Family History.
- Puberty.
- Gender.
- Pregnancy Status.
- Breast feeding status.
- Smoking status.
- Alcohol intake.
- Physical activity (lack of exercise).
- Years since menopause.
- Using of hormone replacement therapy after menopause.

1.5 Expected Outcome

We all know overweight is a source of various physical diseases. People suffer from various types of physical complexity due to obesity and overweight. Improper distribution of fatty tissue cause unwanted expansion. Normally everyone wants a healthy lifestyle. So, it is a must to maintain a healthy weight as well as BMI.
Our goal is to predict the possibility of breast cancer due to overweight and obesity. Also making an awareness among women, so that they get to know about it. Thus they will be able to stay free from the risk of breast cancer which is related to obesity and overweight. If people get to know about their risk factor of cancer they will be able to avoid their future complexity.

1.6 Report Layout

The project paper is divided into six chapters. The full layout of the project is given in here for understanding where which part has been portrayed. The layouts are:

- Chapter 1 – It is the introduction area. It covers the underlying data of the project how it going to begin. The thought process and the goal are described here. This part will give an explicit perspective of the full project.
- Chapter 2 – In this section we added the back ground of the project. We showed related works. We also mentioned the scope of the projects as well as the challenges we faced to complete this research.
- Chapter 3 – Here we mentioned all the research methodology. We added how we collected all the data also we added the process of development. Mentioned some analysis also. For the purpose of the project we have integrated image data analysis and Facial Expression Processing techniques along with Artificial Intelligence.
- Chapter 4 –We tried to show the experimental results so that we can compare this with real time data. We mentioned the summary of the project here.
- Chapter 5 – Here we added the summery, conclusion and the future scope of this research.

CHAPTER 2

BACKGROUND

2.1 Introduction

Breast Cancer is the second leading cause of death of women in today's world. Now a days it has become an alarming issue. If we can identify cancer in very early stage then it will be easier to recover. What we need is awareness. There is some factors works behind of having breast cancer. This research is based on some biomarker of breast cancer. Standing on that factors we can guess or identify the chances of breast cancer, if we can identify the chances in early age there is a great chance to prevent.

2.2 Related works

Miguel Patrício, José Pereira, Joana Crisostomo, Paulo Matafome, Manuel Gomes, Raquel Seiça and Francisco Caramelo performed a research with 166 patients. Measuring their age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin and MCP-1. They applied Machine learning algorithms (logistic regression, random forests, support vector machines). They worked with a specific data set of which sensitivity ranging between 82 and 88% and specificity ranging between 85 and 90% [1].

Steven C. Moore, Mary C. Playdon, Joshua N. Sampson, Robert N. Hoover, Britton Trabert, Charles E. Matthews, Regina G. Ziegler did a survey with 621 post-menopausal women. They made a partial co-relation between metabolites and BMI. They showed that each 5 kg/m2 increase in BMI is associated with a 14% higher risk of breast cancer [2].

Graham A Colditz, Walter C Willett, Frank Speizer their aim was to find out relation between weight gaining and risk of breast cancer. They found postmenopausal women with higher BMI in the risk of breast cancer who never used hormone replacement. In total 2517 cases 21% was in risk of cancer (among them 16% used hormone replacement therapy and rest 5% who did not) [3].

Emily White, Jennifer Hays, Zhao Chen, Lewis Kuller tried to find out the relation between obesity and breast cancer for postmenopausal women of USA. Here they used observational study along 4 years taking anthropometric measures. They analyzed their data using "Cox proportional hazards regression" and measured body fat composition by "dual-energy X-ray absorptiometry (DXA)". Their study showed that if the BMI is more than 31.1 then RR (Relative risk) = 2.52; 95% confidence interval (CI) = 1.62–3.93. The risk is lower for those who've taken HRT (hormone replacement therapy). But after age of 18 years lifetime weight gaining is increases the risk of fatal breast cancer [4].

Manuel Picon-Ruiz, Cynthia Morata-Tarifa, Janeiro J. Valle-Goffin, Eitan R. Friedman, Joyce M. Slingerland established a positive relation between obesity and some types of breast cancers. A pooled analysis of 7 studies (including 337,819 women, 4385 invasive breast cancers) and 9-study meta-analysis shows an inverse relation between premenopausal breast cancer risk and obesity that means postmenopausal obese women are in the higher risk than premenopausal women [5].

Melinda Protani, MichaelCoory, Jennifer H. Martin conducted a systematic search of MEDLINE, EMBASE and CINAHL to identify original data. Total number of their study was 45 (39-observational cohorts, 6-treatment cohorts) including BMI and WHR (waist hip ratio). Their study showed based on the results of their meta-analysis, women are advised to keep their weight within normal limits will have benefits should they develop breast cancer [6].

Maddalena Barba , Patrizia Vici , Laura Pizzuti , Luigi Di Lauro, Domenico Sergi, Anna Di Benedetto, Cristiana Ercolani , Francesca Sperati, Irene Terrenato, Claudio Botti , Lucia Mentuccia , Laura Iezzi , Teresa Gamucci, Clara Natoli , Ilio Vitale, Marcella Mottolese , Ruggero De Maria and Marcello Maugeri-Saccà tried to find out how body mass index modifies the relationship between γ-H2AX, a DNA damage biomarker, and pathological complete response in triple-negative breast cancer. They observed 66 triple-negative breast cancer (TNBC) patients treated with neoadjuvant chemotherapy (NACT). The replication rate of the model in leaner patients was 87%. Conflicting results were reported when BMI was analyzed as a potential prognostic factor. Tait did not observe any effect of BMI and diabetes on survival outcomes, whereas Hao and Cakar observed that overweight is associated with adverse outcomes in TNBC, consistently with the findings reported by Widschwendter in the case of severe obesity (BMI $\geq$ 40) [7].

Ali Montazerl,JilaSadighi, FaranakFarzadi, FarzanehMaftoon, Mariam Vahdaninia, Mariam Ansari, AkramSajadian, Mandana Ebrahimi, ShahparHaghighat and IrajHarirchi conducted a case-control study to assess the relationships between anthropometric variables and breast cancer risk in Tehran, Iran. Cases of Iranian Centre for Breast Cancer (ICBC) was observed along 4years (1996-2000). Medical tests and "logistic regression analysis" were performed to calculate odds ratios and 95% confidence intervals as measures of relative risk. Results showed that women with high BMI had a threefold increased risk of breast cancer [odds ratio (OR) = 3.21, 95% confidence interval (CI): 1.15–8.47] [8].

Kentaro Tamaki, Nobumitsu Tamaki, ShigeharuTerukina, Yoshihiko Kamada, Kano Uehara, Miwa Arakaki, Minoru Miyashita, Takanori Ishida, Keely May McNamara, Noriaki Ohuchi and Hironobu Sasano examined the relation of BMI and the risk of developing breast cancer according to the menstruation status and age, and the correlation between BMI and expression of estrogen receptor (ER) based on Nahanishi Clinic Data Base System. Result of weight and height was self-reported and the presence of ER was

determined by distinctive nuclear immunoreactivity. Statistical analyses were performed using StatMate IV for Windows. Breast cancer risk and ER expression were estimated by computation of the odds ratios and their 95% confidence intervals (CIs) [9].

Steven C. Moore, Mary C. Playdon, Joshua N. Sampson, Robert N. Hoover, Britton Trabert, Charles E. Matthews, Regina G. Ziegler studied nested case–control study of 621 postmenopausal breast cancer case. They applied conditional logistic regression and showed that each 5 kg/m2 increase in BMI increases 14% higher risk of breast cancer [10].

Jennifer M. Petrelli, Eugenia E. Calle*, Carmen Rodriguez & Michael J. Thun used Cox proportional hazards modeling to estimate relative risks between breast cancer and body mass index. They showed 30–50% of breast cancer deaths among postmenopausal women in the US population due to overweight [11].

Zahra Cheraghi, Jalal Poorolajal, TaherehHashem, Nader Esmailnasab, Amin DoostiIrani went through prospective cohort and case-control studies investigating the association between BMI and breast cancer. Their results of meta-analysis showed that body mass index has no significant effect on the incidence of breast cancer during premenopausal period. But overweight and obesity may have a minimal effect on breast cancer, although significant relative risk (RR) = 1.16 (95% CI 1.08, 1.25) [12].

Heather Spencer Feigelson, Carolyn R. Jonas, Lauren R. Teras, Michael J. Thun, and Eugenia E. Calle used Cox proportional hazards models to examine the association of BMI and adult weight gain with breast cancer risk. They found weight gaining of 21–30 pounds was associated with a rate ratio of 1.4 (95% confidence interval 1.1–1.8); rates doubled among women gaining >70 pounds compared with women who maintained their weight within 5 pounds of their weight at age 18 [13].

Giske Ursin, Matthew P. Longnecker, Robert W. Haile and Sander Greenland performed a MEDLINE search from 1966 to April 1992 on body mass and breast cancer. Case-control and cohort studies of breast cancer and BMI were eligible for inclusion.Cohort studies on average showed a 15% more inverse association than case-control studies [14].

Carla Cedolini, Serena Bertozzi, Ambrogio P. Londero, Sergio Bernardi, Luca Seriau, Serena Concina, Federico Cattin, Andrea Risaliti collected data about all women who underwent a breast operation for cancer in our department between 2001 and 2008. Their method of detecting cancer was Screening Imaging Palpable Lesion and analyzed their data by R (version 2.15.2), and $P < .05$ was considered significant [15].

2.3 Research Summary

After researching many papers and journals we can come to a decision that, there is a clear relation between obesity and breast cancer. It is not obvious that if you are obese you will - must have breast cancer but there is a chance. Study shows that at stage I or II who were obese their chances of recovery rate was less than who had normal weight. Menopause is also a vital factor here. Premenopausal women is free from the risk of this according to previous research. But life time weight gaining most specifically after age of 18 increases the risk of breast cancer. Body must have a balance with height. Overweight or obesity is considered as a source of various types' physical complexity. That might turn into fatal breast cancer. Fast growing of fatty tissue can causes swift spreading of cancer cell. Thus growing weight and higher BMI (Body Mass Index) causes invasive breast cancer.

2.4 Scope of the Problem

Actually we are going to predict the risk of breast cancer from body mass index (BMI) in percentage. Prediction of cancer is not an easy task. There is lots of uncertainty factors works behind it. Still research is going on to declare a specific result or determine a specific factor. This field is so vast and it is exploring every day.

2.5 Challenges

- Main challenge was to collect information of patient including all information according to our need.
- Another one is to make a proper data set. Selecting proper attributes from with which we can mark the risk of breast cancer properly.
- Finding the healthy range of our attributes in human body and then calculating the risk factor from collected information.
- Preprocessing the data set and categorized them so that we can execute it.
- Extracting information from data was also big challenge for us.
- Achieving exact percentage of accuracy was a big issue here.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Research methodology is a systematic way to solve a problem. Essentially, the procedures by which researchers go about their work of describing, explaining and predicting phenomena are called research methodology. It contains the theoretical analysis of the physique of strategies and concepts associated with a department of knowledge. It is also defined as the study of methods by which knowledge is gained. Its aim is to give the work plan of research. Here in research methodology we are going to encounter our knowledge about this research. We will get going through subject and instrumentation, data collection, analysis and implementation.

3.2 Research Subject and Instrumentation

3.2.1 Machine Learning

Machine learning is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" form data, without being explicitly programmed. Intelligence is defined as general cognitive problem-solving skills. We have that intelligence quality .We have our brain that we used to thinking, analyzing, taking decision on different situations. That's why we have intelligence. On the other hand intelligence in machines is called artificial intelligence, which is commonly implemented in computer systems using programs and sometimes appropriate hardware. It has become an essential part of the technology industry.
With the help of programming languages or help of some tools machines are taking decisions. Machine can analyzing the data, preprocessing the dataset and also can taking decision. In our research we use machine learning tool for analyzing our dataset and take decision for prediction the risk of breast cancer from BMI, glucose and insulin. In our research we used machine learning software. This software is: WEKA-3.6 version.

Steps of machine learning in our research:

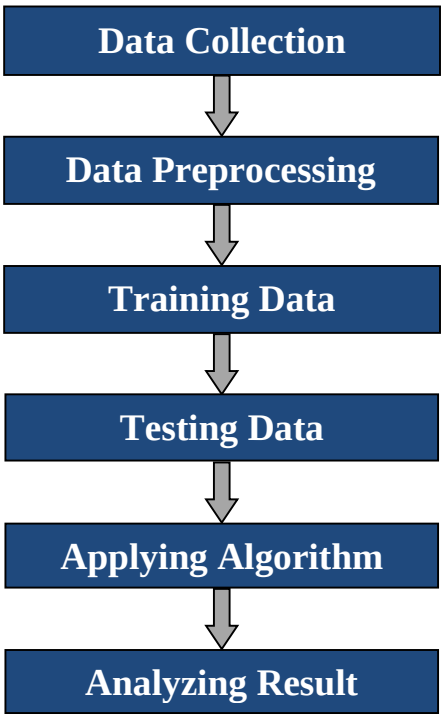


Figure 3.2.1: Data Flow Diagram

3.2.2 Applied Methods

In research purpose there are many algorithms which are used for prediction. They are- decision tree, random forest, support vector machine, k-nearest neighbors, Naïve Bayes classifier, J48, IBK and logistic regression etc. Since our research is prediction breast cancer from body mass index (BMI). In our research we applied four algorithms which can help us to predicted breast cancer form body mass index (BMI), Glucose and Insulin.

They are-

- IBK Classifier Algorithm.
- J48 Classifier Algorithm.
- Random Forest Classifier Algorithm.
- Naïve Bayes Classifier Algorithm.

We choose these algorithms because in our datasets these four algorithms are more suitable than others. This four algorithms gave us better accuracy to represent our work.

3.2.3 Short Description of Applied Algorithms

- **Random Forest Classifier Algorithm:** Random Forest algorithm is a supervised classification algorithm. It is an ensemble learning method for classification, regression and other related tasks. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.
- **Naïve Bayes Classifier Algorithm:** In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem. It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
- **IBK Classifier Algorithm:** It is another approach of K-Nearest Neighbor Algorithm. In WEKA it's called IBK (instance-bases learning with parameter k) and it's in the lazy class folder. KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms. And it can give highly competitive results. KNN is the K parameter. IBK's KNN parameter specifies the number of nearest neighbors to use when classifying a test instance and the outcome is determined by majority vote. WEKA's IBK implementation has the "cross-validation" option that can help by choosing the best value automatically WEKA uses cross-validation to select the best value for KNN.
- **J48 Classifier Algorithm:** The full name is weka.classifiers.trees.J48. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm.

3.3 Data Collection Procedure

We worked to collect real time data. Information of real patient from different medical-center. We collected data from "National Institute of Cancer Research & Hospital (NICRH)", "Ahsania Mission Cancer Hospital "and "BRB Hospital" as their campaign "Solpomulle Breast Cancer Nirnoy" was going on. We took down their age, height, weight, physical condition etc. Findings attributes which plays important role for having breast cancer. After researching we found higher body mass index (BMI), higher range of Glucose and lower range of Insulin has a strong deep connection with physical complexity. Every element has its healthy range of its quantity. We had to collect the information of patient

-including this (BMI, Glucose, and Insulin). This was quiet tough to get all information according to our need. As well as some databases were searched in until September 2018. We collected some dataset from kaggle.com and Github which is related to our research work. We worked both on real time data and searched data to make a comparison among those. Analyzing the attributes which plays important role to make a valid result was a bit challenging. We kept the result of real time data in our work.

3.3.1 Preprocessing the Dataset

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. For achieving better results from the applied model in Machine Learning we need to preprocess our dataset.

3.3.1.1 Balancing the Dataset

Imbalance dataset is very big problem for machine learning when we applied algorithm. For this problem our accuracy result can be decreased. That's why we need to balance our dataset. There are several techniques in WEKA-3.6 to balance the dataset. We use "SMOTE" technique. By using "SMOTE" we can increase the binarity instances in our dataset. That's means positive instances is increased against negative instances. Then positive instances and negative instances are balanced. And then we can get better result.

3.3.1.2 Cross Validation

We used cross validation in our dataset. It is used for training and testing the dataset. Training and testing data we need for preprocessing our dataset. 10 cross validation convert the dataset into 90% for train and 10% for test.

3.3.2 Analyzing the Dataset

Now analyzing the dataset. If the dataset is prepared for applying some classify algorithm and for better accuracy then we applied some algorithm. We applied Naïve Bayes and random forest classifier algorithm.

3.4 Statistical Analysis

Confusion matrix law

Confusion Matrix Law			
		Predicted Result	
		Result = Yes	Result = No
Actual Result	Result = Yes	True Positive	True Negative
	Result = No	False Positive	False negative

Table 3.4: Confusion Matrix Law Table

Precision (P) = $\frac{TP}{TP+FP}$

Sensitivity(R) = $\frac{TP}{TP+FN}$

Specificity = $\frac{TN}{TN+FP}$

Accuracy = $\frac{TP+FN}{TP+FP+TN+FN}$

3.5 Implementation Requirements

We used WEKA-3.6 data mining tool to generate a result from our dataset. WEKA contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. WEKA supports several standard data mining tasks, more specifically-

- Regression.
- Clustering.
- Data preprocessing.
- Classification.
- Visualization.
- Feature selection.

Data format of WEKA-

- CSV
- ARFF
- Database Using ODBC

Here we used ".arff" format to execute our work.

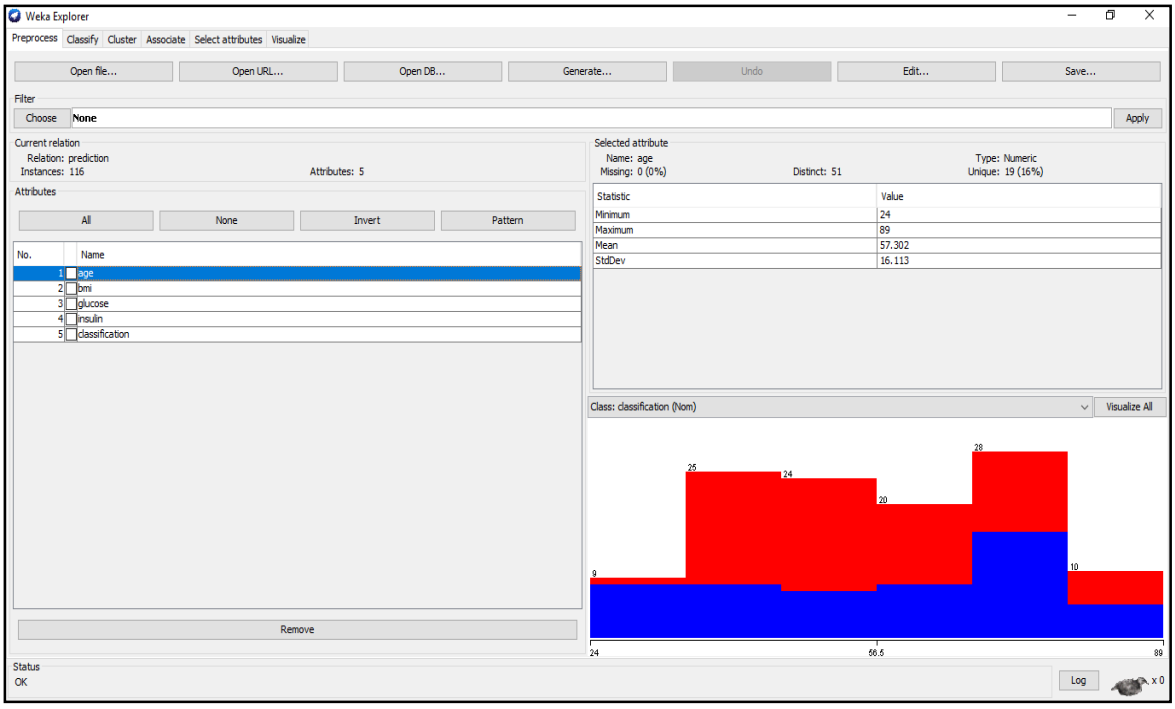


Figure 3.5: Open File in WEKA-3.6

CHAPTER 4

EXPERIMENTAL RESULT

4.1 Introduction

Here we are going to show our experimental result of this research work. The expected outcome of our project is to show the possibility of having breast cancer in adult age due to overweight and obesity. We wanted to show our result or our prediction in percentage. We tried to extract a valid probability of a person from her (we focused only women in this work) weight ration according to height/ BMI, Glucose and Insulin.

4.2 Experimental Results

We applied four algorithms in our dataset and find out the most appropriate one which accuracy was most.

4.2.1 Algorithm's Result

```
Correctly Classified Instances         136               80.9524 %
Incorrectly Classified Instances        32               19.0476 %
Kappa statistic                          0.5784
Mean absolute error                      0.1945
Root mean squared error                  0.4336
Relative absolute error                 41.1953 %
Root relative squared error             89.2611 %
Total Number of Instances              168

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.913     0.359      0.805     0.913     0.856      0.788    yes
                 0.641     0.087      0.82      0.641     0.719      0.788    no
Weighted Avg.    0.81      0.255      0.811     0.81      0.804      0.788

=== Confusion Matrix ===

  a  b   <-- classified as
 95  9 |  a = yes
 23 41 |  b = no
```

Figure 4.2.1(a): Result of IBK Classifier Algorithm

```
Correctly Classified Instances         134               79.7619 %
Incorrectly Classified Instances        34               20.2381 %
Kappa statistic                          0.5405
Mean absolute error                      0.2554
Root mean squared error                  0.3761
Relative absolute error                 54.0804 %
Root relative squared error             77.4297 %
Total Number of Instances              168

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.942     0.438      0.778     0.942      0.852      0.864    yes
                 0.563     0.058      0.857     0.563      0.679      0.864    no
Weighted Avg.    0.798     0.293      0.808     0.798      0.786      0.864

=== Confusion Matrix ===

  a  b   <-- classified as
 98  6 |  a = yes
 28 36 |  b = no
```

Figure 4.2.1(b): Result of Random Forest Classifier Algorithm

```
Correctly Classified Instances         130               77.381  %
Incorrectly Classified Instances        38               22.619  %
Kappa statistic                          0.493
Mean absolute error                      0.2641
Root mean squared error                  0.431
Relative absolute error                 55.939  %
Root relative squared error             88.718  %
Total Number of Instances              168

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.904     0.438      0.77      0.904      0.832      0.775    yes
                 0.563     0.096      0.783     0.563      0.655      0.775    no
Weighted Avg.    0.774     0.307      0.775     0.774      0.764      0.775

=== Confusion Matrix ===

  a  b   <-- classified as
 94 10 |  a = yes
 28 36 |  b = no
```

Figure 4.2.1(c): Result of J48 Classifier Algorithm

```
Correctly Classified Instances         120               71.4286 %
Incorrectly Classified Instances        48               28.5714 %
Kappa statistic                          0.3386
Mean absolute error                      0.2968
Root mean squared error                  0.4645
Relative absolute error                 62.8571 %
Root relative squared error             95.6246 %
Total Number of Instances              168

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.904     0.594      0.712     0.904     0.797      0.833    yes
                 0.406     0.096      0.722     0.406     0.52       0.833    no
Weighted Avg.    0.714     0.404      0.716     0.714     0.691      0.833

=== Confusion Matrix ===

  a  b   <-- classified as
 94 10 |  a = yes
 38 26 |  b = no
```

Figure 4.2.1(d): Result of Naïve Bayes Classifier Algorithm

4.2.2 Receiver Operating Characteristic (ROC) Curve

In the ROC curve we have false positive rate on X-axis and the true positive rate also called recall on the Y-axis. It is an effective method of evaluating the quality or performance of diagnostic tests. This ROC curves are useful to visualize and compare the performance of classifier methods. We see that area under ROC curve value is 0.7885 for IBK Classifier Algorithm, 0.8635 for Random Forest Classifier Algorithm, 0.7749 for J48 Classifier Algorithm and 0.8332 for Naïve Bayes Algorithm. These among result is reasonable AUC. So all the ROC curves are reasonable curve for our dataset.

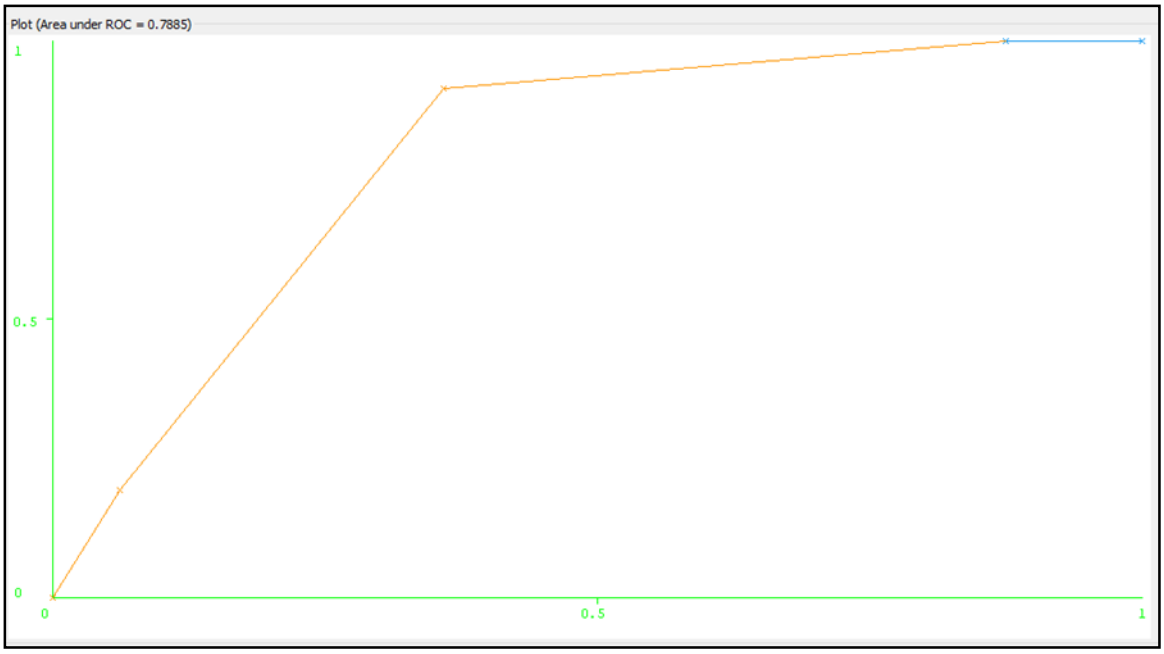


Figure 4.2.2(a): ROC Curve of IBK Classifier Algorithm

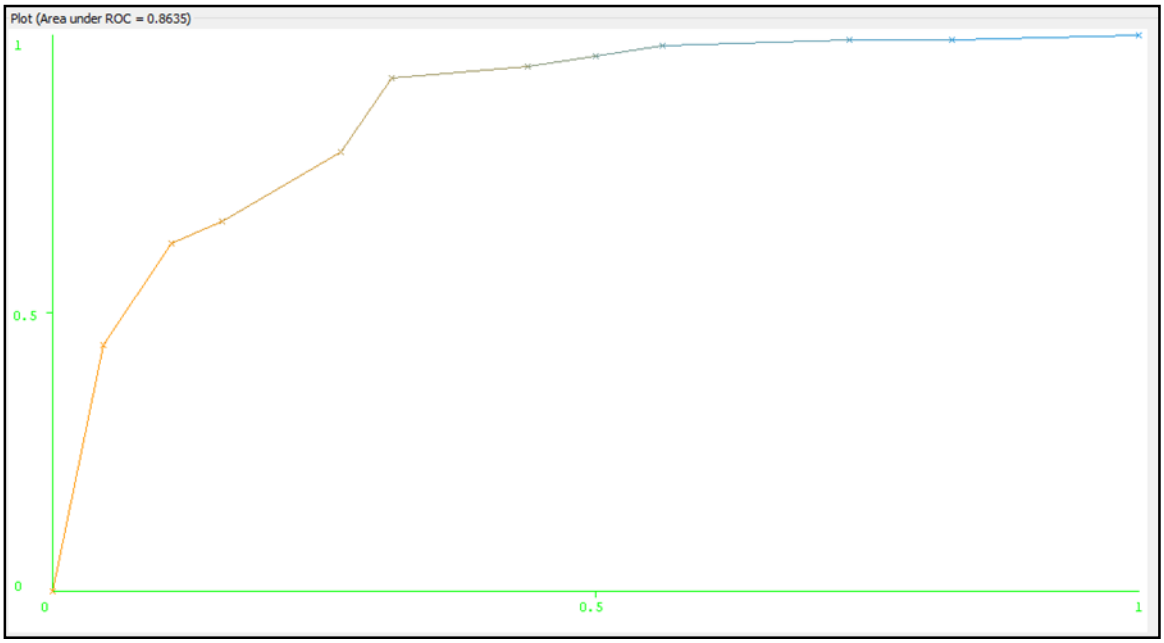


Figure 4.2.2(b): ROC Curve of Random Forest Classifier Algorithm

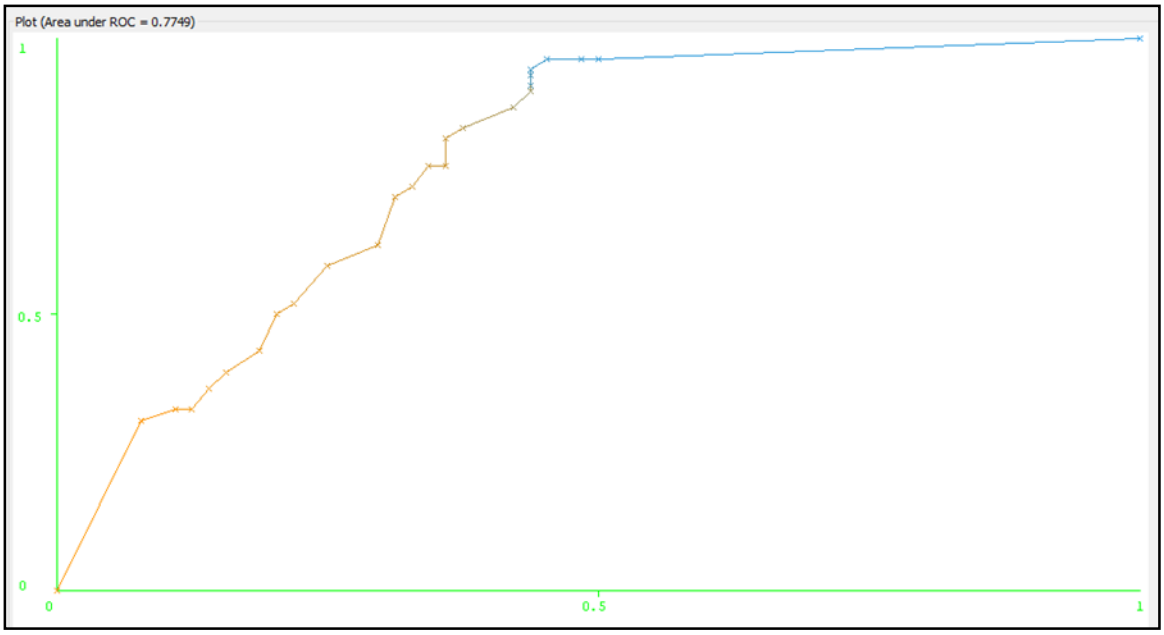
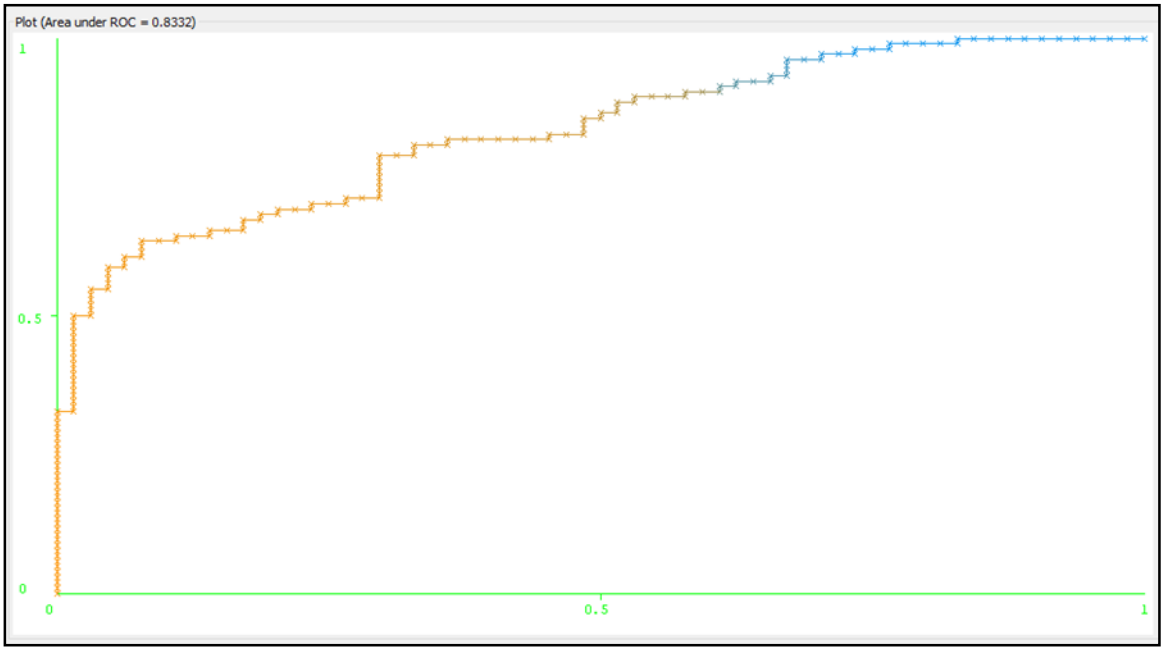


Figure 4.2.2(c): ROC Curve of J48 Classifier Algorithm



4.2.2(d): ROC Curve of Naïve Bayes Classifier Algorithm

4.3 Descriptive Analysis

Based on dataset we applied some types of algorithms. One of this is IBK Algorithm. It is another approach of KNN (K-Nearest Neighbor Algorithm). We have got 80.9524% accuracy for IBK algorithm.
Another one is Random Forest Algorithm. This is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees. We've got 79.7619 % accuracy for Random Forest Algorithm.
Third applied algorithm is J48 classification algorithm. In WEKA data mining tool J48 Classification algorithm is used for accounting for missing values, decision trees pruning, continuous attribute values ranges, derivation of rules etc. Here we have got 77.381% accuracy.
And the last one is Naïve Bayes Classifier Algorithm. In machine learning, Naïve Bayes classifiers are a family of simple "probabilistic classifiers". We've got 71.4286% accuracy for Naïve Bayes Algorithm.
Among all these four algorithms IBK classifier is most accurate for our dataset. Then respectively come Random Forest Classifier, J48 Classifier and Naïve Bayes Classifier Algorithm. And after analyzing our result section we decided that IBK classifier is better than all others algorithms for our dataset.

4.4 Summary

4.4.1 Application of Confusion Matrix

Accuracy				
Expected Outcome		**S2**		**Total**
		Yes	**No**	
	Yes	95	09	104
	No	23	41	64
Total		118	50	

Table 4.4.1(a): Data from IBK Classifier Algorithm for Accuracy

Precision (P) = $\frac{\text{TP}}{\text{TP+FP}} = 95\%\ 118 = 0.8051$
Sensitivity(R) = $\frac{\text{TP}}{\text{TP+FN}} = 95\%\ 136 = 0.6985$
Specificity = $\frac{\text{TN}}{\text{TN+FP}} = 9\%\ 32 = 0.28125$
Accuracy = $\frac{\text{TP+FN}}{\text{TP+FN+TN+FN}} = 136\%\ 168 = 0.809523$

Accuracy				
Expected Outcome		**S2**		**Total**
		Yes	**No**	
	Yes	98	06	104
	No	28	36	64
Total		126	41	

Table 4.4(b): Data from Random Forest Classifier Algorithm for Accuracy

Precision (P) = $\frac{\text{TP}}{\text{TP+FP}} = 98\%\ 126 = 0.7777$

Sensitivity(R) = $\frac{\text{TP}}{\text{TP+FN}} = 98\%\ 134 = 0.73134$

Specificity = $\frac{\text{TN}}{\text{TN+FP}} = 6\%\ 34 = 0.17647$

Accuracy = $\frac{\text{TP+FN}}{\text{TP+FP+TN+FN}} = 134\%\ 168 = 0.79761$

Accuracy				
Expected Outcome		**S2**		**Total**
		Yes	**No**	
	Yes	94	10	104
	No	28	36	64
Total		122	46	

Table 4.4(c): Data from J48 Classifier Algorithm for Accuracy

Precision (P) = $\frac{\text{TP}}{\text{TP+FP}} = 94\%\ 122 = 0.77049$

Sensitivity(R) = $\frac{\text{TP}}{\text{TP+FN}} = 94\%\ 130 = 0.72307$

Specificity = $\frac{\text{TN}}{\text{TN+FP}} = 10\%\ 38 = 0.26316$

Accuracy = $\frac{\text{TP+FN}}{\text{TP+FP+TN+FN}} = 130\%\ 168 = 0.77380$

Accuracy				
Expected Outcome		**S2**		**Total**
		Yes	**No**	
	Yes	94	10	104
	No	38	26	64
Total		132	36	

Table 4.4.1(d): Data from Naïve Bayes Classifier Algorithm for Accuracy

Precision (P) = $\frac{\text{TP}}{\text{TP+FP}} = 94\%\ 132 = 0.71212$

Sensitivity(R) = $\frac{\text{TP}}{\text{TP+FN}} = 94\%\ 120 = 0.7833$

Specificity = $\frac{\text{TN}}{\text{TN+FP}} = 10\%\ 48 = 0.20833$

Accuracy = $\frac{\text{TP+FN}}{\text{TP+FP+TN+FN}} = 120\%\ 168 = 0.71429$

4.4.2. Summary of All Applied Algorithms

In this table we get the accuracy or correctly classified instances for each algorithms. At the same time we get the misclassification or incorrectly classified instances. And also get the sensitivity and specificity.

Accuracy- Accuracy stands for how much appropriate our prediction is. IBK Algorithm gave us higher accuracy than others. In another word IBK indicates that our prediction is 80.9524% correct. And rest are in given table.

Misclassification- Misclassification is opposite of accuracy. It defines how much data we couldn't predict correctly. It is also called incorrectly classified instance. For example, from table we can observe IBK was unable to predict 19.047% data.

Sensitivity- It defines that, how much patients are affected. That means the correct percentage of presence of cancer in patients. As example, sensitivity of IBK 69.85%.

Specificity- It is opposite of sensitivity. It defines the correct percentage of patients who are not affected with cancer. As example, specificity of IBK 28.125%.

Here is the summarized table. We put all numeric values here of our total work to observe it at a glance.

Name of Algorithms	Accuracy Rate	Misclassification	Sensitivity	Specificity
1. IBK	**80.9524%**	19.0476 %	0.6985	0.28125
2. Random Forest	79.7619 %	20.2381 %	0.73134	0..17647
3. J48	77.381%	22.619 %	0.72307	0.26316
4. Naïve Bayes	71.4286%	28.5714%	0.7833	0.20833

Table 4.4.2: Data from All Applied Algorithms

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Summary of the Study

Here we are going to mention the summery of our study. Based on our research we can come to a decision that lifetime weight gaining and unhealthy lifestyle causes various types of physical problem. Several problem leads to some fatal diseases like diabetics and cancer. If Glucose level is constantly higher than 140 mg/dL and Insulin is lower than 25 mIU/L (fasting stage) then there is a risk of physical vulnerability. Our study shows that the percentage of having breast cancer from current condition of body. If we can find out the possibility or the future risk from current condition of body then it will be easier to handle future problems. We took down people's age, weight, height, glucose level, insulin level and then extract information from collected data. We used some algorithms to execute our result. We've got our desire result. While researching we found lots of biomarker to mark our physical condition. This time we worked only with glucose, insulin and BMI. There is also lots of factor with which we can predict physical condition and future also.
Leading a healthy life is a key factor to stay away from physical complexity. To make people aware and decreasing the death ratio of breast cancer is our motto. If we can maintain a healthy lifestyle and healthy weight we will be able to stay away from various disease that leads to cancer.

5.2 Conclusion

At the very end of this research report we can say that, we tried our best to complete this task. We tried to make a valid result and to keep this as simple as possible. Comparatively easy to understand. This field is so vast and everyday it is being explored. Lots of work already done on it. We tried to find out the possibility of having breast cancer in percentage. As cancer is a vital diseases and filled up with huge complexity. Firstly it is not easy to identify this. Secondly sometimes when it gets identified then already the cancer is in stage II or III. It has become too tough to recover from that stage. So we wanted to identify breast cancer from some basic things, such as body mass indexing. That means the proper ratio of weight according to height. If we can predict the chances of cancer at early stage then chances of survive will definitely increase. We tried to get a valid result of "prediction of breast cancer from BMI". We think this will be helpful for further studies.

5.2 Future Scope

We have made our research as feasible as possible though there are some lacking. Some parts of our work can be more improved. Future Studies and research can dig up more features in this field. Let's see some of the aspects that can be implemented in future.

- Here we focused only in women. That doesn't mean men are free from this risk. After a certain age men are also in included in this risk. In future men's risk for breast cancer can also be researched.
- We focused to predict only breast cancer. In upcoming days other types of cancer will be included with our study.
- Here we've taken glucose level, insulin level and BMI as our attributes. In future other attributes can also be included which is related to it.

REFERENCES:

[1] Miguel Patrício, José Pereira, Joana Crisóstomo , Paulo Matafome, Manuel Gomes , Raquel Seiça and Francisco Caramelo "Using Resistin, glucose, age and BMI to predict the presence of breast cancer" Patrício et al. BMC Cancer (2018) 18:29.

[2] Steven C. Moore, Mary C. Playdon, Joshua N. Sampson, Robert N. Hoover, Britton Trabert, Charles E. Matthews, Regina G. Ziegler "A Metabolomics Analysis of Body Mass Index and Postmenopausal Breast Cancer Risk" JNCI J Natl Cancer Inst (2018) 110(6): djx244. First published online January 9, 2018

[3] Graham A Colditz, Walter C Willett, Frank Speizer "Dual Effects of Weight and Weight Gain on Breast Cancer Risk" Article in JAMA The Journal of the American Medical Association · November 1997.

[4] Emily White, Jennifer Hays, Zhao Chen, Lewis Kuller "Obesity, body size, and risk of postmenopausal breast cancer: the Women's Health Initiative (United States)" Article in Cancer Causes and Control · October 2002.

[5] Manuel Picon-Ruiz, Cynthia Morata-Tarifa, Janeiro J. Valle-Goffin, Eitan R. Friedman, Joyce M. Slingerland "Obesity and Adverse Breast Cancer Risk and Outcome: Mechanistic Insights and Strategies for Intervention".

[6] Melinda Protani, Michael Coory, Jennifer H. Martin "Effect of obesity on survival of women with breast cancer: systematic review and meta-analysis" Breast Cancer Res Treat (2010) 123:627–635. Received: 17 January 2010 / Accepted: 5 June 2010 / Published online: 23 June 2010 Springer Science+Business Media, LLC. 2010.

[7] Maddalena Barba, Patrizia Vici, Laura Pizzuti , Luigi Di Lauro , Domenico Sergi , Anna Di Benedetto , Cristiana Ercolani, Francesca Sperati , Irene Terrenat4 , Claudio Botti , Lucia Mentuccia , Laura Iezzi , Teresa Gamucci , Clara Natoli , Ilio Vitale, Marcella Mottolese, Ruggero De Maria and Marcello Maugeri-Saccà "Body mass index modifies the relationship between γ-H2AX, a DNA damage biomarker, and pathological complete response in triple-negative breast cancer" Barba et al. BMC Cancer (2017) 17:101.

[8] Ali Montazerl,JilaSadighi, FaranakFarzadi, Farzaneh Maftoon, Mariam Vahdaninia, Mariam Ansari, AkramSajadian, Mandana Ebrahimi, ShahparHaghighat and IrajHarirchi "Weight, height, body mass index and risk of breast cancer in postmenopausal women: a case-control study". Published: 30 September 2008. BMC Cancer 2008, 8:278. © 2008 Montazeri et al.

[9] Kentaro Tamaki, Nobumitsu Tamaki, Shigeharu Terukina, Yoshihiko Kamada, Kano Uehara, Miwa Arakaki, Minoru Miyashita, Takanori Ishida, Keely May McNamara, Noriaki Ohuchi and Hironobu Sasano "The Correlation between Body Mass Index and Breast Cancer Risk or Estrogen Receptor Status in Okinawan Women" .

[10] Steven C. Moore, Mary C. Playdon, Joshua N. Sampson, Robert N. Hoover, Britton Trabert, Charles E. Matthews, Regina G. Ziegler "A Metabolomics Analysis of Body Mass Index and Postmenopausal Breast Cancer Risk" JNCI J Natl Cancer Inst (2018) 110(6): djx244.

[11] Jennifer M. Petrelli, Eugenia E. Calle, Carmen Rodriguez & Michael J. Thun "Body mass index, height, and postmenopausal breast cancer mortality in a prospective cohort of US women". Cancer Causes and Control 13: 325–332, 2002. 325 /2002 Kluwer Academic Publishers.

[12] Zahra Cheraghi, Jalal Poorolajal, Tahereh Hashem, Nader Esmailnasab, and Amin Doosti Irani "Effect of Body Mass Index on Breast Cancer during Premenopausal and Postmenopausal Periods: A Meta-Analysis".

[13] Heather Spencer Feigelson, Carolyn R. Jonas, Lauren R. Teras, Michael J. Thun, and Eugenia E. Calle "Weight Gain, Body Mass Index, Hormone Replacement Therapy, and Postmenopausal Breast Cancer in a Large Prospective Study" October 17, 2018. © 2004 American Association for Cancer.

[14] Giske Ursin, Matthew P. Longnecker, Robert W. Haile and Sander Greenland " Meta-analysis of Body Mass Index and Risk of Premenopausal Breast Cancer".

[15] Carla Cedolini, Serena Bertozzi, Ambrogio P. Londero, Sergio Bernardi, Luca Seriau, Serena Concina, Federico Cattin, Andrea Risaliti "Type of Breast Cancer Diagnosis, Screening, and Survival" [a] 2014 Elsevier Inc.

[16] Piet A. van den Brandt, Donna Spiegelman, Shiaw-ShyuanYaun, Hans-OlovAdami, Lawrence Beeson, Aaron R. Folsom, Gary Fraser, R. Alexandra Goldbohm, Saxon Graham, Larry Kushi, James R. Marshall, Anthony B. Miller, Tom Rohan, Stephanie A. Smith-Warner, Frank E. Speizer, Walter C. Willett, AlicjaWolk, David J. Hunter "Pooled Analysis of Prospective Cohort Studies on Height, Weight, and Breast Cancer Risk"American Journal of Epidemiology Copyright ©2000 by The Johns Hopkins University School of Hygiene and Public Health.

[17] D. S. M. Chan, A. R. Vieira, D. Aune, E. V. Bandera, D. C. Greenwood A. McTiernan, D. Navarro Rosenblatt, I. Thune, R. Vieira, T. Norat"Body mass index and survival in women with breast cancer—systematic literature review and meta-analysis of 82 follow-up studies" Annals of Oncology 25(10):1901–1914 ©2014, doi:10.1093/annonc/mdu042.

[18] G. Berclaz, S. Li, K. N. Price, A. S. Coates, M. Castiglione-Gertsch, C.-M. Rudenstam, S. B. Holmberg, J. Lindtner, D. Erien, J. Collins, R. Snyder, B. Thürlimann, M. F. Fey, C. Mendiola, I. Dudley Werner, E. Simoncini, D. Crivellari, R. D. Gelber, A. Goldhirsch "Body mass index as a prognostic feature in operable breast cancer: the International Breast Cancer Study Group experience" Annals of Oncology 15(6):875-884 ·©June 2004, doi: 10.1093/annonc/mdh222.

[19] Heather Spencer Feigelson, Carolyn R. Jonas, Lauren R. Teras, Michael J. Thun and Eugenia E. Calle "Weight Gain, Body Mass Index, Hormone Replacement Therapy, and Postmenopausal Breast Cancer in a Large Prospective Study" Cancer Epidemiology, Biomarkers & Prevention 13(2):220-224 •©February 2004, doi:10.1158/1055-9965.EPI-03-0301.

[20] Minouk J. Schoemaker, PhD, Division of Genetics and Epidemiology, The Institute of Cancer Research, 15 Cotswold Rd, London SM2 5NG, United Kingdom. "Association of Body Mass Index and Age with Subsequent Breast Cancer Risk in Premenopausal Women" JAMA Oncol. Published online June 21, 2018. doi:10.1001/jamaoncol.2018.1771

[21] Erica T. Warner, Rong Hu, Laura C. Collins, Andrew H. Beck, Stuart Schnitt, Bernard Rosner, A. Heather Eliassen, Karin B. Michels, Walter C. Willett and Rulla M. Tamimi"Height and Body Size in Childhood, Adolescence, and Young Adulthood and Breast Cancer Risk According to Molecular Subtype in the Nurses' Health Studies" Cancer Prevention Research 9(9):732–738 © 2016 American Association for Cancer. doi;10.1158/1940-6207.CAPR-16-0085.