Ministry of Higher Education and
Scientific Research
University of Diyala
College of Science
Department of Computer Science

Breast Cancer Classification Model Based on Machine Learning Algorithms

A Dissertation

Submitted to the Department of Computer Science\ College of Sciences\ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of master's in computer science

By

Ismail Miteb Hamid

Supervised By

Prof. Dr. Dhahir Abdulhade Abdulah **Ass.Prof.Dr. Salam Abdulkhaleq Noaman**

2020 A.C **1442 A.H**

بِسمِ اللّهِ الرَّحمنِ الرَّحيمِ

﴿يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنكُمْ
وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ﴾

صدق الله العظيم

سورة المجادلة :الآية (١١)

Dedication

This thesis is dedicated to someone,
I will remember his kindness and support in his life

and after…

My parent
And *My Brothers and Sisters*
And My family

Ismael

Acknowledgements

All my thanks first of all are addressed to Almighty ***Allah***, who has guided my steps towards the path of knowledge and without His help and blessing; this thesis would not have progressed or have seen the light.

My sincere appreciation is expressed to my supervisors ***Prof. Dr. Dhahir Abdulhade Abdulah*** *and* ***Ass.prof.Dr.*** **Salam Abdulkhaleq Noaman** for providing me with ideas, inspiration and continuous support me during the period of my study.

I am extremely grateful to all members of Computer Science Department of Diyala University for their general support.

Finally, I would never have been able to finish my thesis without the help from ***friends***, and support from ***my family***.

Thank you all!

ismael

(Supervisor’s Certification)

We certify that this research entitled “***Breast Cancer Classification using Developed Techniques”*** was prepared by ***Ismail Miteb Hamid*** Under our supervisions at the University of Diyala Faculty of Science Department of Computer Science, as a partial fulfillment of the requirement needed to award the degree of Master of Science in Computer Science.

(Supervisor)

Signature:

Name: **Dr. Dhahir Abdulhadi Abdulah**

Date:2/7/2020

(Supervisor)

Signature:

Name: **Ass.Prof.Dr. Salam bdulkhaleq Noaman**

Date:2/7/2020

Approved by University of Diyala Faculty of Science Department of Computer Science**.**

Signature:

Name: **Dr. Taha Mohammad Hassan**

Date: 2/7/2020

(Head of Computer Science Department)

Approved by University of Diyala Faculty of Education for Pure sciences Department of Computer Science.

Linguistic Certification

This is to certify that this thesis entitled " Breast Cancer Classification using Developed Techniques " was prepared by "Ismail Miteb Hamid" under my linguistic supervision. Its language was amended to meet the English style.

Signature :

Name :

Date : / / 20120

Scientific certification

This is to certify that this thesis entitled "Breast Cancer Classification using Developed Techniques" was prepared by "Ismail Miteb Hamid" under my Scientific supervision. It has been evaluated scientifically, therefore, it is suitable for debate by examining committee.

Signature :

Name :

Date : / /2020

Abstract

Breast cancer is one of the leading causes of death among women worldwide. Accurate and early detection of breast cancer can ensure long-term survival for the patients. However, traditional classification algorithms usually aim only to maximize the classification accuracy, and cost failing to take into consideration the misclassification costs between different categories. Furthermore, the costs associated with missing a cancer case (false negative) are much higher than those of mislabeling a benign one (false positive).

To overcome this drawback and further improving the classification accuracy of the breast cancer diagnosis, in this work, present several machine learning algorithms such as Decision Tree (DR) , Random Forest (RF) , Logistic Regression (LR), and Support Vector Machine (SVM) . For all the phases of the work that required data treatment and machine learning techniques are going to use this tool. In technical terms ,the intended output of the work that enables the achievement of the business objectives described before is find the algorithms that can classify more efficiently the different types of breast cancer.

The result of the machine learning by calculate the accuracy of each model obtain , the random forest achieved 0.9857 % accuracy , decision tree achieved 0.9571% accuracy, SVM achieved 0.9714% accuracy, Logistic Regression achieved 0.9643 % accuracy , in an other word the Robust forest archive high accuracy.

List of contents

List of Tables

List of Figure

List of Abbreviations

Abbreviations	Description
2D	Two-dimensional
3D	Three- dimensional
4D	Four- dimensional
ANN	Artificial Neural Network
CAD	Computer-aided Detection
CADX	Computer-aided Detection/Diagnosis
CT	Computed-tomography
DM	Data Mining
DT	Decision Tree
FN	False Negative
FNA	Fine Needle Aspiration
FP	False Positive
KKN	K-Nearest Neighbors
LR	Logistic Regression
MIAS	Mammographic Image Analysis Society
ML	Machine Learning
MRI	Magnetic-resonance-imaging
NB	Naïve Bayesian
NCI	National Cancer Institute
PET	Positron-emission-tomography

RBF	Radial Basis Function
RF	Random Forest
SEE	Span Error Estimate
SOMs	Self-organizing Maps
SRG	Seeded Region Growing
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
US	United State
WBC	Wisconsin Breast Cancer Dataset
WDBC	Wisconsin Diagnosis Breast Cancer
WDBC	Wisconsin Breast Cancer Diagnosis
WPBC	Wisconsin Breast Cancer prognosis

Chapter one

Introduction

Chapter One

Introduction

1.1 Overview

Major public health problem is a cancer, that mortality and cancer incidence has been increased over the past three decades at an accelerated pace globally, [1]. Breast cancer is known as the burden of non-communicable diseases and it is the main reason in deaths in developing countries. The disease GLOBOCAN 2018 diagnosed have affected around (2.2) million new cases in breast cancer around the world, [2]. There are highly chance of causes disease like breast cancer which are still unknown, [3][4].

The distinguish between malignant breast tumors and benign ones by using some of diagnosis techniques. The well-known procedure, Fine Needle Aspiration (FNA). In addition, inexperience or exhaustion cause higher possibility to rise errors, that panic patients when incorrect-positive result happens when incorrect-negative result appears. To help doctors' diagnosis of breast cancer, the evolving well organized diagnosis support system, The research explains that proposing machine learning and data mining approaches for breast tumor diagnosis can obtain lots advantage including high degree of diagnosis accuracy, reducing medical and resource driving down costs, [4][5]. Breast tumor is the popular diagnosed growth among the methods for growth diagnosis, that is treated as breast growth from malignant ones, [6]. However, although, how to obtain better result for common classification issues is still not easy up to now, [7].

Many Machine learning process, like Logistic Regression (LR), Artificial Neural Network (ANN), Support Vector Machine (SVM), Naïve Bayesian (NB) and K-Nearest Neighbors (KNN), are work on principle black box system which are incapable to explain the forecast results. When knowing into a learning model is more important compared with classification outcomes, it is more efficient to employ a system of white box. Although there are numerous assistance systems of diagnosis in the United State (US), that had explained gaining in scientific research clinical experiments and labs , that have not available in clinical practice, or even widely used, [8].

In this thesis , a statistical analysis of the data available in the dataset WBC, Also there are four machine learning algorithms to use classification data and comparison between them.

1.2 Literature Review

Many of researches have been accomplished in the area of classification of Wisconsin Brest Cancer Dataset (WBCD) by use of smart computing types. These can be described briefly below:

- In Martin J. Yaffe, Earlier [9], 2016. Authors have applied the use of the 2D median filtering, image contrast enhancement, algorithm, Seeded Region Growing (SRG) to disconnect the noise, omit radiopaque artefacts and remove the prediction of the Breast from a digitalized mammogram. In addition, the technicality of Artificial Neural Network is used to classify a mammogram like representing a malignant tumor or normal, which is indicating the existence of a benign aggregate.
- Roselina Sallehuddin [10], 2016. Authors have introduced a specific procedure for feature uprooting to distinguish between digital mammograms through using quick limited shear let transform. To enlarge the differences between type delegates, a thresholding method could be put in place as the last scene of feature uprooting.

The categories have been calculated through the optimal property group.

- In Emina Aličković [11] (2017). Applying backup Self-organizing Maps (SOMs), Support Vector Machine (SVM), Radial Basis Function (RBF) networks for breast tumour detection. The (WDBC) dataset can be utilized in the categories experiment. The following classifiers: 1-norm C-SVM (L1-SVM), 2-norm C-SVM (L2-SVM), and u-SVM are utilized, for which the connection search up on span error estimate (GSSEE), gradient descent based on validation error estimate. The SOM–RBF classifier is developed to improve the performance of only the SOM learning procedure, which is based on distance comparison .The SOM–RBF classifier could be effective materials to detect breast tumour with the high detection precision.
- In Peiguang Lin [12] (2016). presented a hybrid system named GRA-SVM that constitute SVM classifier and filter attribute option. Grey Relational Analysis has been suggested and tested against BUPA Disorder the group of data and Wisconsin Breast Cancer Dataset (WBCD). Experiments results show that GRA-SVM get better the SVM precision around 0.48% through use limited two advantages for the WBCD dataset. For BUPA dataset, GRA-SVM gets better the SVM accuracy around 0.97% through utilizing four advantages.
- In Nilashi, Mehrbakhsh., Ibrahim, ..et [13] (2017). Introduced many dissimilar information mining procedures such as: (Support Vector Machine (SMO SVM), Bayesian Network, Multilayer Perceptron and Decision tree (J48)) these methods were put into practice for Wisconsin Diagnostic Breast Cancer (WDBC). The

SVM SMO procedure has accomplished an accuracy around 97.72%.

- In O.N. Oyelade, [14] (2017) . Shows a procedure about the user interface that depends on the ontology. OWL language can be used to describe the context data and the inquiry form data, that in turn gives a better firm for the latter classification and integration of the inquiry interface style. This step can have a great employ of constructing and submitting the inquiry demand. Moreover, the [22] also include the automatically extraction procedure about the impersonation, while the tests display that the procedure can perform greatly to excerpt the attribute information, relationship data and context data of the inquiry interface.
- In S. Wang, Y. Wang, et al [15] (2017) . Introduced Classification and Regression Trees (CART) which create the fuzzy principles to be employed in the knowledge system. The check results on Pima Indian Diabetes, Stat Log, WDBC, Mesothelioma, Parkinson, and Cleveland’s tele monitoring collection of data presents suggested method extraordinarily increases the diseases forecast accuracy. The outcomes displayed that the series of fuzzy principle, CART with limit noise and gathering techniques can have more impacts in diseases forecast from medical datasets.
- In Dalwinder Singh, [16] 2018. Worked on diagnosing breast cancer base on Wisconsin information, firstly, set up an efficient input mechanism that will make it able for the outline to filter through reading and cleaning input from datasets. Secondly, semantic web languages were utilized to set up an ordered principle set and know the outline framework were therefore formed to try and help the causing algorithm. In conclusion, modification of the

structures of Select and Test (ST) can be accommodated to this enhancement

- In S., Birmohan S. [17] 2019. Introduced recovered the method of the Random Forest-Based Rule Extraction Method for Breast Cancer Diagnosis. At beginning, availability of the abundant decision rules could be created by using Random Forest through numbers of decision tree models. Then, the process of a principle of extraction is inventive to detach decision principles which is integrated of trained trees. In the end, multi objective evolutionary algorithm (MOEA) is improved and put in place to find for an optimal principle foreteller that the constituent principle group is the better trade-off in between interpretability and accuracy.
- In Na Liu, Er-Shi Qi, [18] (2019). Introduced merit weighting is applied to evolve an efficient computer as aided diagnosis outline for breast cancer. Merit weighting is appointed because of the enhances the classification achievement more as contrast to merit branch selection. This mostly works when a wrapper procedure employs the Ant Lion Optimization algorithm is showed that finds for better merit weights and values of Multilayer Neural Network at the same time. The option of unobserved backpropagation and neurons training algorithms are applied as variables of nervous networks. The presentation of the suggested method is assessed on three types of breast cancer information. The information of datasets is at first standardized utilizing tanh procedure to expel the impacts of predominant aspect. The outcomes display that suggested wrapper strategies has superior capacity to achieve best precision when contrasted with the existent techniques. The acquired more characterization achievements approves the work

which has the prospect for turning into an option in contrast to another known technicality.

1.3 Statement of the problem

During a literature review of tissue classification, it was noted that breast tissue cancer is the most common malignancy among women and the cause of many endings of life. Early detection of tumours is the best solution to avoid mastectomy, reduces the chances of it coming back and reduces the life death rate. There is no effective way to get rid of this cancer. All the methods used did not carry the problem at all. There are many classification methods used in this field It is not easy to answer the question of the classification approach appropriate for a specific study. Different classification results can be obtained according to the selected works. Different classification methods have their own advantages.

1.4 Aim of Thesis

The aim of this thesis is to solve the problem in the paragraph (1.3) and classification model based on the intelligent computing model to help clinicians or radiologists identify areas of suspicion of the databases used. The main idea to develop a model for classification of Malignant and benign cases in order to reduce the number of erroneous cases in this thesis. And use four algorithms to develop the work which is support vector machine (SVM) , Random Forest (RF) , Logistic Recreation (LR) and Decision Tree (DT) .

1.5 Thesis Outline

This thesis consists four chapters and adding to the chapter one as the following:

Chapter 2: Theoretical and background, it includes the theoretical background of concept that were adopted in the thesis it was reviewed Medical Image and machine learning

Chapter 3: Breast Cancer Classification Model Based on Machine Learning Algorithms: The suggests algorithms and procedures that describe the statistical analysis of the database and explain in detail the tools to use in classification.

Chapter 4: Results and Analysis, this chapter includes a review of the results obtained through the implementation of the proposed system. The results are discussed and analysed with comparisons made with related work.

Chapter 5: Conclusions and Future Work, it includes reviewing the results reached through designing and implementing the proposal as well as the recommendations.

Chapter Two

Theoretical background

Chapter Two

Theoretical Background

2.1 Introduction

This chapter includes the concepts, tools, and techniques used in the practical implementation of the proposed system, as well as the statistical measures used in data analysis and algorithms evaluation and their efficiency.

2.2 Breast cancer

Breasts is a top position of ventral part(human) on each side of females' body, that have each part the anterior area of the females body. Thus, Breasts are located between the second and the sixth rib that contains mammary gland. These glands are started to make milk with stimulation that after born baby. These glands are located as primordial form in male and with some peculiarity in female major. The figure (2.1) shows normal breast structure [19].

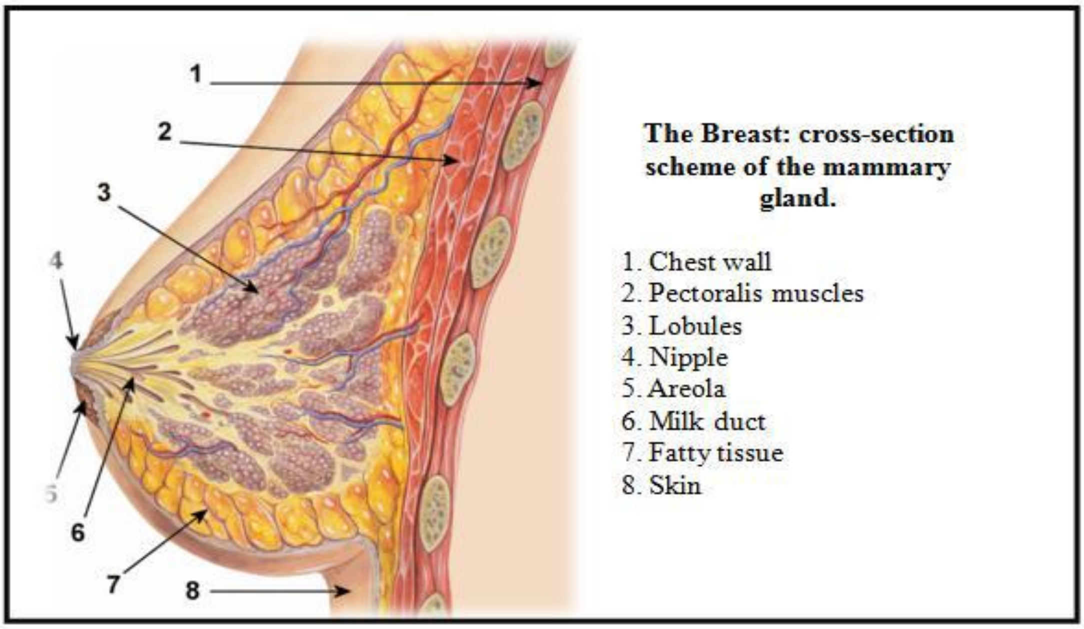


Figure (2.1) The Normal Breast Structure [20].

Cells are dividing and developing length of time and faded away, while the division and developing cells keep creating. The disordered in division of cells led to unregulated illness activity known as cancer as shown in Figure 2.2 in below. In medicine, the disease can be described as a malignant tumefaction that a small bulb of muscle and nipple that closed and open. [21].

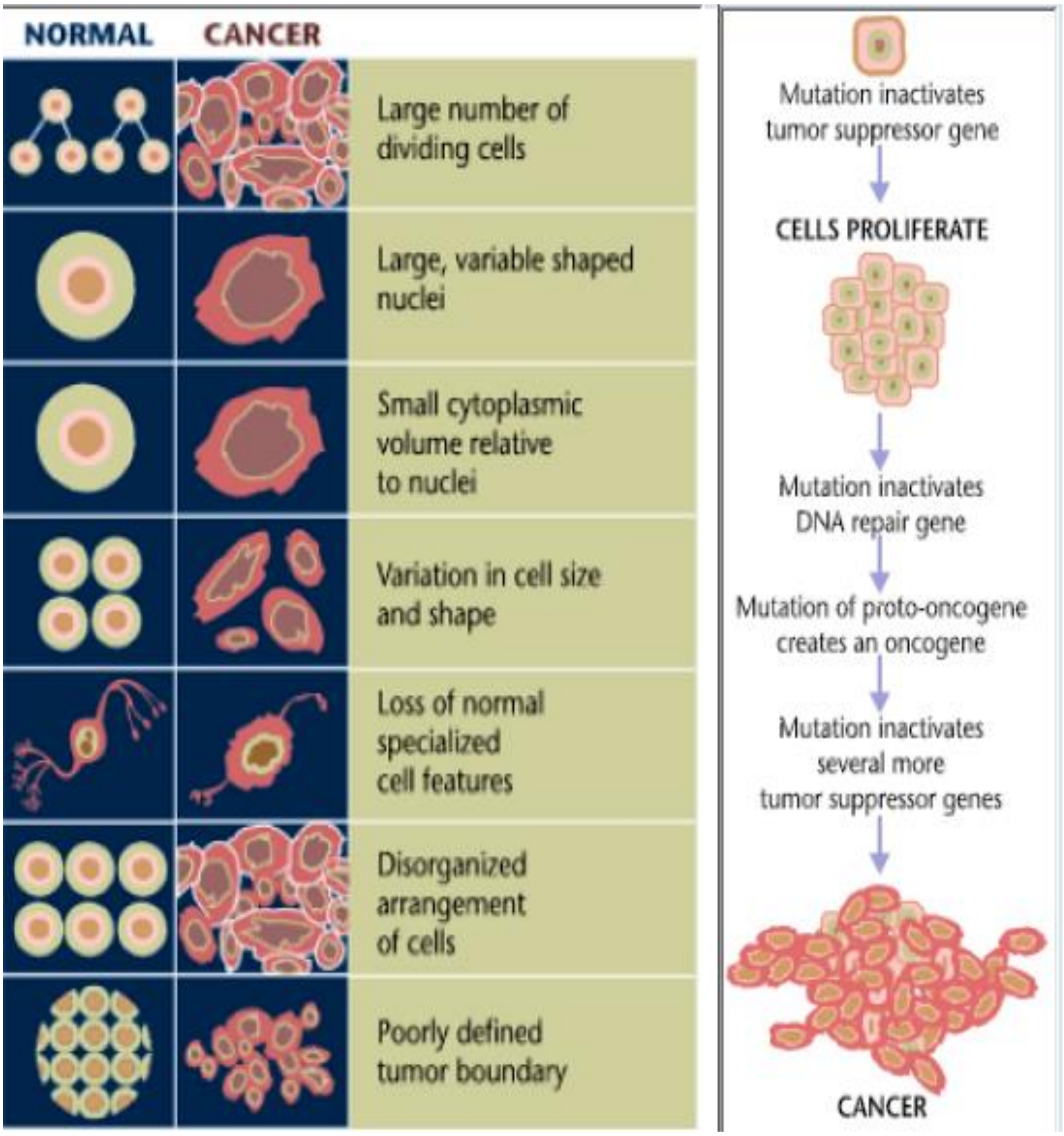

NORMAL
CANCER
Large number of dividing cells
Large, variable shaped nuclei
Small cytoplasmic volume relative to nuclei
Variation in cell size and shape
Loss of normal specialized cell features
Disorganized arrangement of cells
Poorly defined tumor boundary
Mutation inactivates tumor suppressor gene
CELLS PROLIFERATE
Mutation inactivates DNA repair gene
Mutation of proto-oncogene creates an oncogene
Mutation inactivates several more tumor suppressor genes
CANCER

Figure (2.2): Comparison of Normal and Abnormal Cell in Human [21].

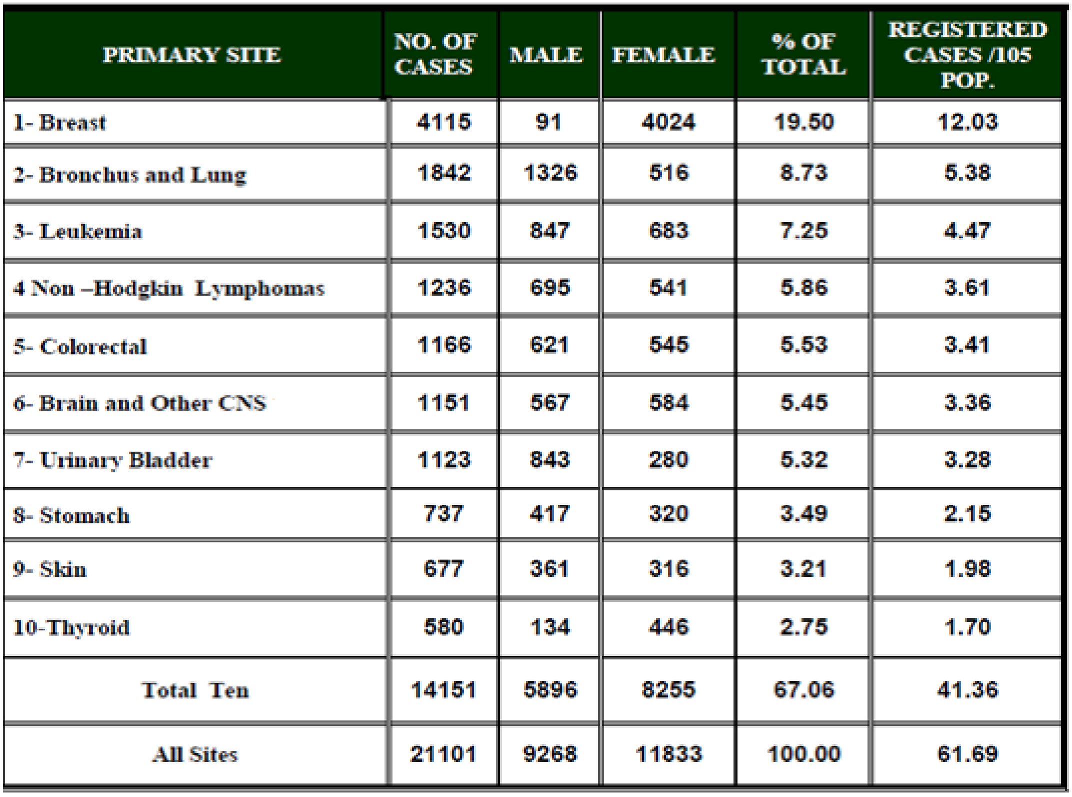
There are differences in the growth of cancer cells and healthy cells, where cancer cells appear irregularly, or the failure to repair is unlike the healthy cells, as in the table (2.1).

Table (2.1): Comparison of Normal and Abnormal Cell in Human [21]

Features and process of cells	**Normal cells**	**Tumor cells**
Growthing	Pick up piece of information from neighbor to stop	Ignorance a piece of information, keep growing
Repairing	Repair or die cell	Neither apoptosis nor repair.
Diffusion	Adhesion to another cell	No sticking, spread widely
Maturation	Grow up different functional kinds	not mature, undifferentiated and undying
Appearance	keep regular both of size and shape	Big nucleus with different in size and shape
Angiogenesis	Blood vessels are controlled by emergency	Store up blood vessels
Control	Controlled by cellular operation	Avoid diagnosis and deactivated enforcement

In 2014, the data of the Table (2.1) was remarked by the Ministry of Health Iraqi Cancer Register, that recorded cases of breast cancer are popular type of cancer in Iraq, as clarify in Table (2.3) and Figure (2.3) [26].

Table (2.2) The Included a Different Case of Cancer in Iraq at 2012[22]



PRIMARY SITE	NO. OF CASES	MALE	FEMALE	% OF TOTAL	REGISTERED CASES /105 POP.
1- Breast	4115	91	4024	19.50	12.03
2- Bronchus and Lung	1842	1326	516	8.73	5.38
3- Leukemia	1530	847	683	7.25	4.47
4 Non –Hodgkin Lymphomas	1236	695	541	5.86	3.61
5- Colorectal	1166	621	545	5.53	3.41
6- Brain and Other CNS	1151	567	584	5.45	3.36
7- Urinary Bladder	1123	843	280	5.32	3.28
8- Stomach	737	417	320	3.49	2.15
9- Skin	677	361	316	3.21	1.98
10-Thyroid	580	134	446	2.75	1.70
Total Ten	14151	5896	8255	67.06	41.36
All Sites	21101	9268	11833	100.00	61.69

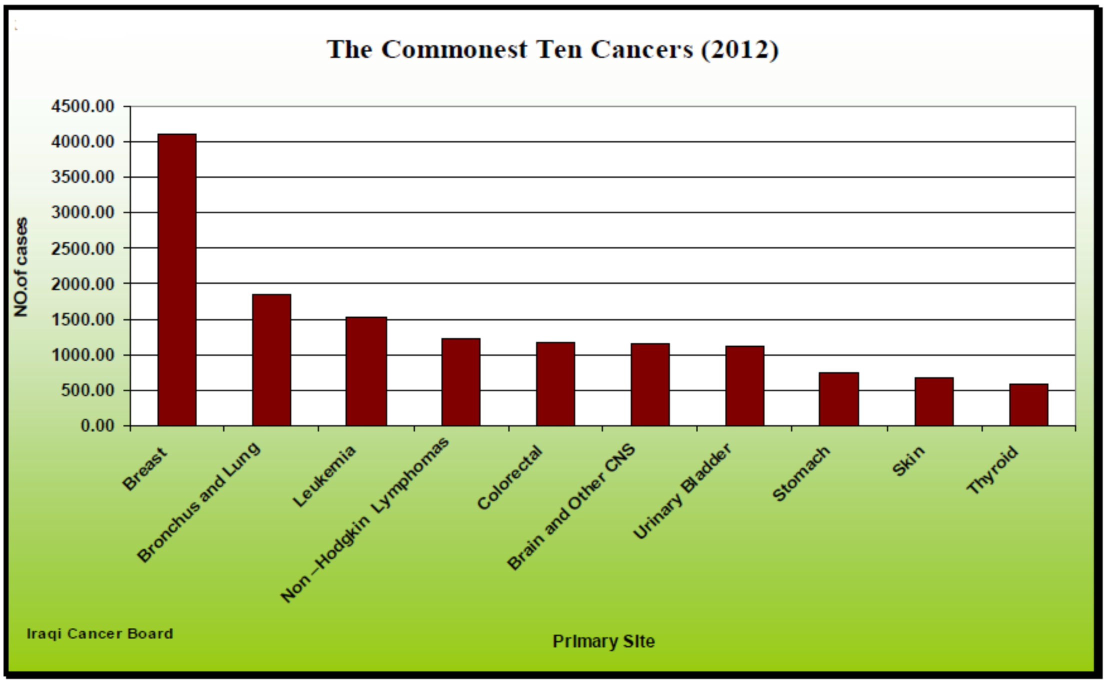
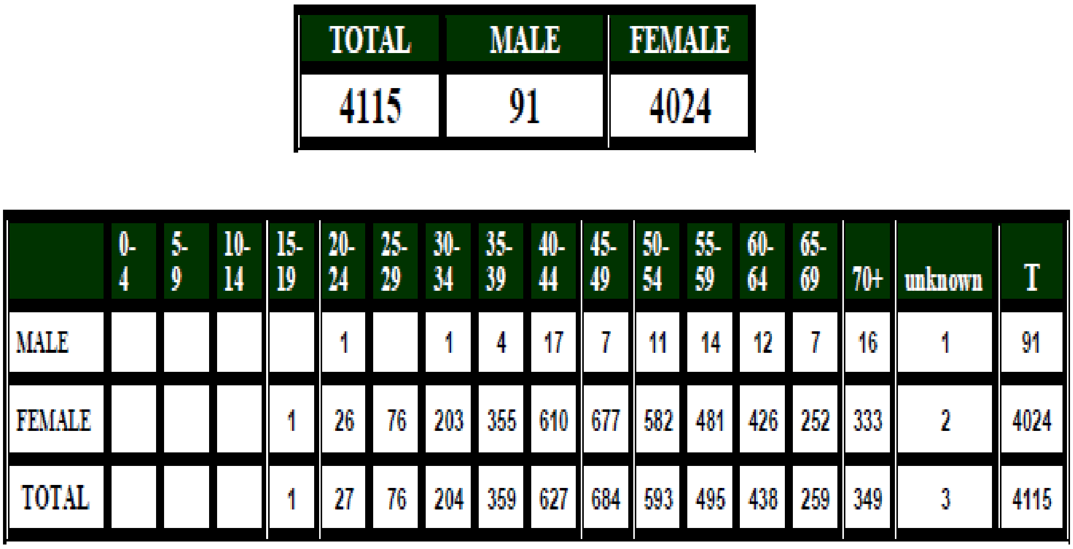


Figure (2.3): Scheme of more Popular Pen Tumors in 2012[22]

In 2012, around 3540 breast tumor cases in women and men are recorded by Iraqi Cancer Board in Ministry of Health, that is accounting about 19.15 percentage of all new diagnosed tumor cases. Thus, there were around 76 cases among men and about 3464 cases for women. Breast cancer ranks the most popular cancer type in Iraq from 1986 to 2010. The comparison between the population of breast cancer among females in Iraq in years 2008 and 2010 shows the least rate (16.65) per 100,000 in 2008 and the most rate about (21.75) per 100,000 in 2010. Other studies, in 2009, of population of breast tumor depends on the specific age of females that recorded high rate around 99.23/100,000 diagnosis in ages between 45 and 49, while 114.79/100000 female in age from 50 to 54, however, in age 65-69 years; the rate was 108.42/100000, followed by ages between 60 and 64 diagnosis around 92.26 /100,000 female. Histopathological distribution of female breast cancers showed that 58.56% % were infiltrating duct carcinoma, 6.47% Intraductal papillary adenocarcinoma, 3.79% were lobularcarcinoma , adeno carcinoma 2.40 %, and the remaining were other types of morphology as shown in table (2.4) [22].

Table (2.3) Cancer Distribution by Age, Sex, and Morphology in 2012[22]



TOTAL	MALE	FEMALE
4115	91	4024

	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70+	unknown	T
MALE					1		1	4	17	7	11	14	12	7	16	1	91
FEMALE				1	26	76	203	355	610	677	582	481	426	252	333	2	4024
TOTAL				1	27	76	204	359	627	684	593	495	438	259	349	3	4115

2.3 Intelligent Computing

Intelligent computing is a section of computer science that is capable to analyses complicated medical information a dataset that is utilized for diagnosis, therapy and forecasting effects in several medical scenarios. Smart artificial models can practical data well-ordered to reach to the right decision. The data provide to the framework can be in medical form or non-medical information. The main aim for computer technology in medicine is to progress more smart tools to assist the doctors or clinical precision regarding their options or decision working. A smart framework can deal with the problems of identifying an area of overlapping symptoms by utilizing human skills and knowledge as shown in figure (2.4) [23].

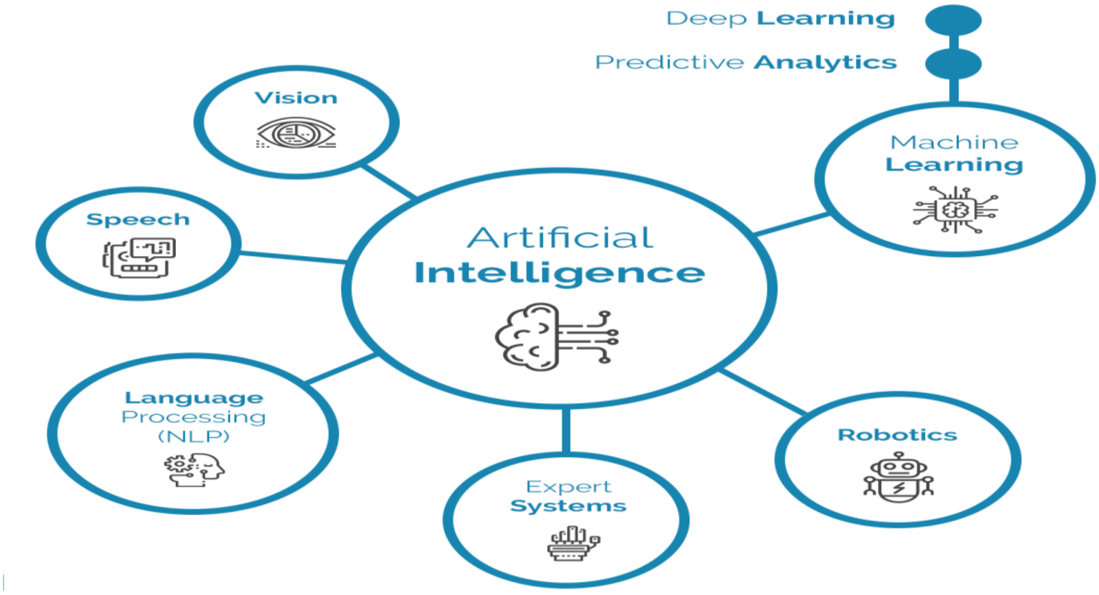


Figure (2.4): Artificial Smart Framework [22].

Practically, human observation includes using of unknown quantity of value is fuzzy group. That support the conception of a linguistic changeable. The fuzzy framework is better accurate while a pretty characterization of the issue in linguistic variables is available as shown in figure (2.5) [24].

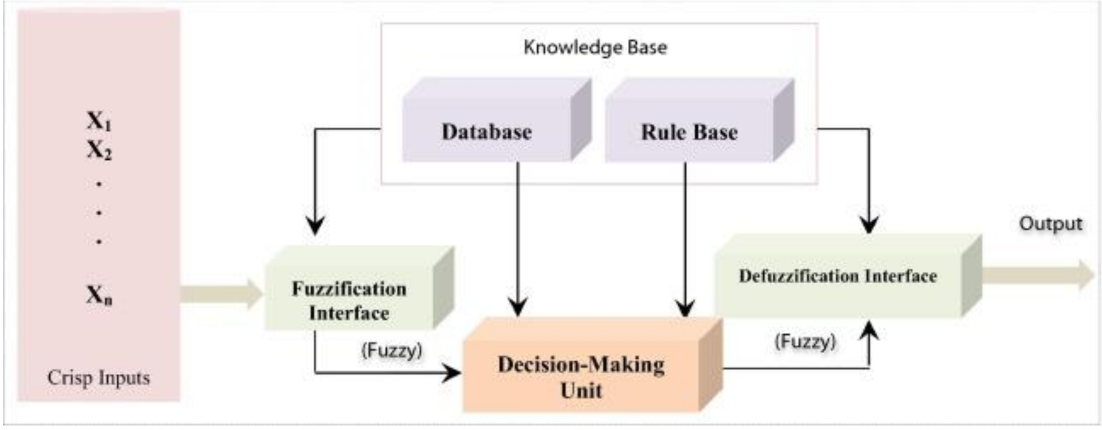


Figure (2.5): Common Fuzzy Framework [24].

Artificial neural networks represent the math structure modelling, which revels human neural such as learning and the ability to generalize this reasoning. Artificial-neural frameworks are portion of the area of machine learning have also been applied to analysis and recognition data as shown in figure (2.6) [25].

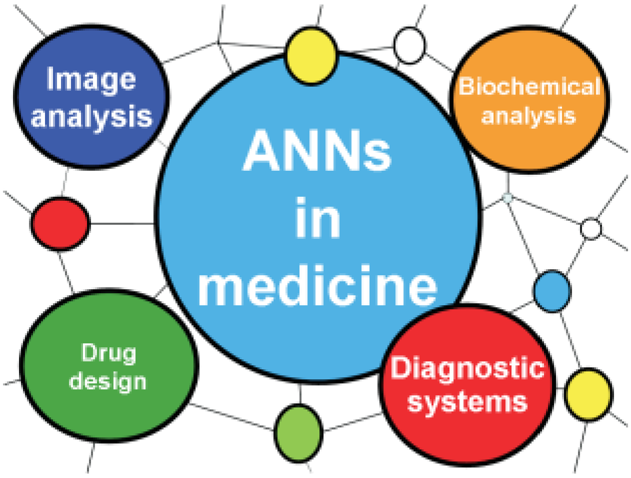


Figure (2.6): The Essential Application of Artificial Neural Framework in Clinic.

In general, the genetic algorithm works on the structure of binary string, identical to biological genomes. These structures evolve with period depends on the rule of survival of the fittest by utilizing unsystematic, structured, data exchange plot. Therefore, in each generation novel group of binary strings is formed, utilizing sections of the fittest members of the ancient group [26].

2.4 Machine Learning techniques

Machine learning is sole of the most important study objects due to the tremendous influence of Alpha Go and other AI applications. The sub-area from AI and computer science is known by Machine learning. It represents an area that uses certain unique algorithms to "learn" computer systems with specific data without programming complex. Particular, it is a mechanism that helps computer systems or computers to perceive, observe, understand, and foresee the world such as a human existence. “Machine learning is the research of computers for the development of new information and abilities and the realignment of established expertise [27]”.The machine learning and people did work to enable computer learning, skill acquisition and automatically building its own world of knowledge. After that, Samuel specifically coined the term "machine learning" in 1959, which originated from several areas of artificial intelligence research, like object recognition. A core concept of a machine learning approach is for encourage an algorithm to gain experiment and adapt itself as without as well much human involvement. It used to solve problems that are complicated to model. In machine learning, data sports a crucial part. The data styles decide learning outcomes and impacts. In the first case, machine learning requires certain data sources, which are often ranked as samples, practice groups, and patterns [28]. A machine, with the aid of the data groups provided, remodel the inner the relationship between them, the result of 'learning' (also known as 'training,' and attends the obtained knowledge byways of particular output forms, such as recognition, classification, and prediction. Specifically, mathematical variable produce by the regression model. Classification models divided into three categories established on data set has labels on its learning attributes: unsupervised-learning, supervised-learning and semi supervised-learning.

2.4.1 Support vector Machine (SVM)

The main idea of SVM training is to find the optimal hyper–plane that separates two classes of training patterns, $\{(x_i, y_i)\}_{i=1}^{n}$, $y_i \in \{+1, -1\}$, where n is the number of training patterns. The optimal hyper plane will be:

$$f(x) = w^t x + b = 0 \tag{2.1}$$

The margin region can be defined as the region between f(x) = 1 and f(x) = −1. Because SVM employs the structural risk minimization principle, the minimum margin of each class of patterns, 1/||w||, is to be maximized in SVM training, and all training patterns should be located on the margin boundary or outside the margin region. However, there can be some training patterns located inside the margin region or even in regions representing other classes of patterns. By introducing slack variables, the SVM soft margin optimization problem can be defined as:

$$Minimize\ 1/2||w||^2 + C\sum_{i=1}^{n}\xi_i \tag{2.2}$$

where $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \ldots, n, \xi_i \geq 0, i = 1, \ldots, n,$

Where C > 0 is the trade–off cost between the empirical error and the margin. By introducing Lagrangian multipliers, the primal optimization problem can be converted into a dual form [28] :

$$Maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \qquad (2.3)$$

Where 0 ≤ αi ≤ C, i = 1, . . ., n,

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \qquad (2.4)$$

The solution of SVM is

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \qquad thus \quad f(x) = \sum_{i=1}^{n} \alpha_i y_i x_i^T x + b. \qquad (2.5)$$

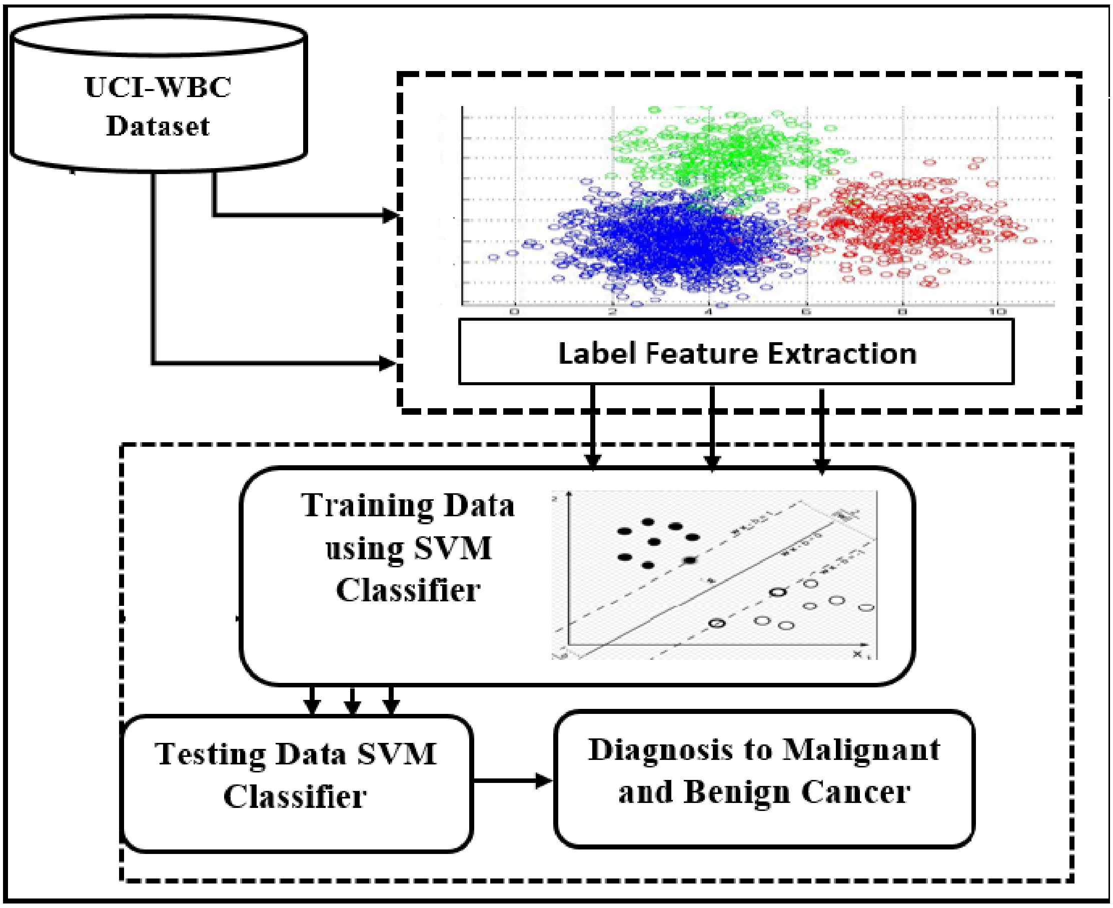


Figure (2.7): The Step Simplified Procedure for Support Vector Machine [28].

2.4.2 Logistic Regression (LR)

In the previous post, a brief overview of methods to analysis data was being explored, extended from the first step which is a fundamental statistic to the other step which is a Machine Learning (ML) and advanced simulations. This by time had a quite elevated-point overview, and beside from the statistics, as mentioned before, with not much detail. In this current post, a machine-learning-driven classification and regression has been taken a deep attention to the equipment in the data analyst's toolbox. Many calculations can do it manually, however this would be taken a great amount of time and extremely tedious. To do these tasks in a simple way by using some data analysis programs [29].

The model of logistic regression is utilized as the interaction between different groups of predictor variables and a categorical outcome variable. Traditionally, here X is a vector includes independent predictor variables x , xi = 1, 2, …, n, while π(x) is the conditional probability of experiencing the event. The model of the dichotomous or binary outcome when Y=1 presents according to the following eq [50].

$$Y = \pi(X) + \varepsilon \qquad (2.6)$$

When ε represents as a random error term that is dependent variable vector and π(x) can be expressed as the following eq.

$$\pi(X) = P(Y = 1|\mathrm{X}) = \frac{e^{X^T\beta}}{1+e^{X^T\beta}} \qquad (2.7)$$

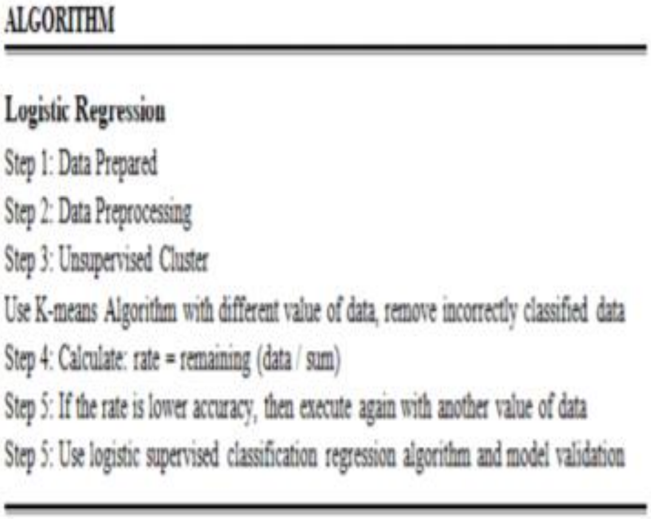
When β represents as the model's parameters vector. Alternatively, (23) can be written as follows.

$$\ln\left(\frac{\pi}{1-\pi}\right) = (\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n) \qquad 2.8$$

The Regression is categorization that attempts to discover the class of object depends on the particular future. This is all there is whilst at the surface level. These incision statistics is accomplished whilst we practice a ML model, in addition we need to experiment it. we sometimes separate this is because data can be expensive, and it take a longer time to gather.

Before build the regression model, dataset should be prepared and split into two parts, the first part represent the training data (with label) usually around (70% or 80%) from the all data , while the second part is testing data (without label) around (30% or 20%). The regression model has been built depends on the training data, then the model tested by using testing data , the high ration accuracy of prediction based on type and number of feature utilized in training phases, the application applied in Least squares regression like: [28-30].

In the LR algorithm process is as follows:



ALGORITHM

Logistic Regression

Step 1: Data Prepared

Step 2: Data Preprocessing

Step 3: Unsupervised Cluster

Use K-means Algorithm with different value of data, remove incorrectly classified data

Step 4: Calculate: rate = remaining (data / sum)

Step 5: If the rate is lower accuracy, then execute again with another value of data

Step 5: Use logistic supervised classification regression algorithm and model validation

Fig. (2.8) The flowchart of algorithm LR

2.4.3 Decision Tree

The class of supervised learning falls under decision tree algorithm. One of the benefits of a decision tree could utilize to solve both problems of classification and degradation. The decision tree utilizes impersonation tree for solving a problem at every terminal node matches with a group naming and appear the characteristics at an internal node of the tree. In this research could exemplify any logical task on separate imputes utilizing the decision tree. Here are several of the suppositions that can use through using the decision: First, in this research consider the entire training group such a root. Feature parameter chooses that be categorical. When values are continuous, they will be estimated before creating a model. This takes into account the values of the attributes and then distributes the records frequently. In this research utilize statistical manners to arrange characteristics such root or an internal node. The decision tree doing within a product aggregate model which is else known as a separate normal model. As shown in the image, the use of computers in people's daily lives [31]. The key challenge in Decision Tree is to define the characteristic of a root node in all steps. The procedure is to know a selection of characteristics. The Information Gain applied the decision tree in training cases to minimal subgroups. Information gain is a gauge for variation in entropy. That known: the likelihood vector has one variable value only (the variable x has only one value) and a variable is known such true. From another side, the impurity level reaches the limit that means all ingredients are equal. Looking at the training group S, the prospect vector for the objective attribute y was determined as shown in equation 2.9 [31].

$$P_y(S) = \left(\frac{|\sigma_{y=c_1} S|}{|S|}, \ldots, \frac{\left|\sigma_{y=c_{|dom(y)|}} S\right|}{|S|} \right) \tag{2.9}$$

Partition quality due to a separate feature identified as a decrease in target feature impurities after S split according to values vi,j∈ dom(ai) as shown in equation 2.10 [31].

$$P_y(S) = \left(\frac{|\sigma_{y=c_1} S|}{|S|}, \ldots, \frac{\left|\sigma_{y=c_{|dom(y)|}} S\right|}{|S|} \right) \tag{2.10}$$

Gini an index is a bug-based standard measures the differences among the prospect allocations of feature values as shown in equation 2.11 [52].

$$Gini(y, S) = 1 - \sum_{c_j \in dom(y)} \left(\frac{|\sigma_{y=c_j} S|}{|S|} \right)^2 \tag{2.11}$$

Thus, the valuation criterion for choosing a feature was defined in equation 2.12:

$$GiniGain(a_i, S) = Gini(y, S) \sum_{v_{i,j} \in dom(a_i)} - \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot Gini(y, \sigma_{a_i=v_{i,j}} S) \tag{2.12}$$

Entropy is an indicator of the suspicion from a variable random since it defines the flaws of a series of instances. When entropy reaches maximum, the more information available, Definition: Assume C is a collection of cases, that E is element [32].

The structure of Decision Tree use Information Gain basics: begin with all training cases aligned with the root node using data for select which characteristic each node should be identified route would include the similar distinct characteristic of frequently create each sub-tree in the sub-group of training cases in order to list along with that rout in the tree [32].

2.4.4 Random Forest

Random forest is the most popular algorithm, namely, it assembles a large amounts of decision trees from training dataset, and it also uses a tool called bagging to perform classification and regression tasks . Each decision tree represents a class prediction, this method collects the votes from these decision trees and the class with most votes is considered as the final class [33].

$$(x_1, y_1), \ldots, (x_n, y_n)$$

Random forest (RF) consists of a set of classifiers, each of which has a tree structure. Assume that a specific random forest has k trees of classifiers that is defined as $h(x, \Theta_n)$ for $n = 1,2, \ldots, k$, where $\{\Theta_n\}_{n=1}^{k}$ is a set of independent identically distributed random vectors and x is the input, and each tree votes for the most popular class at input x [34].

As the RF training ends with k trees, which are unique classifiers models, the test phase will use the popular majority vote among those distinct trees:

$$H(x) = \text{argmax}_Y \sum_{i=1}^{k} I(h_i(x) = Y) \qquad 2.13$$

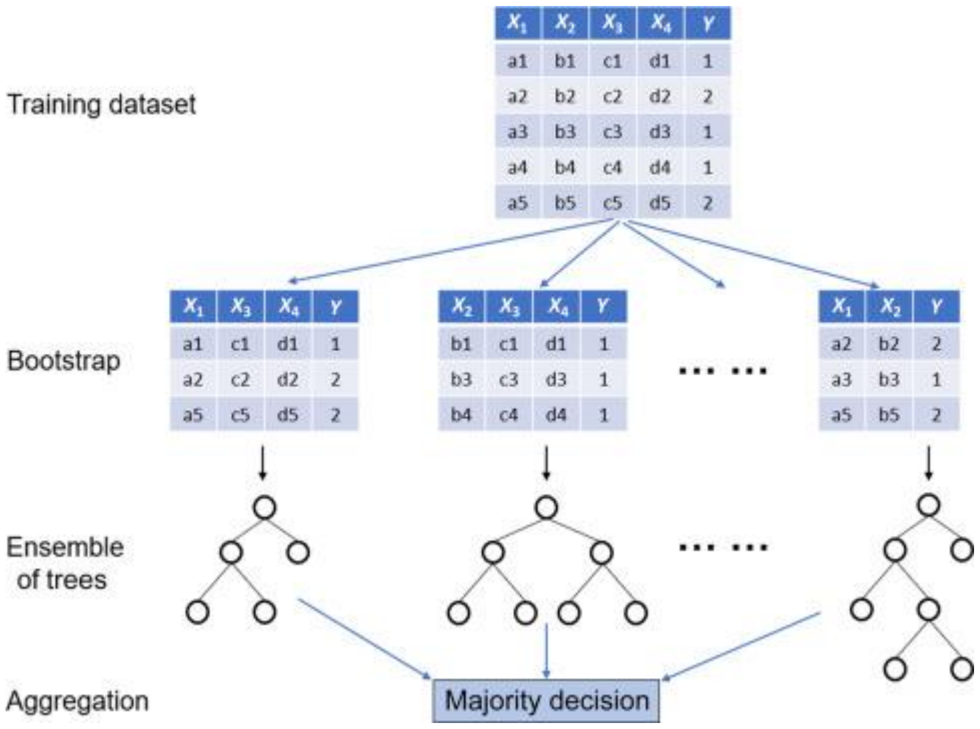
Where

$H(x)$ is combination of classification model,

h_i is a single decision tree model,

Y is the output variable,

$I(.)$ is the indicator function [35].



X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2

Figure (2.9) The Random Forest Model that performs the prediction [36]

RF classifier is an ensemble method that trains several decision trees in parallel with bootstrapping followed by aggregation, jointly referred as bagging (Fig. 2.9). Bootstrapping indicates that several individual decision trees are trained in parallel on various subsets of the training dataset using different subsets of available features. Bootstrapping ensures that each individual decision tree in the random forest is unique, which reduces the overall variance of the RF classifier. For the final decision, RF classifier aggregates the decisions of individual trees; consequently, RF classifier exhibits good generalization [36].

2.5 Performance

Statistical concepts in this paragraph, a set of concepts that were used as tools in designing and implementing the proposed system will be reviewed in the Statistical Analysis of Data section.

Table (2.4) Criterion Performance Metrics

No.	Rubric	Abbreviation	Determination	Characterization
1.	Accuracy		$\boldsymbol{Acc = \frac{A+D}{A+B+C+D} \times 100\%}$	accuracy, while specificity is referred to as true negative rate or negative class accuracy.
2.	Sensitivity	S_n	$\frac{TP}{TP+FN}$	The percentage of positive styles suitable recognized as positive Symbolize by a S_n
3.	Specificity	S_p	$\frac{TN}{TN+FP}$	The ratio of negative styles recognized correctly as negative is measured by a S_p
4.	Confidence	Q(x)	$Q(x) = \begin{cases} 1 & \text{if } uncertainty(x) \geq c \\ 0 & \text{otherwise,} \end{cases}$	Our confidence-based active learning approach is based on identifying and annotating uncertain samples because such samples provide more information to the learner.
5.	No Information	H(Y)	$H(Y) = -\sum_{y} p(y) \log_2 p(y)$	Shannon introduced “entropy” concept as the basis of information theory
6.	Kappa	K	K = PR(*e*),	the kappa statistic is a measure of how closely the instances classified by the machine learning classifier matched the data

				labeled as ground truth, controlling for the accuracy of a random classifier as measured by the expected accuracy.
7.	Mcnemars test P value		$\chi^2 = \frac{(B-C)^2}{B+C}.$	McNemar's test is also referred to as "within-subjects chi-squared test," and it is applied to paired nominal data based on a version of 2x2 confusion matrix
8.	Positive prevalence value PPV	**PPN**	$\frac{TP+TN}{TP+TN+FP+FN}$	A PPV test measures a positive to a total number of styles known as positive.
9.	Negative Prevalence value NPV	**NPV**	$\frac{TN}{TN+FN}$	The NPV a measure of the true negative percentage of the total number of styles as negative.
10.	Detection Rate	**DR**	$(\mathrm{DR}) = \frac{TP}{TP+FP}$	More important than accuracy and times for training are the threat detection rates, is expected to detect as many security problems as possible.
11.	Precision	**P**	$\frac{TP}{TP+FN}$	The percentage is measured by a P.
12.	Detection Prevalence		$\mathbf{D_P} = \frac{\mathbf{A+B}}{\mathbf{A+B+C+D}}$	is defined as the number of predicted positive events (both true positive and false positive) divided by the total number of predictions.
13.	Balanced Accuracy		$\mathbf{B_{ACC}} = \frac{\mathbf{S_V + S_P}}{\mathbf{2}}$	in binary and multiclass classification problems to deal with imbalanced datasets. It is **defined** as

				the average of recall obtained on each class. The best value is 1 and the worst value is 0 when adjusted=False .

These three metrics were used through classification problem PPV, NPV, and P. These percentages represent the correct classification percentages for normal, cancer, and collected types respectively according to the definitions. Wherefore for our eventual classifiers, expect a high rate for all three measurements. It is natural to conclude that execution in cancer types is comparatively more important than execution in normal types. In practice, the wrong classification of cancer type can command to infinitely more dangerous consequences than the wrong classification from a normal type. Thus, we attach additional significance for NPV than PPV while test classifiers [37].

Chapter Three

The Proposed Model

Chapter Three

Breast Cancer Classification using Developed Techniques

3.1 Introduction

This chapter includes the proposed system (the intelligent computing model), where a detailed explanation of the general scheme of the system and the tools and algorithms used in the system was provided.

3.2 The Proposed Models

This work focused on statistical learning and data mining predictive modeling methods. The common steps used for building a predictive model in this work are as follows: first steps are load the data set, second step is preprossing (date initializtion processig) that activated by data clining and nullvlaue in dataset and normaliztion the data set , third step splitting training and testing datasets, fourth step implement statistical learning and data mining technique to training set and obtain the predictive model then perform an evaluation for model using testing dataset repeat the operation using different techniques compare to perform between data extraction and statistical science techniques as shown at the figure (3.1).

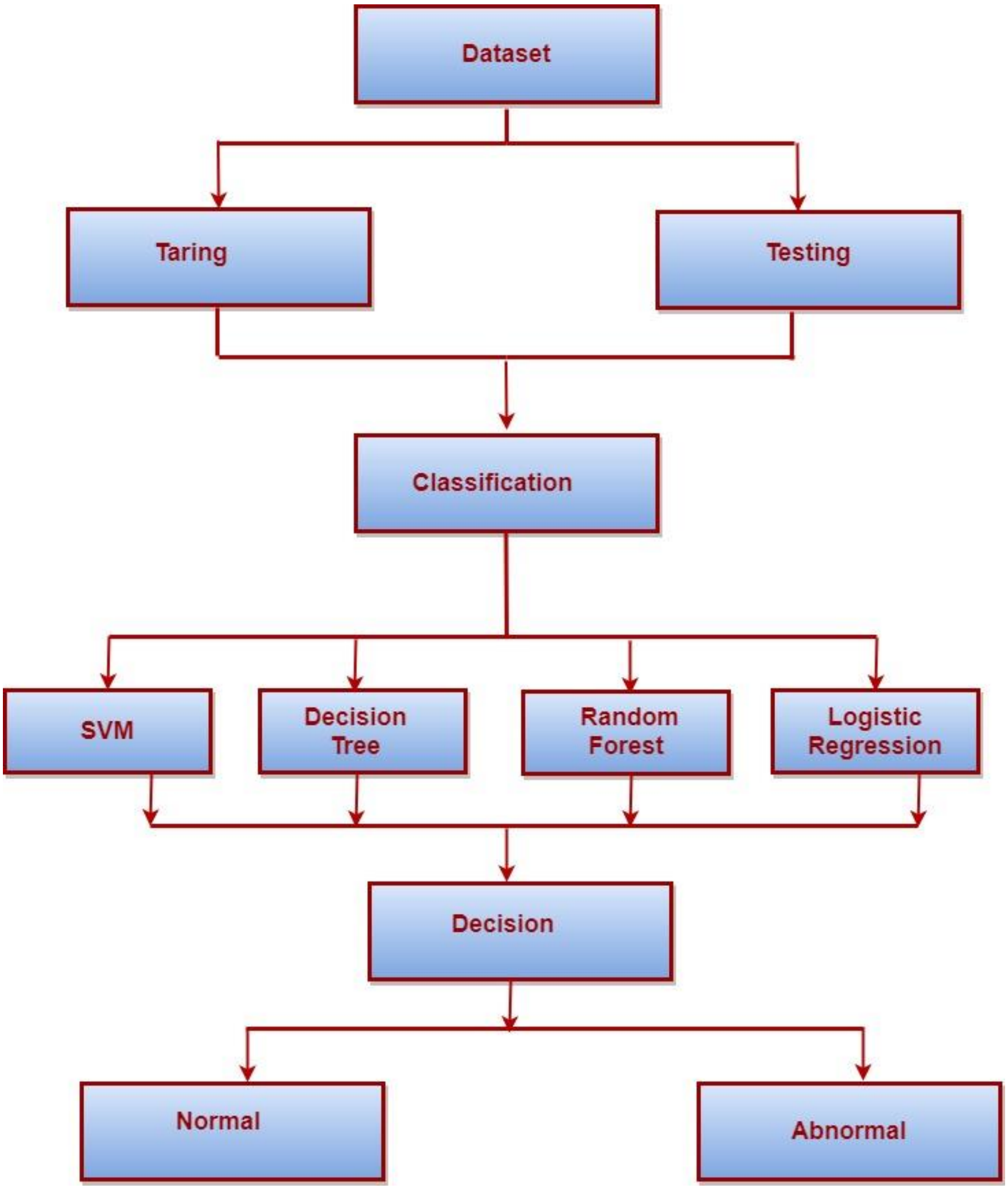


Figure (3.1) Genral Block Daigram For Propose Model.

3.2.1 Data set Properties

In this section, the data set will be treated using statistical analysis .The breast cancer data obtained at UCI are used to train a random forest model was generated from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. A database reflects real clinical cases and grouped in a chronological order. There are 699 instances with the following attributes:

Table3.1: Characteristic of UCI Data

Characteristic	Domain
1. Sample code number	ID Number
2. Clump Thickness	1 – 10
3. Uniformity of Cell Size	1 – 10
4. Uniformity of Cell Shape	1 – 10
5. Marginal Adhesion	1 – 10
6. Single Epithelial Cell Size	1 – 10
7. Bare Nuclei	1 – 10
8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitosis	1 – 10
11. Class	2 for benign, 4 for malignant

Specimens from the UCI repository are used in the thesis study. These Specimens were Wisconsin Breast Cancer Dataset (WBC) original, Wisconsin Breast Cancer Diagnosis (WDBC) and Wisconsin Breast Cancer prognosis (WPBC). WBC is made of 699 registers, every of which has 9 advantages positive the group characteristics. WDBC is made of 569 registers, each one has 32 advantages positive the group characteristics while WPBC included 198 records every of which has 34 advantages positive the group characteristics. The data collection is very good results were realized by the use of a high-quality data; therefore, it was necessary to choose them based on the goodness of the data collection operation. In this thesis study is built at utilizing the UCI online databases that are publicly available.

3.2.2 Data Cleaning

In addition, the raw data in the genuine world has noise m (numbers of ?). The broad dataset registry includes numerous types of anomaly values that affect the results of the study since good models typically need good data for data extraction, databases around the world are not necessarily clean and contain noise, deficient data, redundant information, inexact data, and lost values. Lost data is a popular defect in many factual data groups. Noise decreasing or elimination used for treating lost values. There are 16 cases in WBC and 4 cases in WPBC have one lost characteristic value, referred to as "?" Additionally, there are 9 states in BCD that have two lost values that are usually replaced by the average value of that characteristic-based on statistics. The current study used the approach to building lost characteristic values to satisfy the completion component. The missing feature values were because of inexactness in data containing invalid feature values.

In this step, the results relies on the quality of the data, so a clean dataset has a major impact on the success of the classification results. This process is the first step in classify WDB, where this process reduces the complication of the data and provides better conditions for further analysis. Through this operation, the complexity of the data is known and the processing of the data is done more carefully and effectively. It is unrealistic to anticipate data to be perfect after being extracted. Machine-learning algorithm models generally require good data.. Not only rightness but also a consistency of values is important. Lost data can be a particularly harmful problem. Particularly at the number of lost data is large, not all characteristic with lost values can be deleted from the sample.

3.2.3 Data Transformation and Normalization

Normalization is one of the most common tools used by auto This measurement method is useful when the data group does not contain the extremes achieved in the value after normalization equal machine learning labels to obtain accurate results. Training time is hurried up by the data normalization, where the training time is started to access a function of the same size. Normalization helps to turn the value of the attribute into a limited set. It is the process of sending data to a particular range, such as between 0 and 1 or between -1 and 1. Normalization is needed where there are wide variations in the functionality of different feature ranges. (Value before Normalization – min) / (max - min).

The process of selecting feature is of high importance as it needed to recognize patterns, statistics, and data extract. This characteristic selection is used in this study to reduce the number of characteristics in the data group before starting the extract process. The current thesis just depends on UCI depository.

3.2.4 Dataset Splitting

Divide the dataset to training and testing sets, with 80% training and 20% for test from a whole set. For this purpose, we use the R package Tools. Inspect all parameters ‘columns’ of the dataset using boxplot using ggplot2 library in R. 6: Depending on the boxplot, some of the parameters needed to be adjusted by grouping observations of the same trend into a single group. The threshold that is used to decide which observations are similar is chosen by inspecting each parameter individually verify the selected threshold in of previous step by doing F-test.

3.3 Classification with Machine Learn Technique

There are many analytical methods available in R programming environment. For instance support vector machine (SVM),logistic regression, random forest regression/classification, and decision trees.

3.3.1 Logistic Regression

The following logistic regression is used to classify the Breast dataset. The two-column data of the sample are obtained, corresponding to the attributes of mean radius and mean compactness, and the coordinates of each point are (x, y). First, get the minimum value, the maximum value and the step length from the first column of the two-dimensional array to generate a new array. Then, get the minimum value, the maximum value and the step length from the second column of the two-dimensional array to generate another new array can be described in the following steps. Show the algorithm 3.1

Algorithm (3.1): Logistic Regression

Input: Wisconsin Breast Cancer data set.

Output: benign or malignant

PROCEDURE:

Step 1: Acquire dataset from Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository.

Step 2: Pre-process data.

A: Remove Sample Code Number from attribute list.

B: Numeric to nominal type of data conversion of Class attribute. (2 – Benign, 4- Malignant).

Step 3: Pre-processed dataset uploaded in R studio software toolkit for analysis.

Step 4: Information Gain algorithm applied in R studio software of respective attributes record.

Step 5: Applying Logistic Regression Model without intercept on dataset with constant nonzero column, Spark MLlip outputs zero coefficients for constant nonzero columns) that represent as:

A: Start at the root node.

B: Training phase: For each X, find the set S that minimizes the sum of the node impurities in the two child nodes and chooses the split $\{X^{*} \in S^{*}\}$ that gives the minimum over all X and S. If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn by apply the equation 2.6 , equation 2.7 , and equation 2.8.

C: Testing phase : that achieved after get the training model and apply for all data based on equation 2.6 evaluation by calculate the accuracy of systems.

Step 6: Diagnosis of new patients is achieved by cross referencing new attribute values in the decision tree and following path till the leaf node reached which would either specify benign or malignant tumor.

3.3.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithmic rule, which can be utilized for all regression or classification challenges. It is principally utilized in classification issues. The algorithm rule, plot every data element as a point in the n-dimensional space where n is the number of characteristics one has with the value of each function being the value of a certain coordinate. After that, the classification is performed by discovering the hyper-plane that differentiates the two categories well. Often researchers tend to plot each cognitive element as a limit in a region of n dimensions with the value of each feature with a chosen coordinate value. After that, to do the classification by finding the hyper-plane level that differentiates the two fine categories. It is a non-probabilistic binary linear classifier but is often manipulated in such a way that it also performs non-linear and probabilistic classification, creating a versatile algorithmic program. The SVM model can illustrate states as points in a specific region, so that they would be categorized and divided by a transparent hole. Then, new cases are classified in a specific area, and each part of the difference is assumed to be covered in which group. The clearest advantage of SVM is its high effectiveness in areas with high dimensions. For boot, it is combined for memory because it uses a set of training points within the call process that can be described in the following steps.

Algorithm(3.2) : SVM of WBCD diagnosis
Input: WBC data set is pre-processed to satisfy prerequisites of the data mining technique.
Output: benign or malignant
Procedure: **Step 1:** Fitch dataset from Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository. **Step 2:** The data is pre-processed in order to apply the technique of svm a. Remove Sample Code Number from attribute list ,Numeric to nominal type of data conversion of Class attribute. (2 – Benign, 4-Malignant). **Step 3:** SVM applied as describe bellow

Where Input

D: the WBC dataset,

L: The number of sample

m : the fraction of pattern in a sample

output

S: classy to benign and malignant

procedure of SVM

✓ concord Di for j=1 ...,m by 80% sampling from the dataset

✓ Train SVM with D_j ,$\forall_j$ by apply equation (2.1) and (2.5) where α=0.3 and start with random initial weight

✓ Testing SVM :Evaluate all D by f_j , $\forall_j$ as shown in equation *2.7*

v_{ij} = decision value of x_i evaluated by f_i

p_i = Evaluation accuracy of f_j for D

✓ calculate the estimated margin $M(x_i)$

$m(x_i)=1/l(sum\ (p_j / sum(p_j * v_{ij}))$.

Step 4: where is training is testing, cross-reference the values of new attribute in the decision tree and track the path until finding the leaf node, which would determine if the tumor is malignant or benign.

3.3.3 Decision Tree

Decision tree may be a supervised learning rule that is utilized for regression and classification. It splits the information into two or more subsets support the input variables value. A value cacophonous or operates criterion is employed to indicate the split of most effectiveness among all the split points. The recursive splitting process of info into teams continues until the leaves contain a single sample. Meanwhile, in this model, associate degree optimized version of the CART rule is used for implementing the choice tree classifier. Call trees are straightforward to perceive and interpret, compared to different classification algorithms. Moreover, call trees require a much less preprocessing since the outliers have no effect on the performance. Moreover, they provide no support for the Euclidian distance. Accordingly, feature scaling is not required. In addition, since the values would be modified, the feature scaling might result an incorrect assumptions being tacit. Call trees will deal with all numerical and categorical variables as inputs, therefore, it is acceptable for this model, since the info set contains each variable changes. In this model, the link between feature and target variables is complex and of a high non-linearity. Thus, a call tree contains a greater likelihood of outperforming linear models like provision regression. With all the benefits of the call tree come several disadvantages. One of them is the call trees will result in over fitting by generating a very complicated tree, and thus it will not drive a good prediction one new information. Finally, since call Trees are greedy algorithms, the optimum tree is not essentially came back can be described in the following steps.

Algorithm(3.3) : Decision Tree of WBCD diagnosis

Input: Wisconsin Breast Cancer data set.

Output: Benign Or Malignant

Procedure:

Step 1: Acquire dataset from Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository.

Step 2: Pre-process data for applying decision tree data mining technique.

a. Remove Sample Code Number from attribute list.

b. Numeric to nominal type of data conversion of Class attribute. (2 – Benign, 4- Malignant).

Step 3: Decision Tree algorithm implemented, generating a decision tree with leaf nodes as the class label (benign and malignant).

```
TreeGrowing (S,A,y)
 Where:
        S - Training Set
        A - Input Feature Set
        y - Target Feature
        Create a new tree T with a single root node.
        IF One of the Stopping Criteria is fulfilled   THEN
          Mark the root node in T as a leaf with the most common value
        of y in S as a label.
        .ELSE Find a discrete function f(A) of the input
        attributes values such that splitting S
        according to f(A)'s outcomes (v1,...,vn) gains
        the best splitting metric
        .IF best splitting metric > threshold THEN
        Label t with f(A)
                    FOR each outcome vi  of   f(A)
        Set Subtree  = TreeGrowing (σf(A)=viS,A,y).
            Connect the root node of  tT to Subtreeiwith
            an edge that is labelled as vi
        END FOR
        ELSE
             Mark the root node in T as a leaf with the most
            Common value of y in S as a label.
        .END IF
        END IF
        RETURN T
```

```
Tree Pruning (S, T, y)
          Where:
          S - Training Set
          y - Target Feature
          T - The tree to be pruned
          DO
             Select a node t in T such that pruning it
              maximally improve some evaluation criteria
           IF t=Ø THEN
             T=pruned (T, t)
           UNTIL t=Ø
      RETURN T
```

Step 4: Diagnosis of new patients is achieved by cross referencing new attribute values in the decision tree and following path till the leaf node reached which would either specify benign or malignant tumor.

3.3.4 Random Forests

The training of random forests is performed on various parts of one training dataset. This offers a solution to the problem of over-fitting, which is popular for the trees of decision. The approach that is utilized for generating the forest is illustrated in the following steps:

Algorithm (3.4): Random Forests of WBD Diagnosis
Input: Wisconsin Breast Cancer data.
Output: benign or malignant
Procedure: **Step 1:** Acquire dataset from Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository.

Step 2: Pre-process data for applying decision tree data mining technique. a. Remove Sample Code Number from attribute list b. Numeric to nominal type of data conversion of Class attribute. (2 – Benign, 4- Malignant)

Step 3: Decision Tree algorithm implemented, generating a decision tree with leaf nodes as the class label (benign and malignant) as shown in the steps bellow.

- Randomly sample n cases with replacement from the training data, where the total number of training cases is n.
- At each node:
  - Randomly select k out of a total of m features, where k < m.
  - Split the node on the feature with the best split.
  - At the next node, iterate step (a) and (b).
- Each tree is grown to the largest extent possible.
- Iterate steps 1 to 3 until N number of trees are generated.

Step 4: Diagnosis of new patients is achieved by cross referencing new attribute values in the decision tree and following path till the leaf node reached which would either specify benign or malignant tumor.

After the random forest is crated, predictions are found using the model. To start this process, input features for test data group are used in every forest decision tree. Classifying every tree separately is one vote for the corresponding category. The classification of most votes is the last prediction. It is necessary to make sure the correlation between the N generated trees. The error rate of random forest increases as these correlations increase. Therefore, the trees should be as uncorrelated as potential. Decreasing the value of k, both inter-tree correlation and strength of each individual tree will decrease as well. Hence, the optimal value of k must be determined. This optimal value is typically in the maximum range.

A random forest can handle lost values. In addition, the algorithm can be utilized to engineer features. This means that the algorithm can determine the most important features out of all the available features. A significant difference is noticed between decision tree and random forest algorithm, that difference causes the dissimilarity among the individual trees in the random forest. The process for obtaining an individual tree for a random forest has a small difference that that of decision tree. The nodes will be found in a random method as the steps above show instead of specifying the root node and other decision nodes using the information gain.

Chapter Four

Result and discussion

Chapter Four

Result and discussion

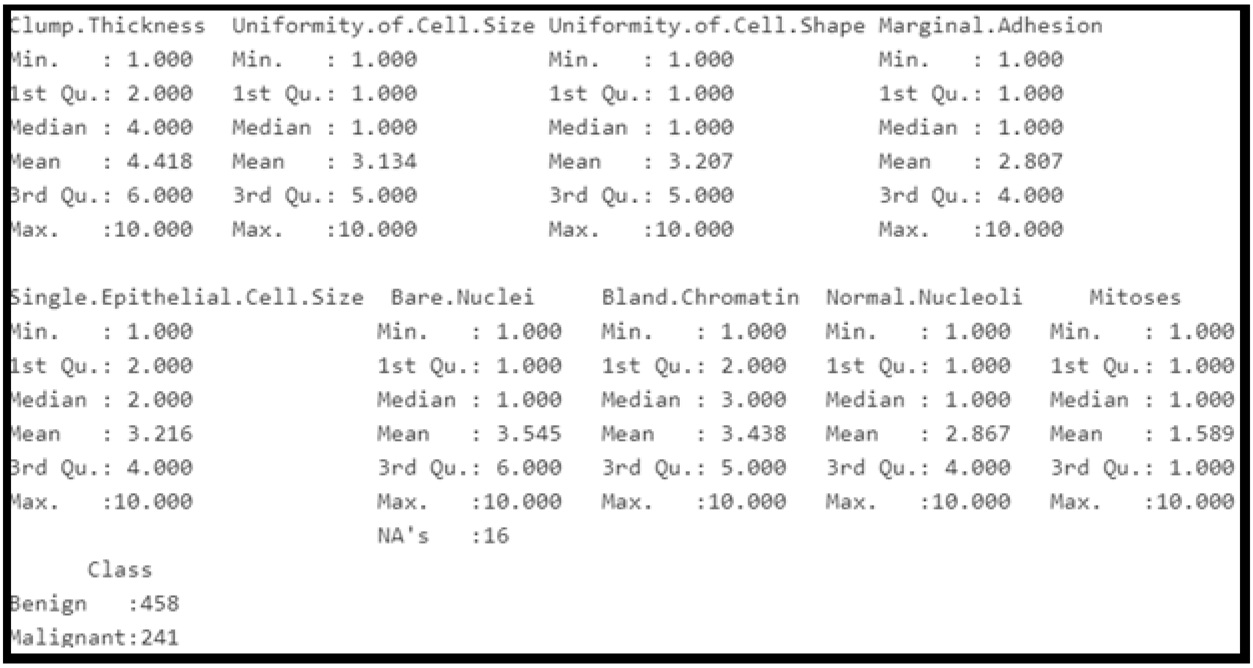
4.1 Introduction

This chapter includes a review and discussion of the results that were reached during the implementation of the proposed system

4.2 Data Set of Breast Wisconsin Diagnostic

In our case, we obtained the data sets from a web page called UCI Machine Learning Repository. In our case, we load the data from a Web URL, HTTP and the data format is “.csv”. For acquire both data sets, we had to do this process for each data set. Data set contains 12column and 698 row of feature breast cancer tissue that defines if the cancer is benign or malign. The benign case represent 458 row and malignant case represent 240 raw from dataset as shown in figure (4.1). The remaining 11 columns describe the following **attributes:** (1) radius, (2) texture, (3) perimeter, (4) area, (5) smoothness, (6) compactness, (7) concavity, (8) concave (9) points, (1 0) symmetry and(1 1) fractal dimension.

First step applied the function for load the dataset of Brest cancer as CSV format. Then apply the function **sum (data)** to display sum feature of data set sum as (min value of each column , max value of each column , and median value of each column) as shown in figure (4.1).



```
Clump.Thickness  Uniformity.of.Cell.Size Uniformity.of.Cell.Shape Marginal.Adhesion
Min.   : 1.000   Min.   : 1.000          Min.   : 1.000           Min.   : 1.000
1st Qu.: 2.000   1st Qu.: 1.000          1st Qu.: 1.000           1st Qu.: 1.000
Median : 4.000   Median : 1.000          Median : 1.000           Median : 1.000
Mean   : 4.418   Mean   : 3.134          Mean   : 3.207           Mean   : 2.807
3rd Qu.: 6.000   3rd Qu.: 5.000          3rd Qu.: 5.000           3rd Qu.: 4.000
Max.   :10.000   Max.   :10.000          Max.   :10.000           Max.   :10.000

Single.Epithelial.Cell.Size  Bare.Nuclei     Bland.Chromatin  Normal.Nucleoli    Mitoses
Min.   : 1.000              Min.   : 1.000  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
1st Qu.: 2.000              1st Qu.: 1.000  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
Median : 2.000              Median : 1.000  Median : 3.000   Median : 1.000   Median : 1.000
Mean   : 3.216              Mean   : 3.545  Mean   : 3.438   Mean   : 2.867   Mean   : 1.589
3rd Qu.: 4.000              3rd Qu.: 6.000  3rd Qu.: 5.000   3rd Qu.: 4.000   3rd Qu.: 1.000
Max.   :10.000              Max.   :10.000  Max.   :10.000   Max.   :10.000   Max.   :10.000
                            NA's   :16
      Class
Benign   :458
Malignant:241
```

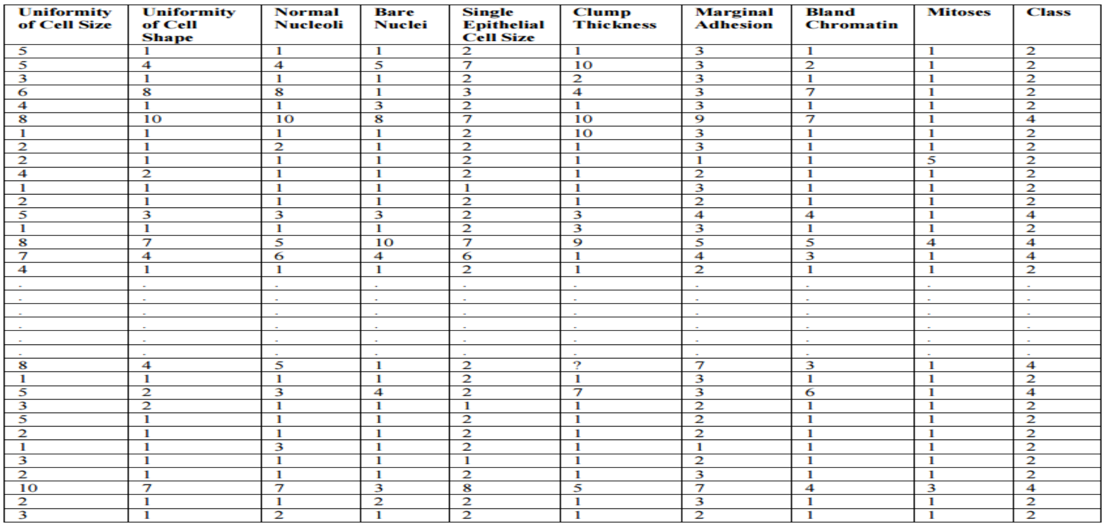
Figure (4.1): Show Some Properties of Data.

4.2.1 Preprocessing result

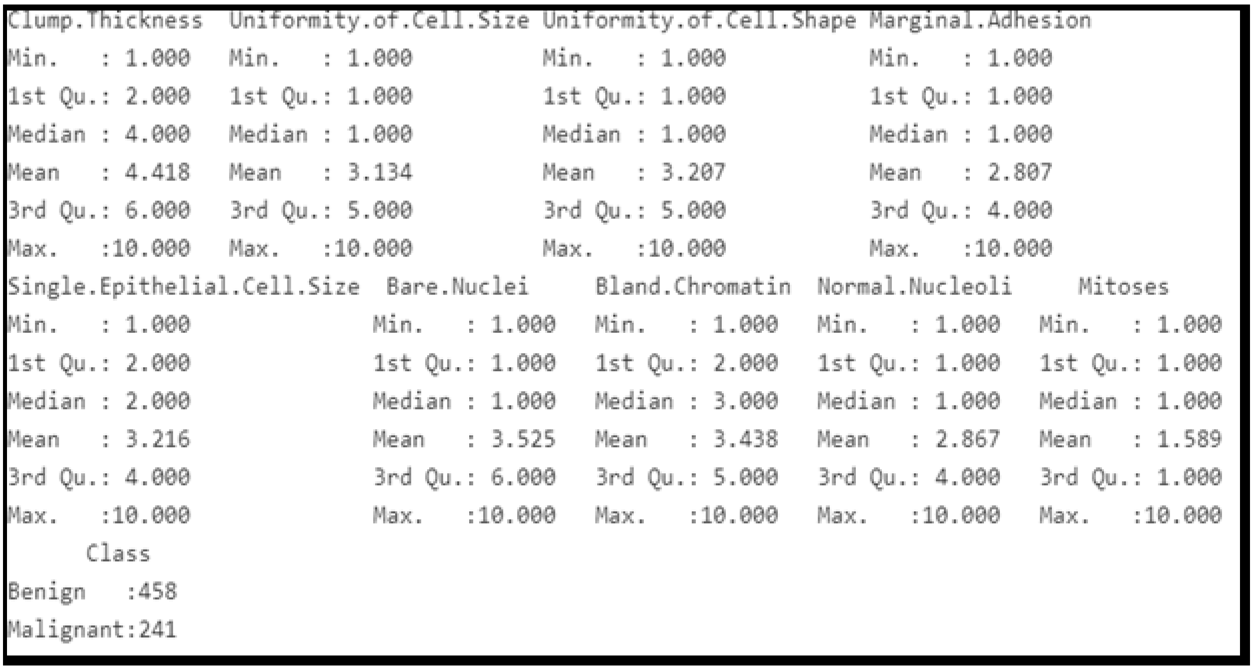
In this paragraph, the results obtained for the database are reviewed after conducting a statistical analysis of them, as shown below. The breast dataset have 16 null cell in column Bare. Nuclei, to fill the missing value apply **impute.missing.data** function . the principle the function calculate the fill the missing value by average value in column Nuclei column in Breast Cancer Wisconsin dataset , then apply the function summary(cancer) to display result after fill missing value as shown in figure (4.1).

The current thesis just depends on UCI depository. Table 4.1 shows a sample of WBC dataset from UCI depository.

Table 4.1: Sample of Wisconsin Breast Cancer Diagnosis Dataset



Uniformity of Cell Size	Uniformity of Cell Shape	Normal Nucleoli	Bare Nuclei	Single Epithelial Cell Size	Clump Thickness	Marginal Adhesion	Bland Chromatin	Mitoses	Class
5	1	1	1	2	1	3	1	1	2
5	4	4	5	7	10	3	2	1	2
3	1	1	1	2	2	3	1	1	2
6	8	8	1	3	4	3	7	1	2
4	1	1	3	2	1	3	1	1	2
8	10	10	8	7	10	9	7	1	4
1	1	1	1	2	10	3	1	1	2
2	1	2	1	2	1	3	1	1	2
2	1	1	1	2	1	1	1	5	2
4	2	1	1	2	1	2	1	1	2
1	1	1	1	1	1	3	1	1	2
2	1	1	1	2	1	2	1	1	2
5	3	3	3	2	3	4	4	1	4
1	1	1	1	2	3	3	1	1	2
8	7	5	10	7	9	5	5	4	4
7	4	6	4	6	1	4	3	1	4
4	1	1	1	2	1	2	1	1	2
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
8	4	5	1	2	?	7	3	1	4
1	1	1	1	2	1	3	1	1	2
5	2	3	4	2	7	3	6	1	4
3	2	1	1	1	1	2	1	1	2
5	1	1	1	2	1	2	1	1	2
2	1	1	1	2	1	2	1	1	2
1	1	3	1	2	1	1	1	1	2
3	1	1	1	1	1	2	1	1	2
2	1	1	1	2	1	3	1	1	2
10	7	7	3	8	5	7	4	3	4
2	1	1	2	2	1	3	1	1	2
3	1	2	1	2	1	2	1	1	2



```
Clump.Thickness  Uniformity.of.Cell.Size Uniformity.of.Cell.Shape Marginal.Adhesion
Min.   : 1.000   Min.   : 1.000          Min.   : 1.000           Min.   : 1.000
1st Qu.: 2.000   1st Qu.: 1.000          1st Qu.: 1.000           1st Qu.: 1.000
Median : 4.000   Median : 1.000          Median : 1.000           Median : 1.000
Mean   : 4.418   Mean   : 3.134          Mean   : 3.207           Mean   : 2.807
3rd Qu.: 6.000   3rd Qu.: 5.000          3rd Qu.: 5.000           3rd Qu.: 4.000
Max.   :10.000   Max.   :10.000          Max.   :10.000           Max.   :10.000
Single.Epithelial.Cell.Size  Bare.Nuclei     Bland.Chromatin  Normal.Nucleoli     Mitoses
Min.   : 1.000               Min.   : 1.000  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
1st Qu.: 2.000               1st Qu.: 1.000  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
Median : 2.000               Median : 1.000  Median : 3.000   Median : 1.000   Median : 1.000
Mean   : 3.216               Mean   : 3.525  Mean   : 3.438   Mean   : 2.867   Mean   : 1.589
3rd Qu.: 4.000               3rd Qu.: 6.000  3rd Qu.: 5.000   3rd Qu.: 4.000   3rd Qu.: 1.000
Max.   :10.000               Max.   :10.000  Max.   :10.000   Max.   :10.000   Max.   :10.000
      Class
Benign   :458
Malignant:241
```

Figure (4.2) Show the Summary Function After Fill Missing Value.

From figure (4.2), observe the fill all missing value of **Bare. Nuclei** has no more NA values. More than where compare the result of summary in figure 4.3 and summary of function in figure (4.2).

4.4 Data Adjustments

Many outliers will certainly affect the statistical model that would be used later in classification. The most critical instance is mitosis as shown in the function **boxplots** of Figure 4.4 below.

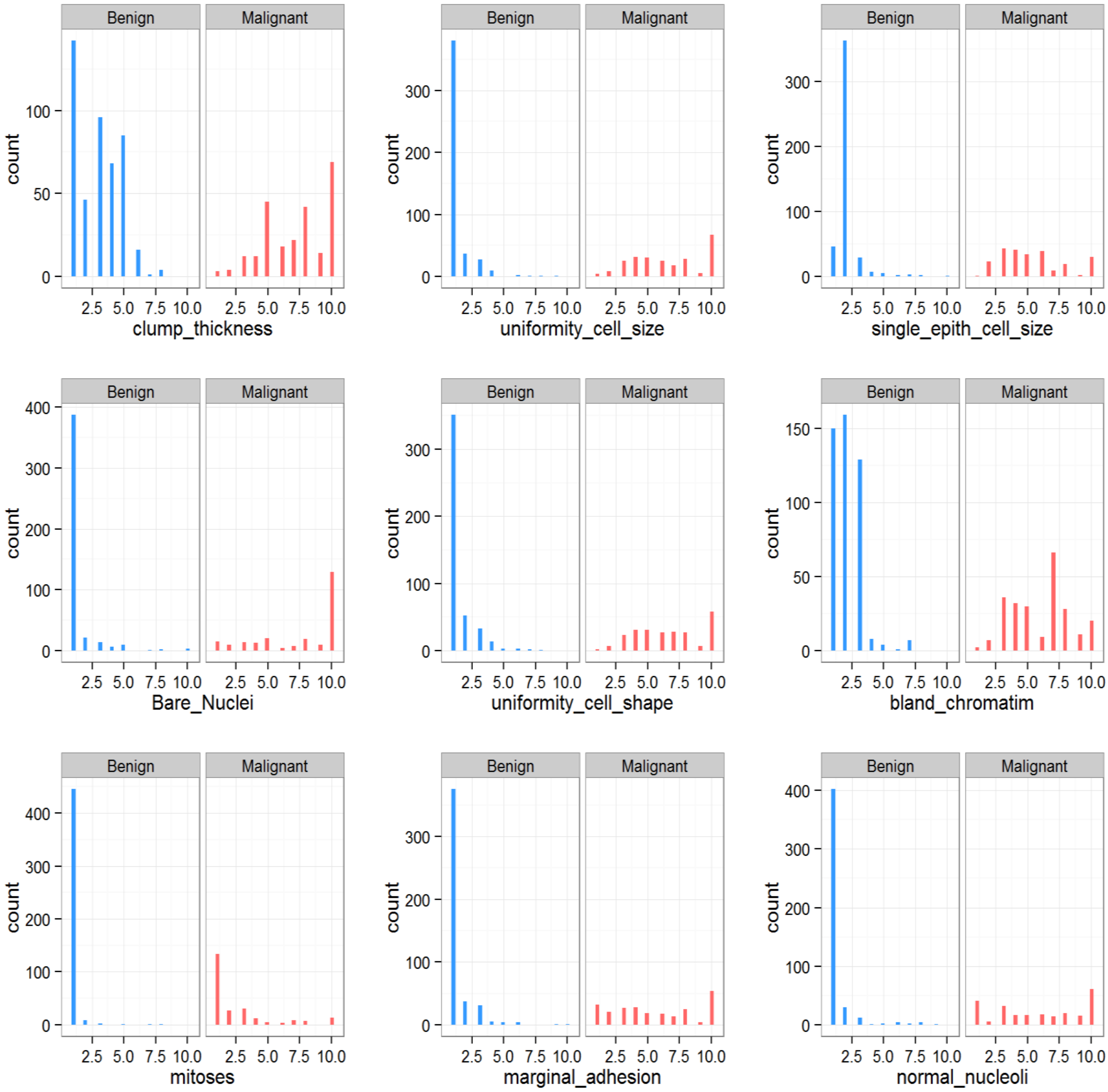


Figure (4.3): Boxplots Of The UCI Breast Cancer Data Set.

From a deep investigation of mitoses in figure 4.3, it could be noticed that the benign and malignant have similar patterns. However, for mitoses of less than 2, the numbers of samples are higher. This type of data malfunction comes either from data collection procedure or due to the nature of the disease. In order to reduce the outlier effect on the classification model, it is more convenient to consider two mitoses categories of mitoses as shown in table (4.1) and figure (4.5).

Table (4.2): Divisions of cells "Mitoses" vs. degree of tumor.

State	No. Of cases							
Benign	356	6	1	0	1	0	1	1
Malignant	111	21	23	11	3	2	6	6

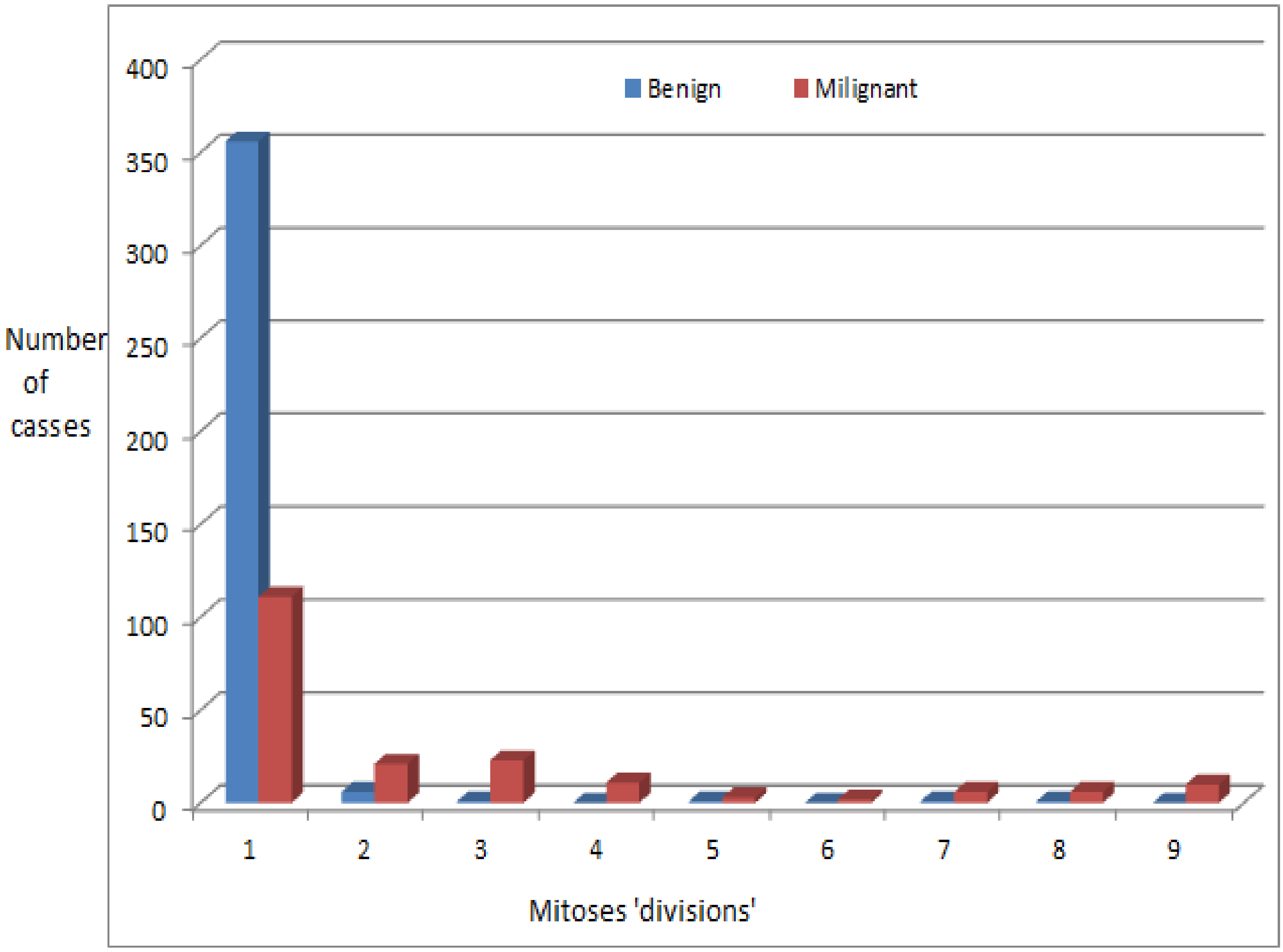


Figure (4.5): Divisions of Cells "Mitoses" Vs. Degree of Tumor.

By doing so, the correlation coefficient will be increased from 0.407 to 0.51, which will increase classification model efficiently. Figure 4.4 shows how classification will be facilitated after the suggested adjustment as shown in table (4.2) and figure (4.6).

Table (4.3): Mitoses or cell divisions after adjusting data.

State	No .of cases	
Benign	356	10
Malignant	111	82

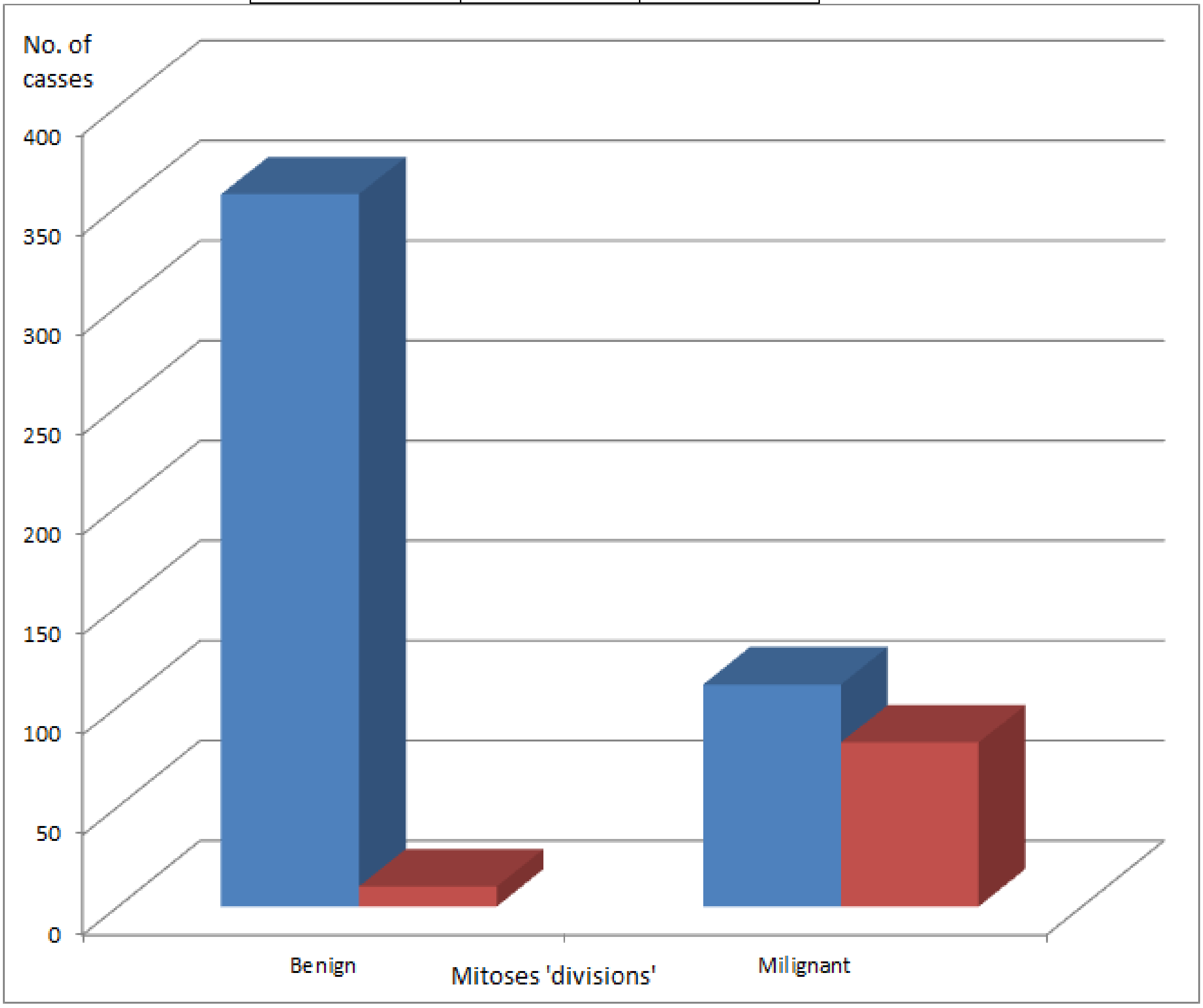


Figure (4.5): Mitoses Or Cell Divisions After Adjusting Data.

To confirm the usefulness and validity of the suggested data adjustment, the following hypotheses are assumed:

	Mitoses	chance
Null Hypothesis	malignant tumor	less or equal to 1
Alternate Hypothesis	malignant tumor	greater than 1

By testing both hypotheses using t-test, the outcome p-value for 95% is 2.2e-16, which has two implications:

1. The mitoses data after the proposed adjustment has a relationship with tumor being malignant, and it did not happen by chance.
2. The small p-value suggests rejecting the null hypothesis, which means the with a tumor of value 2 or more, there will be a great chance of observing malignant state breast cancer.

For epithelial cell size data, the proportion of being malignant or begin tumor seems to be close for sizes 1 and 2. For size above 4, the number of cases is very similar as shown in figure 4.6and table 4.4

Table (4.4): Epithelial Cell Size Vs. Number Of Cancer Cases

State	**No. of cases**							
Benign	37	290	23	5	4	1	3	2
Malignant	1	18	32	33	26	32	9	16

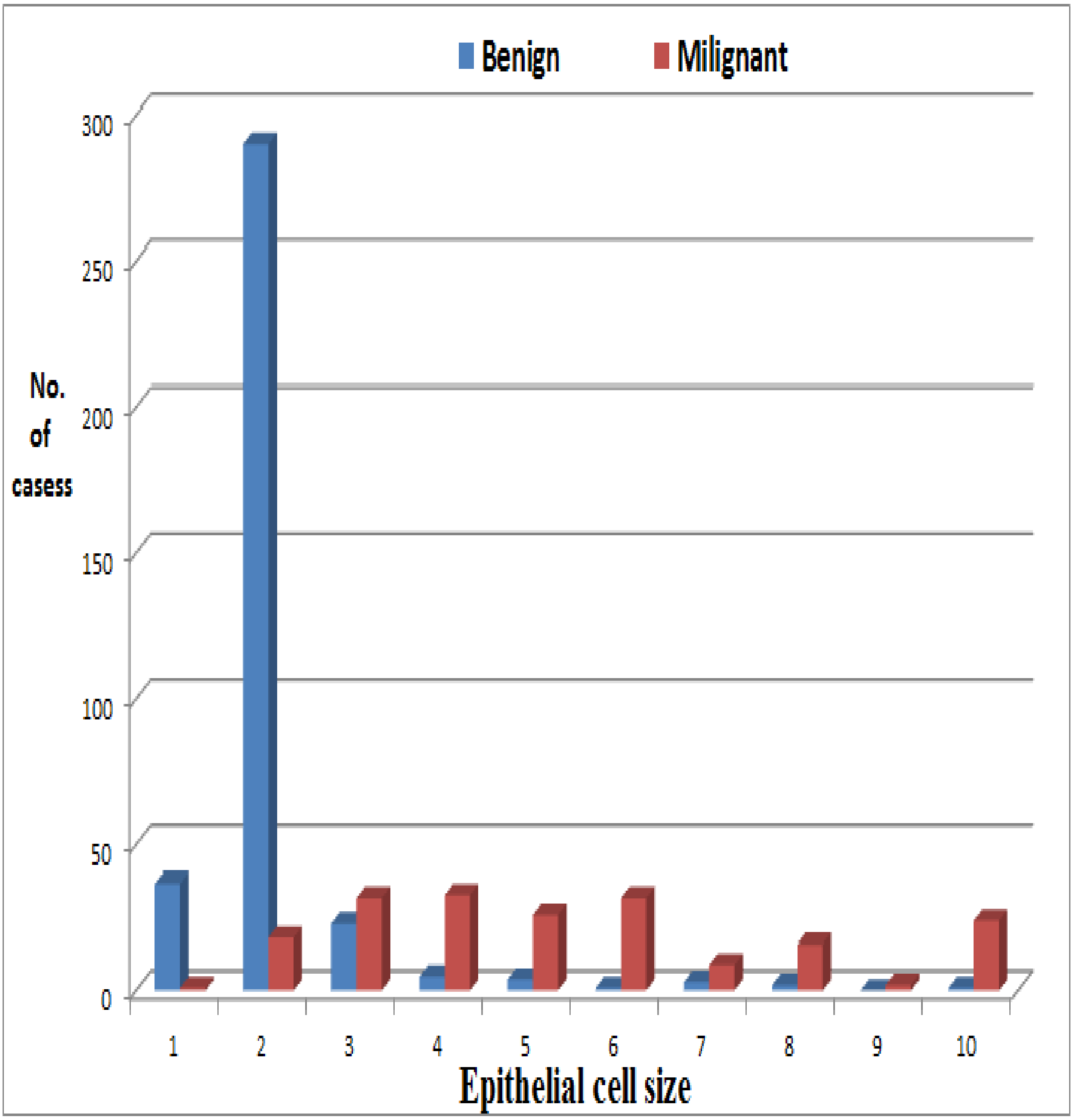


Figure (4.6): Epithelial cell size vs. number of cancer cases.

To make sure that the adjustment procedure is valid, ANOVA test was applied on the modified data. The following table shows the ANOVA test results.

Table 4.5: ANOVA test for epithelial cell size after adjustment

Data	Df	Pr(>F)	Sum Sq.	F value	Mean Sq.
Epithelial Cell Size	3	<2e-15	81.65	491.4	41.32
Residuals	557		46.2		0.1

A very small p-value in table 4.5 as shown in Pr(>F) field, indicates that must exist some difference in the chance of the tumor being Malignant for different values of the adjusted epithelial cell size data. Which means that even after adjustment, the information are still included in the modified data. Figure 4.7 shows the modified epithelial cell size data as shown in table (4.5).

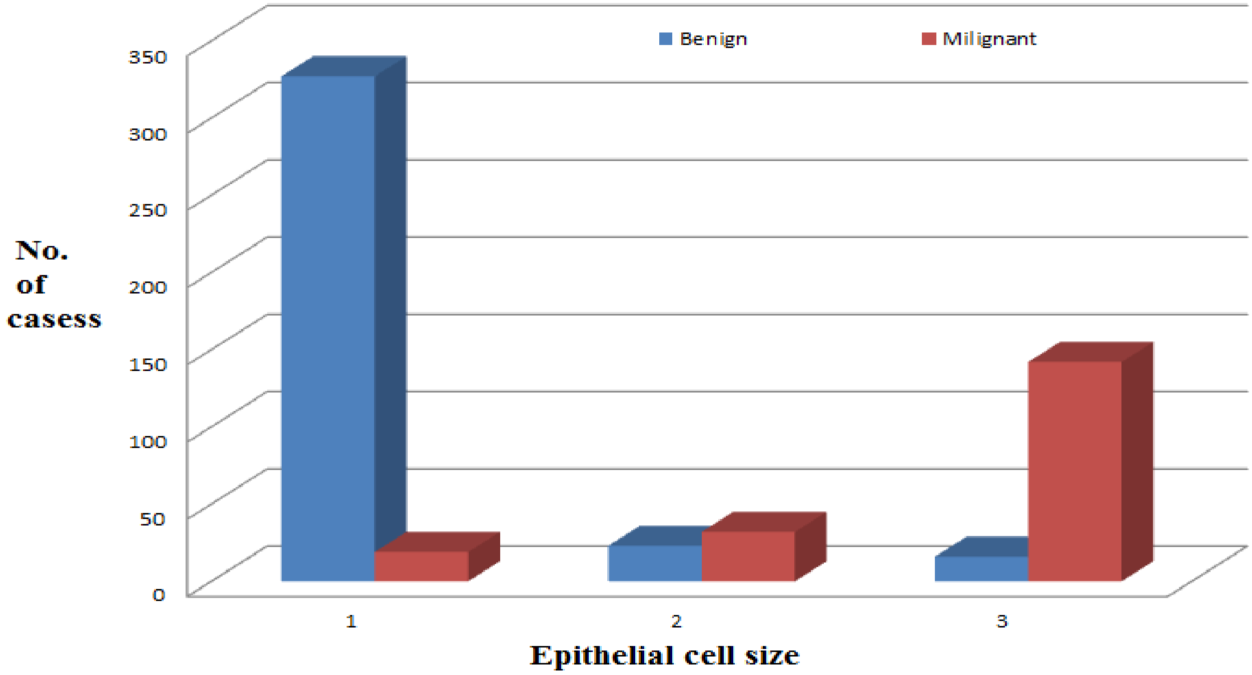


Figure (4.7): Epithelial cell size data after adjustment.

Table (4.6): Epithelial Cell Size Data after Adjustment.

State	No. of cases		
Benign	327	23	16
Malignant	19	32	142

The nucleoli part of the data could be also adjusted by grouping samples of the same behavior. By inspecting table 3, it is clear that the sampled data have similar trend after 4 normal nucleoli. Before 3 nucleoli, the behavior is also similar for the data. When the number of normal nucleoli is 3, the number of patients having benign and malignant tumor is somehow unique. Therefore, 3 nucleoli samples will be kept as a specific group so that no important information will be lost. After grouping the nucleoli data into 3 groups, classification is now more feasible than before. Figure 4.9 and 4.10 shows the number of normal nucleoli, vs. number of cases of begin and malignant tumors, before and after data adjustment shown in table (4.5) and table (4.6).

Table (4.5) Nucleoli Dataset

No of normal Nucleoli	Benign	Malignant
1	324	33
2	22	5
3	9	25
4	1	16
5	1	16
6	4	16
7	1	13
8	3	17

9	1	13
10	0	39

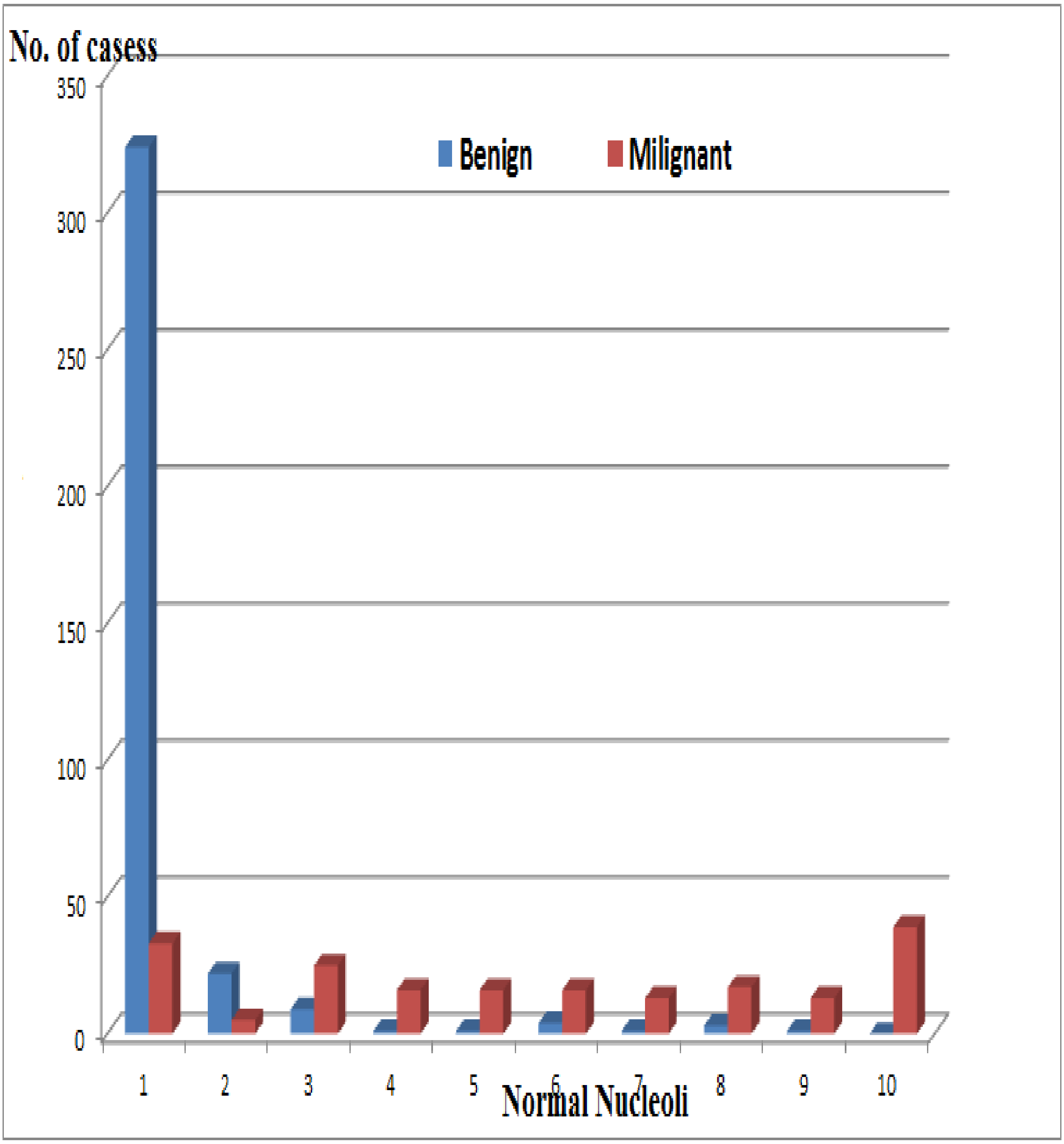


Figure (4.8): No of Nucleoli before adjustment.

Table(4.6): normal nucleoli after data adjustment

State	No. of cases		
Benign	346	9	11
Malignant	38	25	130

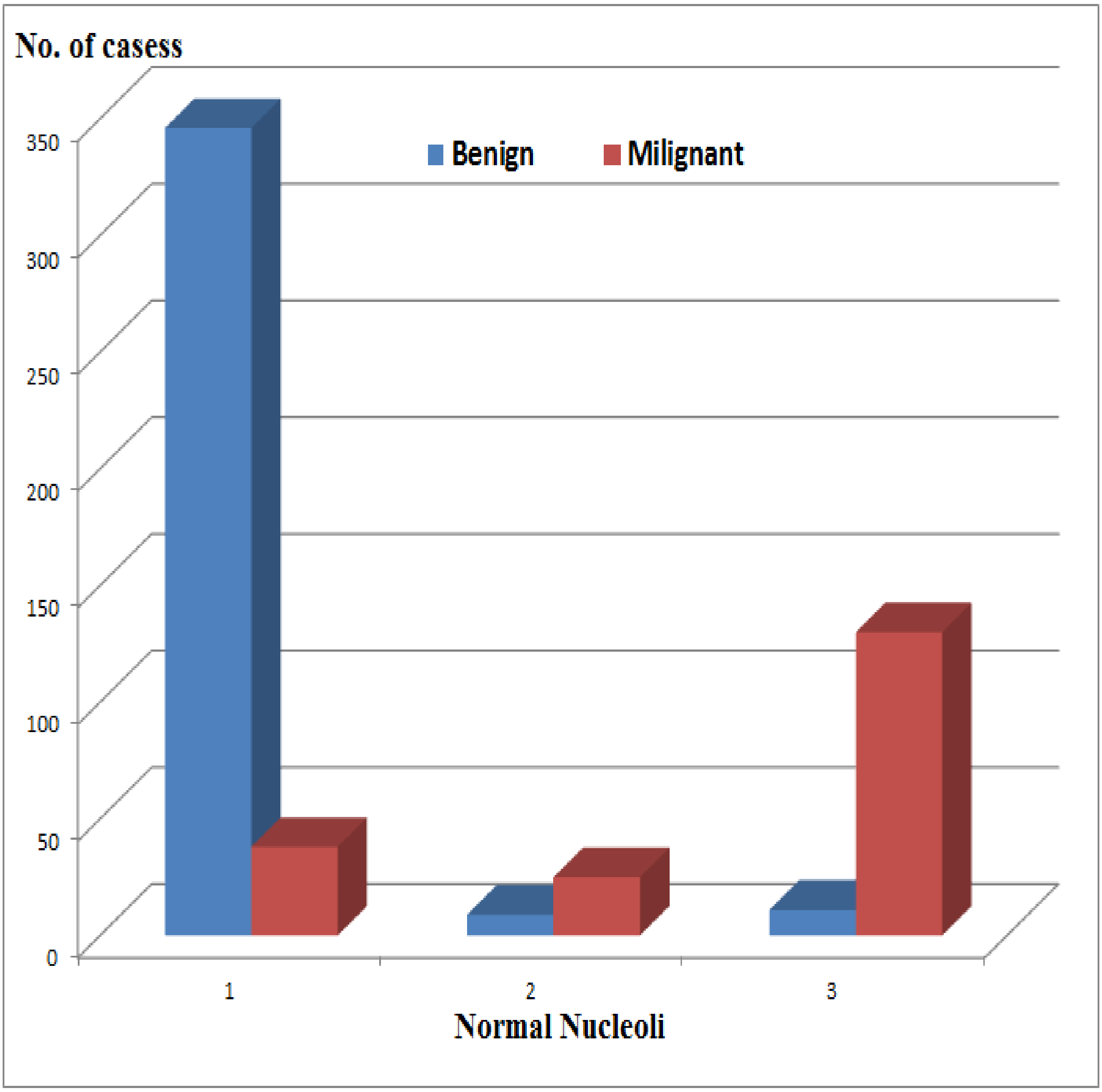


Figure (4.9): No of Normal nucleoli After Data Adjustment.

4.5 Data Selection and Data Splitting

The numeric fields of the UCI breast cancer data are marginal adhesion, clump thickness, uniformity of cell size, shape cell, bare nuclei and bland chromatin. Figure 4.11 shows the inter-parameter correlation of the dataset. The table (4.8) shown in table Data Distribution of the dataset, and accuracy in each model.

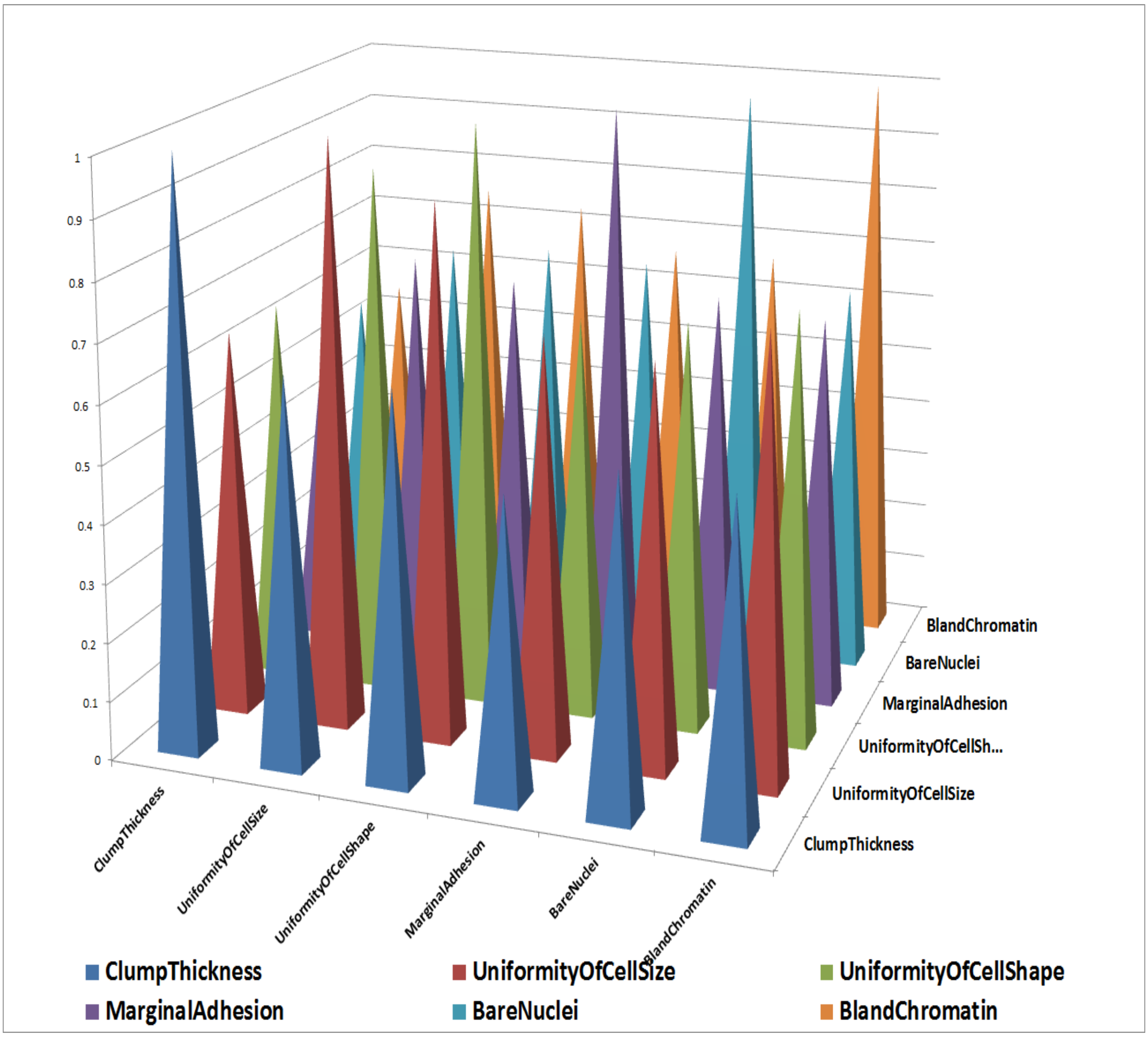


Figure (4.10): Inter-Parameter Correlation Analysis.

Based on the results in Figure 4.10, the uniformity of cell size will not be considered in logistic regression model, as it will have a negative effect on class separation.

Table (4.7) Data splitting

Algorithm	Training (560) 80%	Testing (139) 20%
SVM	100%	0.9714
Logistic Regression	100%	93.912%
Random Forest	100%	0.9857
Decision tree	100%	96.43

4.6 Logistic Regression Classification

In this section describe the simulation of the logistic regression classifier in R software for diagnose the breast cancer , the logistic regression. the dataset are trained on classifier Model for obtained robust model then validated with test data. The results obtained form testing stage are valuated by measures RMSE, Specificity, Accuracy, and Sensitivity, as shown in table (4.8).

Table 4.8 : **Test** Phase Statistic Measures for the

Score	Logistic Regression
Accuracy	0.9643
Test PValue(Mcnemar's)	1
No Information Rate	0.6571
Kappa	0.9203
Confidence Interval	(0.9186, 0.9883)
P-Value	<2e-16
Specificity	0.9783
Positive Prediction Value	0.9574
Detection Rate	0.3214
Prevalence	0.3429
Sensitivity	0.9375
Negative Prediction Value	0.9677
Balanced Accuracy	0.9579
Detection Prevalence	0.3357

4.7 Simulation SVM classification

In SVM classifiers, the dataset is split into two part with the ratio of 80% for training and 20% for testing. The proposed SVM is implemented in phases sequentially, and then followed by SVM classifier . With an accuracy of 96.779%, so this model was selected as the highest performer and the precision and recall of the model were then observed to assess its overall quality.

Table (4.9) Test Phase Statistic Measures for the SVM

score	SVM
Accuracy	0.9714
P-Value [Acc > NIR]	<2e-16
No Information Rate	0.6571
Specificity	0.9783
Mcnemar's Test PValue	0.6831
Kappa	0.9366
Confidence Interval	(0.9285, 0.9922)
Positive Prediction Value	0.9583
Sensitivity	0.9583
Prevalence	0.3429
Balanced Accuracy	0.9683
Detection Prevalence	0.3429
Negative Prediction Value	0.9783
Detection Rate	0.3286

4.8 Building Random Forest Classification

After determine the most important features in the data set, analyzing the feature importance of various selected features, and testing different sizes of random forests, I obtained a final model with an accuracy of 98%. Building a Baseline Model for Comparison Before doing any analysis on the data set, it is usually good to build a baseline classification model that can be used as a benchmark to decide if a machine learning model is effective or not. To build this baseline model, I looked at the distribution of “malignant” (labeled as a 1) and “benign” (labeled as a 0) observations in the data set. Depending on which of these classification categories is more probable (or has the higher occurrence), I predict every observation to be in the more probable category. Here is the bar graph showing the distribution of the two target categories. Building the Random Forest Model, instantiates it with a size of 125 trees (estimators is the number of decision trees that will be constructed to form the random forest object), and fits a random forest to a set of testing data. In this experiment, the data was split into training and testing sets. Which splits the data into training and testing groups based on a desired ratio. When the random forest was training using the partitioned training and testing data, the result was as follows: Random Forest Accuracy. As can be seen above, the random forest of 125 decision trees obtained an accuracy of 96.958% on the data set. However, before the experiment was ended, different numbers of estimators for the random forest were tested to see if the model could be made slightly more accurate. Additionally, the precision and recall of the highest performing random forest was observed to gain another metric on the overall quality of the model. Random forest sizes of 25, 50, 75, 100, 125, and 150 were selected and the accuracy of each of the models was tested to determine which size of random forest was most effective. It can be seen that the most accurate size of forest was 125, with an accuracy of 96.958%, so this model was selected as the highest performer and the precision and recall of the model were then observed to assess its overall quality . The Experiments are carried out on the model , as shown in table (4.10).

Table (4.10) Test Phase Statistic Measures for the Random Forest

Score	**Random Forest**
Accuracy	0.9857
Confidence Interval	(0.9493, 0.9983)
No Information Rate	0.6571

P-Value [Acc > NIR]	<2e-16
Kappa	0.9686
Mcnemar's Test PValue	0.4795
Sensitivity	1.0000
Specificity	0.9783
Positive Prediction Value	0.9600
Negative Prediction Value	1.0000
Prevalence	0.3429
Detection Rate	0.3429
Detection Prevalence	0.3571
Balanced Accuracy	0.9891

4.9 Decision Tree Builds Classification Models

In the experiment, two classes were used and therefore a 2x2-confusion matrix was applied, that achieved accuracy 95.71 as shown in figure (4.15). The Experiments are carried out on the model , the data set is trained well on each classifier and a Model is obtained then validated with test data, results are obtained. The obtained results are calculated and evaluated in terms of measures like Accuracy, RMSE, Specificity, Sensitivity.

Table (4.11) Test Phase Statistic Measures for the Decision Tree

Score	**Decision Tree**
Accuracy	0.9571
No Information Rate	0.6571
Confidence Interval	(0.9091, 0.9841)
Kappa	0.9058
Specificity	0.9565
Sensitivity	0.9583
P-Value [Acc > NIR]	<2e-16
Mcnemar's Test PValue	0.6831
Negative Prediction Value	0.9778
Prevalence	0.3429
Balanced Accuracy	0.9574
Detection Prevalence	0.3571
Positive Prediction Value	0.9200
Detection Rate	0.3286

4.9 Experimental Results

After training four models for the breast cancer data, the classification results are as shown in table 4.12 and 4.13.In table 4.12, a detailed classification of the test samples. The begin and reference columns represent the true situation, while the row values are the predicted ones, the model has to predict 0 malignant as begin, as shown in random forest row 7 in table 4.12.

Table (4.12) Detailed Classification Results

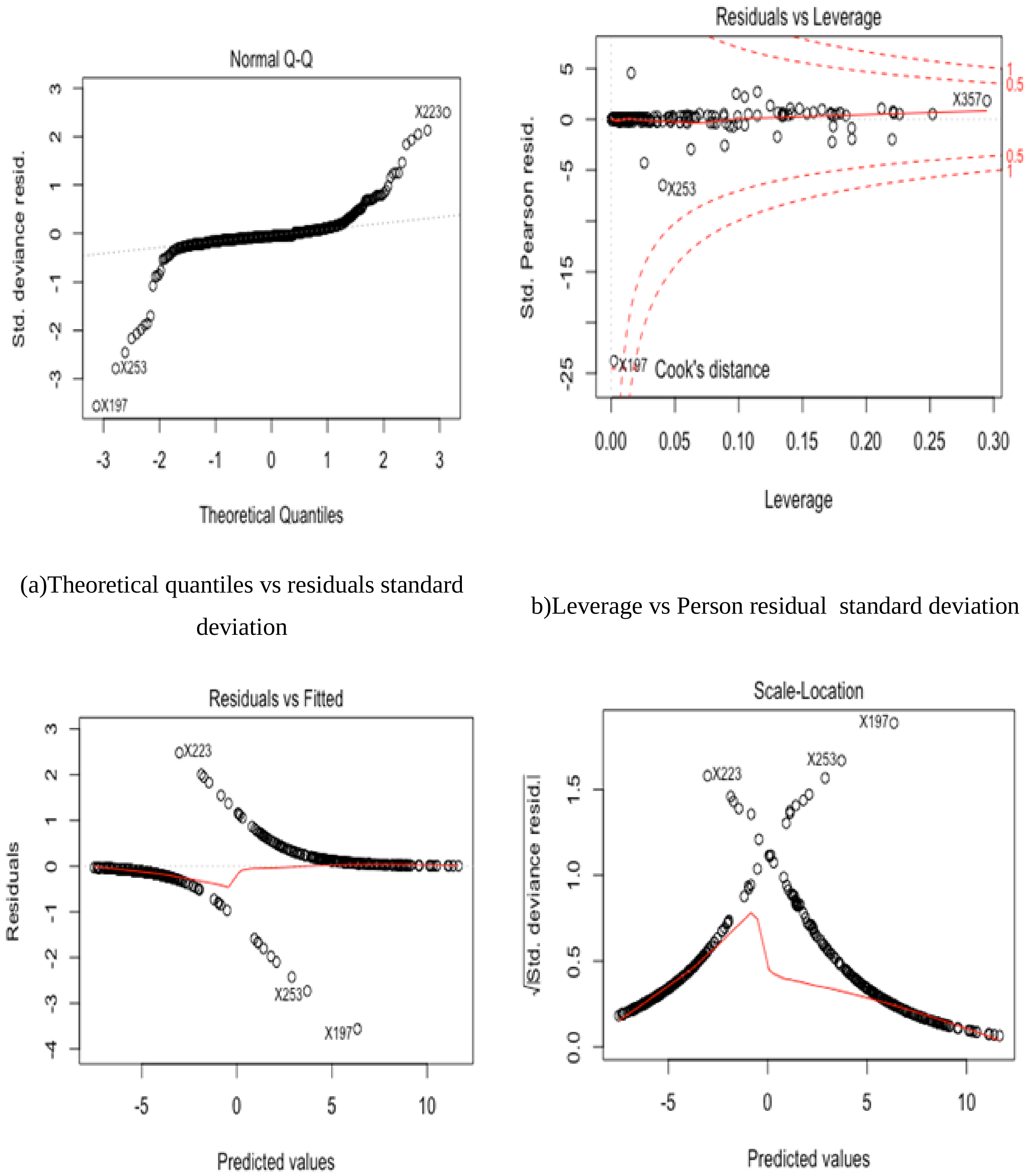
Model Type	Prediction	Reference	
		Begin	Malignant
SVM	Begin	90	2
	Malignant	2	46
Logistic Regression	Begin	90	3
	Malignant	2	45
Decision Tree	Begin	88	2
	Malignant	4	46
Random Forest	Begin	90	0
	Malignant	2	48

On the other hand, Table 4.8 contains the statistical measures of the test phase. Comparing the four models, random forest has a better accuracy and Kappa is close to one.

Table 4.13: Test Phase Statistic Measures

score	SVM	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.9714	0.9643	0.9571	0.9857
Confidence Interval	(0.9285, 0.9922)	(0.9186, 0.9883)	(0.9091, 0.9841)	(0.9493, 0.9983)
No Information Rate	0.6571	0.6571	0.6571	0.6571
P-Value [Acc > NIR]	<2e-16	<2e-16	<2e-16	<2e-16
Kappa	0.9366	0.9203	0.9058	0.9686
Mcnemar's Test PValue	0.6831	1	0.6831	0.4795
Sensitivity	0.9583	0.9375	0.9583	1.0000
Specificity	0.9783	0.9783	0.9565	0.9783
Positive Prediction Value	0.9583	0.9574	0.9200	0.9600
Negative Prediction Value	0.9783	0.9677	0.9778	1.0000
Prevalence	0.3429	0.3429	0.3429	0.3429
Detection Rate	0.3286	0.3214	0.3286	0.3429
Detection Prevalence	0.3429	0.3357	0.3571	0.3571
Balanced Accuracy	0.9683	0.9579	0.9574	0.9891

For Mcnemar’s test measurements, all models except logistic regression ave -value $<\alpha$, where $\alpha = 0.95$ for 95% confidence interval. However, random forest model has the lowest p-value for that test, therefore, it better than the others The random forest model shows better performance for all remaining metrics in



(a)Theoretical quantiles vs residuals standard deviation

b)Leverage vs Person residual standard deviation

(c)Predicted values vs residuals

(d)Predicted values vs the square root standard deviation of residuals

Figure (4.11): Different aspects of error in logistic regression classifier.

Table 4.13. It has the maximum value of 1 for sensitivity and negative prediction value (NPV). For other metrics, it showed to have higher

Confidence than the other tested models. In McNamara's test, the logistic regression Model has a p-value > 0.95, which means that the model has a 50% chance of being incorrect. Therefore, a deeper analysis is necessary. By testing the logistic regression model deeply, the residuals was analyzed from different points of view, as shown in figure 4.11 The residuals are the error between the true value of the input versus the predicted one, or:

$$residual = real\ value - predicted\ value$$

Figure 4.11 (a) shows the normalized quantiles of classifier PDF. While the dotted line represent the theoretical relationship, the test samples shows many outliers scattered away from the theoretical line. The outlier problem is clearer in figure 4.11(b), which depicts the Leverage using Cook's distance measure. The Leverage metric id used to analyze the distances among observations, and it is clear that they have scattered pattern. In figure, 4.11 (c) and (d) show the fitted values vs the residual, in which some highly incorrect predictions are noticed. All the above error analysis explain why logistic regression had McNamara's p-value =1 for the tested observations. Therefore, McNamara's test could not reject the null hypothesis of being 50% incorrect despite the fact that it has 96.43% accuracy. Although random forest has 500 trees, the error from each tree is relatively small if compared with the other tested models. The proposed system result has been comparing that results with six previous related works and show this work more accurate than the compared works, as shown in table (4.14).

Table (4.14) Comparison with Some Related Work

Authors	Accuracy	tools
The Purposed Work	0.9714	SVM
	0.9857	Random Forest
	0.9643	Logistic Regression
	0.9571	Decision Tree
In [11](2015)	0.95	SVM
In [12](2016)	0.97%	SVM
Hiba Asri 2016[56]	97.13%	SVM
Vivek Kumar[57]	0.9715	Random Forest
Abdelghani Bellaachia[58]	86.5	ANN
Shelly Gupta Et Al[59]	88.8%	ANN

Chapter Five

Calculation and Future Work

Chapter Five

Calculation and Future Work

5.1 Calculation

This study focused on the predictive ability of data mining and statistical learning techniques to identify breast cancer patient survivability. The ability of a medical practitioner to effectively pinpoint how breast cancer patients survived during treatment, may lead to better evaluation of the treatment and design of personalized medicine or drugs to breast cancer patients.

1- This study provides a detailed account of a data mining process applied to the prediction of breast cancer survivability. The data mining process followed for this study began with the inclusion of a large set of factors that was reduced manually through standard statistical analysis, The R statistical learning toolsets were used for this study and the following classifiers were used, Random Forest did outperform decision tree, SVM, Logistic Regression based on accuracy.

2- Performed data analysis of the breast cancer dataset using statistical learning and data mining algorithms and found that the null hypothesis could not be rejected.

3- Findings indicate that none of the data mining and statistical learning algorithms applied to the breast cancer dataset outperformed the others in such a way that it could be declared the optimal algorithm.

4- None of the algorithms performed poorly as to be eliminated from future prediction models in breast cancer survivability tasks. The most important results obtained for the algorithms used were: -

prediction models in breast cancer survivability tasks. The most important results obtained for the algorithms used were :-

1- **Logistic Regression**

Accuracy	0.9643

2- SVM classification

Accuracy	0.9714

3- **Random Forest Classification**

Accuracy	0.9857

4- **Decision Tree Classification**

Accuracy	0.9571

5.2 Future Work

Future work can be described as follows. The current research resided mainly on classification accuracy as the main criteria for measuring the performance of proposed approaches.

The future work can summarized as follows:

1- Improve the accuracy and reliability of the established system by broadening the databases and expanding the criteria for measuring the performance of established systems.

2- Collection data from breast cancer patents form Iraqi hospital.

3- Applied the deep learn algorithm for classify the dataset.

References

References

[1] Breast cancer network Australia,” Current Breast Cancer Statistics in Australia” Australian Institute of Health and Welfare. Cancer in Australia 2017.

[2] Faozia A. S. Alsarori," Automatic Detection Of Breast Cancer in Mammogram Images”, thesis, M.S. Department of Computer Engineering, Cankaya ,university , 2013.

[3] Henry J. Mankin, MD, et al.," Diagnosis, Classification, and Management of Soft Tissue Sarcomas”, paper, Dorothy Fox. Fifth Avenue Springtime. Watercolor, Vol. 12, No. 1,2005.

[4] Noor Thamer, "Medical Diagnosis Using Bayesian neural network" thesis, M.S. Department of Computer science, Bagdad university, 2014

[5] Ram Meghe ,and Research Badnera," Artificial Intelligence in Medical Diagnosis", paper , International Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11,2012,

[6] William B. Gevarter,” An Overview of Artificial Intelligence and Robotic”,book, National Aeronautics and Space Administration, Scientific and Technical Information Branch ,1983

[7] Tanusree Dutta, et al.,"An Intuition Based Fuzzy Logic Driven Approach for Designing Symptomatic Medical Diagnostic Expert System ", paper, International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181,2015.

[8] S. Govinda Rao,” Fever Diagnosis Rule-Based Expert Systems”, paper, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 8, August – 2013

[9] Martin J. Yaffe,Earlier," Detection and Diagnosis of Breast Cancer"Director, Cancer Screening, Cancer Control Sciences Department, American Cancer Society ,2016

[10] Roselina Sallehuddin ,"An Improvement In Support Vector Machine Classification Model Using Grey Relational Analysis For Cancer Diagnosis",Computer Science Department, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, MalaysiaArticle history Received 29 November 2015 Received in revised form 20 March 2016 Accepted 29 February 2016

[11] Emina Aličković ,"Data Mining Techniques for Medical Data Classification", The International Arab Conference on Information Technology (ACIT) 2017

[12] Peiguang Lin,"Research on the Method for WDB's Characteristics Extraction based on Independent Data Samples1877-7058 © 2011 Published by Elsevier Ltd. doi: 10.1016/j.proeng.2011.08.735"

[13] Nilashi, Mehrbakhsh., Ibrahim, Othman bin., Ahmadi, Hossein., & Shahmoradi, Leila., An Analytical Method for Diseases Prediction Using Machine Learning Techniques.Computers and Chemical Engineering http://dx.doi.org/10.1016/j.compchemeng.2017.06.011

[14] O.N. Oyelade,"ST-ONCODIAG: A semantic rule-base approach to diagnosing breast cancer base on Wisconsin datasets",Received 6 October 2017; Received in revised form 19 December 2017; Accepted 20 December 2017 Available online 28 February 2018 2352-9148/© 2017 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

[15] S. Wang, Y. Wang, D. Wang et al., An improved random forest-based rule extraction method for breast cancer diagnosis, Applied Soft Computing Journal (2019), doi: https://doi.org/10.1016/j.asoc.2019.105941

[16] Dalwinder Singh ,"Simultaneous feature weighting and parameter determination of Neural Networks using Ant Lion Optimization for the classification of breast cancer",0208-5216/© 2019 Published by Elsevier B.V. on behalf of Nalecz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences.

[17] S, Birmohan S. Simultaneous feature weighting and parameter determination of Neural Networks using Ant Lion Optimization for the classification of breast cancer. Biocybern Biomed Eng (2019), https://doi.org/10.1016/j. bbe.2019.12.004

[18] Na Liu , Er-Shi Qi,"A novel intelligent classification model for breast cancer diagnosis",1877-0509 © 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/) Peer-review under responsibility of the scientific committee of the 13th International Symposium "Intelligent Systems" (INTELS'18). 10.1016/j.procs.2019.02.085

[19] Na Liu , Er-Shi Qi,"A novel intelligent classification model for breast cancer diagnosis",1877-0509 © 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/) Peer-review under responsibility of the scientific committee of the 13th International Symposium "Intelligent Systems" (INTELS'18). 10.1016/j.procs.2019.02.085

[20] Kamil Pohlodek, "An Introduction to breast diseases", Comenius University Bartislava, Faculty of Medicine, ISBN978-80-223-3766-3 ,2014.

[21] Rick Alteri et al, "Breast Cancer Facts & Figures "American Cancer Society 2015-2016, No. 861017, 2015.

[22] Republic of Iraq ministry of health cancer board, "Iraqi Cancer Registry 2012", Baghdad – Iraq,2015.

[23] Tanusree Dutta, et al.,"An Intuition Based Fuzzy Logic Driven Approach for Designing Symptomatic Medical Diagnostic Expert System ", paper, International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181,2015.

[24] Tanusree Dutta, et al.,"An Intuition Based Fuzzy Logic Driven Approach for Designing Symptomatic Medical Diagnostic Expert System ", paper, International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181,2015.

[25] S. Govinda Rao," Fever Diagnosis Rule-Based Expert Systems", paper, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 8, August – 2013

[26] Qeethara Kadhim Al-Shayea," Artificial neural networks in medical diagnosis", paper, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[27] Achi I. I.,Prof. Inyiama H. C.,Prof. Bakpo F. F.,Agwu C. O, "Machine Learning Based on Intelligent Tutoring System" International Journal of Engineering Trends and Technology (IJETT) Volume 26 Number 4- August 2015

[28] Demidova LA, Egin MM, Tishkin RV. Novel modifications of the multiobjective genetic algorithm for SVM classifier

development. International Journal on Information Technologies and Security (IJITS) 2018; 10(2):89-100

[29] Peterek T., Dohnalek P., Gajdos, P., & Smondrk M. (2013, December). Performance evaluation of Random Forest regression model in tracking Parkinson's disease progress. In Hybrid Intelligent Systems (HIS), 2013 13th International Conference on (pp. 83-87). IEEE.

[30] Alireza 0, Bita S. 2010. Machine Learning Techniques To Diagnose Breast Cancer. In: Fifth International Symposium on Health Informatics and Bioinformatics. 114-120

[31] Mahmoodian H, Ebrahimian L. Using support vector regression in gene selection and fuzzy rule generation for relapse time prediction of breast cancer. Biocybern Biomed Eng 2016;36:466–72.

[32] Y.I. Chang, S.C. Lin, Synergy of logistic regression and support vector machine in multiple-class classification, in: Proceedings of the Fifth International Conference on Intelligent Data Engineering and Automated Learning, Exeter, UK, 2004, pp. 25–27

[33] Y.Wanga , Y. Zhanga , Y. Lua , Xi. Yua, “ A Comparative Assessment of Credit Risk Model Based on Machine leaning “ , 2019 International Conference on Identification, Information and Knowledge in the Internet of Thing , 2019

[34] Breiman, L.: Using adaptive bagging to debias regressions, Technical Report 547, Statistics Dept. UCB (1999).

[35] S. K., K. Shankar, M. Ilayaraja, Abdul Wahid Nasir, V. Vijayakumar, and Naveen Chilamkurti. "Random forest for big data classification in the internet of things using optimal features."

International Journal of Machine Learning and Cybernetics (2019): 1-10.

[36] S. Misra, Hao Li, “Machine Learning for Subsurface Characterization” , The University of Oklahoma, Norman, OK, United States 2020

[37] Sadaaki Miyamoto, et al., "Algorithms for Fuzzy Clustering: “Methods in c-Means Clustering with Applications", ISBN 10:3540787364, Hardcover; Springer.