SUDAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
COLLEGE OF GRADUATE STUDIES

COLLEGE OF COMPUTER SCIENSE AND INFORMATION TECHNOLOGY

A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS TO PREDICT BREST CANCER

دراسة مقارنه بين خوارزميات تعلم الاله للتنبؤ بسرطان الثدي

SEP 2018

PREPARED BY

Mino Assad Eltieb

SUPERVISED BY

Dr.Hwaida Ali Abdalgadir

THEESIS SUBMITTED AS A PARTIAL REQUIREMENTS OF MASTER DEGREE IN COMPUTER SCIENCE

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God Almighty for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

At this moment of accomplishment I am greatly indebted to my research supervisor, Dr. Hwaida Ali Abdalgader, who accepted me as her student and offered me her mentorship, motherly love and care.

I acknowledge the people who mean a lot to me, my family, for showing faith in me and giving me liberty to choose what I desired. I salute you all for the selfless love, care, pain and sacrifice you did to shape my life.

I owe thanks to a very special person, my husband, Mohammed for his continued and unfailing love, support and understanding during my pursuit of master degree that made the completion of thesis possible. You were always around at times I thought that it is impossible to continue, you helped me to keep things in perspective. I greatly value his contribution and deeply appreciate his belief in me.

Mino Assad

LIST OF ACRONYMS

- AI= Artificial intelligence
- ANN= Artificial neural network
- DT= Decision trees
- KDD= knowledge discovery in databases
- KNN=K-nearest neighbor
- ML=machine learning
- NB=Naïve bayes algorithm
- NN= Neural network
- OLS=Ordinary Least Squares
- RF= Random Forest
- SVM=Support vector machine algorithm
- SARSA=State-Action-Reward-State-Action
- WDBC= Wisconsin (Diagnostic) Breast Cancer Data Set
- WHO =World Health Organization
- WPBC= Wisconsin Prognostic Breast Cancer

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

According to World Health Organization (WHO), breast cancer is the top cancer in women both in the developed and the developing world. The incidence of breast cancer is increasing in the developing world due to increase life expectancy, increase urbanization and adoption of western lifestyles. About one in eight women are diagnosed with breast cancer during their lifetime. There's a good chance of recovery if it's detected in its early stages.

This research intended to achieve a feature subset with minimum number of features providing efficient classification accuracy. Sequential forward selection algorithm used to find the subset of features that can ensure highly accurate classification of breast cancer as either benign or malignant and to measure the goodness of these selected feature sets.

Then a comparative study on different cancer classification approaches viz. Naïve Bayes, K-nearest, Gradient Boosting and AdaBoost, with and without feature selection, the different algorithms almost find different feature sets by using Sequential forward selection algorithm.

Here, Gradient Boosting classifier is concluded as the best classifier for both mammography dataset and Wisconsin dataset, with and without feature selection.

KEYWORDS

Breast Cancer, Naïve Bayes, K-nearest neighbors, Gradient Boostin, AdaBoost.

مستخلص

وفقا لمنظمة الصحة العالمية (WHO) ، فإن سرطان الثدي يمثل أعلي نسبة سرطان لدي النساء في كل من البلدان المتقدمة والنامية. تتزايد حالات الإصابة بسرطان الثدي في البلدان الناميه بسبب زيادة التوقعات المعيشيه ، وزيادة التحضر وتبني أنماط الحياة الغربية.
واحده من كل ثمانيه سيدات تصاب بسرطان الثدي لذلك فإن الكشف المبكر يؤدي لتحسين نتائج سرطان الثدي وفرص النجاه و لا يزال حجر الزاوية في مكافحة سرطان الثدي.
يهدف هذا البحث إلى العثور على تصنيف دقيق للغاية لسرطان الثدي اعتمادًا على مميزات مختلفة من مجموعة بيانات مختلفة تم رصدها من حالات سابقه لسرطان الثدي . والعثور على اقل مجموعة فرعية من المميزات يمكن أن تضمن أدق تصنيف للاورام المكتشفه إما خبيثه أو حميده.
ثمم مقارنة تصنيفات اربعه خوارزميات من خوارزميات تعلم الاله وهي Gradient Boosting و Naïve Bayes و K-nearest و AdaBoost مرة مع اختيار مميزات فرعيه من مجموعة البيانات ثم مرة اخري مع كل المميزات .
وقد أظهرت الدراسه ان Gradient Boosting هي أفضل خوارزميه لكل من مجموعتي البيانات المستخدمتين في الدراسه ، حتي بعد استخدام خوارزمية Sequential forward selection التي استخدمت لاختيار اقل مجموعه فرعيه من مميزات البيانات التي تضمن أدق تصنيف للأورام المكتشفه.

CHAPTER ONE

INTRODUCTION

1.1 Background

Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012 (second most common cancer overall). This represents about 12% of all new cancer cases and 25% of all cancers in women. It is the fifth most common cause of death from cancer in women. Breast cancer risk doubles each decade until menopause, after which the increase slows. However, breast cancer is more common after menopause [1].

Routine breast cancer screening allows the doctors to detect early breast cancer when treatment can be most successful. Early detection means finding and diagnosing a disease earlier than waiting for symptoms to start [2].

The process of early detection involves examining the breast tissue for abnormal lumps or masses. If a lump is found, a fine-needle aspiration biopsy is performed, which uses a hollow needle to extract a small sample of cells from the mass. A clinician then examines the cells under a microscope to determine whether the mass is likely to be malignant or benign.

This mass is called a tumor. A tumor can be benign or malignant. A benign tumor is not cancer and will not spread to other parts of the body. A malignant tumor is cancer. Cancer cells divide and damage tissue around them. When breast cancer spreads outside the breast, cancer cells are most often found under the arm in the lymph nodes. In many cases, if the cancer has reached the lymph nodes, cancer cells may have also spread to other parts of the body via the lymphatic system or through the bloodstream. This can be life-threatening .

1.2 Machine Leaning

Machine learning is a technique that can discover previously unknown regularities and trends from diverse datasets, in the hope that machines can help in the often tedious and error-prone process of acquiring knowledge from empirical data, and help people to explain and codify their knowledge [3].

Machine learning is a fast growing trend in the health care industry and helps medical experts to analyze data and identify trends. It was born from pattern

recognition and the theory that computers can learn without being programmed to perform specific tasks. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. Machine learning also allows computers to find hidden insights without being explicitly programmed where to look, using algorithms that iteratively learn from data. A machine learning healthcare application that detects the percentage growth or shrinkage of a tumor over time based on image data from dozens or hundreds of X-ray images from various angles. While machine learning might help with "suggestions" in a diagnostic situation, a doctor's judgments would be needed in order to factor for the specific context of the patient. Therefore, it leads to improved diagnoses and treatment specially using image analysis which is a process of extracting meaningful and important information from a digital image [4].

1.3 Research importance

Breast cancer is the most common type of cancer in women, Because of early detection, intervention, and postoperative treatment, breast cancer mortality has been decreasing.

A mammogram is a specific type of breast exam used to aid in the early detection and diagnosis of breast diseases in women. This quick medical exam uses a noninvasive X-ray targeted to each breast, producing pictures that the doctor can use to identify and treat any abnormal areas, possibly indicating the presence of cancer.

Annual mammograms can detect cancer early , when it is most treatable. In fact, mammograms show changes in the breast up to two years before a patient or physician can feel them. Mammograms can also prevent the need for extensive treatment for advanced cancers and improve chances of breast conservation.

1.4 Problem Statement

The exact cause of breast cancer — that is, what causes breast cells to start to grow out of control — is not known, thus early detection of breast cancer would have a very important role in getting decision of doctors to apply the methods of accurate treatment and rescue the life of people.

The main problem of this research the difficult diagnosis of breast cancer that faces doctors using mammogram. Then most previous studies compare between several classification algorithms to define the differences between the algorithms are real or random, and define which one is better performance and accuracy to help

doctors to diagnosis breast cancer perfectly. But they didn't establish procedure for comparing classifiers over multiple data sets with multiple features. This study is using K-nearest neighbors algorithm, Gradient Boosting algorithm, AdaBoost algorithm and Naive Bayes algorithm as classifiers and compare between them with different datasets.

1.5 Research questions

- Which subset of features in Wisconsin data set and Mammography mass dataset , that can ensure highly accurate classification of breast cancer ?
- Which algorithm (K-nearest neighbors algorithm or Naive Bayes algorithm or Gradient Boostin or AdaBoost) can achieve high accuracy and high performance with two datasets depending of the chosen subset of features for both datasets?

1.6 Objective of the Study

This study formulates the following specific objectives:

i- To achieve a feature subset with minimum number of features providing efficient classification accuracy.
ii- To apply K-nearest neighbors and Naive Bayes and Gradient Boosting and AdaBoost algorithms.
iii- To compare classifiers over multiple data sets.
iv- To evaluate the results to decide which algorithm is better in terms of accuracy.

1.7 Scope and Limitations

The qualitative methodology used in this thesis does not claim to give a completely implementation for All classification algorithms. It just uses K-nearest neighbors algorithm and Naive Bayes algorithm and Gradient Boostin and AdaBoost algorithms. This study focuses on compare between these algorithms, the proposed work bases on the available two breast cancer datasets (Mammography mass) and (Wisconsin).

1.8 Layout of the Thesis

The remaining part of the thesis is organized as follows:

The basic theory and concepts of Data mining and other relevant topics of classification that are required for better understanding of the research domain are reviewed in chapter two. A general overview characteristic of breast cancer is also

provided in the same chapter and related research works that has been done on the classification of breast cancer. Chapter three gives a detail description of methodology used in this research.

Chapter four presents the implementation of the classification model and experiment results and the experiment results of KNN and Naïve Bayes and Gradient Boostin and AdaBoost algorithms will be compared.

Finally, the conclusions drawn from the study and possible future works will be pointed out in chapter five.

CHAPTER TWO

LITERATURE REVIEW

2.1 Artificial Intelligence

Artificial Intelligence is one of approach that can train computers to think like human, where it can learn through experience, recognize patterns from large amount of data and also decision making process based from human knowledge and reasoning skills. According from an AI text book titled AI: Structures and Strategies for Complex Problem Solving, an AI can be defined as the branch of computer science that is concerned with the automation of intelligent behavior . It is combination of science and engineering field in order to make an intelligent machines, especially intelligent computer programs [5].

Artificial intelligence (AI) systems are designed to adapt and learn. The first definition of AI is based on the Turing test. Alan Turing undertook a test of a machine‘s ability to demonstrate intelligence. It proceeds as follows: a human judge engages in a natural language conversation with one human and one machine, each of which tries to appear human. The aim of the judge is to distinguish human from machine, only on the basis of conversation (without visual or other help). When the judge cannot distinguish between human and machine, than the machine may be considered as intelligent [6].

2.2 Data

Is a set of values of qualitative or quantitative variables . While the concept of data is commonly associated with scientific research, data is collected by a huge range of organizations and institutions, including businesses (e.g., sales data, revenue, profits, stock price), governments (e.g., crime rates, unemployment rates) and non-governmental organizations (e.g., censuses of the number of homeless people by non-profit organizations) .

There is a great increase in data owned as people come from past to present, and so, to control and manage those rapidly increasing data get harder evenly. When the calculations that are kept on papers did not suffice to store data and also when to find a data got harder, the need for easy manageable and relatively big systems appeared. It is started to keep rapidly increasing data in computer hard discs through the proliferation of computer usage. Although the usage of only computer hard discs seems to be solution at first glance, difficulties in some operations like

accessing data that takes up large spaces in memories and making changes in some data directed people to the idea of database management systems. The facility to make the operations on stored data easily is provided by those systems. The operations that normally take a lot of time to achieve are realized in a short period of time with error rate minimization thanks to database management systems. However, current database management systems become insufficient when the needs require obtaining more information from data. The need of gathering more and more information from data is felt in about all areas of life and the methods are considered to satisfy these needs. Realistic predictions for the future are made by analyzing the data on hand with the developed methods. The process of obtaining information from data is then called as “data mining”.

2.3 Data mining

Data mining refers to extracting useful information from vast amounts of data. Many other terms are being used to interpret data mining, such as knowledge mining from databases, knowledge extraction, data analysis, and data archaeology. Nowadays, it is commonly agreed that data mining is an essential step in the process of knowledge discovery in databases, or KDD. based on a broad view of data mining functionality, data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories [7].

Data mining is defined as the process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions [8].

Data mining is considered as the key process of knowledge discovery in databases (KDD) [9].

The main data mining techniques are classification and clustering analysis, time-series mining, and association rules mining. Data mining techniques are mostly based on statistics, as well as machine learning while the patterns may be inferred from different types of data. Methods used in data mining, such as machine learning, belong to the field of artificial intelligence.

2.4 Machine learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch ofartificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention [10].

Machine learning focuses on the development of computer programs that can access data and use it learn for themselves [11].
The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly [11].

2.4.1 Types of Machine Learning Algorithms

Machine learning algorithm can model each problem differently based on the input data, so before getting into algorithms we should briefly view various kinds of learning styles broadly used. This way of organizing machine learning algorithms forces us to choose the right algorithm to tackle a given problem based on the available input dataset and model preparation process and achieve efficient results. We can divide machine learning algorithms into three different groups based on their learning style:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

2.4.1.1 Supervised learning

The supervised machine learning algorithms are those algorithms which needs external assistance. The input dataset is divided into train and test dataset. The train dataset has output variable which needs to be predicted or classified. All algorithms learn some kind of patterns from the training dataset and apply them to the test dataset for prediction or classification [12].

The supervised approach is further divided into two: Classification and Regression. Supervised machine learning common algorithms:

According to [13], the supervised machine learning algorithms includes the following:

Linear Classifiers, Logistic Regression, Naïve Bayes Classifier , Perceptron , Support Vector Machine; Quadratic Classifiers, K-Means Clustering, Boosting, Decision Tree, Random Forest (RF); Neural networks, Bayesian Networks and so on.

2.4.1.1 .1 Linear Classifiers

Linear models for classification separate input vectors into classes using linear (hyperplane) decision boundaries [14]. The goal of classification in linear classifiers in machine learning, is to group items that have similar feature values, into groups. [15] stated that a linear classifier achieves this goal by making a classification decision based on the value of the linear combination of the features. A linear classifier is often used in situations where the speed of classification is an issue, since it is rated the fastest classifier [13].Also, linear classifiers often work very well when the number of dimensions is large, as in document classification, where each element is typically the number of counts of a word in a document. The rate of convergence among data set variables however depends on the margin. Roughly speaking, the margin quantifies how linearly separable a dataset is, and hence how easy it is to solve a given classification problem [16].

2.4.1.1 .2 Logistic regression

This is a classification function that uses class for building and uses a single multinomial logistic regression model with a single estimator. Logistic regression usually states where the boundary between the classes exists, also states the class probabilities depend on distance from the boundary, in a specific approach. This moves towards the extremes (0 and 1) more rapidly when data set is larger. These statements about probabilities which make logistic regression more than just a classifier. It makes stronger, more detailed predictions, and can be fit in a different way; but those strong predictions could be wrong. Logistic regression is an approach to prediction, like Ordinary Least Squares (OLS) regression. However, with logistic regression, prediction results in a dichotomous outcome [17]. Logistic regression is one of the most commonly used tools for applied statistics and discrete data analysis. Logistic regression is linear interpolation[18].

2.4.1.1 .3 Naive Bayesian (NB) Networks

These are very simple Bayesian networks which are composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent [19].Thus, the independence model (Naive Bayes) is based on estimating [20]. Bayes classifiers are usually less accurate that other more sophisticated learning algorithms (such as ANNs).However, [21] performed a large-scale comparison of the naive Bayes classifier with state-of-the-art algorithms for decision tree induction, instance-based learning, and rule induction on standard benchmark datasets, and found it to be sometimes superior to the other learning schemes, even on datasets with

substantial feature dependencies. Bayes classifier has attributeindependence problem which was addressed with Averaged One-Dependence Estimators [22].

2.4.1.1 .4 Multi-layer Perceptron

This is a classifier in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non convex, unconstrained minimization problem as in standard neural network training [15].Other well-known algorithms are based on the notion of perceptron [23]. Perceptron algorithm is used for learning from a batch of training instances by running the algorithm repeatedly through the training set until it finds a prediction vector which is correct on all of the training set. This prediction rule is then used for predicting the labels on the test set [24].

2.4.1.1 .5 Support Vector Machines (SVMs)

These are the most recent supervised machine learning technique [25].Support Vector Machine (SVM) models are closely related to classical multilayer perceptron neural networks. SVMs revolve around the notion of a margin either side of a hyperplane that separates two data classes. Maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error [24].

2.4.1.1 .6 K-means

According to [26] and [27] K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. K-Means algorithm is be employed when labeled data is not available [28]. General method of converting rough rules of thumb into highly accurate prediction rule. Given weak learning algorithm that can consistently find classifiers (rules of thumb) at least slightly better than random, say, accuracy 55%, with sufficient data, a boosting algorithm can provably construct single classifier with very high accuracy, say, 99% [29].

2.4.1.1 .7 Decision Trees

Decision Trees (DT) are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values [24]. Decision tree learning, used in data mining and machine learning, uses a decision

tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees [30]. Decision tree classifiers usually employ post-pruning techniques that evaluate the performance of decision trees, as they are pruned by using a validation set. Any node can be removed and assigned the most common class of the training instances that are sorted to it [24].

2.4.1.1 .8 Neural Networks

Neural Networks (NN) that can actually perform a number of regression and/or classification tasks at once, although commonly each network performs only one [26]. In the vast majority of cases, therefore, the network will have a single output variable, although in the case of many-state classification problems, this may correspond to a number of output units (the post-processing stage takes care of the mapping from output units to output variables).Artificial Neural Network (ANN) depends upon three fundamental aspects, input and activation functions of the unit, network architecture and the weight of each input connection. Given that the first two aspects are fixed, the behavior of the ANN is defined by the current values of the weights. The weights of the net to be trained are initially set to random values, and then instances of the training set are repeatedly exposed to the net. The values for the input of an instance are placed on the input units and the output of the net is compared with the desired output for this instance. Then, all the weights in the net are adjusted slightly in the direction that would bring the output values of the net closer to the values for the desired output. There are several algorithms with which a network can be trained [31].

2.4.1.1 .9 Bayesian Network

A Bayesian Network (BN) is a graphical model for probability relationships among a set of variables (features). Bayesian networks are the most well-known representative of statistical learning algorithms [24]. The most interesting feature of BNs, compared to decision trees or neural networks, is most certainly the possibility of taking into account prior information about a given problem, in terms of structural relationships among its features [24]. A problem of BN classifiers is that they are not suitable for datasets with many features [32].

2.4.1.2 Unsupervised learning

Occurs when an algorithm learns from input data without any labels and does not have a definite result, leaving the algorithm to determine the data patterns on its own. A model is prepared by learning the features present in the input data to extract general rules. It is done through a mathematical process to reduce

redundancy or to organize data by similarity. Unsupervised learning is again majorly used in two different formats: Clustering in which we group similar items together and density estimation which is used to find statistical values that describe the date.

Unsupervised machine learning common algorithms:

2.4.1.2.1 K-Means Clustering

There are multiple ways to cluster the data but K-Means algorithm is the most used algorithm. Which tries to improve the inter group similarity while keeping the groups as far as possible from each other.

Basically K-Means runs on distance calculations, which again uses "Euclidean Distance" for this purpose. Euclidean distance calculates the distance between two given points using the following formula:

Euclidean Distance = $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$ ----------(1)

Above formula captures the distance in 2-Dimensional space but the same is applicable in multi-dimensional space as well with increase in number of terms getting added. "K" in K-Means represents the number of clusters in which we want the data to divide into. The basic restriction for K-Means algorithm is that the data should be continuous in nature. It won't work if data is categorical in nature.

2.4.1.2.2 Hierarchical Clustering

Unlike K-mean clustering Hierarchical clustering starts by assigning all data points as their own cluster. As the name suggests it builds the hierarchy and in the next step, it combines the two nearest data point and merges it together to one cluster.

1. Assign each data point to its own cluster.
2. Find closest pair of cluster using euclidean distance and merge them in to single cluster.
3. Calculate distance between two nearest clusters and combine until all items are clustered in to a single cluster.

2.4.1.3 Reinforcement learning

Allows machines to determine automatically its behavior within a specific context to maximize its performance. Simple reward feedback is required to learn its behavior known as reinforcement signal. This learning occurs when you present the algorithm with examples that lack labels, as in unsupervised learning Reinforcement learning machine learning common algorithms:

2.4.1.3.1 Q-Learning algorithm

Q-Learning is an off-policy, model-free RL algorithm based on the well-known Bellman Equation:

$$v(s) = \mathbb{E}[R_{t+1} + \lambda v(S_{t+1}) | S_t = s] \quad \text{----------(2)}$$

E in the above equation refers to the expectation, while λ refers to the discount factor. When re-write it in the form of Q-value:

$$Q^{\pi}(s,a) = \mathbb{E}[r_{t+1} + \lambda r_{t+2} + \lambda^2 r_{t+3} + \ldots | s, a]$$

$$= \mathbb{E}_{s'}[r + \lambda Q^{\pi}(s', a') | s, a] \quad \text{----------(3)}$$

The optimal Q-value, denoted as Q* can be expressed as:

$$Q^*(s,a) = \mathbb{E}_{s'}[r + \lambda \max_{a'} Q^*(s', a') | s, a] \quad \text{----------(4)}$$

The goal is to maximize the Q-value. Before diving into the method to optimize Q-value, I would like to discuss two value update methods that are closely related to Q-learning.

2.4.1.3.2 State-Action-Reward-State-Action (SARSA)

SARSA very much resembles Q-learning. The key difference between SARSA and Q-learning is that SARSA is an on-policy algorithm. It implies that SARSA learns

the Q-value based on the action performed by the current policy instead of the greedy policy.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad \text{---------(5)}$$

The action a_(t+1) is the action performed in the next state s_(t+1) under current policy**.**

The core function of Machine learning attempts is to tell computers how to automatically find a good predictor based on past experiences and this job is done by good classifier [36].

Data mining and machine learning depend on classification which is the most essential and important task.

2.4.2 Classification

Classification is the process of using a model to predict unknown values (output variables), using a number of known values (input variables). The classification process is performed on data set D which holds following objects:

- Set size → $A = \{A1, A2, \ldots, A|A|\}$, where $|A|$ denotes the number of attributes or the size of the set A.
- Class label→ C: Target attribute; $C = \{c1, c2, \ldots, c|C|\}$, where $|C|$ is the number of classes and $|C| \geq 2$.

Given a data set D, the core objective of ML is to produce prediction/classification function to relate values of attributes in A and classes in C [33].

Classification technique can solve several problems in different fields like medicine, industry, business, science. Basically it involves finding rules that categorize the data into disjoint groups. There are several classification discovery models and these are: the decision tree, neural networks, genetic algorithms and some statistical models [34].

2.4.2.1 Classification algorithms

Classification algorithms are widely used in various medical applications. Data classification is a two phase process in which first step is the training phase where

the classifier algorithm builds classifier with the training set of tuples and the second phase is classification phase where the model is used for classification and its performance is analyzed with the testing set of tuples.

2.5 Characteristics of medical data

The data gathered in medicine is generally collected as a result of patient-care activity to benefit the individual patient and research is only a secondary consideration. As a result, medical data contain many features that create problems for the data mining techniques and they might be in a format which is not suitable for the direct application of those techniques [35].

In general, medical collections, diagnoses and treatments are subject to error rates, imprecision and uncertainty [36].

As with any large databases and due to the collection method, medical databases may contain missing values and can introduce noisy, redundant, incomplete or inconsistent data [37].

2.5.1 Medical data classification:

As medicine plays a great role in human life, automated knowledge extraction from medical data sets has become an immense issue. Research on knowledge extraction from medical data is growing fast [38].

All activities in medicine can be divided into six tasks: screening, diagnosis, treatment, prognosis, monitoring and management. As the healthcare industry is becoming more and more reliant on computer technology, machine learning methods are required to assist the physicians in identifying and curing abnormalities at early stages. Medical diagnosis is one of the important activities of medicine.

The prevention, diagnosis, treatment and cure of the disease have remained most challenging for the medical fraternity. Here, medical data classification a key role. The extraction of knowledge from medical data assists medical experts, medical decision support system and in discovery of new drugs.

Medical data classification is a challenging task in the field of medical research. The medical record is very important for a patient as well as the doctor. Generally the medical record will help the doctor to classify the diseases, diagnose and give an appropriate treatment to the patient.

Machine learning algorithm can significantly help in solving the healthcare problems by developing classifier systems that can assist physicians in diagnosing and predicting diseases in early stages.

2.6 Feature selection

The medical dataset are complex to handle. The preprocessing is important for real time medical data. The medical data with large attribute need efficient feature selection algorithm for predicting it into binary and multiclass data with improve accuracy.

The feature selection techniques are categorized into three Filter method, Wrapper method, and Embedded method.

2.6.1 Filter Method

The filter attribute selection method is independent of the classification algorithm. Filter method is further categorized into two types

- Attribute evaluation algorithms
- Subset evaluation algorithms

The algorithms are categorized based on whether they rate the relevance of individual features or feature subsets. Attribute evaluation algorithms rank the features individually and assign a weight to each feature according to each feature's degree of relevance to the target feature. The attribute evaluation methods are likely to yield subsets with redundant features since these methods do not measure the correlation between features. Subset evaluation methods, in contrast, select feature subsets and rank them based on certain evaluation criteria and hence are more efficient in removing redundant features [39].

The FSDD, RFS, CFS are some of the feature selection algorithm which uses the Filter methodology. The relevance score is calculated for the features to check the correlation between the features. The calculated score is high with some threshold value then the particular feature is selected for further classification. When the ranking is low those feature are removed. This method is very simple, fast and also independent of classification algorithm. The following are the basic filter feature selection algorithm

- $\chi 2$ test
- Euclidian distance
- T-test

- Information gain
- CFS-correlation based feature selection method
- MBF- Markov blanket filter
- FCBF-fast correlation based feature selection
- Sequential Forward Selection (SFS)

2.6.2 Wrapper Method

The wrapper method is slower and expensive than the filter method. The interaction between the feature subsets and the maintaining the dependencies between the features are the main advantageous of wrapper method. Wrapper methods are better in defining optimal features rather than simply relevant features and that they do this by allowing for the specific biases and heuristics of the learning algorithm and the training set. The wrapper method uses the backward elimination to remove the insignificant features from the subset. The SVM-RFE is one of the feature selection algorithms which use the Wrapper method. The Wrapper method need some predefine learning algorithm to identify the relevant feature. It has interaction with classification algorithm. The over fitting of feature is avoided using the cross validation. But it takes much time comparing to the filter method.

2.6.3 Embedded method

The embedded method has the interaction with the classification algorithm The KP-SVM is the example for embedded method. It consumes less time than the wrapper method. The embedded method uses the support vector mechanism to select the feature.

2.7 Related work:

Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques. A good amount of research on breast cancer datasets is found in literature. Many of them show good classification accuracy.

In the following some of the relevant work which has already made its mark in this field of study:

In paper [40] author used SVM with feature selection to diagnose the breast cancer. For training and testing experiments the WDBC dataset has been taken from the University of California at Irvine (UCI) machine learning repository .It was spotted that the proposed method produced the highest classification accuracies (99.51%, 99.02% and 98.53% for 80–20% of training-test partition, 70–

30% of training-test partition and 50–50% of training-test partition respectively) for a subset that carried five features.

In paper [41] author applied the Naive Bayes classifier to the Wisconsin Prognostic Breast Cancer (WPBC) dataset, containing a number of 198 patients and a binary decision class: non-recurrent-events having 151 instances and recurrent-events having 47 instances. The testing diagnosing accuracy, that was the main performance measure of the classifier, was about 74.24%, in compliance with the performance of other well-known machine learning techniques.

In paper [42] author have proposed the Naïve Bayes classifier gives the maximum accuracy with only five dominant features and time complexity is least compared to other two classifiers.

In paper[43] prediction of the breast, cancer disease was done through Artificial Neural Network (ANN), Logistic regression, Naive Bayes techniques. The target of the research aims at giving the following results; firstly, it evaluates medical data set in terms of quality grammatically and secondly, it evaluates data mining methods with respect to their applicability to the data. Eventually, the knowledge drawn out from the data set is used for disease prediction by applying Artificial Neural Network (ANN), Logistic Regression, Naïve Bayes. It is found that these methods had highest lifting factor for most of the class values.

In paper [44] authors found the smallest subset of features from Wisconsin Diagnosis Breast Cancer (WDBC) dataset by applying confusion matrix accuracy and 10-fold cross validation method that can ensure highly accurate ensemble classification of breast cancer as either benign or malignant. For classification, the breast cancer data were first classified by Support Vector Machine (SVM) and Naïve Bayes classifiers and then finalize the classification process.

In paper[45] authors demonstrate the comparison of different classification techniques like Bayes Network, Radial Basis Function, Pruned Tree and Nearest Neighbors algorithm using Waikato to Environment for Knowledge Analysis (WEKA) on large dataset. The data utilized in their research is the breast cancer data. It holds a total of 6291 data and a dimension of 699 rows and 9 columns. In this 75% of overall data is used training and the remainder is used for testing the accuracy of classification technique. Agreeing to the simulation result, highest accuracy is 89.71% which owe support to bayes network with the minimum time taken to build the model is 0.19 seconds and lowest average error is 0.2140 compared to others.

In paper [46] authors have proposed the results of both NN and SVM were compared on the basis of accuracy and precision. It was observed that classification implemented by Neural Network technique in this paper is more efficient compare to SVM as seen in the accuracy and precision. Bayesian classification provides practical learning algorithms and prior knowledge on observed data.

In paper [47] author analyzed Brest cancer using data mining technique classification. They use machine learning technique like Decision Tree (C4.5), Artificial Neural Network and support vector machine for predicting a breast cancer. That process in WEKA tool kit with the Iranian center of breast cancer .this work targeted to analyzed the performance to achieve higher accuracy specificity and sensitivity result by SVM technique indicated promising level of accuracy level of 95.7%,97.1% and 94.5%.

In paper [48] author used prediction on data mining technique on heart disease , Diabetes, Breast Cancer in heart disease data collected data from a hospital information system the heart disease prediction in that machine learning algorithms namely naïve bayes, K-NN, Decision List. In all technique classification accuracy of the naïve bayes algorithm is better when compared to other algorithm. Second one breast cancer as per survey of united state prediction data mining technique like C4.5, ANN and Fuzzy decision tree. ANN conducts better accuracy and good performance. Third one about diabetes as per base on the American diabetes association perdition by using homogeneity based algorithm genetic algorithm predicts batter accuracy. For feature work enhance they predates diff type of disease prediction using data mining technique.

In paper [49] author focus on use three different type of machine learning technique for predicting of breast cancer. They use Iranian canter for breast cancer (ICBC) data set and implement machine learning technique like decision tree (C4.5) Support vector machine (SVM) and Artificial Neural network (ANN) compared the performance of technique and find sensitivity, specificity , accuracy. As per a conclusion SVM provide better performance with highest accuracy rate.

CHAPTER THREE

METHODS AND TECHNIQUES

The methodology for breast cancer classification is beginning with data collection from UCI Machine Learning Repository and cleaning the data from missing values then feature selection algorithm used to select subset the minimum subset of features that can ensure highly accurate classification of breast cancer and finally four classification algorithms applied to datasets to compare accuracy. Fig 1 workflow diagram represents breast cancer cell detection as follow.

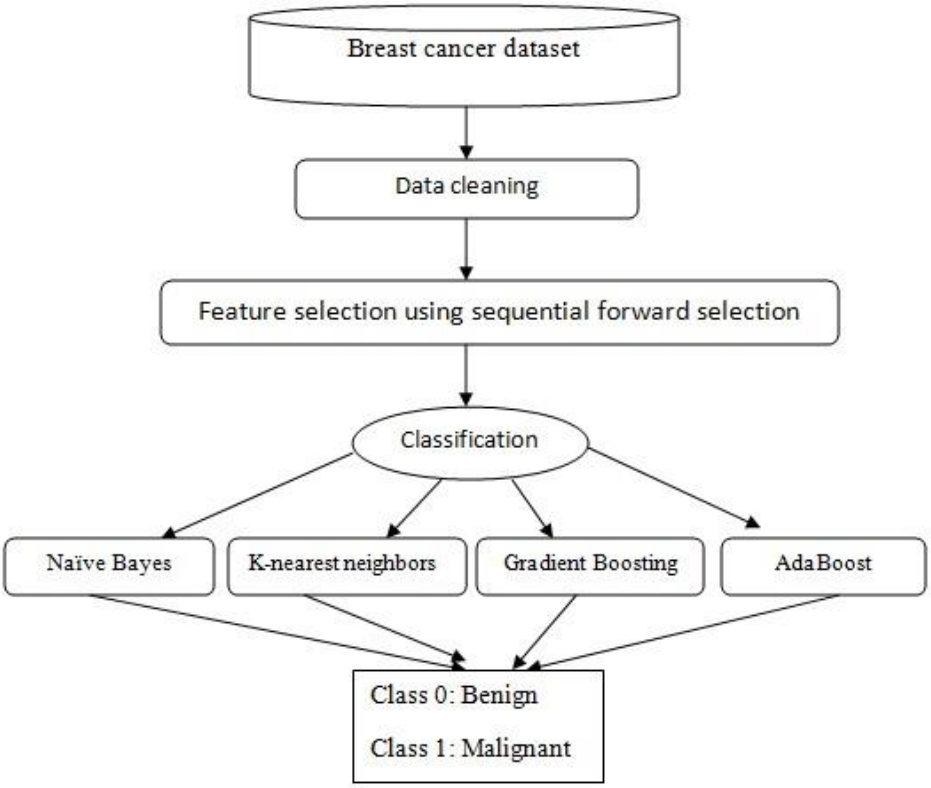

Fig 1: Workflow diagram for breast cancer cell detection.

3.1. Data collection

The breast cancer data used for this research (Wisconsin Breast Cancer Diagnostic dataset and Mammographic Mass Data Set) was obtained from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml).

3.1.1 Wisconsin Breast Cancer Diagnostic dataset

Wisconsin Breast Cancer Diagnostic dataset was donated by researchers of the University of Wisconsin and includes the measurements from digitized images of fine-needle aspirate of a breast mass. It includes 569 instances of cancer biopsies. Each record has 32 attributes. One attribute is an identification number, another is the cancer diagnosis, and 30 are numeric-valued laboratory measurements. The diagnosis is coded as "M" to indicate malignant or "B" to indicate benign. The values represent the characteristics of the cell nuclei present in the digital image [50].

The 30 numeric measurements comprise the mean, standard error, and worst (that is, largest) value for 10 different characteristics of the digitized cell nuclei. The 10 real valued features calculated for each cell nucleus are:

1) Radius (mean of distances from center to points on the perimeter)

2) Texture (standard deviation of gray-scale values)

3) Perimeter

4) Area

5) Smoothness (local variation in radius lengths)

6) Compactness (perimeter^2 / area - 1.0)

7) Concavity (severity of concave portions of the contour)

8) Concave points (number of concave portions of the contour)

9) Symmetry

10) Fractal dimension ("coastline approximation" - 1) All feature values are recoded with four significant digits. Furthermore, there are no missing values.

3.1.2 Mammographic Mass Data Set

The database contains a BI-RADS assessment, the patient's age and three BI-RADS attributes (mass shape, mass margin, mass density) and is based on digital mammograms collected at the Institute of Radiology of the University of Erlangen-Nuremberg between 2003 and 2006 [51]. It consists of 961 records and each record in the database has one dependent and five independent variables. There are 516 benign and 445 malignant masses. The database does not reflect all variables that are collected by radiologists during mammography practice, hence, it is one of the limitations of this work.

Attribute Information:

6 Attributes in total (1 goal field, 1 non-predictive, 4 predictive attributes)

1) BI-RADS assessment: 1 to 5 (ordinal, non-predictive!)
2) Age: patient's age in years (integer)
3) Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4) Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
5) Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
6) Severity: benign=0 or malignant=1 (binominal, goal field!)

3.2 Preprocessing method

Today's real-world databases are highly vulnerable to noisy, missing and inconsistent data due to their typically massive size and their likely origin from multiple, miscellaneous sources. Hence data preprocessing is a necessary phase for classification purposes. Data preprocessing includes data cleaning, data dimensionality reduction, data transformation (data normalization, data binning) followed by classification [52].

The data cleaning technique includes removing the missing values if present. with the mean of the attributes. Data normalization brings the range of all attribute values between 0 and 1.

3.3 Feature selection

Feature selection is an important issue in building classification systems. It is advantageous to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model [48].

In this research one feature selection technique used , which is sequential forward selection.

3.3.1 Sequential forward selection algorithm

Sequential forward selection (SFS) algorithm is a bottom-up search procedure which start from an empty set and gradually add features selected by some evaluation function. At each iteration, the feature to be included in the feature set, it is selected among the remaining available features of the feature set, which have not been added to the feature set. So, the new extended features set should produce a minimum classification error compared with the addition of any other feature. SFS is widely used for their simplicity and speed.

The SFS consists of a forward step which is as follows: starting from an initially empty set of features $\mathcal{Z}_0$, at each forward (inclusion) step at the level l we seek the feature $X^+ \in (\mathcal{X} - \mathcal{Z}_{l-1})$ such that for $\mathcal{Z}_l = \mathcal{Z}_{l-1} \cup \{X^+\}$ the probability of correct classification achieved by the Bayes classifier $J(\mathcal{Z}_l)$ is maximized. In addition to the aforementioned inclusion step the SFFS algorithm [53]

3.4 Implementation

3.4.1 Tools

For the experimentation of the proposed detection and classification model, python on windows platform is used for dataset processing and classification.

3.4.2 development method

32 features for 569 instances from WDBC breast cancer dataset and 6 features for 961 instances from Mammographic Mass Data Set have been taken and applied over four classification techniques Naïve Bayes , KNN, Gradient Boosting and AdaBoost to obtain the classification of dataset under two categories, either malignant or benign.

3.4.2.1 Naive Bayes algorithm

Naive Bayes Classifier is the simple Statistical Bayesian Classifier [54]. It is called Naive as it assumes that all variables contribute towards classification and are

mutually correlated. This assumption is called class conditional independence. It is also called Idiot's Bayes, Simple Bayes, and Independence Bayes.They can predict class membership probabilities, such as the probability that a given data item belongs to a particular class label. A Naive Bayes classifier considers that the presence (or absence) of a particular feature(attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given.

The Naïve Bayes algorithm is one of the most important supervised machine learning algorithms for classification. This classifier is a simple probabilistic classifier based on applying Bayes' theorem as follows:

Let X is a data sample whose class label is not known and let H be some hypothesis, such that the data sample X may belong to a specified class C.

Bayes theorem is used for calculating the posterior probability P(C|X), from P(C), P(X), and P(X|C). Where

P(C|X) is the posterior probability of target class.

P(C) is called the prior probability of class.

P(X|C) is the likelihood which is the probability of predictor of given class.

P(X) is the prior probability of predictor of class.

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \qquad \text{----------(6)}$$

The Naive Bayes classifier works as follows:

1. Let D be the training dataset associated with class labels. Each tuple is represented by n-dimensional element vector, X=(x1, x2, x3,.....,xn).

2. Consider that there are m classes C1, C2, C3...., Cm. Suppose that we want to classify an unknown tuple X, then the classifier will predict that X belongs to the class with higher posterior probability, conditioned on X. i.e., the Naive Bayesian classifier assigns an unknown tuple X to the class Ci if and only if :

P(Ci|X) > P(Cj|X)

For $1 \leq j \leq m$, and $i \neq j$, above posterior probabilities are computed using Bayes Theorem.

Advantages **:**

1. It requires short computational time for training.
2. It improves the classification performance by removing the irrelevant features.
3. It has good performance.

Disadvantages:

The Naive Bayes classifier requires a very large number of records to obtain good results. Less accurate as compared to other classifiers on some datasets.

3.4.2.2 The K-Nearest Neighbor Algorithm

The K-Nearest Neighbor Algorithm is the simplest of all machine learning algorithms. It is based on the principle that the samples that are similar, generally lies in close vicinity [55].K-Nearest Neighbor is instance based learning method. Instance based classifiers are also called lazy learners as they store all of the training samples and do not build a classifier until a new, unlabeled sample needs to be classified [56]. Lazy-learning algorithms require less computation time during the training phase than eager learning algorithms (such as decision trees, neural networks and bayes networks) but more computation time during the classification process[57][58].

Nearest-neighbor classifiers are based on learning by resemblance, i.e. by comparing a given test sample with the available training samples which are similar to it. For a data sample X to be classified, its K-nearest neighbors are searched and then X is assigned to class label to which majority of its neighbors belongs to. The choice of k also affects the performance of k-nearest neighbor algorithm [58]. If the value of k is too small, then K-NN classifier may be vulnerable to over fitting because of noise present in the training dataset. On the other hand, if k is too large, the nearest-neighbor classifier may misclassify the test sample because its list of nearest neighbors may contain some data points that are located far away from its neighborhood. K-NN fundamentally works on the belief that the data is connected in a feature space. Hence, all the points are considered in order, to find out the distance among the data points. Euclidian distance or Hamming distance is used according to the data type of data classes used[59].

In this a single value of K is given which is used to find the total number of nearest neighbors that determine the class label for unknown sample. If the value

of K=1, then it is called as nearest neighbor classification. The K-NN classifier works as follows:

1. Initialize value of K.

2. Calculate distance between input sample and training samples.

3. Sort the distances.

4. Take top K- nearest neighbors.

5. Apply simple majority.

6. Predict class label with more neighbors for input sample.

Advantages:

1. Easy to understand and implement.

2. Training is very fast.

3. It is robust to noisy training data.

4. It performs well on applications in which a sample can have many class labels[58].

Disadvantages:

1. Lazy learners incur expensive computational costs when the number of potential neighbors which to compare a given unlabeled sample is large [58].
2. It is sensitive to the local structure of the data [59].
3. Memory limitation.
4. As it is supervised lazy learner, it runs slowly.

3.4.2.2 Gradient Boost Algorithm

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

The gradient boosting method assumes a real-valued *y* and seeks an approximation $\hat{F}(x)$ in the form of a weighted sum of functions $h_i(x)$ from some class $\mathcal{H}$, called base (or weak) learners:

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const.} \qquad \text{----------(7)}$$

In accordance with the empirical risk minimization principle, the method tries to find an approximation $\hat{F}(x)$ that minimizes the average value of the loss function on the training set, i.e., minimizes the empirical risk. It does so by starting with a model, consisting of a constant function $F_0(x)$, and incrementally expands it in a greedy fashion:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma), \quad \text{----------(8)}$$

$$F_m(x) = F_{m-1}(x) + \arg\min_{h_m \in \mathcal{H}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h_m(x_i)), \quad \text{----------(9)}$$

where $h_m \in \mathcal{H}$ is a base learner function.

3.4.2.3 AdaBoost algorithm

AdaBoost algorithm creates a set of poor learners by maintaining a collection of weights over training data and adjusts them after each weak learning cycle adaptively. The weights of the training samples which are misclassified by current weak learner will be increased while the weights of the samples which are correctly classified will be decreased [60]. The final equation for classification can be represented as

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)), \quad \text{----------(10)}$$

where f_m stands for the m_th weak classifier and theta_m is the corresponding weight. It is exactly the weighted combination of M weak classifiers. The whole procedure of the AdaBoost algorithm can be summarized as follow.

Given a data set containing n points, where

$$x_i \in \mathbb{R}^d, y_i \in \{-1, 1\}. \quad \text{----------(11)}$$

Here -1 denotes the negative class while 1 represents the positive one.

Initialize the weight for each data point as:

$$w(x_i, y_i) = \frac{1}{n}, i = 1, ..., n. \quad \text{----------(12)}$$

For iteration m=1,…,M:

(1) Fit weak classifiers to the data set and select the one with the lowest weighted classification error:

$$\epsilon_m = E_{w_m}[1_{y \neq f(x)}] \quad \text{---------- (13)}$$

(2) Calculate the weight for the m_th weak classifier:

$$\theta_m = \frac{1}{2} ln(\frac{1 - \epsilon_m}{\epsilon_m}). \quad \text{----------(14)}$$

For any classifier with accuracy higher than 50%, the weight is positive. The more accurate the classifier, the larger the weight. While for the classifer with less than 50% accuracy, the weight is negative. It means that we combine its prediction by flipping the sign.

(3) Update the weight for each data point as:

$$w_{m+1}(x_i, y_i) = \frac{w_m(x_i, y_i) exp[-\theta_m y_i f_m(x_i)]}{Z_m}, \quad \text{----------(15)}$$

where Z_m is a normalization factor that ensures the sum of all instance weights is equal to 1. If a misclassified case is from a positive weighted classifier, the “exp” term in the numerator would be always larger than 1 (y*f is always -1, theta_m is positive). Thus misclassified cases would be updated with larger weights after an iteration. The same logic applies to the negative weighted classifiers. The only difference is that the original correct classifications would become misclassifications after flipping the sign.After M iteration, then can get the final prediction by summing up the weighted prediction of each classifier.

3.5 Evaluation Procedures

After training the model using 80%of the training dataset, the proposed algorithms will be measured using 20%of the test dataset respectively for overall accuracy. then will calculate a confusion matrix by calclating number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data and then calculate accuracy for different classifires.

CHAPTER FOUR

EXPERIMENTAL RESULTS

4.1 Measures and Metrics

To evaluate the effectiveness of our methods, experiments on WBCD and Mammography Mass dataset is conducted. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

4.1.1 Confusion matrix: The confusion matrix contains four classification performance indices: true positive, false positive, false negative, and true negative as shown in Table1. These four indices are also usually used to evaluate the performance the two-class classification problem. The four classification performance indices included in the confusion matrix is shown in Table

Actual Class	**Predicted Class**	
	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negative(TN)

Table 1: Confusion matrix

4.1.2 Accuracy Measures: Accuracy measure represents how far the set of tuples are being classified correctly.TP refers to positive tuples and TN refers to negative tuples classified by the basic classifiers. Similarly FP refers to positive tuples and FN refers to negative tuples which is being incorrectly classified by the classifiers. The accuracy measures used here are sensitivity and specificity.

$$\text{Accuracy}=\frac{TP + TN}{TP+FP+FN+ TN} \quad \text{----------(16)}$$

4.2 Results

The classification algorithms has been implemented using python (Version 3.6). Feature selection is used to eliminate the unimportant and redundant features. The classification algorithms are used to classify Wisconsin Breast Cancer Diagnostic dataset and Mammographic Mass dataset with all the features and with optimum features.

The results are shown in Tables 2,3. To get four accuracy of a prediction model, optimal parameter setting play a crucial role. The research evaluated the proper algorithmic parameters of all the mentioned four classification algorithms and uses 80-20 training-testing partition of the data.

No	Classification algorithm	Considering all the features. (Accuracy %)	With features selected subset (Accuracy %)
1.	Naïve Bayes	92.98%	96.49%
2.	K-nearest	93.86%	94.74%
3.	Gradient Boosting	96.49%	97.37%
4.	AdaBoost	95.61%	96.49%

Table 2: Classification Accuracy on WDBC dataset

S.No	Classification algorithm	Considering all the features. (Accuracy %)	With features selected subset (Accuracy %)
1.	Naïve Bayes	80.72%	82.53%
2.	K-nearest	77.71%	81.93%
3.	Gradient Boosting	83.13%	84.34%
4.	AdaBoost	81.33%	82.53%

Table 3: Classification Accuracy on Mammographic Mass dataset

No	Algorithm name	Selected features id's for **WDBC dataset**	Selected features id's for **Mammographic Mass dataset**
1.	KNN	0, 3, 13, 23	0, 2
2.	Naïve Bayes	11, 21, 22, 24	0, 2
3.	AdaBoost	0, 2, 20, 21, 27	0,2
4.	Gradient Boosting	0, 1, 2, 4, 23, 26	0, 2

Table 4: Features selected subsets from both datasets

This comparative study shows that the classification accuracy of Gradient Boosting is higher than other classification algorithms in the diagnosis of predictive breast cancer data for both datasets Wisconsin and Mammography mass. Gradient Boosting is the most accurate algorithm having classified the samples with 96.49%for Wisconsin and 83.13% for Mammography mass with considering all the features. And 97.37% for Wisconsin and 84.34% for Mammography mass with features selected by sequential forward selection algorithm.

CHAPTER FIVE

CONCLUSION AND FUTURE WORK

5.1 Conclusions

There is no doubt that breast cancer is a dangerous disease for women throughout the world, therefore the effective solution is desired. Nowadays, the early detection is common and effective method to cure and save patients, which is generally using mammography. However, there exists the lack of analysis of the mammogram images, the low accuracy of classifying images of benign or malignant often affects the practitioner make more reasonable choices and manage patients. Hence, the problems of breast cancer diagnosis are in the scope of the more general and widely discussed data classification problems.

In this research, we compared the classification accuracy of four Machine Learning algorithms – KNN, NB, AdaBoost and Gradient boosting on UCI Wisconsin Breast Cancer dataset and Mammography Mass dataset. The aim of this comparative study was to find the most accurate machine learning tool that can act as a tool for diagnosis of breast cancer.

5.2 Discussion

Two datasets obtained from UCI Machine Learning Repository, Mammographic Mass Data Set has more than 100 missing values from various features. In real world data, there are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. Deleting Rows method used to handling missing values in Mammographic Mass Data Set by deleting any row has null value for a particular feature. The method has advised because there are enough samples in the data set .

Feature selection is an important issue in classification, because it may have a considerable effect on accuracy of the classifier. It reduces the number of dimensions of the dataset, so the processor and memory usage reduce; the data becomes more comprehensible and easier to study on.

The research investigated the relationship between feature selection and the resulting classification accuracy .

The results demonstrate that the feature selection using sequential forward selection algorithm improves the classification accuracy of all the classification algorithms for both two datasets used as shown in Tables 2, 3 in the previous section.

Since the input features may be mutually dependent, the different algorithms may find different feature sets as shown in table 4 in the previous section. To measure the goodness of these selected feature sets and to find feature subset with minimum number of features providing efficient classification accuracy, sequential forward selection algorithm used.

When an algorithm as AdaBoost used , which uses Boosting technique to select features subsets for minimize misclassification error, feature selection also can improve the classification accuracy as shown in tables 2,3 in the previous section .

Finally, the highest classification accuracy become 97.37% as shown in Table 2, Gradient Boosting has highest accuracy for the given datasets using feature selection .

5.1 Further work

For successful breast cancer treatment, accurate mammography image segmentation, detection and classification are an important task in medical diagnosis before the tumor grow and spreads.
In order to achieve this goal perfectly, the following future works are pointed out for further research and improvement on the current work:

- Using other preprocessing techniques for these datasets.
- Compare between these preprocessing techniques to find the best one that can increase accuracy especially for Mammogram mass dataset by using same algorithms.

REFRENCES

[1] World Cancer Research Fund. (2018). " *Breast cancer statistics*". [online] Available at: https://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics [Accessed 21 Nov. 2018].

[2] American Cancer Society, "Detailed Guide: Breast Cancer", cancer.org 201[Online]Available:www.cancer.org/Cancer/BreastCancer/DetailedGuide/index[Accessed 21 Nov. 2018].

[3] Sivagami, P., 2012. Supervised learning approach for Breast cancer classification. *International Journal of Emerging Trends & Technology in Computer Science*, *1*(4).

[4] SAS. (2018). *Evolution of machine learning*. [online] Available at: https://www.sas.com/en_us/insights/analytics/machine-learning/ [Accessed 21 Nov. 2018].

[5] Luger, G.F., 2005. *Artificial intelligence: structures and strategies for complex problem solving*. Pearson education.

[6] Hi'ovská, K. and Koncz, P., 2012. Application of Artificial Intelligence and Data Mining Techniques to Financial Markets. *Economic Studies & Analyses/Acta VSFS*, *6*(1).

[7] Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

[8] Simoudis, E., 1996. Reality check for data mining. *IEEE Intelligent Systems*, (5), pp.26-33.

[9] Seifert, J.W., 2004, December. Data mining: An overview. Congressional Research Service, Library of Congress.

[10] Sas.com. (2018). *Machine Learning: What it is and why it matters*. [online] Available at: https://www.sas.com/en_us/insights/analytics/machine-learning.html [Accessed 21 Nov. 2018].

[11] Expertsystem.com. (2018). *What is Machine Learning? A definition - Expert System*. [online] Available at: http://www.expertsystem.com/machine-learning-definition [Accessed 21 Nov. 2018].

[12] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, pp.3-24.

[13] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, pp.3-24.

[14] Ayodele, T.O., 2010. Types of machine learning algorithms. In *New advances in machine learning*. InTech.

[15] Elder, J. (n.d), Introduction to Machine Learning and Pattern Recognition . Available at LASSONDE University EECS Department York website: http://www.eecs.yorku.ca/course_archive/2011-12/F/4404-5327/lectures/01%20Introduction.pd

[16] Verlinde, P., Maitre, G. and Mayoraz, E., Decision fusion using a multi-linear classifier.

[17] Setiono, R. and Leow, W.K., 2000. FERNN: An algorithm for fast extraction of rules from neural networks. *Applied Intelligence*, *12*(1-2), pp.15-25.

[18] Newsom, I.,Data Analysis II: Logistic Regression,2015 . Available at: http://web.pdx.edu/~newsomj/da2/ho_logistic.pdf

[19] Logistic Regression pp. 223 – 237. Available at: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12 .pdf

[20] Good, I.J., 1950. Probability and the Weighing of Evidence.

[21] Nilsson, N.J., 1965. Learning machines.

[22] Domingos, P. and Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, *29*(2-3), pp.103-130.

[23] Rosenblatt, F., Principles of Neurodynamics (Spartan, New York, 1962).

[24] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, pp.3-24.

[25] Hormozi, H., Hormozi, E. and Nohooji, H.R., 2012. The classification of the applicable machine learning methods in robot manipulators. *International Journal of Machine Learning and Computing*, *2*(5), p.560.

[26] Bishop, C.M., 1995. *Neural networks for pattern recognition*. Oxford university press.

[27] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., 2004. A local search approximation algorithm for k-means clustering. *Computational Geometry*, *28*(2-3), pp.89-112.

[38] Alex and Vishwanathan, S.V.N,2008. ,Introduction to Machine Learning.

[29] Rob Schapire (n.d) ,2006.Machine Learning Algorithms for Classifrication.

[30] Friedman, J., Hastie, T. and Tibshirani, R., 2001. *The elements of statistical learning* (Vol. 1, No. 10). New York, NY, USA:: Springer series in statistics.

[31] Neocleous, C. and Schizas, C., 2002, April. Artificial neural network learning: A comparative review. In *Hellenic Conference on Artificial Intelligence* (pp. 300-313). Springer, Berlin, Heidelberg.

[32] Cheng, J., Greiner, R., Kelly, J., Bell, D. and Liu, W., 2002. Learning Bayesian networks from data: an information-theory based approach. *Artificial intelligence*, *137*(1-2), pp.43-90.

[33] Muhammad, I. and Yan, Z., 2015. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing*, *5*(3).

[34] Deulkar, M.D.S. and Deshmukh, R.R., 2016. Data Mining Classification. *Imperial Journal of Interdisciplinary Research*, *2*(4).

[35] Delen, D., Walker, G. and Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), pp.113-127.

[36] Pattaraintakorn, P., Cercone, N. and Naruedomkul, K., 2005, July. Hybrid intelligent systems: Selecting attributes for soft-computing analysis. In *Computer Software and Applications Conference, 2005. COMPSAC 2005. 29th Annual International* (Vol. 1, pp. 319-325). IEEE.

[37] Delen, D., Walker, G. and Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), pp.113-127.

[38] Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E. and Tabar, V.K., 2014. Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, *41*(9), pp.4434-4463.

[39] Pitt, E. and Nayak, R., 2007, December. The use of various data mining and feature selection methods in the analysis of a population survey dataset. In *Proceedings of the 2nd international workshop on Integrating artificial intelligence and data mining-Volume 84* (pp. 83-93). Australian Computer Society, Inc..

[40] Akay, M.F., 2009. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, *36*(2), pp.3240-3247.

[41] Dumitru, D., 2009. Prediction of recurrent events in breast cancer using the Naive Bayesian classification. *Annals of the University of Craiova-Mathematics and Computer Science Series*, *36*(2), pp.92-96.

[42]. Hazra, A., Mandal, S.K. and Gupta, A., 2016. Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. *International Journal of Computer Applications*, *145*(2).

[43] Shiny, K., Swaminathan, M., Kumar, N.S. and Thiyagarajan, L., 2015. Implementation of Data Mining Algorithm to Analysis Breast Cancer. *International Journal for Innovative Research in Science and Technology*, *1*(9), pp.207-212.

[44] Kathija, S.N., 2016. Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, *4*(12).

[45] bin Othman, M.F. and Yau, T.M.S., 2007. Comparison of different classification techniques using WEKA for breast cancer. In *3rd Kuala Lumpur International Conference on Biomedical Engineering 2006* (pp. 520-523). Springer, Berlin, Heidelberg.

[46]. Ali, E.E.E. and Feng, W.Z., 2016. Breast Cancer Classification using Support Vector Machine and Neural Network. *International Journal of Science and Research (IJSR) ISSN (Online)(????)*.

[47] Al-Qarzaie, S. and Al-Odhaibi, S., Bedoor Al-Saeed and Dr. *Mohammed Al-Hagery—Using the Data Mining Techniques for Breast Cancer Early Prediction*.

[48] Vijiyarani, S. and Sudha, S., 2013. Disease prediction in data mining technique–a survey. *International Journal of Computer Applications & Information Technology*, *2*(1), pp.17-21.

[49] Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M., Razavi, A.R. and Ahmad, L.G., 2013. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *J Health Med Inform*, *4*(124), p.2.

[50] Mangasarian, O.L., 1990. Cancer diagnosis via linear programming. *SIAM news*, *23*(5), pp.1-18..

[51] Elter, M., Schulz-Wendtland, R. and Wittenberg, T., 2007. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical physics*, *34*(11), pp.4164-4172.

[52] Hazra, A., Mandal, S.K. and Gupta, A., 2016. Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. *International Journal of Computer Applications*, *145*(2).

[53] Pudil, P., Ferri, F.J., Novovicova, J. and Kittler, J., 1994, October. Floating search methods for feature selection with nonmonotonic criterion functions. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)* (Vol. 2, pp. 279-283). IEEE..

[54] Duda, R.O. and Hart, P.E., 1973. Pattern classification and scene analysis. *A Wiley-Interscience Publication, New York: Wiley, 1973*.

[55] Cover, T. and Hart, P., 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), pp.21-27.

[56] Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

[57] K. P. Soman, 2006, Insight into Data Mining Theory and Practice, New Delhi: PHI.
[58] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, pp.3-24.
[59] Soundarya, M. and Balakrishnan, R., 2014. Survey on classification techniques in data mining. *International Journal of Advanced Research in Computer and Communication Engineering*, *3*(7), pp.7550-7552.

[60] Li, X., Wang, L. and Sung, E., 2005, July. A study of AdaBoost with SVM based weak learners. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on* (Vol. 1, pp. 196-201). IEEE.