A Data Mining Approach for Forecasting Cancer Threats

Benard Nyangena Kiage

A thesis submitted in partial fulfillment for the degree of Master of Science in Computer Systems in the school of computing in the Jomo Kenyatta University of Agriculture and Technology

2015

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature: _____ Date: _____

Benard Nyangena Kiage

This thesis has been submitted for examination with my approval as university supervisor.

Signature: _____ Date: _____

Dr. George Okeyo

JKUAT, Kenya

Signature: _____ Date: _____

Dr. Wilson Cheruiyot

JKUAT, Kenya

DEDICATION

I dedicate this thesis to my beloved wife, Mrs. Joyce Moraa, Son Felix Maranga and Daughter, Christine Kemunto. My Dad John Kiage, My late beloved Mother Callen Nyatugah, My uncle Micah Bundi and aunt Sarah Omariba for bringing me up as a potential academician.

ACKNOWLEDGEMENT

This thesis is as a result of inspirational assistance and guidance of my Dad, mentors, lecturers, professionals, and the administrative staff at the university as well as at the Machakos university college, my work place.

First and foremost, I am grateful to my supervisors; Dr. George Okeyo and Dr. Wilson Cheruiyot, for their invaluable and continuous guidance during the conception of this thesis. My other regards sincerely goes to Dr. Kimwele, Dr. Musau, and Prof. Kanali, who did provided me with an undisputed considerable help in every way possible while carrying out this research.

I owe you all!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS/ACRONYMS

ADALINE	Adaptive linear Element
ANFIS	Fuzzy Inference System
ANN	Artificial Neural Network
CAD	Computer Aided Diagnosis
CART	Classification and Regression Tree
CES	Consistency Based Subset Evaluation
CFS	Correlation based feature selection
DM	Data Mining
EHealth	Electronic Health
EHR	Electronic Health Record
ERR	Error Rate
FIS	Fuzzy Inference System
GA	Genetic Algorithm
HIS	Hybrid Intelligent System
IG	Information Gain
IGANFIS	Information Gain Adaptive Neuro-Fuzzy Inference
K-NN	*k*- nearest Neighbors
LSE	Least Square Estimate
ML	Machine learning
PCA	Principle Components Analysis
UCI	University of California Irvine
WBC	Wisconsin Breast Cancer Dataset
WDBC	Wisconsin Diagnosis Breast Cancer Dataset
WPBC	Wisconsin Prognosis Breast Cancer Dataset
WEKA	Waikato Environment for Knowledge Analysis

ABSTRACT

Healthcare facilities have at their disposal vast amounts of cancer patients' data. The analysis of available data can lead to more efficient decision-making. The challenge is how to extract relevant knowledge from this data and act upon it in a timely manner. To turn into knowledge, efficient computing and data mining tools must be used. This data can aid in developing expert systems for decision support that can assist physicians in diagnosing and predicting some debilitating life threatening diseases such as cancer. Expert systems for decision support can reduce the cost, the waiting time, liberate medical practitioners for more research and reduce errors and mistakes that can be made by humans due to fatigue and tiredness. The process of utilizing health data effectively however, involves many challenges such as the problem of missing feature values, data dimensionality due to a large number of attributes, and the course of actions to determine features that can lead to more accurate diagnosis. Effective data mining tools can assist in early detection of diseases such as cancer. This research proposes a new approach called Information Gain Artificial Neuro-network Fussy Inference System (IG-ANFIS). This approach optimally minimizing the number of features using the information gain (IG) algorithm, then applies the new reduced features dataset to the Adaptive Neuro Fuzzy Inference system (ANFIS). The research also proposes a new approach for constructing missing feature values based on iterative k-nearest neighbours and the distance functions.

Key words: Data Mining, Clustering, Classification, Neural networks, Fuzzy Inference system, Information gain.

CHAPTER ONE
INTRODUCTION

1.1 BACKGROUND

Medical Databases today can range in size into hundreds of millions of terabytes. Within these masses of data lies hidden information of strategic importance. Due to these vast amounts of data, it then begs the question, “How do you draw meaningful conclusions about this data?” Data mining answers this question.

Although computational, the utility of data mining algorithms can be used as a qualitative tool to analyze quantitative data, particularly the large, complex databases being created by the health informatics community, Young (2012). Lloyd-Williams (2013), Data stored in hospital warehouses range from quantitative to analog to qualitative data; however well structured, these data conceal implicit patterns of information which cannot readily be detected by conventional analysis techniques. The formats data warehouses also vary and amounting to information explosion within the health care field. The problem however, is finding the right methodological tools to mine this new data given its enormous variety, size, and complexity.

The advancement of information technology, software development, and system integration techniques have produced a new generation of complex computer systems. These systems have presented challenges to information technology researchers. The major challenge is how to benefit from the existing resources and data.

These complex systems include the healthcare system. In recent times, there has been an increased interest in the utilization and advancement of data mining technologies and communication in healthcare systems and in this respect, a global healthcare system is getting adopted by many countries by setting healthcare standardization in communication and building electronic health records (EHR).

Gunter (2005), EHR is a systematic collection of electronic health data about individual patients or populations and is capable of being shared across healthcare providers in a certain state or country. Health records may include a range of data such as general medical records, patient examinations, patient treatments, medical history, allergies, immunization status, laboratory results, radiology images, and some useful information for examination. This rich information may help researchers in examining and diagnosing diseases using computer techniques.

The shift of many countries moving fast toward electronic healthcare information systems has produced huge EHRs for health related information. This data can be a valuable asset for populations and healthcare providers. In this respect the aim of this research is to investigate the aspects of utilizing health data for the benefit of humans by using novel data mining techniques.

The current research focuses on diagnosing cancer based on machine intelligence and previous history. The approach develops a new technique known as Information Gain Artificial Neuro Inference System (IG-ANFIS). This uses a combination of an Adaptive Network based Fuzzy Inference System (ANFIS) and the Information Gain method (IG). The purpose of ANFIS is to build an input-output mapping using both human knowledge and machine learning ability and the purpose of IG method is to reduce the number of input features to ANFIS. The IG method approximates the quality of each attribute using the entropy by estimating the difference between the prior entropy and the post entropy. IG is one of the attribute ranking methods often applied in text categorization. In text categorization, it is used to measure the number of bits of information obtained for category prediction. This is done by knowing the presence or absence of a term in a document.

1.2 Statement Of The Problem

Data mining methods used for diagnosing diseases based on previous data and information have been improving over the years. The data mining methods used currently particularly for disease diagnosis use various feature selection techniques which includes Correlation based Feature Selection (CFS), Relief (R), Principle Components Analysis (PCA), Consistency based Subset Evaluation (CSE), Information Gain (IG), and symmetrical uncertainty (SU). These techniques have no doubt improved disease diagnosis. However there are several problems associated with effectively utilizing this previously acquired patient data, which can make any electronic healthcare system problematic and less efficient i.e. the problem of missing values and how to process them, huge features and attributes and how to select the most beneficial features, the problem of extracting accurate diagnostic markers that can predict the early onset of the disease and monitoring of different stages of the disease.

Based on the power of the current data mining methods and the previous evidence or data, this research tries to investigate feature selection techniques, and a novel hybrid method (IGANFIS) for diagnosing diseases (in this case cancer) has been developed. IGANFIS combines IG method and ANFIS method for Cancer Diagnosis. The IG will be used for selecting the quality of attributes. A set of features with high ranking values will be the output

of applying IG method. These high ranking values will constitute the input for ANFIS and Odeh (2008).

1.3 Objectives

1.3.1 Broad Objective

The general objective of this research thesis is:-

To develop a data mining approach that will combine information gain algorithm and adaptive neural fuzzy inference system to analyze large data obtained from healthcare databases.

1.3.2 SPECIFIC OBJECTIVES

The specific objectives of this research thesis were to:-

i. To identify the current data mining algorithms used in healthcare sector for cancer diagnosis.
ii. To identify the significance of diagnostic features that best describe cancer data using data mining techniques.
iii. To describe how missing feature values improve prediction in determining the performance achieved by data mining algorithms.
iv. To develop a hybrid data mining model from the existing techniques that can improve classification accuracy and missing values.
v. To test the developed hybrid data mining model for classification accuracy and missing values.

1.4 Research Questions

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

1. What are the data mining algorithms used currently in the healthcare sector for cancer diagnosis?
2. How can the diagnostic features that best describe data for the purpose of differentiating malignant and benign form of cancer be identified using data mining techniques?
3. How do missing feature values improve prediction in determining the performance achieved by data mining algorithms?
4. Does hybridization model of the existing data mining algorithms produce better approaches for cancer in terms of classification accuracy and missing values?

5. How can the developed hybrid data mining model be tested for classification accuracy and missing values?

1.5 Justification Of The Study

The medical industry has been slow to adopt new, efficient and timely data mining techniques which ideally lower the cost of information and accelerate information access. These are the things that healthcare practitioners want i.e. integrated historical data, easy and fast information access.

In a global perspective, the limited medical resources and long waiting times to receive medical services has magnified people's suffering. The World Health Organization (WHO) ranks Kenya at 140 out of 190 countries in their report of the year 2000. A study shows that all African countries including Kenya had fewer practicing physicians and limited care beds per one thousand people than the median of some countries. This is according to the Organization for Economic Cooperation and Development (OECD) (Source: OECD Health Data, 2010).

The available medical resources and infrastructure force Health organizations and state governments to set procedures, plans, manage, and cope with the challenges of medical personnel and equipment. This helps them in delivering decent healthcare services for residents however there still exists shortage of innovative e-Health technologies. IG-ANFIS could be the solution for this suffering.

1.6 SCOPE OF THE STUDY

In this thesis, EHRs have been used as data sources for developing automatic data mining techniques, so as to produce useful patterns and decision support logic for automatic computer aided diagnosis. The study has used Wisconsin Breast Cancer (WBC) datasets from the University of California Irvine (UCI). This is a machine learning repository available publicly for research purposes. The research will combine Naïve Bayes and *k*-NN as one classifier for constructing missing feature values to find the most suitable feature values that satisfy classification accuracy.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Data mining (DM) is the process of discovering meaningful correlation, patterns, and trends by sifting through large data, using recognition technologies. DM emphasizes on making and testing algorithms that can assist the process of classification, prediction, and pattern recognition. This process uses computer models obtained from existing data (previous data) with limited human interaction. The idea is to increase accuracy and reduce human biases by using automatic pre-programmed methods. As a result, a solid and reliable functional data mining algorithms can be developed to classify objects or predict new cases of diseases.

2.2 Data Mining

Frawley (2012) describes DM as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Baxt (1990) defines DM as the process of automating information that has been discovered. Moxon (2012) states that data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Han and Kamber (2012), argues that DM techniques can be considered to be descriptive (summarize data and to highlight their interesting properties) or predictive (build models to forecast future behaviours).

2.3. 3 Machine Learning (Ml)

ML is a scientific discipline responsible for recognizing complex patterns and making intelligent decisions based on data. Emphasizing on making and testing algorithms, ML can assist the process of classification, prediction, and pattern recognition using computer models. ML provides limited human involvement and uses the automatic pre-programmed methods that reduce human biases. The process of proposing the algorithm and its functionality to classify objects or predict new cases are to be built on solid and reliable data, Mitchell, 1997. The database contains a collection of instances (records or case). Each instance used by ML algorithms is formatted using same set of fields (features, attributes, inputs, or variables). When the instances contain the correct output (class label) then the learning process is called the supervised learning. Whilst the process of ML without knowing the class label of instances is called unsupervised learning.(Ozgür, 2004), clustering is a common unsupervised learning

method (some clustering models are for both). The goal of clustering is to describe data. However, classification and regression are predictive methods. This research will focus on supervised machine learning.

2.4 Classification

Classification is the process of learning the target function that maps between a set of features (inputs) and a predefined class labels (output) i.e. it puts data in single groups that belongs to a common class, inferring the defining characteristics of a certain group done by Regression algorithms which attempt to map input to domain values. For instance, a regressor can forecast certain goods sales by considering the goods features. At the same time, classifiers can map the input space into pre-defined classes. Consequently, a classifier can predict a new case of patient whether benign (healthy) or malignant (suffer from a certain disease**).**

Kotsiantis et al, 2007, describes supervised ML as the search for algorithms that reason from externally supplied instances to produce general hypotheses; the general hypotheses are then used to make predictions about future instances. The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, and the value class label is unknown.

The input data for the classification is a set of instances. Each instance is a record of data in the form of ($\boldsymbol{x}$,y) where $\boldsymbol{x}$ is the features set and $\boldsymbol{y}$ is the target variable (class label). Classification model is a tool that is used to describe data (Descriptive Model) or a tool to predict the target variable for a new instance (Predictive Model). The decision tree, artificial neural network, Naïve Bayes, and k-nearest neighbour's classifier are some of the examples of classification models.

The general approach for solving classification problem is shown in Figure 2.1. The training data consists of instances whose class labels are known. The classification model can be built based on the training data. The model then can be evaluated and tested by using the testing data which contains records with missing class labels. The evaluation of model performance is based on the number of testing instances that are correctly forecasted. The result of performing the model on the testing data produces the confusion matrix.

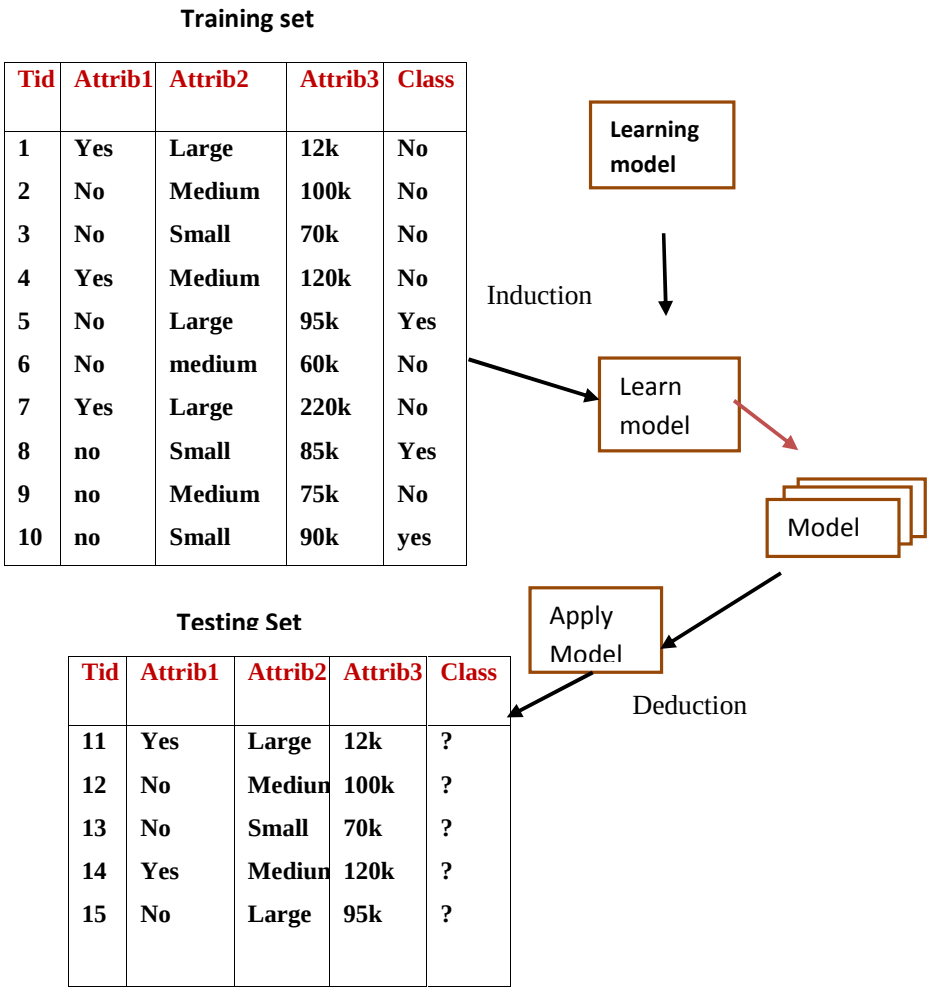

Figure 2.1: General approach for building a classification model

(Source: Review of Classification Techniques by Kotsiantis, 2007).

A Classification problem can be solved using the following illustration under;

Suppose the goal is to classify some objects $i=1,\ldots,n$ into k predefined classes, where k represent the number of classes, i.e. if the aim of classification is to diagnose a patient whether or not suffering from cancer then the value of k will be 2 corresponding to either benign or malignant. The database (available data) can be organized as $n \; x \; p$ matrix X, where xij represent the feature value j in the record i. Every row in the matrix X is represented by a vector xi with p features and a class label yi. The classifier can then be denoted as (x).

One method to evaluate the classifier is by calculating the error estimation based on the confusion matrix. Error estimation can be explained by considering an example as follows; suppose the aim of a certain classifier ($\boldsymbol{x}$) is to train and test input vectors $\boldsymbol{x}$ into two possible classes benign and malignant. Suppose the result of classification of the classifier ($\boldsymbol{x}$) on vectors $\boldsymbol{x}$ is as shown in the confusion matrix in Table 2.1

Table 2.1: The confusion matrix for classifier c(x) on matrix X that contains 160 records

	Predicted	
	Benign	Malignant
Benign	70	15
Malignant	5	90

The error rate (Er) of algorithm is computed as the total number of incorrectly classified samples divided by the total number of records in the matrix X. In the example above,

Er = (15 + 5) / 160 = 0.125. Classification accuracy of the model can be calculated as:-

$$Acc = 1 - Er = 0.875 \quad \text{(2.1)}$$

2.4.1 K-Nearest Neighbors Algorithm (K-Nn)

K-NN is an instance based machine learning algorithm that classifies feature space based on the closest training cases. *K-NN* finds the *k* closest instances to a predefined instance and decides its class label by identifying the most frequent class label among the training data that have the minimum distance between the query instance and training instances.

The distance metric determines the distance i.e. it minimizes the distance between similar instances and maximizes the distance between different instances. Larose (2013), provides an illustration for *k*-NN implementation as shown in the following pseudo-code to define this metric distance. Euclidean and Manhattan methods are some amongst several of the approaches that are used for distance determination.

Procedure K-NN-Learner (Testing Dataset)

For each testing instance

{Find the k most nearest instances of the training set according to a distance metric (Euclidean distance or Manhattan distance)

Resulting Class = most frequent class label of the k nearest instances}

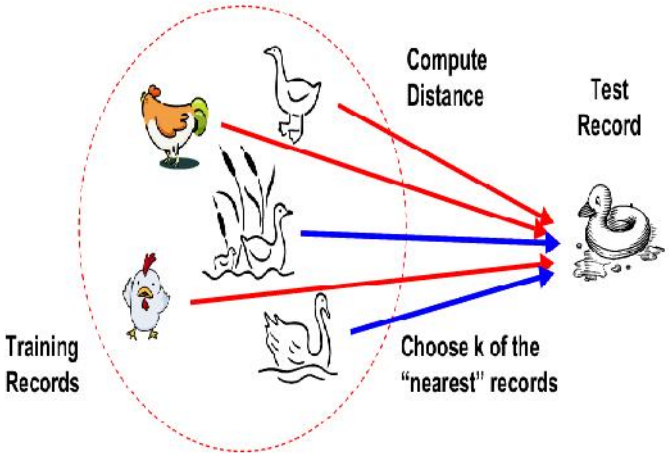

Figure 2.2: Example of *k*-NN

(Source: A Detailed Introduction to *K*-Nearest Neighbor (*K*-NN) Algorithm By Saravana Thirumuruganathan, 2010)

Advantages of K-NN :

i. It has a very efficient pattern recognition method and can be easily carried out
ii. Simple to use
iii. Strong against noisy data
iv. Can be used for large and small datasets
v. Suitable for linear and nonlinear functions
vi. Has the ability to add additional instances with no need to train the data set
vii. Its weight is used to measure features significance
viii. Missing values can be easily imputed using *k-NN*
ix. Has excellent flexibility (nonparametric model except the value of *k*)

Disadvantages of using k-NN

It requires that the distance between the query instance and all other instances calculated

i. It requires the use of a huge memory
ii. It is not useful for multidimensional dataset because of high error rate
iii. It has the option of using many distance functions which may lead to different accuracy level

2.4.2 Distance Functions

Young M., et al. 2004, in their Distance Metrics Overview, describes various distance metrics used to determine the distance between two data points. These are:-

i. Euclidean distance

The Euclidean distance is most regularly used metric to compute the distance between data points. The square root of the sum between two points is Euclidean distance. For n-dimensional data, the distance is given by the formula;

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \text{..... (2.2)}$$

Where d denote to distance, x and y are two different cases in the dataset, n is the total number of cases in the dataset.

ii. Manhattan distance

Another of the well-known function for measuring distance is Manhattan distance. Manhattan distance is calculated by summing the absolute value of the difference of data points. Manhattan distance is less costly to calculate in comparison to Euclidean distance. Manhattan distance is given by formula;

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad \text{..... (2.3)}$$

Where d denote to distance, x and y are two different cases in the dataset, n is the total number of cases in the dataset.

iii. Minkowski distance

Minkowski function is a geometric distance between two points and uses a scaling factor, r. The main use is to find the similarity between objects. When r=2 then it becomes the Euclidean distance. When r=1 then it become the Manhattan distance. The distance is given by the formula:-

$$d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|)r^{\frac{1}{r}} \quad \text{..... (2.4)}$$

Where d denote the distance, x and y are two different cases in the dataset, n is the total number of cases in the dataset.

iv. Chebyshev distance

Chebyshev distance function calculates the absolute differences between the coordinates of two points. Example of common application for using Chebyshev distance commonly used in Fuzzy C-means Clustering.

$$d(x,y) = max_i|x_i - y_i| \quad \ldots\ldots (2.5)$$

Where d denote to distance, x and y are two different cases in the dataset.

v. Canberra distance

Canberra distance is the sum of absolute values of the differences between ranks divided by their sum, thus it is a weighted version of the Manhattan distance function, where d denote to distance, x and y are two different cases in the dataset, n is the total number of cases in the dataset.

$$d(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad \ldots\ldots (2.6)$$

2.5 Feature Selection Techniques

Feature selection is the process of identifying as much irrelevant redundant information as possible. This reduces the dimensionality of the data and allows learning algorithms to operate faster and more effectively. The current approaches used in feature selection methods are correlation feature selection (CFS), principal components analysis (PCA), symmetrical uncertainty (SU), relief (R) and information gain (IG)

2.5.1 Correlation Feature Selection (Cfs)

CFS removes redundant or irrelevant features from the data set as they can lead to a reduction of the classification accuracy or clustering quality. This reduction leads to an unnecessary increase of computational cost (Blum & Langley, 1997). Koller and Sahami, 1996, With dimensionality reduction techniques the size of the attribute space can often be decreased strikingly without losing a lot of information of the original attributes space. There are three types of feature subset selection approaches: filters, wrappers, and embedded approaches which perform the features selection process as an integral part of a machine learning (ML) algorithm.

2.5.1.1 Wrapper Feature Selection Technique

The wrapper approach was proposed by Kohavi & Paeger, 1997 in Stanford university AI lab. In wrapper method, the feature selection algorithm was located as a wrapper around the learning algorithm. The process starts with a search for relevant subset of attributes by using the learning

algorithm. The learning algorithm itself is used to evaluate the feature subset which was obtained by the search.

Figure 2.3 illustrates how the wrapper approach performs on the training set and the evaluation process. The learning algorithm is treated as a black box with no modification to the learning algorithm itself. The learning algorithm assesses the subsets of features obtained by the search method. The learning algorithm obtains a hypothesis about the quality and the relevance of a certain feature subset. Features subset with the highest estimated value is chosen as the final set on which to run the learning algorithm. The final step is to evaluate the model on new dataset (not used by the search) to ensure the independency between the training process and the testing process. The result is an estimated accuracy by using the highly relevant features subset on the desired learning algorithm.

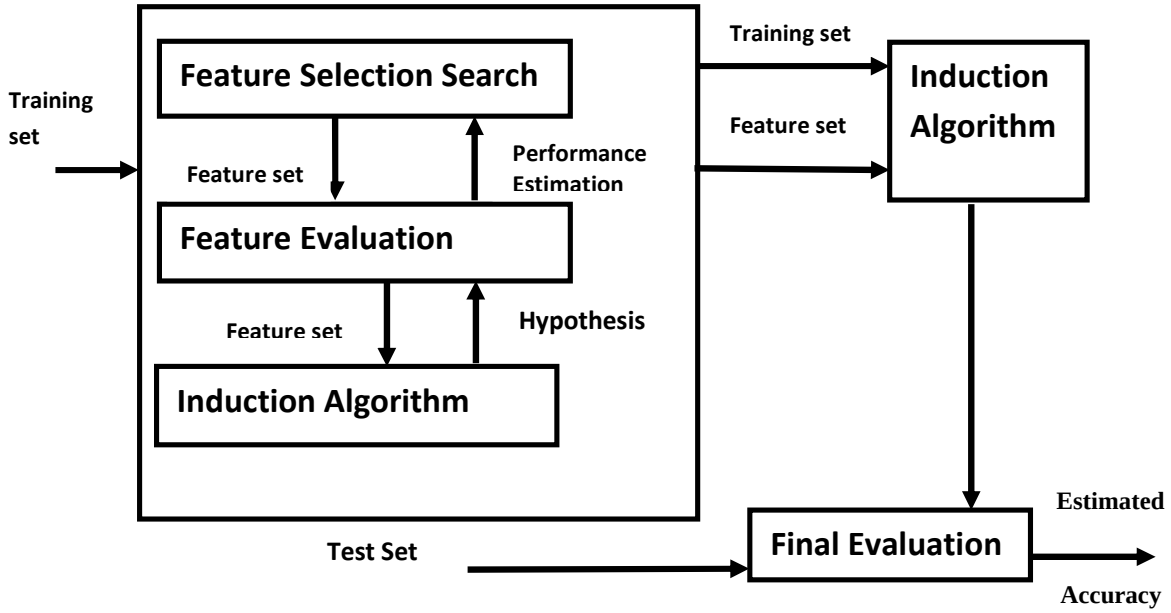

Figure 2.3: The Wrapper approach for features subset selection

(Source: "Wrappers for feature subset selection by Kohavi Ron, 1997)

The main advantages and disadvantages of using wrapper as a feature selection method, and examples of existence methods that utilize the wrapper approach are shown in Table 2.2

Table 2.2: Examples, advantages and disadvantages of Wrapper approach

Advantages	Disadvantages	Examples

• Simple to use and easy to implement • Interactive with learning classifier • Models feature dependencies	• The risk of overfitting • Computationally intensive	• Sequential forward selection by forward pass • Sequential backward elimination by backward pass

2.5.1.2. Filters Feature Selection Techniques

Filter techniques examine the significance of features by investigating the real characteristics of the data. In most cases feature rank is calculated, and low ranking features are ignored during the learning process. Afterwards, the high ranking subset of features is used as training set to the classification algorithm. The main difference of filter in comparison with wrapper is that filter ignores the learning algorithm during features subset search. Figure 2.4 shows the filter approach; it shows that features subset extraction is totally independent from the learning classifier.

Figure 2.4: The filter approach

(Source: A review of feature selection techniques by Saeys, 2007)

Some advantages of filter techniques include: - they are able to be performed on large databases that contain large number of attributes and cases, simple computation, fast in comparison to wrapper and embedded methods, and they are independent of the classification algorithm. The aim behind the independency between filters and learning classifier is that feature selection needs to be performed only once and then different classifiers can be used to evaluate the subset. On the other hand, the independency between filter methods and learning algorithms may cause low level of classification accuracy. Table 2.3 summarizes the main advantages and challenges of filter methods and some examples of popular filter methods.

Table 2.3: Examples, advantages and disadvantages of filter feature selection

Advantages	Disadvantages	Examples
• Relatively fast • Scalable • Independent classifier	• Ignores feature dependencies • Ignores interactive with classi	• Correlation based featur selection(CFS) • Relief

2.5.1.3. Embedded Feature Selection Techniques

Embedded Methods (EM) vary from other feature selection methods in how classification methods and feature selection cooperate. In filter methods, there is no corporation between learning classifiers and feature selection. In wrapper methods, the learning classifier is used to measure the quality of subsets of features without intervening in with the structure of the classification. In contrast to filter and wrapper approaches, EM feature selection methods and learning process cannot be taken apart. The process of finding the optimal subset of features is combined into the classifier construction. EM computational cost is less than wrapper methods and the fact that there is interaction between the classifier and EM is significant. Table 2.4 shows some advantages and disadvantages of using such a method along with examples.

Table 2.4: Examples, advantages, and disadvantages of embedded feature selection

Advantages	**Disadvantages**	**Examples**
• Interactive with learning classifier • Better computational complexity than wrapper	• Classifier is dependent on Selection method	• Decision tree • Weighted naïve Bayes

2.5.2 Principal Components Analysis (PCA)

PCA is a Dimensional reduction algorithms and techniques that create new attributes as combinations of the original attributes in order to reduce the dimensionality of a data set, Liu and Motoda, 1998. PCA produces new attributes as linear combinations of the original. Jolliffe, 2002, explains that the goal of PCA is to find a set of new attributes

which meets some criteria i.e. linear combinations of the original attributes, orthogonal to each other, and capture the maximum amount of variation in the data.

PCA can be represented mathematically as the covariance of two attributes and measures how strongly the attributes vary together. The covariance of two random variables x and y of a sample with size n and mean x, y can be calculated as

$$\mathrm{Cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y - \overline{y}).$$ ………………………….. . (2.7)

Where x and y are normalized by their standard deviations δ_x and δ_y then the covariance of x and y is equal to the correlation coefficient of x and y,

Corr(x, y) = Cov(x, y)/$\delta_x \delta_y$, which indicates the strength and direction of a linear relationship between x and y.

2.5.3 Symmetrical Uncertainty (SU)

SU is a probabilistic model of a nominal valued feature Y that can be formed by estimating the Individual probabilities of the values $y \in Y$ from the training data. If this model is used to estimate the value of Y for a novel sample (drawn from the same distribution as the training data), then the entropy of the model (and hence of the attribute) is the number of bits it would take, on average, to correct the output of the model. Entropy is a measure of the uncertainty or unpredictability in a system. The entropy of Y is given by:-

$$H(Y) = -\sum_{y \subset Y} p(y) \log 2\ (p(y)$$ …………………………………………….(2.8)

If the observed values of Y in the training data are partitioned according to the values of a second feature X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y prior to partitioning, then there is a relationship between features Y and X.

2.5.4 Relief (R)

R is a feature weighting algorithm that is sensitive to feature interactions Kononenko notes that R attempts to approximate the following difference of probabilities for the weight of a feature X:

W_X= P$(different\ value\ of\ X | neares\ instance\ of\ different\ class)$

- p$(different\ value\ of\ X|nearest\ instance\ of\ same\ class)$

This can be reformulated as-:

$$Relief_X = \frac{Gini'\ X \sum_{x \in X} p(x)^2}{(1-\sum_{c \in C} P(c)^2) \sum_{c \in C} P(c)^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.9)$$

Gini' is a modification attribute quality measure, p is the probability of the instance occurring, c is the class attribute, X is the feature in consideration and is the relation between the in distances between the nearest features.

2.5.5 Consistency Subset Evaluation (CSE)

CSE is an algorithm that exhaustively searches the space of feature subsets until it finds the minimum combination of features that divides the training data into pure classes (that is, where every combination of feature values is associated with a single class). CSE algorithm consists of forward selection search coupled with a heuristic to approximate the min-features bias. CSE is computationally feasible on domains with many features. **CSE** algorithm evaluates the features using the formula:-

Entropy (Q) = - $\sum_{i=0}^{2|Q|-1} \frac{pi+[ni}{|sample|} \left[\frac{pi}{pi+Ni} log_2 \frac{pi}{pi+Ni} + \frac{ni}{pi+ni} log_2 \frac{ni}{pi+ni}\right]$…….. (2.10)

Where, for a given feature subset Q, there is 2|Q| possible truth value assignments to the features. And for a given feature set Q divides the training data into groups of instances with the same truth value assignments to the features in Q. Where pi and ni denote the number of positive and negative examples in the $i-th$ group respectively

2.5.6 Information Gain (IG)

IG is a feature selection technique used to reduce the number of input features to ANFIS. It uses ranking method and is often used in text categorization. If x is an attribute and c is the class, the following equation gives the entropy of the class before observing the attribute:

$$H(x) - -\sum_x p(x) \log_2 p(x) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.11)$$

Where (p) is the probability function of variable c and the conditional entropy of c given x (post entropy) is given by:

$$H(c/x) = -\sum_x p(x) \sum_c p(c/x) \log_2 p(c/x) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.12)$$

The information gain (the difference between prior entropy and postal entropy) is given by the following equations:

$$H(c, x) = H(c) - H\,c|x) \text{…….....(2.13)}$$

$$H(c, x) - -\sum_c p(c)\log_2 p(c) - \sum_x(-p(x)\sum_c p(c/x)\log_2 p(c/x) \text{.....(2.14)}$$

2.6. Machine Learning Methods

A machine learning algorithm also called an induction, forms concept descriptions from a sample data. The concept descriptions are often referred to as the knowledge or the model that the learning algorithm has induced from that data. The machine learning algorithms are used for comparison in this thesis. These are artificial Neuro network, Naïve Bayes and decision tree.

2.6.1 Artificial Neural Network (Ann)

An artificial neuron (AN) is a computer simulated model that is stimulated from natural neurons. Natural neurons receive signals from synapses located on the surface of the neuron. When the neuron starts to work it sends a signal through the axon once the signal extend to a certain threshold. This signal then transfers through to other neurons and may get to the control unit (the brain) for a proper action. Priddy, 2005, ANN dates back to the nineteenth century when William James and Alexander Bain realized the ability of constructing a man-made system based on neural models. Widrow and Hoff 1960s developed the Adaptive linear Element (ADALINE) that was used to eliminate the echoes in telephone systems based on adaptive signal processing.

In 1974, Paul Werbos had developed a learning rule based on error minimization approach in which the error is propagated in reverse by adjusting the weights using the Gradient descent model. Paul's technique is the back propagation error algorithm which is the most used artificial neural networks model that spread widely in mid 1980s by a group of researchers.

During 1980s and 1990s, computers had extended in speed about hundred times quicker since the beginning of the research, academic programs appeared, new courses were introduced, and funding became available. All the mentioned factors encouraged researchers to concentrate on neural networks application, development, and new approaches for prediction, forecasting, and diagnoses. Many studies demonstrated the potential applications of ANN for clinical decision making. Figure 2.5 shows a representation of the human neuron.

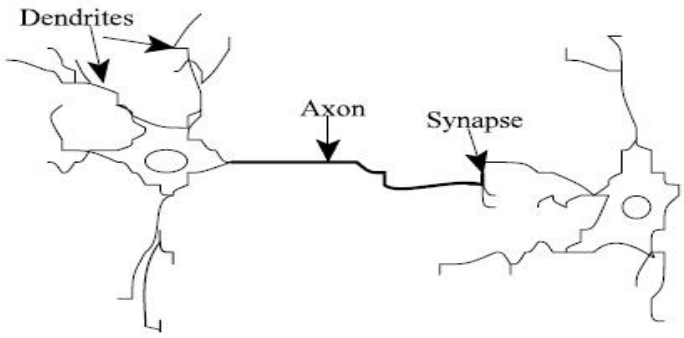

Figure 2.5: Human neuron

AN simulates the functionality of real neuron and has a set of inputs associated with weights. Inputs and weights are calculated by a mathematical equation to control it when the AN is activated. ANN is a combination of artificial neurons that process information. Figure 2.6 shows a simple artificial neuron

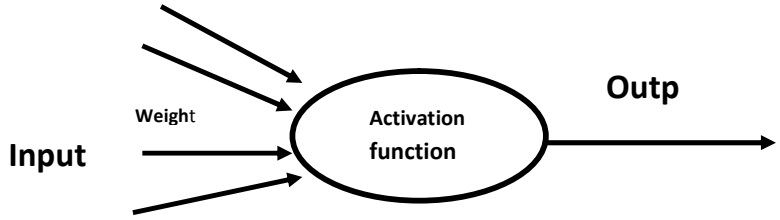

Figure 2.6: A simple artificial Neuron

(Source: Artificial neural networks for beginners by Gershenson Carlos, 2003)

In a general sense, the AN operation can be modeled by use of the data flow diagram as shown in Figure 2.7

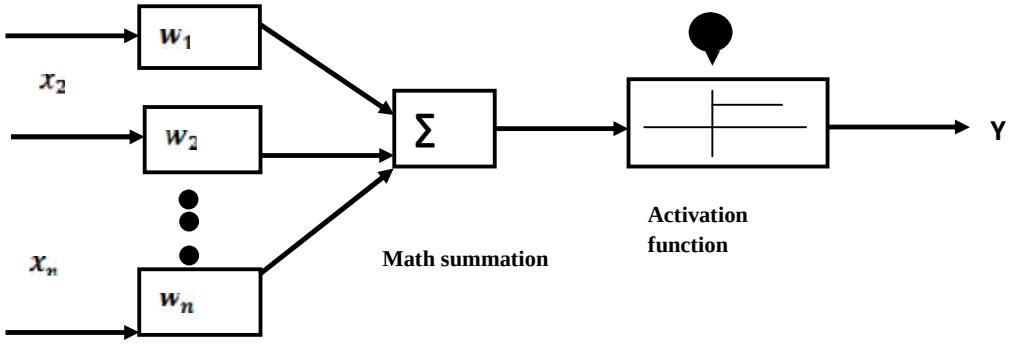

Figure 2.7: Simplified neuron operation

(Source: Artificial neural networks for beginners by Gershenson Carlos, 2003)

ANN is a set of connected artificial neurons. The most used ANN model is the Feed Forward Networks. Figure 2.8 shows a three layer topology of Feed Forward Networks.The outcome of ANN is subject to input and the value of the weight.

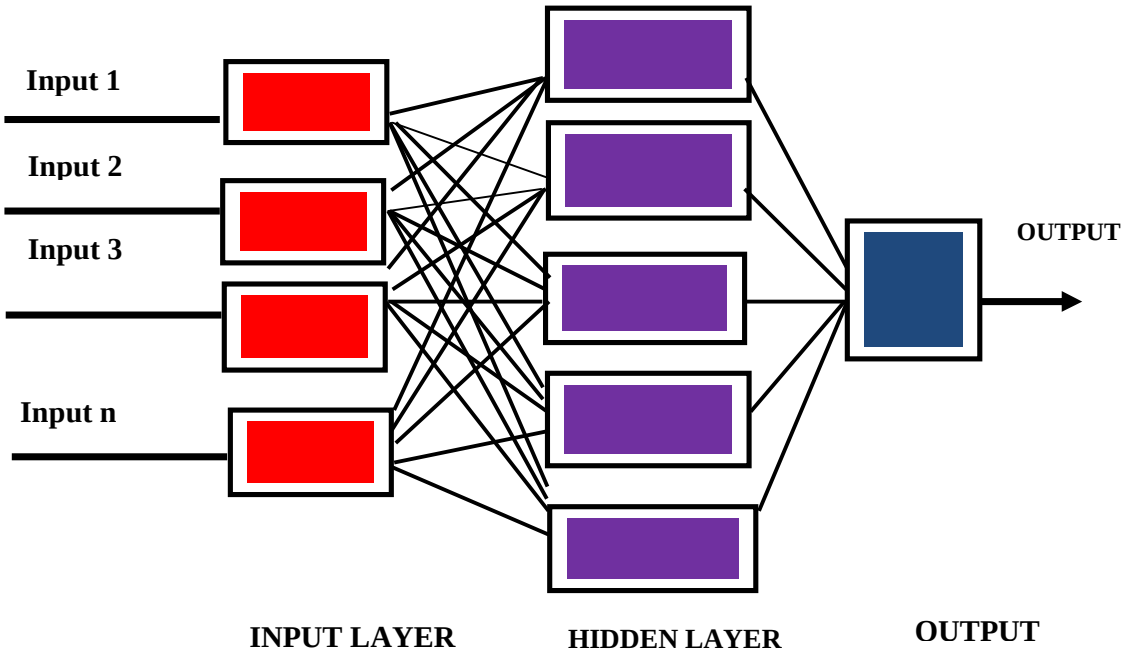

Figure 2.8: ANN architecture

(Source: Artificial neural networks for beginners by Gershenson, 2003)

The learning features for Artificial Neural Network includes; accuracy in general, speed of learning, speed of classification, tolerance to missing values, tolerance to irrelevant attributes, tolerance to redundant data, tolerance to noise, dealing with overfitting, and explanation ability.

2.6.2 Decision Tree (DT)

DT is a classification method which contains nodes, branches, and leafs. The first node on the tree or the top node is called the root node. Each node in the tree is connected with one or more nodes using branches, the last node in the tree that contains no outgoing branches is called leaf node. The leaf node indicates to termination or the outcome value. Figure 2.7 shows an example of how a real time problem is solved based on making questions and answers about attributes in the testing records. The terminology of such classification method is to keep asking question until conclusion is reached. The set of questions and answers could form a decision tree with set of nodes: first, root node having a zero or more outgoing nodes and no incoming nodes, as well as containing the testing condition that separate the records; second, Normal nodes, those nodes are internal nodes and each has one and only one incoming node and two or more outgoing edges. It also contains the testing condition that separate records and thirdly, Leaf nodes, those nodes hold the class labels, have no outgoing edges, and only one incoming edge.

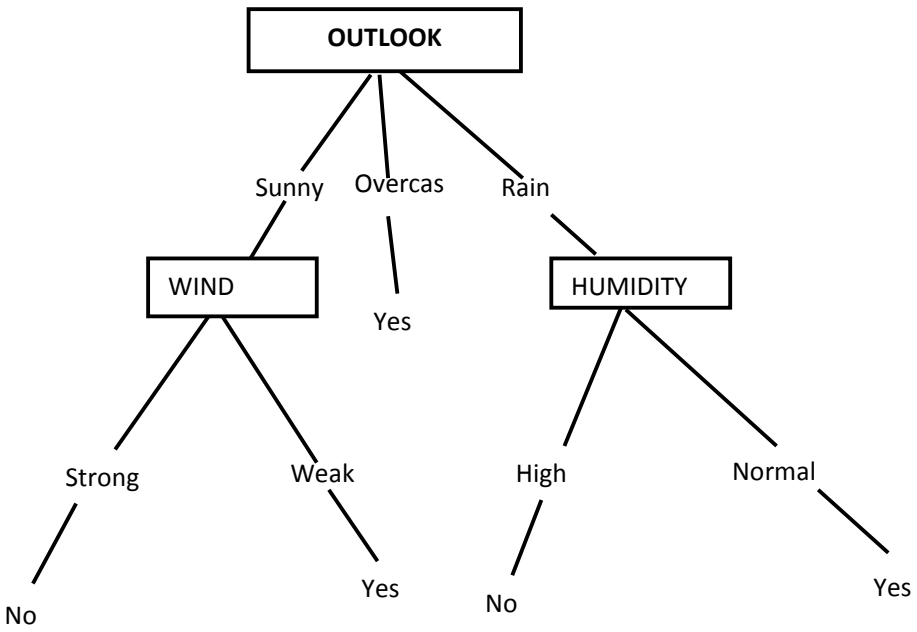

Figure 2.9: Simple Decision Tree

(Source: Data mining with decision trees; by Rokach L., 2007)

2.6.3 Naïve Bayes Classifier (NB)

NB classifier is a mathematical classifier based on independency and probability (Bayes theorem). The Naïve Bayes classifier adopts the idea that the existence of a certain feature of an object is unrelated to the existence of any other feature, given the class variable. E.g. an animal may be considered to be a cat if it can hunt, play with kids, has four legs, has a head, and weighs about 3kgs. Naïve Bayes algorithm treats all features independently and how they make a prediction with no feature depending on other features values. Its advantages includes: it's easiness to construct, it requires no parameter estimation, it's easiness to interpret, it can be performed by expert and inexpert data mining developers and performs well in comparison with other data mining methods.

The two types of Naïve Bayes in literature include Multinomial model and Multivariate. In these models, the classification is performed by the following Naïve rule:

$$P(c_j|x_i) - \frac{P(cj).P((x_i|c_j)}{P(x_i)} \quad \text{..... (2.15)}$$

Where c_j is the instance class label, x_i is the test attribute, $P(c_j|cx_i)$ is the posterior probability of the class label c_j given the attribute x_i, (c_j) is the prior probability of class label c_j, $(xi|cj)$ is the likelihood which is the probability of attribute x_i given the class label c_i. Assume that each attribute x_i is conditional independent of every other attribute x_j then the conditional distribution over the class variable c is:

$$P(c|x_i) = p(c)\prod_{i=1}^{n} p(x_i|c) \quad \text{(2.16)}$$

The advantage of Bayesian classifier over other classification methods is its opportunity of considering the prior information about a given problem. The main disadvantages of Bayesian classifier are (1) the numerical attributes require discretization in most cases; (2) it is not suitable for large data sets which contain many attributes.

2.7 ANFIS Structure

Adaptive Neural Fuzzy Inference System (ANFIS) exploit the advantages of NN and FIS by combining the human expert knowledge (FIS rules) and the ability to adapt and learn. Three major components constitute Fuzzy Inference systems (FIS). This includes: a rule base which is made up of a selection of fuzzy rules; a database that defines the membership functions and a reasoning mechanism that is a way of inferring a reasonable output or conclusion.

Our approach applies Sugeno fuzzy rules which can be illustrated as follows; for a first-order Sugeno fuzzy inference system with two inputs, a common rule set with two fuzzy if-then rules is the following:

Rule 1： if x is A_1 and y is B_1, then $f_1 = p_{1x} + q_{1x+r_1}$ ..... (2.17)

Rule 2： if x is A_1 and y is B_2, then $f_2 = p_{2x} + q_{2x+r_2}$ ..... (2.18)

For understanding purposes, these rules can be described as follows:

Letting the membership functions of fuzzy sets: A_i, B_i i=1, 2, to be μA_i μB_i

We can evaluate the rule premises which results in:-

$w_i = \mu_{A_i}(x) \times \mu_{B_i}(y)$, i=1, 2 ..... (2.19)

Evaluating the implication:

Rule 1: $\mu_{A_1}(x)\mu_{B_1}(y)\ f_1(x, y) = w_1(x, y)\ f_1(x, y)$ .....(2.20)

Rule 2: $\mu_{A_2}(x)\mu_{B_2}(y)\ f_2(x, y) = w_2(x, y)\ f_2(x, y)$ .....(2.21)

Evaluating the rule consequences and aggregating them becomes

$$f(x, y) = \frac{w_1(x,y) f_1(x,y) w_2(x,y) f_2(x,y)}{w_1(x,y)+w_2(x,y)} \quad \text{.....(2.22)}$$

Simplifying this further, it becomes:

$$f=\frac{w_1 f_1 - w_2 f_2}{w_1 + w_2} \quad \text{..... (2.23)}$$

Separating this computation into two phases with notation they become:

$$f = \overline{W_i} = \frac{w_i}{w_1+w_2} \quad \text{..... (2.24)}$$

Then

$$f = \overline{w_1} f_1 + \overline{w_2} f_2 \quad \text{..... (2.25)}$$

Where w_1 and w_2 are the FIS inference rule, x,y are input and output member functions and f is the class label.

Figure 2.10 (a) shows the fuzzy reasoning and (b) shows the corresponding structure of ANFIS.

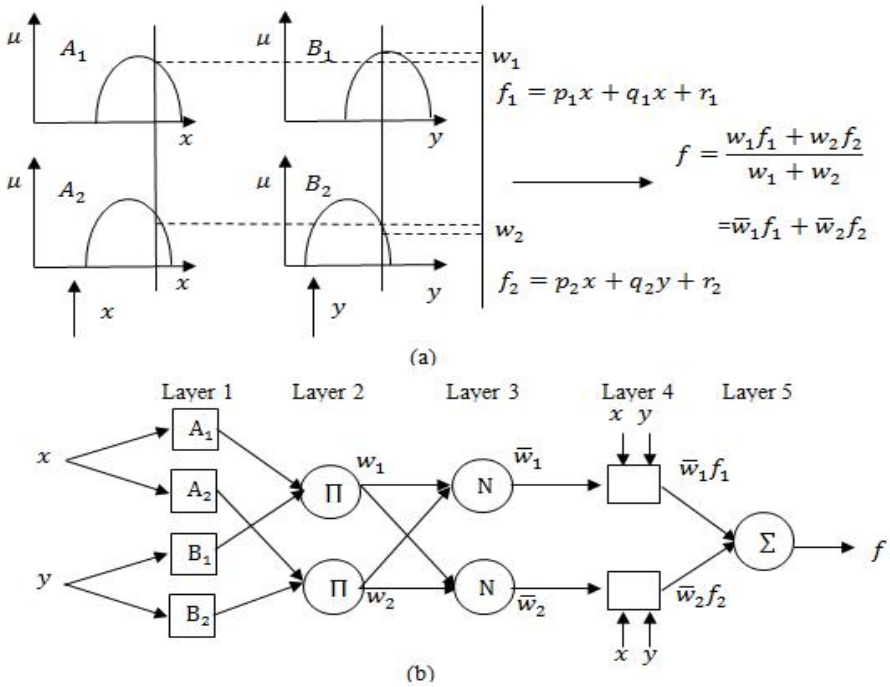

Figure 2.10: (a) A two-input first-order Sugeno FIS with two rules,

(b) An equivalent ANFSI Architecture.

An ANFIS network of five layers is demonstrated with the equivalent Sugeno fuzzy inference system in Figure 2.10.

The Learning of ANFIS applied consists of structure - learning in the first place and then parameters-learning. Structure-learning includes space classifying of fuzzy input and rule-extracting. Accordingly clustering is done by extracting a set of rules that models the data

behavior to classify the training sample space. If the space is clustered into n_i classes, then there will be corresponding n_i fuzzy rules. Hence, initial input parameters of membership functions for each class are determined by the clustered center coordinates and its radius length. In Figure 2.10 (b), the node function in each layer can be described as follows:

i. Layer1:

Each node i (represented as a square) in this layer accepts input and computes the membership $\mu_{A_i}(x)$

$$O_i^1 = \mu_{A_i}(x) \quad \ldots\ldots (2.26)$$

Where x is the input to node i, and A_i is the label (small, large, etc.) associated with this node. In other words, O_i^1 is the membership function of A_i and it specifies the degree to which the given x satisfies the quantifier.

$\mu_{A_i}(x)$ is chosen to be bell-shaped with values between 0 and 1, such as the generalized bell function:

$$\mu_{A_i}(x) = exp\left[\left(\frac{x - c_i}{a_i}\right)^2 \right] \quad \ldots\ldots (2.24)$$

Where a_i and c_i are two parameters called premises

ii. Layer2:

Every node in this layer (represented by a circle) takes the corresponding outputs from Layer 1 and multiplies them to generate a weight:

$$w = \mu_{A_i}(x) \times \mu B,\ i = 1, 2 \quad \ldots\ldots (2.28)$$

The output of this node represents the firing strength of the rule.

iii Layer3:

Every node in this layer is a circle node labeled *N*. This layer normalize the weight of a certain node in comparison to the sum of other nodes weights (The ratio of weight) then compute the implication of each output member function.

$$\overline{w_i} = \frac{w_i}{\sum_j w_j},\ \ i = 1, 2.\ j = 2 \quad \ldots\ldots (2.29)$$

iv. Layer 4:

Every node in this layer is illustrated with a square. Based on Sugeno inference system, the output of a rule can be written on the following linear format:

$$O_i^4 = w_i f_i = w_i(p_i x + q_i y + r_i) \quad \ldots\ldots (2.30)$$

v. Layer 5:

This layer called the aggregation layer, which computes the summation of rules, the proposed algorithm produce a single output (centroid):

$$0_i^5 = finaloutput = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad \text{..... (2.31)}$$

Note that the Output is linear in consequent parameters p, q, and r are linear.

2.8 Breast Cancer Diagnosis Based On IG-ANFIS

ANFIS was first proposed by Jang in 1993. ANFIS can be easily implemented for a given input/output task. This characteristic makes it attractive for many application purposes. ANFIS is a combination of two machine learning approaches; NN and Fuzzy Inference System (FIS). The ANFIS model integrates the ANN and FIS tools into a "compound", meaning that there are no boundaries to differentiate the respective features of ANN and FIS. ANFIS data mining technique with a pre-processing stage involving IG method enhances breast cancer dataset's classification accuracy.

2.8 Related Work

2.8.1 Data Mining Applications in Medical Diagnosis

Meesad and Yen (2003) proposed a hybrid Intelligent System (HIS) which integrated the Incremental Learning Fuzzy Network (ILFN) with the linguistic knowledge representations. The linguistic rules were determined based on knowledge embedded in the trained ILFN or been extracted from real experts. In addition, the method also utilized Genetic Algorithm (GA) to reduce the number of the linguistic rules to sustain high accuracy and consistency.

After being completely constructed, the system could incrementally learn new information in both numerical and linguistic forms. The proposed method was evaluated using Wisconsin Breast Cancer (WBC) Dataset. The results showed that the proposed HIS performed better than some well-known methods.

Setiono (2006) proposed a method to extract classification rules from trained neural networks and discussed its application to breast cancer diagnosis. He explained how the pre-processing of datasets improve accuracy of the neural network and the accuracy of the rules since some rules could be extracted from human experience, and may be erroneous. The data pre-processing involved the selection of significant attributes and the elimination of records with missing

attribute values from Wisconsin Breast Cancer Diagnosis WBCD dataset. The rules generated by Setiono's method were more brief and accurate than those generated by other methods mentioned in the literature.

Song, 2010 presented an automatic breast cancer diagnosis, a hybrid system for diagnosing new breast cancer cases in collaboration between GA and Fuzzy Neural Network. They showed that many problems having high complexity and strong non-linearity with huge data to be analyzed, can use inputs reduction i.e. feature selections methods.

Arulampalam and Bouzerdoum, 2011, proposed a method for diagnosing breast cancer named Shunting Inhibitory Artificial Neural Networks (SIANNs). This was a neural network stimulated by human biological networks in which the neurons interact among each other's via a nonlinear mechanism called shunting inhibition. The feed forward SIANNs have been applied to several medical diagnosis problems and the results were more favourable than those obtained using Multilayer Perceptions (MLPs). SIANNs showed reduction in the number of inputs.

Liao, 2010, proposed a hybrid features selections method along with *k*-NN and support vector machine (SVM). This was used to identify the most significant genes that demonstrate the highest capabilities of discrimination between sample classes. First they ranked the genes in terms of their expression difference using filter method and then a clustering method based on *k*-NN principles for clustering gene expression data. SVM was applied to validate the classification performance of candidate genes. The experimental results demonstrated the effectiveness of their method in addressing the problem.

Vijayasankari and Ramar, 2012, proposed a novel hybrid features selections method to select relevant features and cast away irrelevant and redundant features from the original dataset using C4.5 and Naïve Bayes classifier. The efficiency and effectiveness of the proposed method was demonstrated through extensive comparisons with other methods using real world data of high dimensionality. Experimental results on datasets revealed that the algorithm increased classifier accuracy with less error rate.

Hall and Holmes, 2003, presented a benchmark comparison of several attribute selection methods for supervised classification. Attributes selections is achieved by cross-validating the attribute rankings with respect to a classification learner C4.5 and Naïve Bayes. The results concluded that features selections methods can enhance the performance of some learning algorithms. The findings also concluded that Correlation based feature selection method has produced the best result among six different feature selections methods, However increasing the number of features led to a drop of performance.

Saeys, 2007 reviewed the importance of feature selection approach in a set of well-known bioinformatics applications. They focused into two the large input dimensionality, and the small sample sizes. The results showed that the features selection applications are fundamental in dealing with high dimensional applications.

Donald Rubin, 2008 classified the missing feature values from the literature into three types: missing completely at random, missing at random, and missing not at random.

i. Missing Completely At Random (MCAR)

MCAR describes how the missing values occurred. Here the probability that a feature value is missing is unrelated to the feature value or to the value of any other features in the dataset e.g. data may be missing because equipment malfunctioned, the weather was terrible and could not record the observation for a certain day, people got sick, or the data were not entered correctly.

ii. Missing At Random (MAR)

MAR is the case when the existence of missing feature value does not depend on the feature value itself and may depend on other features values in the dataset e.g. a depressed person is more likely not to report income just due to depression.

iii. Missing Not At Random (MNAR)

MNAR is the case when the missing feature value is not missing at random or completely at random e.g. if a person suffers depression and a person who suffer depression is more likely not report his mental status, then the data are not missing at random. Respectively, if a person refuses to tell the age, then the missing data are not random.

Howell David 2010 (2009) explains that the most popular methods for dealing with missing feature values are omitting instances, imputation, and expectation maximization. All these methods can be applied in conjunction with any classifier that operates on complete data. These methods are:-

a) Omitting Instances

In this method, any record of data that contains missing features values is deleted from the data set. After omitting instances that contain missing features values, classification process run on the remaining instances. The main drawback of this method is discarding important information in some cases. This is not a common method. However, it could be used in cases of a small amount of missing data.

b) Features Imputation

This is a well-known method for constructing missing features values in the datasets for learning purposes. The imputation method can be divided into two major types: single imputation and multiple imputations. In single imputation, the missing features values are substituted by the

correspondence features values according to certain rules such as the features values means, mode, median, or learning algorithm e.g. the mean imputation calculate the mean of feature *f* in the dataset that contain values which is then used to fill the features *f* that has missing values. The scenario for constructing missing features values in multiple imputations is similar to the scenario for single imputation. However, the multiple imputation use more than one value to fill missing features values in the dataset, such as mean of observed feature values, the mode of observed feature values, and regression method. However multiple imputations approach has a number of drawbacks include the computational cost being higher than in single imputation. However, the classification performance (accuracy) is higher than single imputation.

c) Expectation Maximization (EM)

Expectation Maximization is one of the most effective methods for handling missing data. To perform Expectation Maximization; the mean, variance, and covariance are estimated from instances whose data is complete, Moss S and Hancoock E, 2009. Expectation Maximization uses maximum likelihood procedures to estimate regression equations to calculate the relationships between variables.

2.9 The Proposed Approach

2.10 IG-ANFIS

IG-ANFIS data mining method for Cancer Diagnoses has been used. The approach uses the advantages of ANFIS and IG method. The output of IG became the input for ANFIS.

2.10.1 Treating Missing Feature values

The approach for constructing missing feature values was based on iterative nearest neighbors' and distance metrics. This approach employed weighted k-nearest neighbors' algorithm and propagated the classification accuracy to a certain threshold. Classification accuracy in the constructed dataset was computationally compared with original dataset containing some missing feature values.

Missing feature values that matters but still missing creates a challenge for researchers in data mining applications. Handling unknown attributes values with the most appropriate values is a common concern in data mining and knowledge discovery. It was important to Construct missing values most supervised and unsupervised data mining they affect the quality of learning and performance of classification algorithm.

2.10.2 Training ANFIS Model

The method to train ANFIS is the hybrid learning algorithm. This algorithm uses the gradient descent method and Least Square Estimate (LSE). Each cycle of the hybrid learning consists of a forward pass and a backward pass. In the forward pass the signal travels forward until Layer 4 and the consequent parameters are identified using the LSE method. In the backward pass the errors are propagated backward and the premise parameters are updated by gradient descent.

The process is repeated until it achieves the lowest error or a predefined threshold. In other words; the total parameter set is split into three: S = set of total parameters, S_1 = set of premise (nonlinear) parameters, S_2 = set of consequent (linear) parameters. So, ANFIS uses a two pass learning algorithm: where S_1 is unmodified and S_2 is computed using a LSE algorithm. In Backward Pass, S_2 is unmodified and S_1 is computed using a gradient descent algorithm such as back propagation. So, the hybrid learning algorithm uses a combination of steepest descent and least squares to adapt the parameters in the adaptive network. The simple process followed by ANFIS is;-

Forward pass: present the input vector - calculate the node outputs layer by layer - repeat for all data $\rightarrow A$ and y formed - identify parameters in S_2 using least squares - compute the error measure for each training pair.

Backward Pass: Use steepest descent algorithm to update parameters in S_1 (back propagation)

For given fixed values of S_1 the parameters in S_2 found by this approach are guaranteed to be the global optimum point. Based on the approach of error correction learning, Back propagation method systematically trains and provides a computationally efficient method for changing the synaptic weights in the neural network with differentiable activation function units. This is an error back propagation algorithm that uses method of supervised learning. In our case, we provide the algorithm with the recorded set of observations or training set i.e. examples of the inputs and the desired outputs that we want the network to compute, and then the error (difference between actual and expected results) is computed.

These differences in output are back propagated in the layers of the neural networks and the algorithm adjusts the synaptic weights in between the neurons of successive layers such that

overall error energy of the network, E is minimized. The idea of the back propagation algorithm is to reduce this error, until the ANN learns the training data. Training of network i.e. error correction is stopped when the value of the error energy function has become sufficiently small and as desired in the required limits. Total error for p^{th} observation of data set and j^{th} neuron in the output layer can be computed as:

$$E_i = t_i - y_i \quad \text{..... (2.32)}$$

Where t_i, represents the desired target output and y_i represents the predicted output from the system.

Summary

This chapter has presented a background study of the main data mining technologies used in the current research. These problems posed by the current techniques have been identified. The problem were missing feature values and how to process them, huge features and attributes and how to select the most beneficial ones, extracting accurate diagnostic markers that can predict the early onset of the disease and monitoring of different stages of the disease. IG-ANFIS approach reduced the number of features to the optimal using the IG and the output was then fed as input to ANFIS.

CHAPTER THREE
METHODOLOGY

3.1 Research Design

This Research Was A Positivist Research (Scientific) And Was Based On Utilizing The Principles Of Prediction Upon Previous History And Data Manipulations (Manipulating In This Regard, Does Not Involve Change In Data Structure Or Values. However, Manipulating Data Is The Process Of Filling Missing Feature Values, Treating Noisy Data, Data Normalization Etc.)
This research design was based on survey, observation and reasoning as a tool for understanding a certain problem or behavior, i.e. this was an experimental research design. The design involved manipulations to variables and predictions on the basis of previous observation or history. The study was concerned with what could be the cause of a particular relationship and what the effects of that relationship could be.

3.2 Data Mining Methodology

Knowledge discovery or DM refers to extracting useful relationships and patterns from large databases. Due to the vast amount of data, and to obtain useful outcomes, a systemic method that was applied in the research was represented in the figure 3.1.

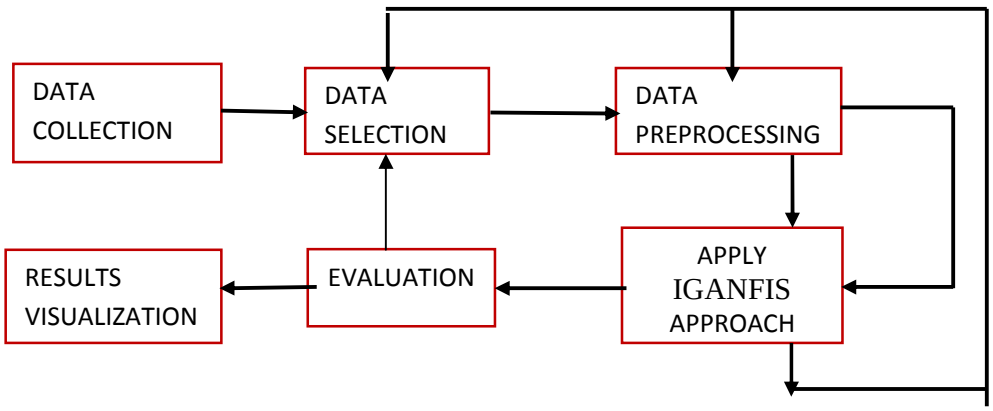

Figure 3.1: IGANFIS Conceptual framework

3.2.1 Population and Sample

The research study used samples from UCI repository. These samples were Wisconsin Breast Cancer Dataset (WBC) original, Wisconsin Breast Cancer Diagnosis (WDBC) and Wisconsin Breast Cancer prognosis (WPBC). WBC contained 699 records with each record having 9

features plus the class attributes. WDBC contained 569 records with each record having 32 features plus the class attribute while WPBC contained 198 records with each record having 34 features plus the class attribute.

3.2.2 Data Collection

A high quality data was required for realizing best results, it was important therefore that its acquisition be highly reliant on the quality of the data collection process.

The study relied on the utilization of UCI online databases available publicly for research purposes. Data in these databases were collected from clinical environment, and have undergone proper organizational ethics approval processes.

3.2.3 Feature Selection

Feature selection was important in this research since it required pattern recognition, statistics, and data mining. The aim behind feature selection was to select a subset of record variables by ignoring features that possessed little or less importance. For example, a physician can make a decision based on some features i.e. whether a dangerous surgery is necessary for treatment or not. The study used feature selection methods to minimize the number of features in the dataset before the mining process started. This research solely relied UCI repository. Table 3.1 shows a sample of WBC dataset from UCI repository.

Table 3.1: 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

3.2.4 Data Preprocessing and Analysis

This study used a new approach for constructing missing feature values to satisfy the completeness element. This was done by using the weighted iterative *k*-nearest neighbour's algorithm. The missing feature values were as a result of inaccuracies in data that had incorrect feature values. This was probably brought about by data entry errors, faulty data collection tools, errors in data transmission, and users who had submitted incorrect feature values just to fill mandatory fields during surveying; this data resulted in inconsistencies since records conflicted with other records on the dataset**;** some data were also incomplete. This occurred either due to some feature values not being important during data entry or some features values were not always available.

3.2.5 Applying IG-ANFIS Approach

This stage had data that was ready for the mining process with no or little data pre-processing. The processed data was used to evaluate IGANFIS approach. IG algorithm was used and reduced the number of features through ranking; the output of IG was then fed as the input for ANFIS. ANFIS build an input-output mapping using both human knowledge and machine learning reasoning. The experimental results were investigated to ascertain the classification accuracy that underlined the capability of the proposed algorithm.

The study investigated the improvement of classification accuracy in the constructed dataset comparing it to the original dataset which contained missing features values. The maximum classification accuracy on k=1 will also investigated. Comparisons between various features selection techniques were made. This was done to cover the whole aspects of DM because it was a comprehensive approach branching into many areas.

The study compared benchmark cancer feature; WBC, WDBC and WPBC with three well-recognized machine learning algorithms KNN, NB and DT. The study made a comparison to ascertain if multiple feature selection methods can satisfy all datasets.

3.2.6 Evaluation

This phase involved data mining experts to test and assess the proposed model.

This study evaluated IG-ANFIS by comparing its results with the real data values (class features) i.e. the classification accuracy and error rate were calculated. The error rate (Err) of the classifier is the average number of misclassified samples divided by the total number of records in the dataset. Classification accuracy of the model was also calculated as one minus the error rate. The study evaluated the results by making a comparison between the results obtained by the proposed methods and previous methods in the literature. This was done by using the same dataset used in literature by other methods. This was to ensure that a competitive method has been obtained.

3.2.7 Analysis of Results

The research propos**es** the use of WEKA and MATLAB. WEKA (Waikato Environment for Knowledge Analysis) - written in JAVA language, is an open source machine learning software that provides the environment to calculate information gain and contains some data mining and machine learning methods for data pre-processing, classification, regression, clustering, association rules, and visualization. MATLAB is a fourth generation language and interactive environment for numerical computation, visualization, and programming. MATLAB is used to

analyze data, develop algorithms, and create models and applications. Therefore, its users come from various backgrounds of engineering, science, and economics.

CHAPTER FOUR
RESEARCH RESULTS AND DISCUSSION

4.1 Chapter Overview

Cancer is a debilitating disease and has over the years given the medical practitioners' sleepless nights trying to find effective, accurate and reliable ways of diagnosing it. To this end therefore, it has become difficult in providing prompt response to cancer patients when it newly emerges.

This study tried to identify significant diagnostic features that best described cancer data using IG-ANFIS technique. Missing feature values that improved prediction in determining the performance achieved by DM algorithms were also investigated. A hybrid DM model was developed from the existing techniques and finally tested to ascertain improvement in classification accuracy and missing values. This approach showed improvement in classification accuracy and missing feature values.

4.2 Data Presentation

The study used UCI machine learning repository. This repository was created by William Wolberget, 1991, from the University of Wisconsin-Madison, USA. The WBC database attributes were collected from digital fine needle aspirate (FNA) of breast mass. A summary of the WBC datasets from UCI that were used in this study are shown in table 4.1.

Table 4.1: Wisconsin Breast Cancer dataset (WBC)

Attribute	**Domain**
Clump Thickness	1-10
Uniformity of Cell Size	1-10
Uniformity of Cell Shape	1-10
Marginal Adhesion	1-10
Bare Nucleoli	1-10
Single Epithelial Cell Size	1-10
Bland Chromatin	1-10
Normal Nucleoli	1-10
Mitoses	1-10
Class	(2 for benign, 4 for malignant)

This work used WBC (Original), Wisconsin Diagnosis Breast Cancer (WDBC), and Wisconsin Prognosis Breast Cancer (WPBC) from UCI repository to find the best classifiers. Our intention here was to come up with the best combination of classifiers that best classifies cancer patient's data; these data sets were represented as in the table 4.2

Table 4.2: WBC (Original), WDBC, and WPBC datasets

Dataset	Number of features	Number of Instances	Number of Classes
Wisconsin Breast Cancer (Original)	11	699	2
Wisconsin Diagnosis Breast Cancer (WDBC)	32	569	2
Wisconsin Prognosis Breast Cancer (WPBC)	34	198	2

4.3 Discussions of Results

The selected features were applied to ANFIS to train and test the proposed approach. The structure of the proposed approach is shown in Figure 4.1, where X= $\{x_1, x_2,\ldots,x_n\}$ are the original features in dataset, Y=$\{y_1, y_2,\ldots,y_k\}$ are the features after applying the information gain (features selection), and Z denote the final output after applying Y on ANFIS The information gained is then fed to ANFIS.

$$X=\{x_1, x_2,\ldots,x_n\} \rightarrow H(c,) = H(c) - H(c|x) \rightarrow Y=\{y_1, y_2,\ldots,y_k\} \rightarrow \text{ANFIS} \rightarrow Z$$

Figure 4.1 structure of the proposed approach

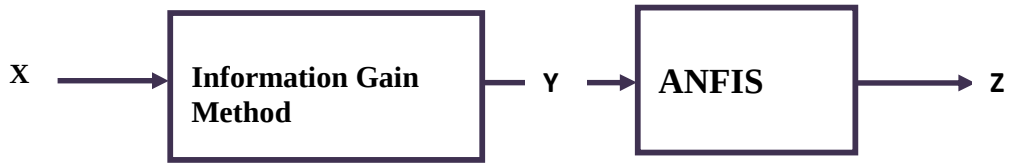

Figure 4.2: architecture for the general approach for IGANFIS.

This process involved a number of stages: The first stage was to select the most important features leading to more accurate results. The second stage was the construction of the FIS. This study used the most known fuzzy inference system, Sugeno-FIS (MATLAB Type fuzzy) method. Sugeno FIS was used to map feature to feature membership functions, feature membership functions to rules, rules to a set of output, output to output membership functions, and the output membership function to a single-valued output. This process was as shown in Figure 4.3.

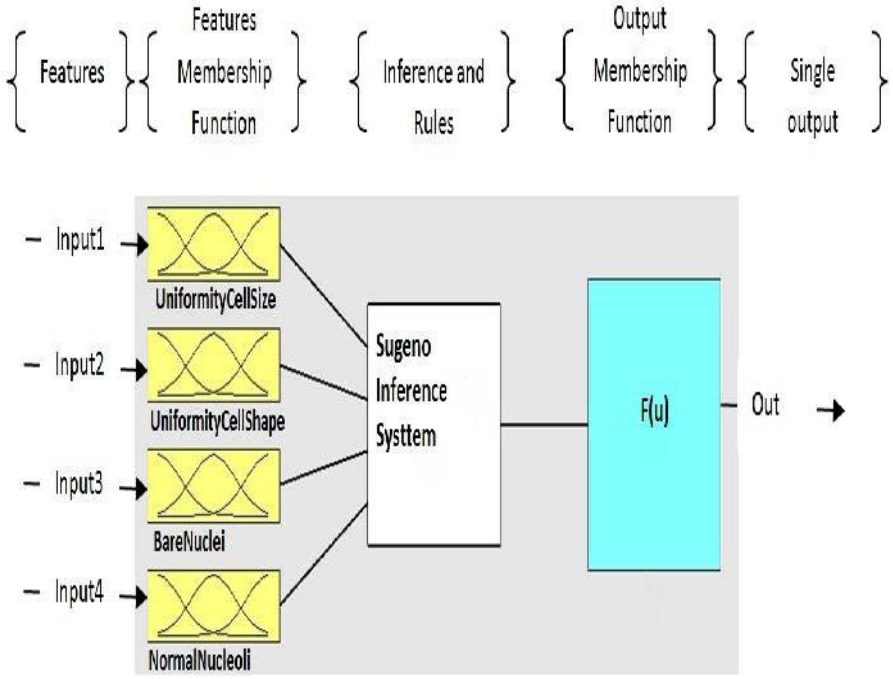

Figure 4.3: Sugeno FIS with four features input and single output

The Membership functions that were used for the sugeno FIS were Poor, Average and High. In addition to the membership functions, FIS contained rules that added human reasoning capabilities to machine intelligences. These rules were based on Boolean logic. In this approach, the rules were defined from the real data and they expressed the weight of each feature by giving higher priority for features that have the highest rank. The proposed approach contained 81 rules (Number of rules = $x^y$where x is the Number of member functions and y is the number of features i.e. (3^4=81 rules). The following were two examples of the rules that were used in the proposed approach:

> IF AND(UniformityCellSize is poor, UniformityCellShape is Avg, BareNuclei is poor, NormalNucleoli is poor)THEN (output is OK)

IF AND(UniformityCellSize is poor, UniformityCellShape is high, BareNuclei is poor, NormalNucleoli is avg)THEN (output is NOT_OK)

The ANFIS Graphical User Interface Editor shown in figure 4.4 was used to select appropriate functions for the purpose of editing or viewing data presented to ANFIS. This was ANFIS structure on MATLAB version 8.10 that was used to implement the approach.

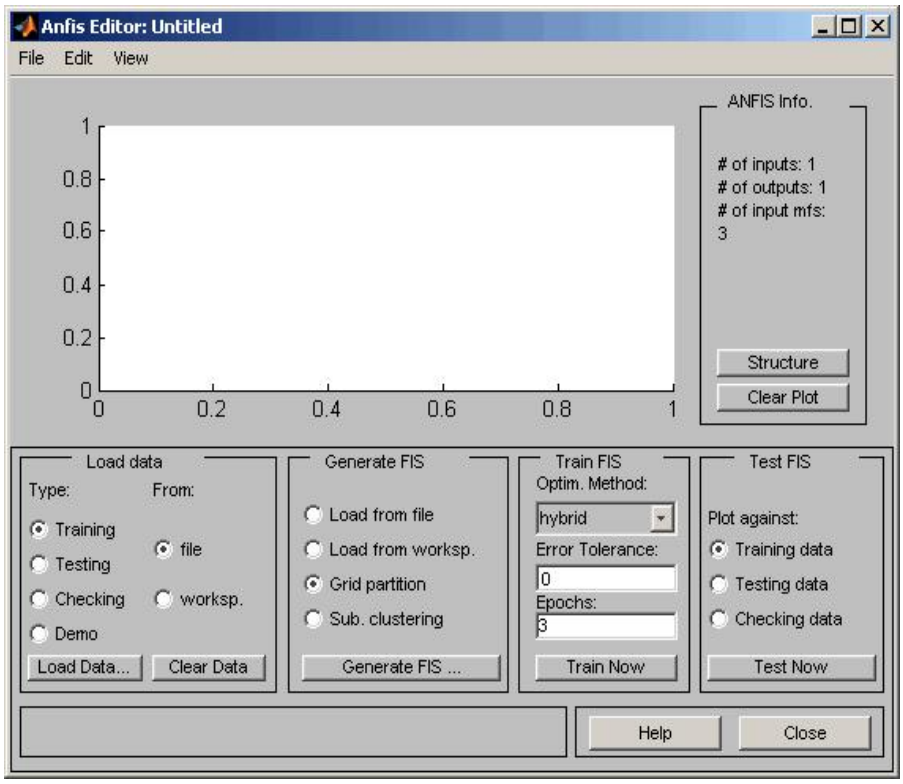

Figure 4.4: ANFIS Editor GUI

The built FIS architecture in Figure 4.4 demonstrated the visualization of the training results of the ANFIS model as in figure 4.5. The dataset were clustered into 4 member functions and the total fuzzy rules were 81.

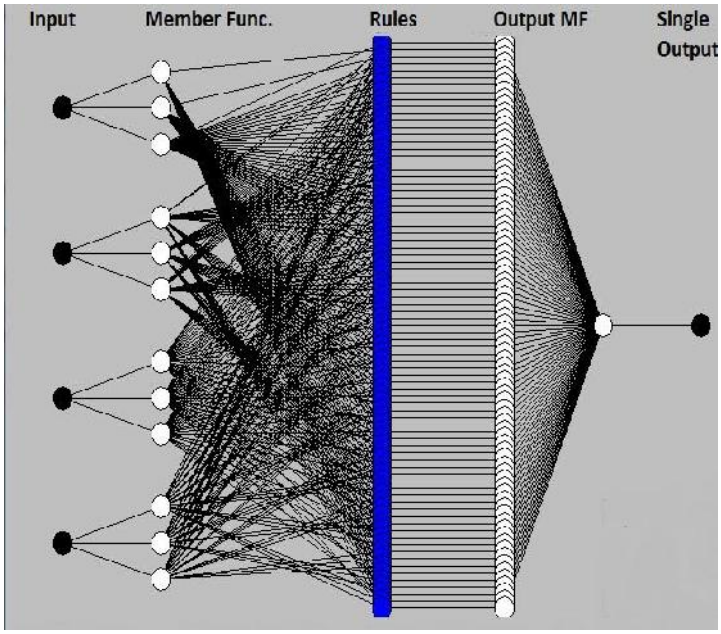

Figure 4.5: ANFIS Structure on MATLAB

4.3.1 IG-ANFIS EXPERIMENTAL RESULTS

The study divided the database into training and testing datasets. 341 records were used for training and 342 records for testing. Records which contained missing values (16 records) were ignored. The class features were normalized to [0=Benign, 1=Malignant]. The IG method was used to select the quality of features. Table 4.3 showed the ranking of features after InfoGainAttributeVal (the attribute evaluator) and the searching method Ranker-T-1 using WEKA on WBC dataset was applied.

Table 4.3: Information Gain Ranking Using WEKA on WBC

Attribute Name	**Rank**
Uniformity of Cell Size (UCSize)	0.636
Uniformity of Cell Shape (UCSshape)	0.633
Normal Nucleoli (NN)	0.555
Bare Nuclei(BN)	0.538
Single Epithelial Cell Size (SECS)	0.421
Clump Thickness (CT)	0.411
Marginal Adhesion (MA)	0.394
Bland Chromatin (BC)	0.316
Mitoses(MI)	0.278

In determining the number of features that were used in the experiment, a certain number of features based on features rank were selected i.e. a point where the rank dropped significantly. The feature ranking was as represented by the graph in figure 4.6. This showed a drop in the ranks and that was the most significant change in the graph (the slope point). The slope point then, gave us an indication to choose the first four top ranking features located above the slope point as the recommended number of features to be used later as inputs to ANFIS. The biggest drop was just after the feature number 4 (BN) as shown in the graph. Respectfully, features; Uniformity of Cell Size (UCSize), Uniformity of Cell Shape (UCShape), Normal Nucleoli (NN), and Bare Nuclei (BN) were selected to train and test the model. At this stage, the features were deducted and the recommended number of features set to 4.

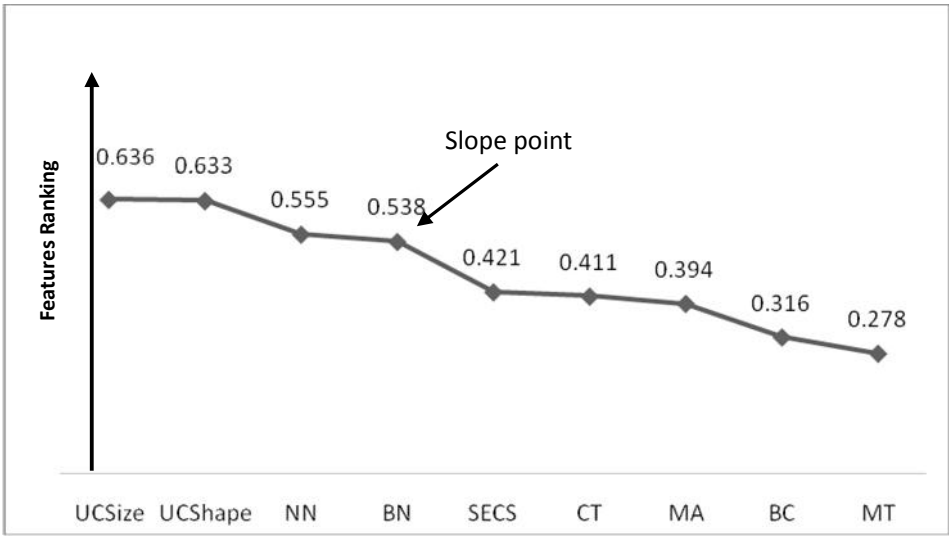

Figure 4.6: Information Gain Ranking on WBC

In the third and final stage, the constructed FIS and the new features set were loaded to ANFIS for training and testing. Figure 47 represents the structure of IG-ANFIS approach that was used.

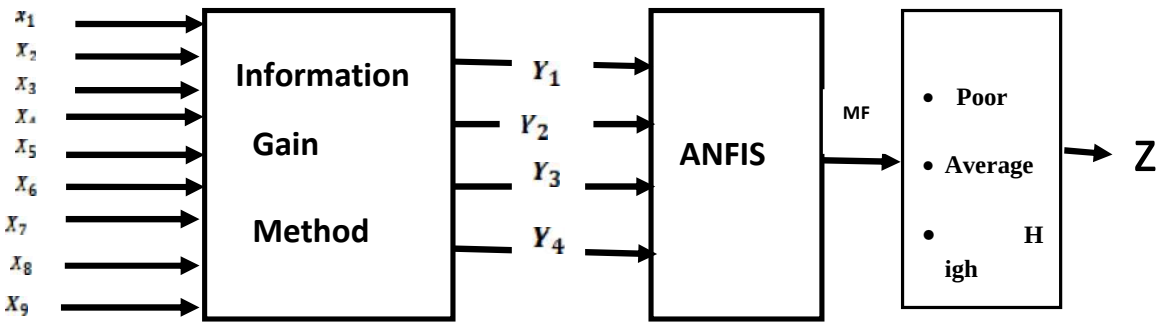

Figure 4.7: The structure for IG-ANFIS approach

The visual implementation for the feature Cellsize was shown as in the figure 4.8.

This rule contained three member functions: Poor, Average, and High.

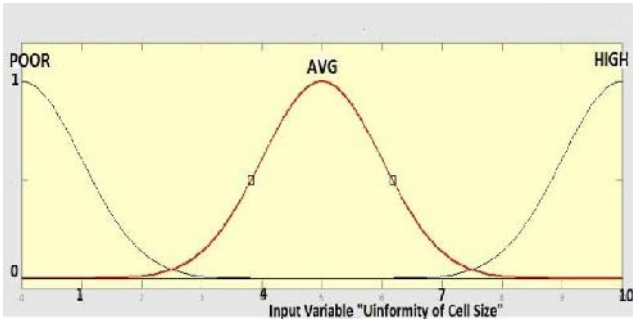

Figure 4.8: Input Membership Function for the feature "Uniformity of Cell Size"

The results obtained from IGANFIS were then compared with some previous work to ascertain classification accuracy of our method. These results were tabulated as in the table 4.4 .

Table 4.4: Comparison of classification accuracy between IG-ANFIS and previous work

The approach	**Accuracy**
AdaBoost	57.60%
ANFIS	59.90%
SANFIS	96.07%
FUZZY	96.71%
FUZZY- GENETIC	97.07%
ILFN	97.23%
NNs	97.95%
ILFN and FUZZY	98.13%
IG-ANFIS	**98.24%**
SIANN	100.00%

The results of Table 4.4 were then represented as shown in figure 4.9.

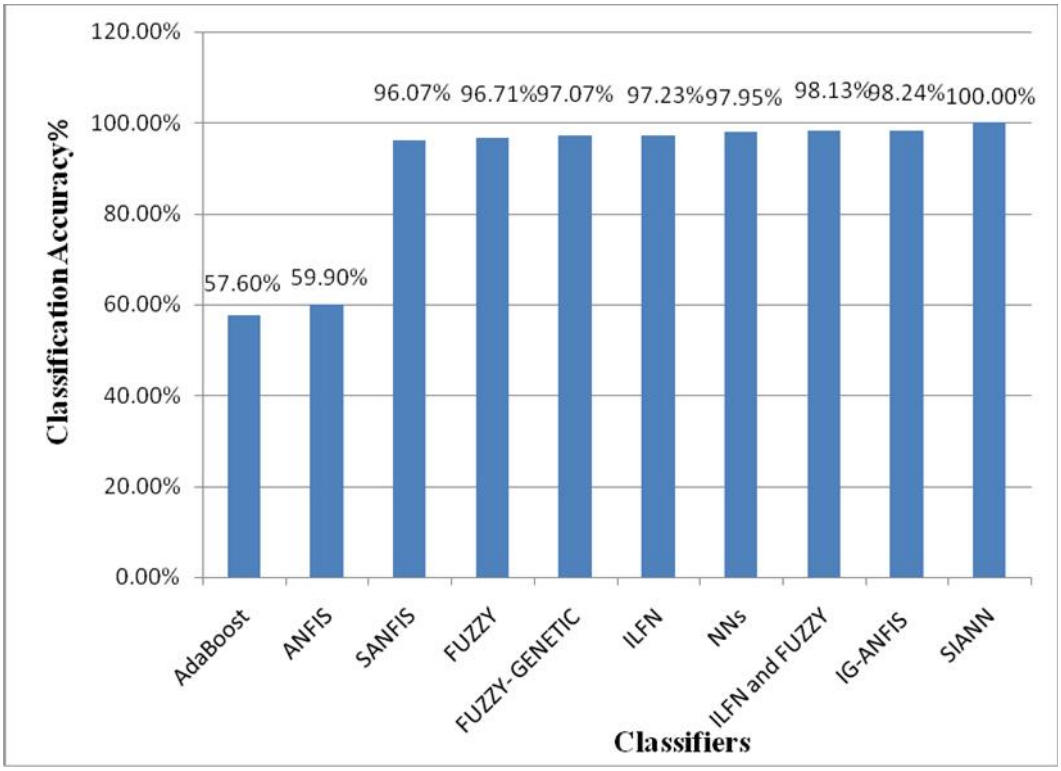

Figure 4.9: Comparison of classification accuracy between IG-ANFIS and some previous work

4.4 Filling Missing Feature Values

This study integrated weighted K-NN algorithm to find the closest neighbours ($n_1 \ldots n_k$). Euclidean and Minkowski distance functions were used. This approach found the most similar instance to (x_i) from ($n1\ldots$) where (x_i), is an instance that contains missing feature values using the formula:

$$P(c_j|x_i) = \frac{P(cj).P((x_i|c_j)}{P(x_i)} \quad \text{..... (4.22)}$$

By finding the distances values (cn_i) the formula below will be applied;

$$cn_i = \frac{\sum_{j=1}^{k} {}^{n_{ij}}/_{d(x_{j,n_j})}}{\sum_{j=1}^{k} {}^{1}/_{d(x_{j,n_j})}} \quad \text{..... (4.23)}$$

Where cn_i denote the closest neighbours to the instance x_i, (x_j,n_j) is the distance between the instance x_j and the neighbour n_j, and n_{ij} denote the feature i of the neighbour n_j. After finding

the closest neighbour (the smallest value of cn_i call it cn', the missing feature values in $x_i$were filled by the equivalent features values in n_i having cn' distance to x_i. The process of filling missing features values produced a new training dataset (*NT*) that contained no missing features values.

To verify the accurateness of the constructed missing features values, the new training dataset was applied to *k*-NN and the accuracy recorded. If classification accuracy was less than a threshold then the algorithm stepped back to fill the missing features values until the desired classification accuracy was reached. Figure 4.10 represented the flowchart for constructing missing feature values used by the study.

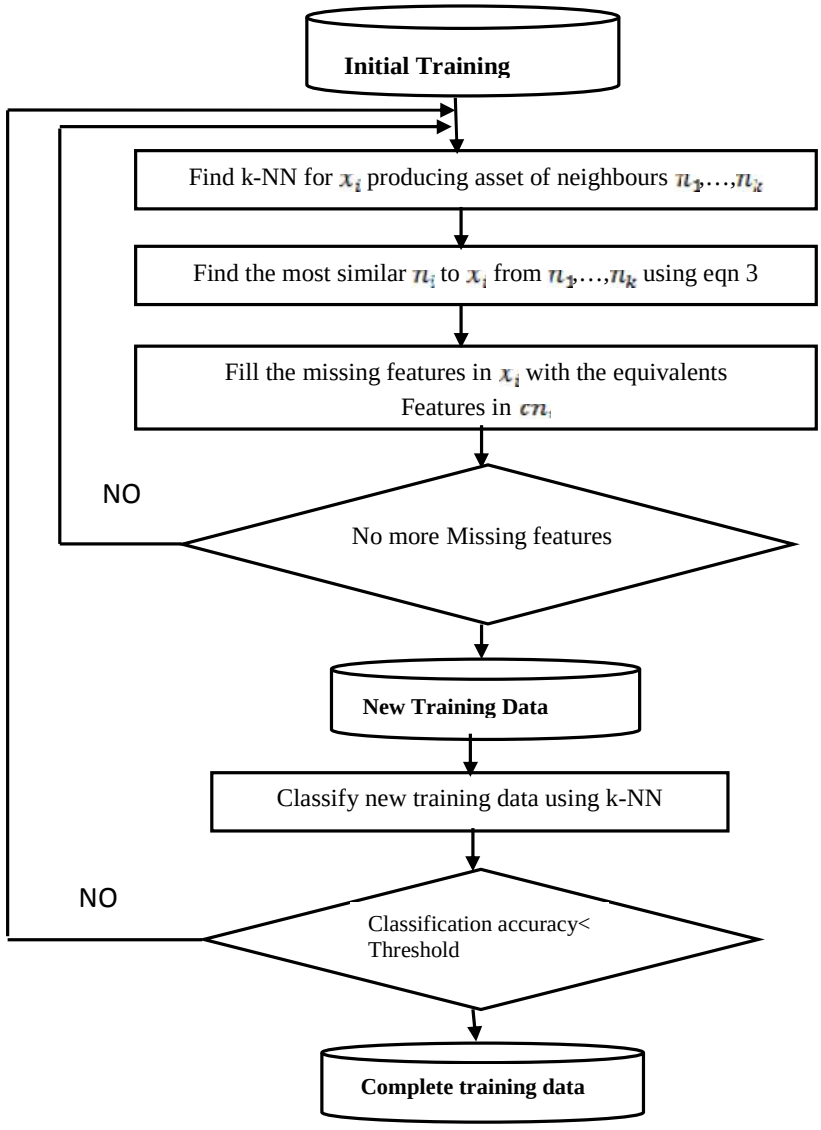

Figure 4.10: The Flowchart for Constructing Missing Features

4.5.1. The Experimental Results for Missing Feature

The dataset (WBC) was randomly divided into two parts; training dataset and testing dataset to avoid unfairness the dataset separation was random. The training dataset contained 500 cases where 16 of them contained missing features values.

Using Euclidean and Minkowski distance functions metrics were computed. Constructing the missing features values using the proposed method through iterative *k*-NN classifier with the Euclidean distance function will showed classification accuracy enhanced. Varying k between k=1 to k=3 the iteration showed maximum classification accuracy when k=3.

Linear graph tabulation was used to compare the classification accuracy when the missing feature values were not treated and when treated. This tabulation also presented various classification accuracies dependent on the number of neighbours, (*k*) in *k*-NN.

constructing the missing feature values using *k*-NN classifier with the Minkowski distance function showed classification accuracy enhanced by 0.005 when *k*=3 and r=1.5 from the first iteration and a maximum classification accuracy of 0.9698. Figure 4.11 was a comparison of classification accuracy when the missing features values were not treated and when treated using Minkowski/*k*-NN distance metrics.

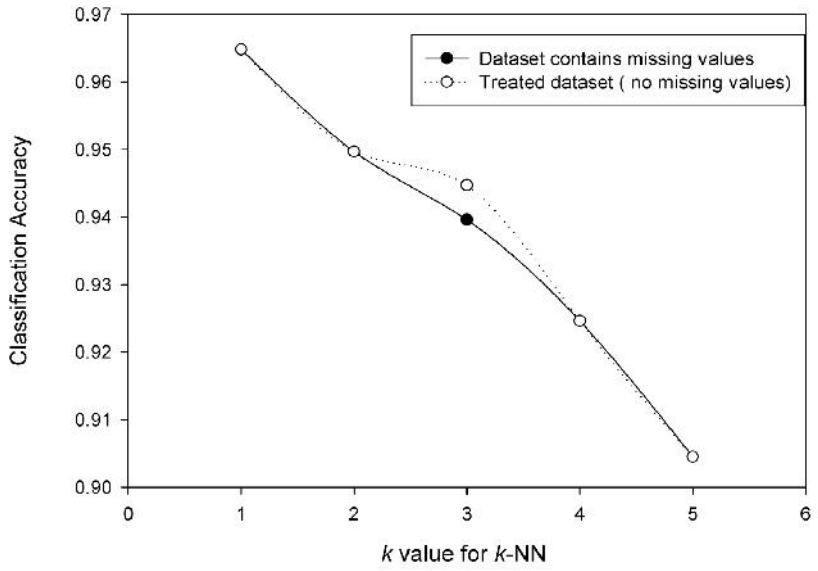

Figure 4.11: A comparison of classification accuracy for our method through Euclidean/*k*-NN

Enhancing the experiment showed that Manhattan, Chebychev, and Canberra distance metrics were not suitable for constructing the missing feature values, this was because the classification accuracy after treating the missing values remained lower than the classification accuracy for the original dataset.

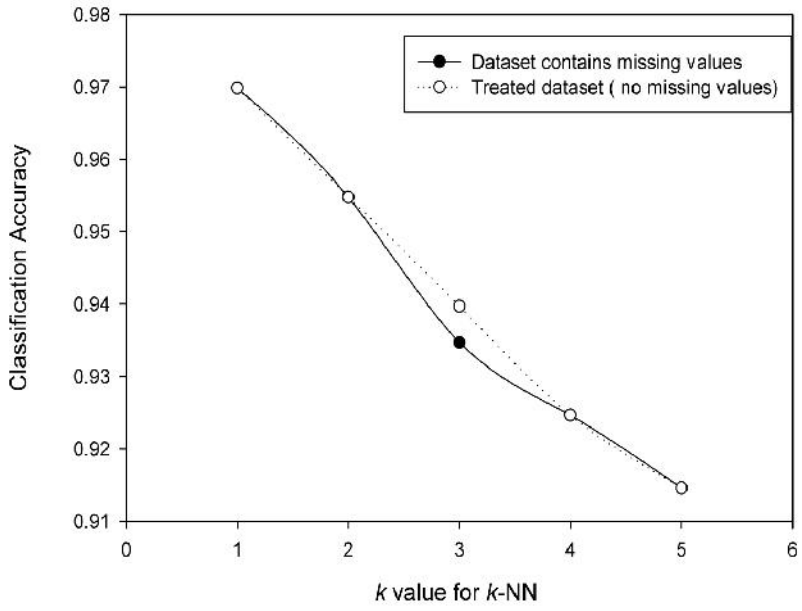

Figure 4.12: A comparison of classification accuracy for our method through Minkowski/*k*-NN

4.5 Feature Selection

In evaluating the benchmark features selection methods, experiments were carried out on WBC to come up with a deductive judgment on a satisfactory feature selection method to be applied in our study.

This work considered three machine learning algorithms from three categories of learning methods. The purpose for this was to arrive at a fair deduction between the features selection methods used.

K-NN algorithm, an instance-based classifier from lazy learning category was the first to be used. Here, the class of a test instance is based upon the class of those training instances alike to it. Distance functions are common to find the similarity between instances.

WEKA Experimenter made it easy to compare the performance of different learning schemes where the Results were written into file. The Evaluation options in the experimenter cross-validation, learning curve, etc. iterated over different parameter

settings with in-built Significance-testing. The figure 4.13 illustrates a window of the WEKA experimenter environment that was used.

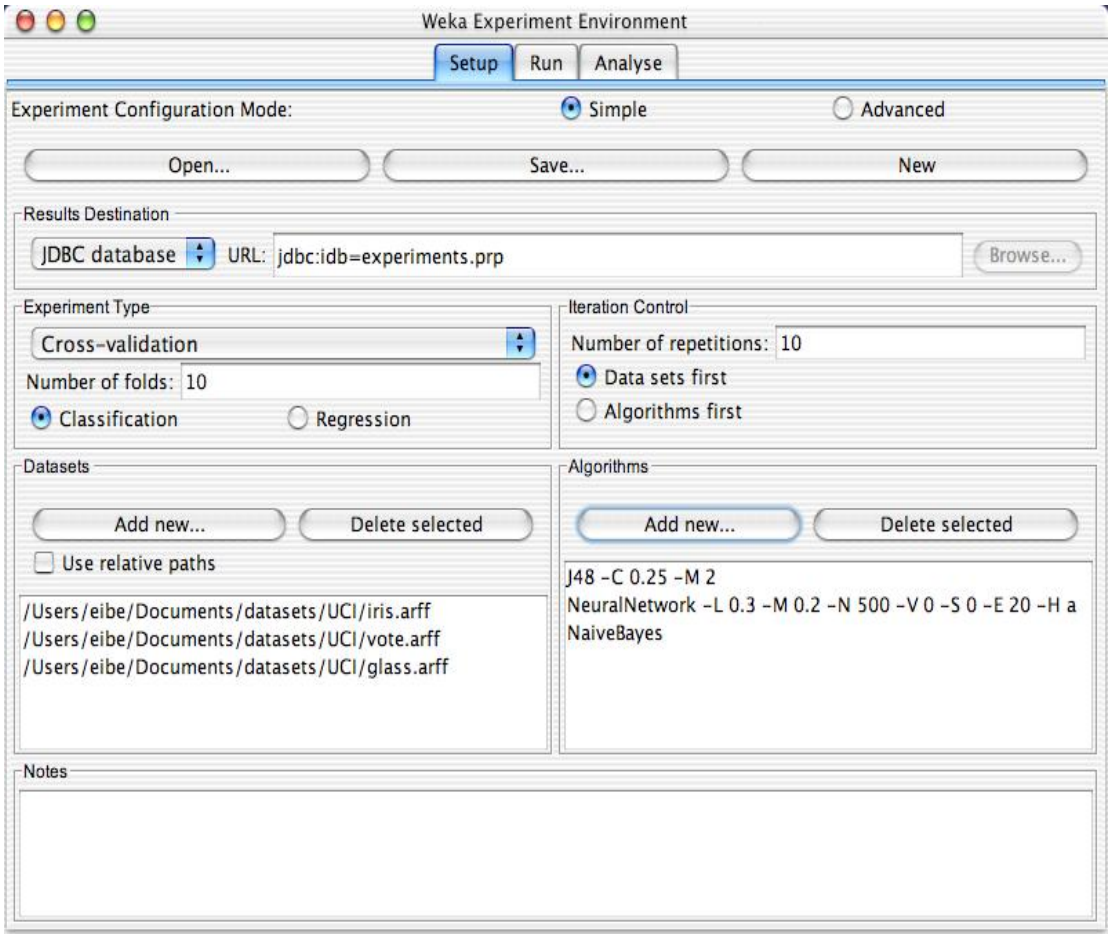

Figure 4.13: WEKA experimenter environment

Naïve Bayes classifier (NB) from Bayes category was the second algorithm used. Random Tree (RT) or decision tree was the third and last machine learning algorithm used. RT was used to classify an instance to a predefined set of classes based on their attributes values.

After applying features selections techniques and the learning algorithms on the dataset and obtaining classification accuracy results, a hybrid method was constructed. This combined the advantages of the best performing feature selection technique and the advantages of best performing learning algorithm. The process of carrying out this was as represented in the Figure 4.14.

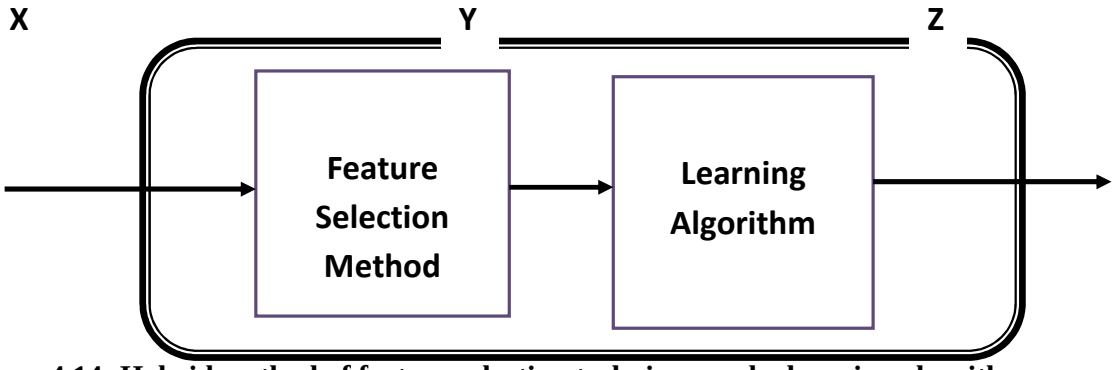

Figure 4.14: Hybrid method of feature selection technique and a learning algorithm

4.5.1 Feature Selection Experimental Results

The notations "+", "-", and "=" were used to show the feature selection methods classification performance in comparison with the original dataset (before performing feature selection methods); where "+" denoted improvement, "-" denoted degradation, and "="denoted unchanged. The table 4.5 showed the experimental results of using Naïve Bayes (NB) as a machine learning algorithm on WBC dataset.

Table 4.5: WBC dataset on Naïve Bayes learning method and

Some features Selections techniques

Feature selection Technique	**WBC**
NB Original Dataset	95.99%
Correlation feature selection (CFS)	95.99% =
Principal Components Analysis (PCA)	96.14% +
Symmetric Uncertainty (SU)	95.99% =
Consistency Subset Evaluation (CSE)	96.28% +
Relief (R)	95.99% =
Information Gain (IG)	95.99% =

The features selection methods performances with Naïve Bayes in table 4.5 were illustrated figure 4.15.

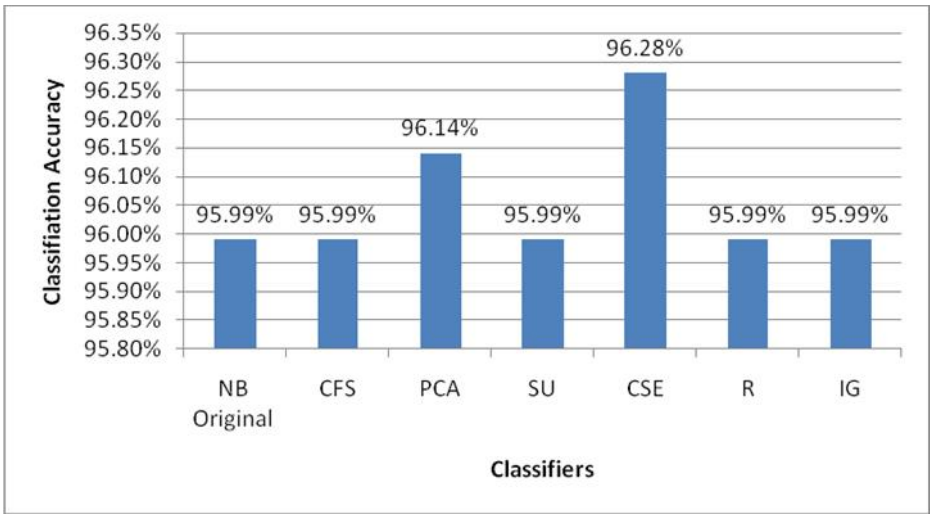

Figure 4.15: Feature selection methods performance with Naïve Bayes

Naïve Bayes on original WBC dataset showed a classification accuracy of 95.99%, however when using Principle Components Analysis (PCA) and Consistency based Subset Evaluation (CSE) features selection methods, the classification accuracy of 96.28% and 96.14% respectively was realized. With the application of correlation based feature selection (CFS), Information gain (IG), Relief(R), and Symmetrical Uncertainty (SU), the classification accuracy did not change.

Using *k*-NN as our second machine learning algorithm on WBC, the experimental results were tabulated as in table 4.6.

Table 4.6: WBC dataset on K-NN learning method and some features Selection techniques

Feature selection Technique	**WBC**
K-NN Original Dataset	95.42%
Correlation feature selection (CFS)	95.42% =
Principal Components Analysis (PCA)	96.42% +
Symmetric Uncertainty (SU)	95.42% =
Consistency Subset Evaluation (CSE)	96.85% +
Relief (R)	95.42% =
Information Gain (IG)	95.42% =

The Table showed that the classification accuracy of using *k*-NN on the original WBC is 95.42%, however when applying Consistency based Subset Evaluation (CSE) feature selection method, the results improved. Irrespective of this, other features selections methods produced the same classification accuracy as the original dataset. Figure 4.16 illustrates the results on Table4.6.

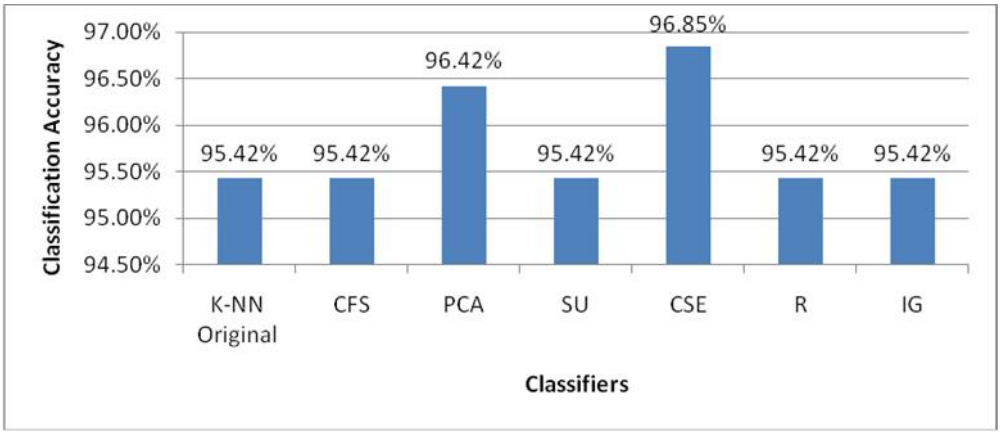

Figure 4.16: Results for feature selection methods with *k*-NN

The final machine learning classifier in our experiment was the Decision Tree (DT). The rules generated from the DT were described in Figure 4.17., Where UCSIZE, UCSHAPE, CT, BN and MA were the best features.

1. If (UCSIZE < =2,BN<=3) Then Diagnosis = **Benign**
2. If (UCSIZE <= 2,BN>3,CT<=3) Then Diagnosis = **Benign**
3. If (UCSIZE <= 2.5,BN>3,CT>3,BC<=2,MA<=3) Then Diagnosis = **Malignant**
4. If (UCSIZE <= 2.5,BN>3,CT>3,BC<=2,MA>3) Then Diagnosis = **Benign**
5. If (UCSIZE <= 2.5,BN>3,CT>3,BC>2) Then Diagnosis = **Malignant**
6. If (UCSIZE >2,UCSHAPE<=3,CT<=5) Then Diagnosis = **Benign**
7. If (UCSIZE >2,UCSHAPE<=3,CT>5) Then Dia = **Malignant**
8. If (UCSIZE >2,UCSHAPE>2,UCSIZE<=4,BN<=2,MA<=3) Then Diagnosis = **Benign**
9. If (UCSIZE >2,UCSHAPE>2,UCSIZE<=4,BN<=2,MA>3) Then Diagnosis = **Malignant**

Th

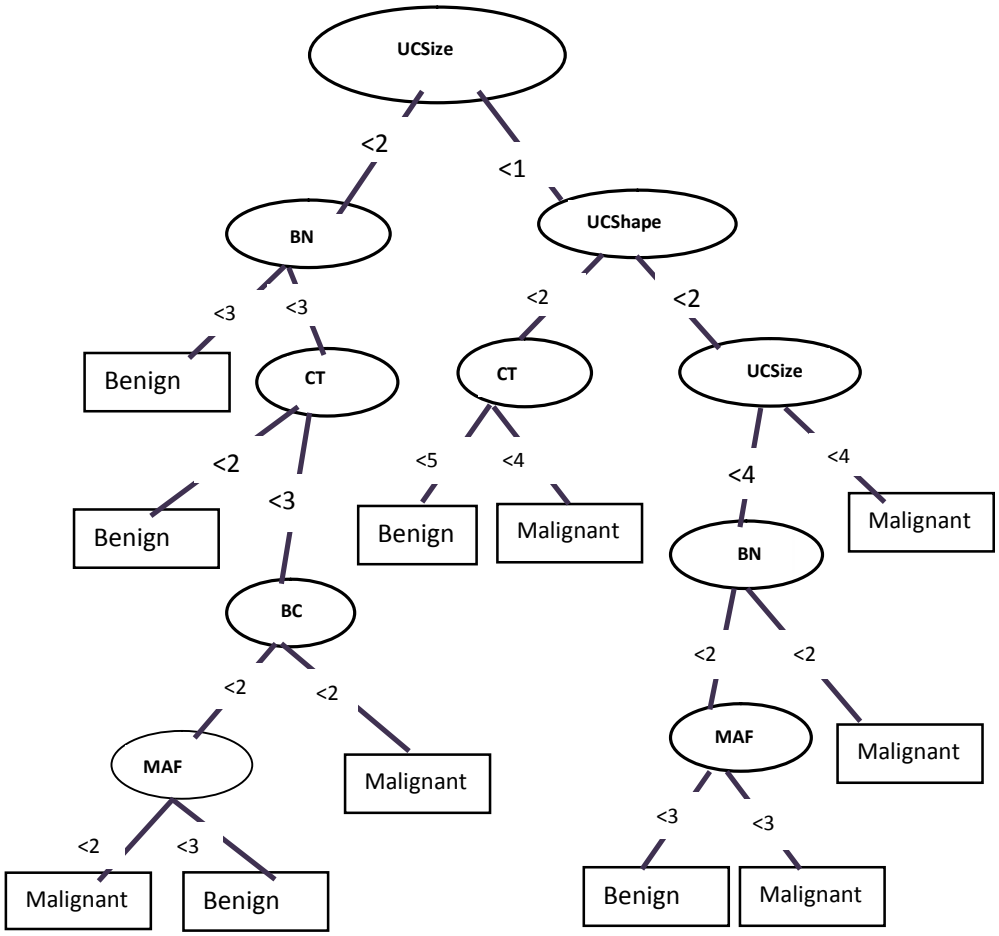

Figure 4.18: Decision Tree using Random tree algorithm

The experimental results of using DT machine learning algorithm on WBC were shown in Table 4.7.

Table 4.7: Results for Attributes Selection Methods with Decision Tree

Feature selection Technique	WBC
DT Original Dataset	94.56%
Correlation feature selection (CFS)	94.56% =
Principal Components Analysis (PCA)	94.85% +
Symmetric Uncertainty (SU)	94.56% =
Consistency Subset Evaluation (CSE)	93.56% -
Relief (R)	94.56% =
Information Gain (IG)	94.56% =

The results showed an improvement in classification accuracy by applying PCA feature selection technique. However a decline was noticed in classification accuracy by using CSE. However classification accuracy did not change when CFS, IG, R, and SU Were used. This results were illustrates in Figure 4.19

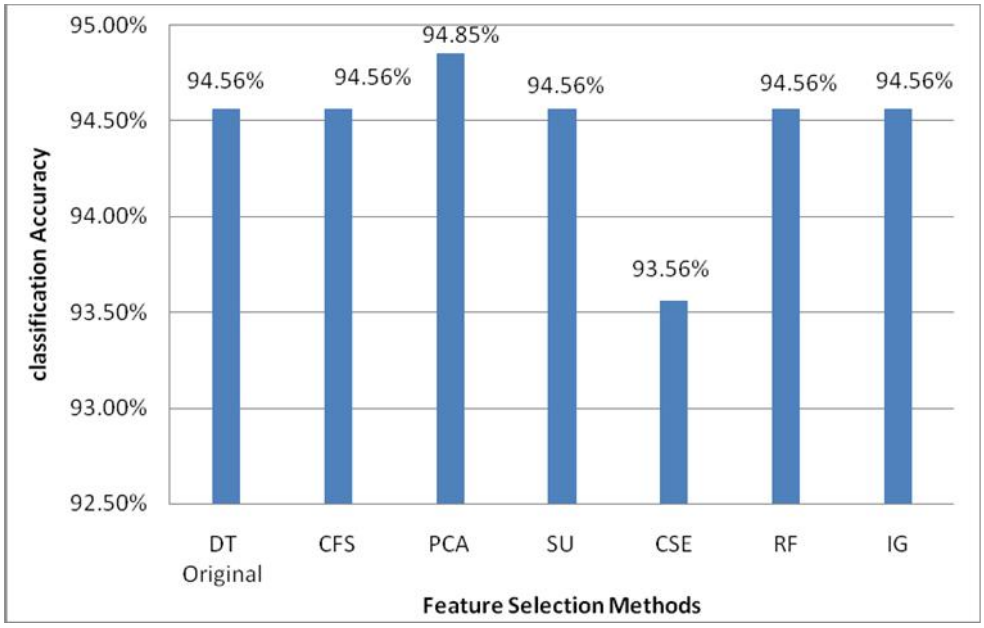

Figure 4.19: Results for Features selection methods with Decision Tree

4.6 Classifier Selection

This study employed a Multi-classification approach where combinations of two or more classifiers were done. The classification approach was divided into two parts namely; classifier selection and classifier fusion.

This was to evaluate two or more classifiers on the training dataset and then make use of the best performing classifiers on the testing dataset. To do this, WBC (Original), WDBC and WPBC were used. This study considered the use of k-NN, NB and RT classifiers from three machine learning categories. The study used k-fold cross validation technique to separate the training set from test set with k=10. WEKA was used as the experimental environment. Table 4.9 shows WBC (Original), WDBC, and WPBC datasets.

Table 4.8: WBC (Original), WDBC, and WPBC datasets.

Dataset	**Number of Attributes**	**Number of Instances**	**Number of Classes**
Wisconsin Breast Cancer (Original)	11	699	2
Wisconsin Diagnosis Breast Cancer (WDBC)	32	569	2
Wisconsin Prognosis Breast Cancer (WPBC)	34	198	2

4.6.1 Classifier Selection Experimental Results

Using the three different datasets of breast cancer, three experiments were performed. First, it was done on a single classifier model. The purpose for this was to set a base line of classification accuracy and how enhancements were to be made. Secondly it was done using a combination of two classifiers while the last experiment was done after the fusion of the three classifiers. Table 4.9 below shows the results of a Single Classifier on three datasets WBC, WDBC, and WPBC.

Table 4.9: Single Classifier on three datasets WBC, WDBC, and WPBC

Classifier	**Classification Accuracy**
NB-WBC	0.9599
NB-WDBC	0.9297
NB-WPBC	0.6667
KNN-WBC	0.9542
KNN-WDBC	0.9473
KNN-WPBC	0.65515
RT-WBC	0.9456

RT-WDBC	0.9244
RT-WPBC	0.6768

The results indicated that NB performed the best in classification accuracy on WBC (0.9599). K-NN and RT had better results on WDBC and WPBC respectively.

The Single Classifier on three datasets WBC, WDBC, and WPBC were then tabulated as in Figure 4.20.

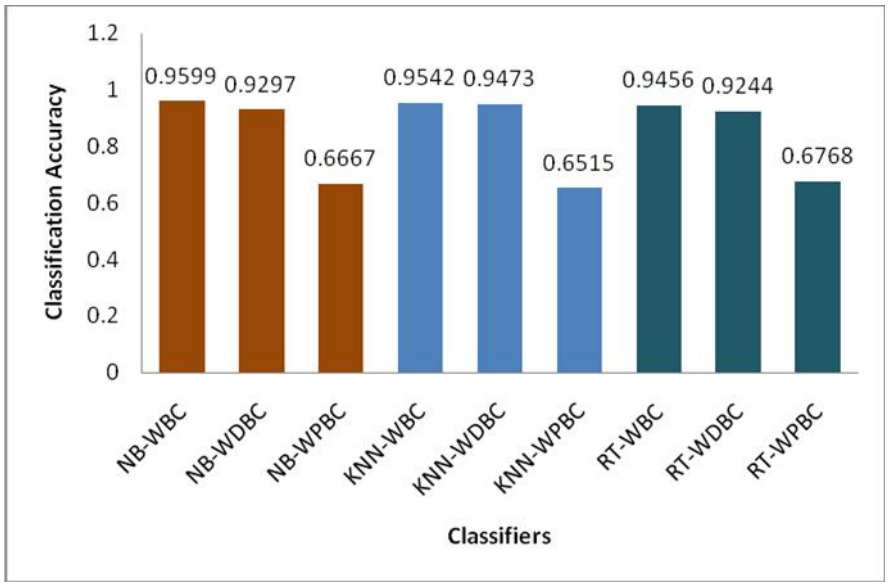

Figure 4.20: Single Classifier on three datasets WBC, WDBC, and WPBC.

On combining two classifiers, (Naïve Bayes and *k*-NN, Naïve Bayes and Random Tree, and *k*-NN and Random Tree), the results were recorded in the table 4.11.

Table 4.11: Two Classifiers on three datasets WBC, WDBC, and WPBC

Classifier	**Classification Accuracy**
NB,KNN-WBC	0.9642
NB,RT-WBC	0.9456
KNN,RT-WBC	0.9485
NB,KNN-WDBC	0.9508
NB,RT-WDBC	0.9279
KNN,RT-WDBC	0.9244
NB,KNN-WPBC	0.6869
NB,RT-WPBC	0.6768
KNN,RT-WPBC	0.6768

The results indicated that fusing Naïve Bayes and *k*-NN produced the best classification accuracy of 0.9642 On WBC, 0.9508 on WDBC, and 0.6869 on WPBC). This ideally draws a conclusion that Naïve Bayes and *k*-NN may produce better results when they combined together. A bar graph in figure 4.21 below was then drawn using the results of table 4.9 to show comparisons of classification accuracies' of various combinations of classifiers.

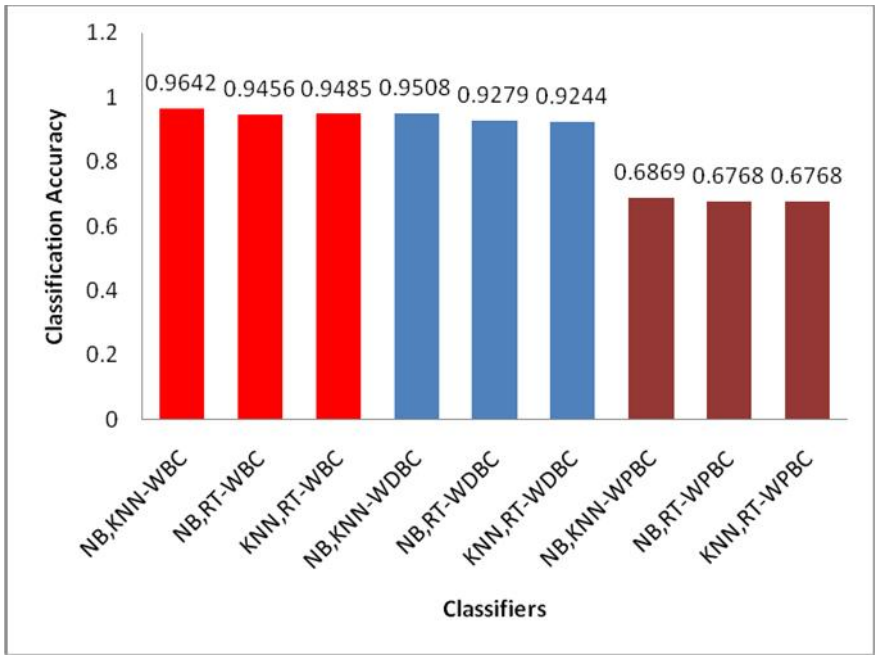

Figure 4.21: Two Classifiers on three datasets WBC, WDBC, and WPBC.

The fusion of three classifiers, (Naïve Bayes, *k*-NN, and Random Tree) showed that in combining the three classifiers for all the three datasets had high classification accuracy of (0.9585) on WBC and (0.9473) on WDBC. However there was a noticeable improvement of (0.7323) in classification accuracy on WPBC dataset. The tabulated results from the fusion of three classifiers; WBC, WDBC, and WPBC were shown in table 4.11.

Table4.11: Results of the fusion of three classifiers on three datasets; WBC, WDBC, and WPBC

Fused classifiers	**Classification accuracy**
NB,KNN,RT-WBC	0.9585

NB,KNN,RT-WDBC	0.9473
NB,KNN,RT-WPBC	0.7323

From these experiments, the study deduce that Naïve Bayes and *k*-NN produced better results when combined as one classifier with maximum classification accuracy obtained on WBC dataset (0.9642). The results of were represented in the figure 4.22.

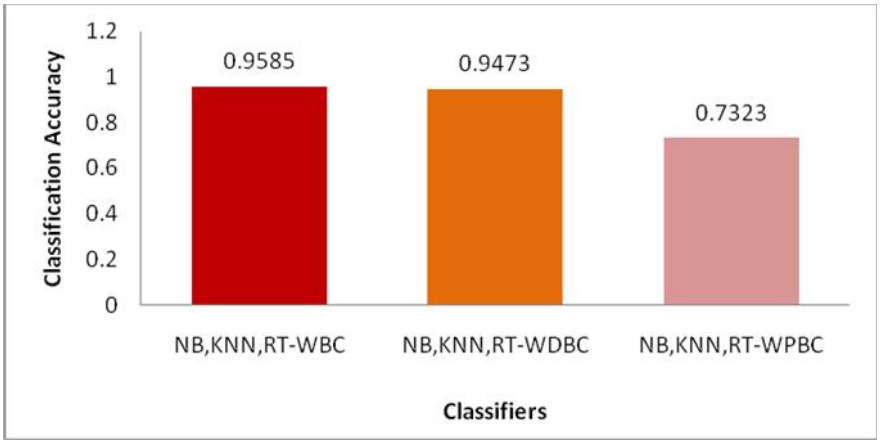

Figure 4.22: The Fusion of three classifiers on three datasets WBC, WDBC, and WPBC.

Summary

In this chapter, a new approach IG-ANFIS for diagnosing breast cancer was put into test. IG was used to minimize the number of features. The reduced numbers of features were then applied as the new dataset to ANFIS. Further, *K-NN* and the distance functions were computed. The process iterated until it found the most suitable feature values that satisfied classification accuracy. The resulted computed indicated that ideally, No single features selection method best satisfies all datasets and learning algorithms.

CHAPTER FIVE
CONCLUSIONS, RECOMMENDATIONS AND FUTURE WORK

5.1 Conclusions

The research used IGANFIS data mining technique on UCI cancer data sets to provide the diagnosis results. The results from the approach were so promising. If further attempts are engaged in the application of Information Technology in diagnosing various diseases such as cancer; then efficient, timely and decent healthcare services will be realized.

Large databases that used in the medical sector still have a concern of Missing features values brought about by many factors as discussed early. IGANFIS approach had considerably good results i.e.

1. An improvement of classification accuracy of 0.005 on the constructed dataset was realized with the proposed approach than the original dataset on both Euclidean and Minkowski distance functions.
2. Further study showed lower classification accuracy on the new dataset than the original dataset when using Manhattan, Chebychev, and Canberra distance functions. Classification accuracy according in this study depended greatly upon the number of neighbours (k). To be specific; the maximum classification accuracy was on k=1 which was 0.9698 i.e. the less the number of neighbours, the more the classification accuracy and vice versa.
3. Our study showed that in an overall view, CES features produced better results compared to IG, SU, R, CFS and PCA. On WBC, NB was at the top in classification accuracy. However *k*-NN and DT performed just better on the dataset after applying feature selection methods in comparison with the original dataset having no feature selection techniques.
4. A hybrid approach showed that NB learning algorithm and CSE had higher classification accuracy (0.9628) in comparison to other classifies used in this study. This showed its capability on WBC Dataset.
5. Classifier fusion was introduced in this study on the three well-known machine learning classifiers on WBC to enhance their ability by combining their advantages in a single algorithm. However, fusions Classification approach depended on the classifiers

characteristics that were involved. In this study, NB and KNN performed well when combined as a single classifier with maximum classification accuracy on WBC dataset of 0.9642.

Performing different experiments using various machine learning algorithms on WBC Dataset, the study concluded that hybridization of the existing machine learning algorithms can produce better approaches for medical diagnosing.

5.3 Recommendations

Future work should focus on the cost of computation. This is because if a computational approach is cheaper and can have the ability to produce the best results, then the better the approach. The future study should focus on broadening disease options since Clinical practice is a complex endeavor.

REFERENCES

Arulampalam, G., & Bouzerdoum, A. (2001, November). Application of shunting inhibitory artificial neural networks to medical diagnosis. In *Intelligent Information Systems Conference, The Seventh Australian and New Zealand 2001* (pp. 89-94). IEEE.

Bech, A. G. (2012). *Breast Cancer in Australia: An Overview* (No. 71). AIHW.

Moxon, B. (1996). Defining Data Mining: The Hows and Whys of Data Mining, and How it Differs from Other Analytical Techniques. DBMS Data Warehouse Supplement, Miller Freeman. *Inc., San Francisco, CA*.

Daniel T. L. (2013). *Discovering Knowledge in Data*, Uniqueness of medical data mining. *Artif. Intell. Med,* 26(1-2), 1-24.

Duda, R.O., Hart, P.E. And.Stork, D.G.(2003). Pattern Classification An Introduction To Variable And Feature Selection. *J. Mach. Learn. Res,*3, 1157-1182.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery In Databases: An Overview. *AI Magazine*, *13*(3), 57.

Baxt, W. G. (1990). Use Of An Artificial Neural Network For Data Analysis In Clinical Decision-Making: The Diagnosis Of Acute Coronary Occlusion. *Neural Computation*, *2*(4), 480-489.

Grzymala-Busse, J. W., & Grzymala-Busse, W. J. (2010). Handling missing attribute values. In *Data mining and knowledge discovery handbook* (pp. 33-51). Springer US.

Gunter, D.T. and Terry,P.N. (2005). The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. *Juornal of Medical Internet Research, 7(1)*

Hall, M., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on*, *15*(6), 1437-1447.

Han, J. (2012). *Data Mining Concepts and Technique.* Vol. 3. San Franscisco: Morgan Kaufmann.

Hertz, J. Arzucan, O, (2007). *Supervised and Unsupervised Machine Learning Technique for Text Document Categorization*

Ozgür, A. (2004). *Supervised and unsupervised machine learning techniques for text document categorization* (Doctoral dissertation, Bogaziçi University).

Howell, D. C. (2007). The treatment of missing data. *The Sage handbook of social science methodology*, 208-224.

Jolliffe, I.T. (2002). *Principal Component Analysis*. London: Springer: NY.

Kotsiantis, S., (2007).Supervised Machine Learning: *a Review of Classification Techniques. Informatica,* 31249-268.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1), 273-324.

Larose, D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining.* New Jersey: John Wiley & Sons, Inc.

Liao, B.A. (2012). Novel Hybrid Method for Gene Selection of Microarray Data. *Journal of Computational and Theoretical Nanoscience*, 9(1), 5-9.

Lazarou, J., Pomeranz,B. and. Corey,*P. (2008).* Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *Journal of the American Medical Association, 279(15).*

Lloyd , W. (2013).Empirical *studies of the knowledge discovery approach to health information analysis. Informatica*, 31, 249-253.

Marlin, B.,(2008). Missing Data Problems in Machine Learning, in Department of Computer Science. Canada: University of Toronto.

Meesad, P. and Yen, G. (2003). Combined numerical and linguistic knowledge representation and its application to medical diagnosis. *Component and Systems Diagnostics, Prognostics, and Health Management* II. 4733,. 98-109.

Mitchell T. (2005). *Machine Learning:* London:McGraw Hill.

McCulloc and Pitts : *Introduction to the theory of neural computation*. Retrieved from http://www.learnartificialneuralnetworks.com/.

Moss, S.(2009). Expectation maximization--to manage missing data.

Odeh, S. (2008). *A computer aided diagnosis systems: using genetic algorithm with classifier of the k-nearest neighbors, The International Arab Conference on Information Technology, Tunis.*

Organization for Economic Cooperation and Development (OECD).(2009). Health Data.

Organization for Economic Cooperation and Development (OECD), (2010). Health Data.

Rokach, L.(2007). Data mining with decision trees: *theory and applications. Vol. 69.* Washington: *World scientific Publishing*

Ross, T. J.,(2012). *Fuzzy Logic with Engineering Applications,* in Graduate Program in Computer Engineering. Bogazici: Bogazici University.

Rubin, D.B.,(1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Saeys, Y., Inza,I. and P. Larrañaga, A. (2007). Review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.

Setiono, R., (2006). Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis, Artificial *Intelligence in Medicine*, 18(3), 205-219.

Song, H., seun, L.,, Dongwon, k.and Gwitae, P. (2010). New methodology of computer aided diagnostic system on breast cancer, *in Proceedings of the Second international conference on Advances in Neural Networks* - Volume Part III. Chongqing: Springer-Verlag. 780-789.

Singapore cancer registry factsheet, (2012). *Most Frequent Cancers in Men and Women,* Retrieved from http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900.

Vijayasankari, S. and Ramar,K. (2012). Enhancing Classifier Performance Via Hybrid Feature Selection and Numeric Class Handling- A Comparative Study. *International Journal of Computer Applications*, 41(17), 30-36.

Widrow, B. and Hoff,M. (19899). Adaptive Switching Circuits, *in WESCON Conference Record*. 709-717.

WHO, (2010). *Assesses the World's Health Systems.* Geneva: World Health Organization, Retrieved from http://www.who.int/whr/2000/media_centre/press_release/en/index.html.

Young,*T., Abel, R., Kim, B., Berne, B.J. and Friesner, R., (2013)* : *Distance Metrics Overview. Retrieved from* http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Distance_Metrics_Overview.htm.

Koller and Sahami 1996: efficient attributes dimensionality techniques / *http://www.cs.binghamton.edu/~lyu/SDM07/DR-SDM07.pd*

Blum and Langley 1997: correlation feature selection techniques. *http://airccse.org/journal/cnc/6314cnc15.pdf*

Kohavi and Peager 1997: wrapper feature selection techniques *http://dl.acm.org/citation.cfm%3Fid%3D1756013*

Moss S and Hancoock E.R.: rationale underpinning expectation maximization. *http://www.academia.edu/1786470/Teacher_beliefs_about_feedback_within*

Howell David 2010, Missing feature values *http://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing.html*

Saravana Thirumuruganathan, 2010, A Detailed Introduction To *K*-Nearest Neighbor.: *https://wiki.eecs.yorku.ca/course_archive/2014-15/F/4412/_media/a_detailed_*

Gershenson Carlos, 2003,. Artificial neural networks for beginners :

*http://www.ijarcsse.com/docs/papers/10_October2012/Volume_2_issue_10_*

Appendices

Appendix 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

Appendix 2: Wisconsin Breast Cancer dataset (WBC)

Attribute	**Domain**
Clump Thickness	1-10
Uniformity of Cell Size	1-10
Uniformity of Cell Shape	1-10
Marginal Adhesion	1-10
Bare Nucleoli	1-10
Single Epithelial Cell Size	1-10
Bland Chromatin	1-10
Normal Nucleoli	1-10
Mitoses	1-10
Class	(2 for benign, 4 for malignant)

Appendix 3: WBC (Original), WDBC, and WPBC datasets.

Dataset	**Number of Attributes**	**Number of Instances**	**Number of Classes**
Wisconsin Breast Cancer (Original)	11	699	2
Wisconsin Diagnosis Breast Cancer (WDBC)	32	569	2
Wisconsin Prognosis Breast Cancer (WPBC)	34	198	2