EXTRACTING USEFUL INFORMATION AND BUILDING PREDICTIVE MODELS FROM MEDICAL AND HEALTH-CARE DATA USING MACHINE LEARNING TECHNIQUES

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Md Faisal Kabir

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Computer Science

June 2020

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

EXTRACTING USEFUL INFORMATION AND BUILDING PREDICTIVE MODELS FROM MEDICAL AND HEALTH-CARE DATA USING MACHINE LEARNING TECHNIQUES

By

Md Faisal Kabir

The supervisory committee certifies that this dissertation complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Simone A. Ludwig
Chair

Anne M. Denton
Co-Chair

Saeed Salem

Pratap Kotala

María de los Ángeles Alfonseca-Cubero

Approved:

June 20, 2020
Date

Kendall E. Nygard
Department Chair

ABSTRACT

In healthcare, a large number of medical data has emerged. To effectively use these data to improve healthcare outcomes, clinicians need to identify the relevant measures and apply the correct analysis methods for the type of data at hand. In this dissertation, we present various machine learning (ML) and data mining (DM) methods that could be applied to the type of data sets that are available in the healthcare area.

The first part of the dissertation investigates DM methods on healthcare or medical data to find significant information in the form of rules. Class association rule mining, a variant of association rule mining, was used to obtain the rules with some targeted items or class labels. These rules can be used to improve public awareness of different cancer symptoms and could also be useful to initiate prevention strategies.

In the second part of the thesis, ML techniques have been applied in healthcare or medical data to build a predictive model. Three different classification techniques on a real-world breast cancer risk factor data set have been investigated. Due to the imbalance characteristics of the data set various resampling methods were used before applying the classifiers. It is shown that there was a significant improvement in performance when applying a resampling technique as compared to applying no resampling technique.

Moreover, super learning technique that uses multiple base learners, have been investigated to boost the performance of classification models. Two different forms of super learner have been investigated - the first one uses two base learners while the second one uses three base learners. The models were then evaluated against well-known benchmark data sets related to the healthcare domain and the results showed that the SL model performs better than the individual classifier and the baseline ensemble.

Finally, we assessed cancer-relevant genes of prostate cancer with the most significant correlations with the clinical outcome of the sample type and the overall survival. Rules from the RNA-sequencing of prostate cancer patients was discovered. Moreover, we built the regression model and from the model rules for predicting the survival time of patients were generated.

ACKNOWLEDGEMENTS

First and foremost, all praise is due to Allah (God) alone. Thanks and appreciation to Allah for blessing me the strength to overcome all challenges during my Ph.D. dissertation.

I want to give my heartfelt thanks to my advisor, Dr. Simone A. Ludwig, for her endless support, advice, and encouragement during my Ph.D. study. Her continuous guidance helped me succeed in my exciting area of research. She has also been a great help in organizing and writing this research.

Besides, I would like to express my special thanks to my co-advisor, Dr. Anne M. Denton, for her support and guidance, especially during the early days of my Ph.D. study.

Moreover, I am grateful to the committee members, Dr. Saeed Salem, Dr. Pratap Kotala, and Dr. María de los Ángeles Alfonseca-Cubero, for their valuable comments and suggestions. I enjoyed discussing many aspects of this research with them and look forward to future discussions. They dedicated their valuable time on supporting my graduate study and helping on my professional developments.

I want to thank Dr. Kendall Nygard, chair department of computer science, for his valuable suggestions and support throughout my Ph.D. study. Likewise, I would like to thank the faculty and staff of the computer science department for their support.

I would also like to thank the graduate center for writers for helping me to improve my writing. Also, I want to thank Graduate school, college of science and mathematics for their support of my travel to attend conferences. Furthermore, I want to thank Boston Medical center for allowing me to work as an intern in the Summer of 2016, and learn by working with medical and public health professionals.

Finally, I thank my family members and friends here in the US and my home country Bangladesh. I would also like to express undescribable thanks to my wife, Humaira Rahman (Mouri), for her support and patience.

DEDICATION

I dedicate this research to my parents, family members, teachers, students, and my friends. I also love to dedicate this research to my newborn baby girl “Sahira Mahnisa Faisal”.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Machine Learning (ML) and data mining (DM) have become an integrative part of modern scientific methodology, providing insights about data and offering prediction based on historical observations. The use of DM and ML techniques require a reasonable understanding of their mechanisms, properties and constraints in order to better understand and interpret their results. Researchers have used ML and DM techniques in various fields of science, technology, and humanities, as in biology, meteorology, healthcare or finance [1].

Cancer has become one of the most devastating diseases worldwide, with more than 10 million new cases every year, according to the World Health Organization (WHO) [2]. The causes and types of cancer vary in different geographical regions, however, nearly every family in the world is touched by cancer. The disease burden is enormous, not only for affected individuals but also for their family as well as society. Detecting cancer early saves lives. According to WHO, 8.8 million people die from cancer each year, mostly in low- and middle-income countries. One problem is that many cancer cases are diagnosed too late. In addition, detecting cancer early also greatly reduces economic cost: not only is the cost of treatment much less in cancer's early stages, but people can also continue to work and support their families if they can access effective treatment in time.

DM has been widely used in the healthcare domain to extract knowledge in the form of rules. Public awareness of various disease/cancer symptomps can be taken and different preventation strategies could also be initiated by using these rules. ML and DM have been widely used to build predictive models from historical observations [3], [4], [5], [6]. These models can predict whether new patients are vulnerable to particular diseases or cancers. Performance of the ML model is very important and researchers are trying to use an appropriate model for a particular problem. However, choosing the best ML or DM model for a specific problem is a complex task. Due to this researchers are trying to use multiple models to obtain better performance. Effective use of DM and ML can contribute in early detection of many diseases including various cancers. For that, a detailed analysis needs to be performed before selecting a model for a specific task. Ultimately, by early detecting disease or cancer cases accurately, economic costs can be reduced and most importantly human lives can be saved.

Building an integrative models considering both clinical and genomic data simultaneously is a challenging task, however, it can provide vital information that is present in both data sets. In most of the cases, the goal is bio-marker discovery which is to find the clinical and genomic factors related to a particular disease phenotype such as cancer vs. no cancer, tumor vs. normal tissue samples, or continuous variables such as the survival time after a particular treatment. These models can help in the design of effective diagnostics, and novel drugs, which can lead us one step closer to personalized medicine.

The following sections briefly describe the research conducted. Brief description of the background are presented in Sections 1.1-1.4. The motivation of the work is discussed in Section 1.5. The contributions of the work is described in Section 1.6, and an overview of the dissertation is listed in Section 1.7.

1.1. Data Mining and Machine Learning Techniques

Machine Learning (ML) or Data Mining (DM) algorithms [7], [8] can be classified into supervised or unsupervised learning depending on the data. Supervised methods are used when there is a variable whose value has to be predicted. Such a variable is referred to as a response or output variable. For an unsupervised method, the data is not labeled and there is no value to predict or classify.

Supervised learning algorithms generate a function that is able to map the input/output values. In these algorithms the data provides examples about the kind of relationship between the input and output variables that has to be learnt. In unsupervised learning, there is no output value, but instead just a collection of input values. Supervised learning algorithms can be further divided into classification and regression algorithms.

Unsupervised is a learning method, in which an (output) unit is trained to respond to clusters of patterns within the input. In this paradigm the system is supposed to discover statistically salient features of the input population. Unlike the supervised learning paradigm, there is no a priori set of categories into which the patterns are to be classified; rather the system must develop its own representation of the input stimuli. Unsupervised algorithms seek out similarity between pieces of data in order to determine whether they can be characterized as forming a group. Apriori algorithm is another classic unsupervised algorithm which extracts association rules from a data set [9].

1.2. Knowledge Discovery

Knowledge in the form of rules can be extracted from various data sets using the association rule mining technique. The idea of association rules originated from the market basket analysis where a rule is sought to be like "When a customer buys a set of products what is the probability that he or she buys another product?" Mathematically, an association rule is defined as $A \Rightarrow B$ where A (antecedent) and B (consequent) are logical predicates constructed by Boolean predicates. A logical predicate in an association rule consists of one or more Boolean conditions and they are connected by the logical AND ($\wedge$) operator. In a transactional data set (e.g., sales database of a supermarket), an association rule appears as $(item = milk) \wedge (item = bread) \Rightarrow (item = butter)$, which means when a customer buys milk and bread it is most likely that he or she also buys butter. The likelihood of an association rule is measured by many values, e.g., support, confidence, lift, and so on.

Association rule mining [9] has been introduced in 1993, and since then it has attracted considerable attention. The discovery of association rules is an important component of data mining [10]. Association Rule Mining (ARM) has been widely used by the retail industry under the name "market-basket analysis". However, the concept of association rules is general and has wide applicability also in the medical domain [11], [12], [13], [14]. ARM can be applied to a cancer risk factors data set to discover hidden but significant rules that could be useful not only for medical professionals but also for health organizations. Rules can also be generated from data sets having specified target classes as their consequences under the name of class association rule mining.

1.3. Classification Techniques

Classification is the task of learning a target function $f(x)$ that maps each attribute set x into one of the pre-defined class labels y [15], [16]. The target function is also informally known as the classification model. The goal of classification is to predict quantitative or categorical outputs that assume values in a finite set of classes (e.g. Yes/No, Disease/No-disease or Green/Red/Blue etc.) without an explicit order. Categorical variables are also called Factors. In regression the output to predict is a real-valued number.

The problem of classification can be stated as "given a set of training data points along with associated training labels, determine the class label for an unlabeled test instance". Classification

algorithms typically contain two phases namely training and testing. In the training phase, a model is constructed from the training instances while in the testing phase, the model is used to assign a label to an unlabeled test instance. The different techniques that are commonly used for classification are tree based, rule-based methods, probabilistic methods, SVM methods, instance-based methods, neural networks, and so forth [17].

Decision trees (DT) create hierarchical partitioning of the data, which relates the different partitions at the leaf level to the different classes. Some of the methods for decision tree construction include c4.5, C5.0, ID3, and CART [18], [19] [20], [21]. Rule-based methods are closely related to decision trees, except they do not create a strict hierarchical partitioning of the training data. Here, overlaps are allowed in order to create greater robustness for the training model.

Probabilistic methods are the most fundamental among all data classification methods that use statistical inference to find the best class for a given example. In addition to simply assigning the best class, probabilistic classification algorithms will output a corresponding posterior probability of the test instance for each of the possible classes. Example of probabilistic algorithms for classification include Naive Bayes, logistic regression, Bayesian network construction [22], [23].

Instance-based learning [24], the training phase is omitted entirely, and the classification is performed directly from the relationship of the training instances and the test instances. This method also refereed as lazy learning because the knowledge of the test instance is acquired first in order to create a locally optimized model, which is specific to the test instance. An example of a very simple instance-based method is the nearest neighbor classifier.

The SVM (Support Vector Machine) classifier uses linear conditions in order to separate out the classes from one another. The idea is to use a linear condition that separates the two classes from each other as far as possible [25]. Neural Networks (NN) [17] attempt to simulate biological systems, corresponding to the human brain. In the human brain, neurons are connected to one another via points, which are referred to as synapses. In biological systems, learning is performed by changing the strength of the synaptic connections, in response to impulses. This biological analogy is retained in an artificial neural network (ANN). The basic computation unit in an ANN is a neuron or unit. These units can be arranged in different kinds of architectures by connections between them. The most basic architecture of a NN is a perceptron, which contains a

set of input nodes and an output node. Other variations of NN are multi-layer feed-forward NN, back-propagation, deep neural network (DNN) [26], and so forth.

Another classifier named ensemble learning generates multiple models for robustness, or combining the results of the same algorithm with different parts of the data. The general goal of the algorithm is to obtain more robust results by combining the results from multiple training models either sequentially or independently. Examples of ensemble classifiers are Bagging, Boosting, Random Forests, stacking, and so forth [27], [28], [29], [30].

Researchers are using these classification models or learning algorithms in various fields such as healthcare, network security, business, and so on. Researchers are trying to find which algorithm will perform well for a particular research problem and the available data at hand.

1.4. Machine Learning and Data Mining Techniques in Healthcare

As stated above, cancer has become one of the most devastating diseases worldwide, with more than 10 million new cases every year, according to WHO [2]. In general, healthcare institutions are becoming more and more dependent on advances in technology, and the use of DM and ML techniques can provide useful support to assist physicians. In the last decade, ML contributed to healthcare by improving not only service quality and care but also saving human lives by detecting diseases/cancer cases early.

In the healthcare and bio-medical domain, current technologies are generating and collecting large volumes of data and extracting useful information from these huge data sets is the key. Rules are very natural for knowledge representation since people can understand and interpret them easily. In several studies, knowledge in the form of rules has been extracted from the medical domain [11], [12], [13], [14].

In addition, in most cases the data sets that are available for analysis contain irrelevant features, noise in the data, imbalance characteristics, that makes the data too complex to be analyzed using traditional methods. In some cases, particularly in the healthcare and bio-medical domain, the lack of qualitative training data is also a common problem in ML since the training data is a critical resource to build classifiers. As a result, the scarcity of qualitative training data is also the most common problems which leads to poor classification accuracy. For that reason, proper data analysis is necessary before applying a classification model to enhance the model's performance.

In ML, classification is applied to most of the application areas. ML approaches have been widely applied in many domains including healthcare. Several prediction models have been extensively investigated and have been successfully deployed in clinical practice [26]. Clinical data refers to a broad category of a patient's pathological, behavioral, demographic, familial, environmental, medication history, and so forth. The choice of the model to be used for a particular healthcare problem primarily depends on the outcomes to be predicted. In addition, understanding the problem clearly and the domain knowledge is the key of selecting the best algorithm by which better performance of the classification model can be obtained.

Human diseases are inherently complex in nature and are usually governed by a complicated interplay of several diverse underlying factors, including different genomic, clinical, behavioral, and environmental factors. It is essential to build integrative models considering both genomic and clinical data simultaneously so that they can combine the vital information that is present in both clinical and genomic data [31]. Such models can help in the design of effective diagnostics, new therapeutics, and novel drugs, which will most likely lead us one step closer to personalized medicine. This opportunity has led to an emerging area of integrative predictive models that can be built by combining clinical and genomic data, which is called clinico–genomic data integration. Clinical data refers to a broad category of a patient's pathological, behavioral, demographic, familial, environmental and medication history, while genomic data refers to a patient's genomic information including SNPs (single nucleotide polymorphisms), gene expression, protein and metabolite profiles. In most of the cases, the goal of the integrative study is biomarker discovery which is to find the clinical and genomic factors related to a particular disease phenotype such as cancer vs. no cancer, tumor vs. normal tissue samples, etc.

1.5. Motivation and Problem Definition

Prevention of major types of cancer through a quantified assessment of risk is a major concern in order to decrease its impact on our society. Identifying risk factors of various cancers is important whereby physicians can inform the patients about the potential cancer risks from the risk factors and suggest preventive measures. It is also more important to extract important knowledge from these available risk factors in the form of rules.These rules could be useful for better healthcare as medical professionals or other health related organizations can develop policies to identify and prevent its impact in early stage.

In the field of ML, classification is a common problem and has been widely applied to various application domains including healthcare, cyber-security, geographic information system, businesses, and so forth. However, the performance of learning algorithms strongly depend on sufficient qualitative training data to build an accurate model and make prediction on future or unseen data. Nonetheless, in real-life settings, particularly in the healthcare and bio-medical domain, obtaining useful training data (as generally data contains missing values, irrelevant feature, and so forth) has been a major bottleneck of making effective prediction models that can be applied in practice. To overcome these issues, it is more important to understand each problem carefully and select the appropriate technique accordingly. In addition to selecting the appropriate model, searching/choosing the best hyper-parameter setting for a particular problem is the key for achieving better performance.

Another challenge of obtaining better classification performance is the imbalance characteristics of data, meaning that there are significantly more samples for one category than the other. For that, data needs to be analyzed and preprocessed correctly before appropriate ML algorithms can be applied for better performance. Finally, building integrative models considering both clinical and genomic data simultaneously can provide vital information that is present in both data sets, is a challenging task. These models can help in the design of effective diagnostics, and novel drugs, which can lead us one step closer to personalized medicine. For that, a prediction model of survival for prostate cancer patients is proposed, thus, the findings of this study can be used to determine predictors of survival outcomes for other cancer types.

The main motivations of this research can be summarized as follows.

- Discovery of association rules is an important component of data mining. Association Rule Mining (ARM) has been widely used by the retail industry under the name "market-basket analysis". However, the concept of association rules is general and has wide applicability also in the medical domain [11], [12], [13], [14].Rules provide a concise statement of potentially important information that is easily understood by end users. By using these rules medical professionals or other health related organizations can develop strategies to identify and prevent its impact in the early stage. For this purpose, rule discovery from breast cancer risk

factors using association rule mining is developed. The techniques can be used to find rules for other cancer types.

- Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as this has direct influence in their daily practice and clinical service. A reliable prediction will help oncologists and other clinicians in their decision-making process and allow clinicians in choosing the most reliable and evidence-based treatment and prevention strategies for their patients. The data set for this study has imbalance characteristics, meaning there are significantly more samples for one category than the other. For this purpose, several resampling techniques were used before developing a classification model of breast cancer risk factors to obtain better performance.

- Achieving better performance for a ML model on the available data sets is the key and researchers are generally using appropriate single classifiers.However, selecting the best data mining or machine learning model for a specific problem is complex. Due to this, researchers are also using multiple different models for a particular problem to obtain better performance. Super learning or stacked ensemble is a ML technique that finds the optimal weighted average of diverse learning models and generally provides better performance compared to individual base learners. For this reason, enhancing the performance of classification using super learning is developed.

- Early detection of cancer produces an increase in survival rate, and consideration of clinical variables along with RNA-Sequencing data can be utilized to increase efforts at the early detection of cancer. The genes with the greatest survival correlation can be useful for analysis. The main motivation is to determine which clinical variables and RNA-Sequencing expression levels predict clinical outcomes, such as survival for various types of cancer. This study focused on prostate cancer patients, but the findings of this study have implications for determining predictors of survival outcomes for various cancer types.

1.6. Contributions

This dissertation makes several contributions towards better healthcare using data mining, and machine learning techniques. Useful knowledge in a form of rules using DM techniques has

been discovered from healthcare and medical data. Building predictive models using several ML techniques has been proposed to enhance the performance of classification models. Finally, extracting knowledge from bio-medical data and building predictive models is proposed. The contributions can be summarized as follows:

1. Hidden but important rules from a breast cancer risk factors data set was discovered using an association rule mining technique. In addition, the logit model was used to check the statistical significance of all risk factors or predictors. Rules or knowledge for both breast cancer and non-breast cancer patients were discovered to understand and compare the characteristics of both groups. These rules can be useful for developing strategies to prevent its impact in the early stages. Details of this work is discussed in Chapter 2.

2. Three different classification techniques were applied to a breast cancer risk factors data set to attain better performance. The data set is highly imbalanced meaning that the data has an unequal distribution between the classes. For this purpose, several resampling methods were used before applying different classifiers to achieve better performance. Details of this work is discussed in Chapter 3.

3. To enhance the performance of a classification model super learning (SL) or stacked-ensemble technique was used on four benchmark data sets that are related to healthcare. SL uses two or more machine learning algorithms as base learners that finds the optimal combination of a collection of prediction algorithms. Three supervised learning algorithms were selected as base learners and a meta learner was used. The performance of the proposed technique was compared to the individual base learners and the baseline ensemble. Details of this work is discussed in Chapter 4.

4. Cancer relevant genes of prostate cancer with the most significant correlations with the clinical outcome of the sample type (cancer / non-cancer) and the overall survival (OS) were assessed. Rules from the RNA-sequencing of prostate cancer patients was discovered from a decision tree classifier. Moreover, the regression model was built using a decision tree regressor and from the model rules for predicting or estimating the survival time of patients were generated. Details of this work is described in Chapter 5.

1.7. Dissertation Overview

This dissertation is a paper-based version, where each chapter has been derived from the papers published during the Ph.D. work. This is an overview of the remaining chapters of this dissertation.

In Chapter 2, rule discovery from breast cancer risk factors using association rule mining is discussed. The chapter is derived from the publication:

- Md Faisal Kabir, Simone A. Ludwig, and Abu Saleh Abdullah. "Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining." IEEE International Conference on Big Data (Big Data), 2018.

In Chapter 3, the classification of breast cancer risk factors using several resampling approaches is described. The chapter is derived from the publication:

- Md Faisal Kabir and Simone A. Ludwig. "Classification of Breast Cancer Risk Factors Using Several Resampling Approaches." 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018.

In Chapter 4, enhancing the performance of classification using super learning is discussed. The chapter is derived from the publication:

- Md Faisal Kabir and Simone A. Ludwig. "Enhancing the Performance of Classification Using Super Learning." Data-Enabled Discovery and Applications 3.1 (2019):5, Springer International Publishing.

In Chapter 5, the classification models and survival analysis for prostate cancer using RNA sequencing and clinical data is discussed. The chapter is derived from the publication:

- Md Faisal Kabir and Simone A. Ludwig. "Classification Models and Survival Analysis for Prostate Cancer Using RNA Sequencing and Clinical Data." IEEE International Conference on Big Data (Big Data), 2019.

In Chapter 6, the conclusion and future research is presented.

2. RULE DISCOVERY FROM BREAST CANCER RISK FACTORS USING ASSOCIATION RULE MINING

Breast cancer is the most common cancer in women worldwide. Prevention of breast cancer through risk factors reduction is a significant concern to decrease its impact on the population. Attaining or detecting significant information in the form of rules is the key to prevent breast cancer. Our objective is to find hidden but important knowledge of the form of rules from the risk factors data set of breast cancer. Mining rules is one of the vital tasks of data mining as rules provide concise statement of potentially important information that is easily understood by end users. In this chapter, we use association rule mining, a data mining technique to attain information in the form of rules from breast cancer risk factors data that could be useful to initiate prevention strategies. We discovered rules of both breast cancer and non-breast cancer patients so that we can understand and compare the characteristics of both breast cancer and non-breast cancer individuals. The experimental results show that generated or mined rules hold the highest confidence level.

The rest of the chapter is structured as follows. The related work is discussed in Section 2.1. The preliminaries including data description, data pre-processing, and the problem statement are described in Section 2.2. The analytical workflow are discussed in Section 2.3. In this section, the binary logit model and association rule mining is discussed. In Section 2.4, experiments and results are shown. The outputs obtained from the logit model is discussed and presented. Also, the rule generation using the association rule mining technique is also shown in this section. Moreover, important rules along with their interpretation are listed in this section. In Section 2.5, discussion is presented. Finally, summary and possible future directions are discussed in Section 2.6.

2.1. Related Work

Researchers have developed different models for breast cancer risk prediction and association between risk factors [32], [33], [34], [35]. In [32], authors applied statistical methods to show a positive association between Hormone Replacement Therapy (HRT) and breast cancer risk, although this relationship varies according to race/ethnicity, BMI (Body Mass Index), and breast

density. The Gali model is used to estimate the number of expected breast cancers for white females who are examined annually [33]. In [34], the authors used commonly identified risk factors such as race/ethnicity, breast density, BMI, and use of hormone therapy, type of menopause, and previous mammographic results to improve the model. In [35], the Breast cancer risk score is determined using a data mining approach called k-nearest-neighbor (KNN) to improve readability for physician and patients. In addition, authors [35] tried to get higher risk detection performances and impact levels of each risk factor.

Association rule mining has been used in the medical domain to find useful information from the data. In [11], authors used the ARM technique for generating the rules for heart disease patients. Based on the rules they discovered the factors which cause heart problems in men and women. In [12], the authors implemented the ARM based concept for finding co-occurrences of diseases carried by a patient using a healthcare repository. The authors extracted data from a patients' healthcare database and from that they generated association rules. Class association rule mining has also been used in the literature to discover the characteristics features [36]. A class association rule set is a subset of association rules with the specified classes as their consequents [37]. In traditional association rule mining, if the support value is kept too low, the class association rule mining will generate overfitting rules for frequent or majority classes; while keeping support value high will not generate sufficient rules for infrequent or minority classes. In class association rule mining this is not the case since mining is done according to the class, the algorithm is not influenced by the unequal distribution between the classes (imbalanced class).

In this research, we used a risk factors data set from the Breast Cancer Surveillance Consortium (BCSC) [38] to examine significant rules of breast cancer and non-breast cancer patients. Rules of breast cancer patients can be useful for physicians to make informed decision as they have to inform patients about risk factors and alert patients about the potential risks of developing breast cancer (if any). This way, a prevention program or process can be initiated in the early stage of disease progression.

2.2. Preliminaries

2.2.1. Data Description

The data set includes information from $6,318,638$ mammography examinations obtained from the Breast Cancer Surveillance Consortium (BCSC) database collected from January 2000 to

December 2009 [38]. Data for this study was obtained from the BCSC Data Resource and more information is available at http://www.bcsc-research.org.

2.2.2. Data Pre-processing

The data is aggregated such that the total number of instances or records is $1,144,565$, with 13 attributes or columns. The data set also contains missing or unknown values denoted by 9. To build a reliable model, we discarded the records containing at least one missing or unknown value. We also removed the attribute year that represents the calendar year of the observation. After discarding these records and one attribute, there are $219,524$ available records with 12 attributes. In the data set, there is an attribute named count, representing the number of records that have the combination of variable-values shown in the row. For instance, the value of the count column for the particular row is 12. It indicates that there were 12 similar records; the same as that particular row in the original data. For that reason, we created the number of rows or records the same as the count value in the original data set, and discarded the count column after that. Finally, there are a total of $1,015,583$ records with 11 attributes for building the model. Among $1,015,583$ records, $60,800$ individuals have prior breast cancer, and $954,783$ are non-breast cancer individuals. Among the 11 attributes, "prior breast cancer" values yes or no is considered as response or class variable and the remaining 10 attributes are considered as explanatory or predictors or independent variables. The distribution of all features are shown in Table 2.1 through Table 2.10. Bar plots of the age group, age first birth, BMI group, and breast cancer history are shown in Fig. 2.1, Fig. 2.2, Fig. 2.3, and Fig. 2.4 respectively.

Table 2.1. Distribution of race/ethnicity.

Race/Ethnicity	**Count**
Non-Hispanic-White	902736
Asian_or_Pacific Islander	39139
Hispanic	35451
Other_or_Mixed	20972
Non-Hispanic-Black	14389
Native American	2896

Table 2.2. Distribution of hormone replacement therapy (HRT)

HRT	Count
No	849225
Yes	166358

Table 2.3. Distribution of age group

Age_group_range	Count
age_55_59	168659
age_50_54	168158
age_45_49	146665
age_60_64	127459
age_40_44	115237
age_65_69	93919
age_70_74	72315
age_75_79	53983
age_80_84	29750
age_35_39	21841
age_greater_equal__85	12557
age_30_34	4113
age_18_29	927

Table 2.4. Distribution of menopausal status

Menopaus	Count
Post_menopausal	687566
Pre_or_peri_menopausal	292699
Surgical_menopause	35318

Table 2.5. Distribution of body mass index (BMI)

BMI_range	Count
10-to-lessThan_25	430102
25-to-lessThan_30	310555
30-to-lessThan_35	161785
35-or-above+	113141

Table 2.6. Distribution of BI-RADS breast density

BIRADS_breast_density	**Count**
Scattered_fibroglandular_densities	429488
Heterogeneously_dense	414732
Almost_entirely_fat	90005
Extremly_dense	81358

Table 2.7. Distribution of age first birth

Age_first_birth	**Count**
Age_20_24	331615
Age_25_29	216877
Nulliparous	166180
Age_less_20	157723
Age_greater_equal_30	143188

Table 2.8. Distribution of first degree relative

First_degree_relative	**Count**
No	824472
Yes	191111

Table 2.9. Distribution of previous breast biopsy

biopsy	**Count**
No	724364
Yes	291219

Table 2.10. Distribution of prior breast cancer diagnosis

breast_cancer_history	**Count**
No	724364
Yes	291219

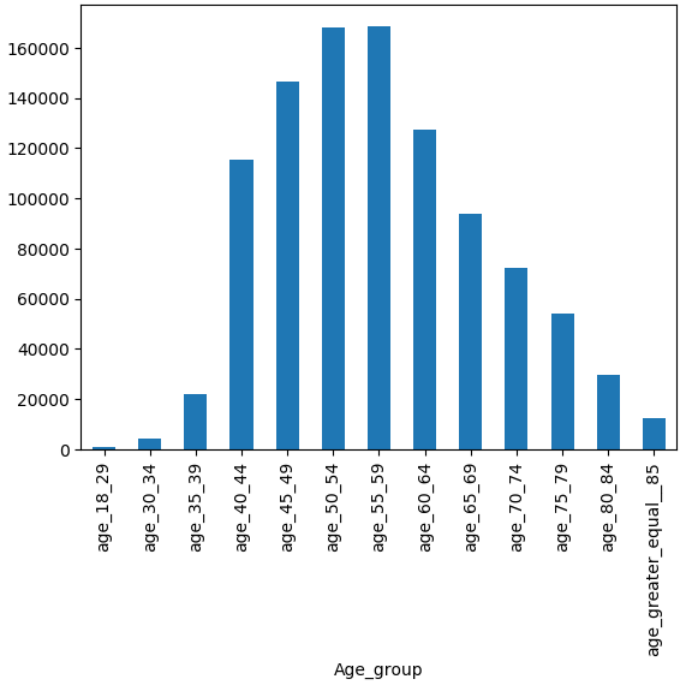

Figure 2.1. Bar graph of age group for BCSC risk factors data.

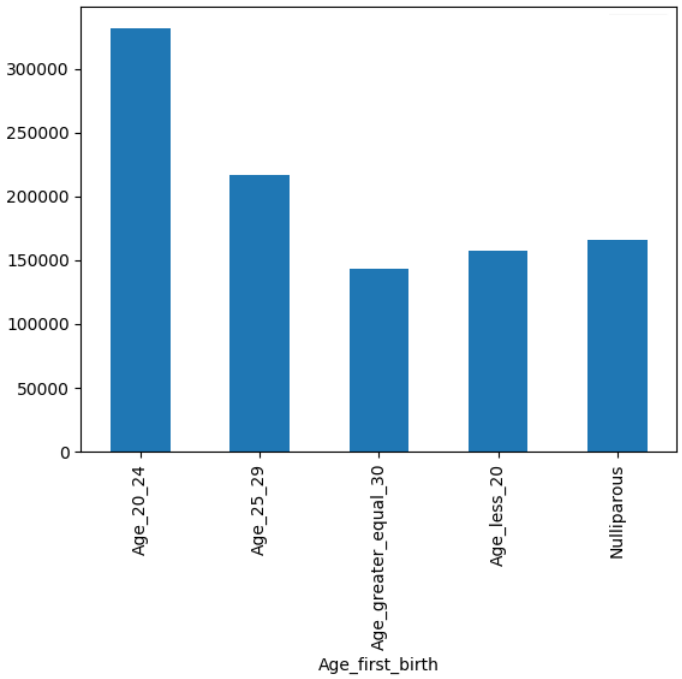

Figure 2.2. Bar graph of age first birth for BCSC risk factors data.

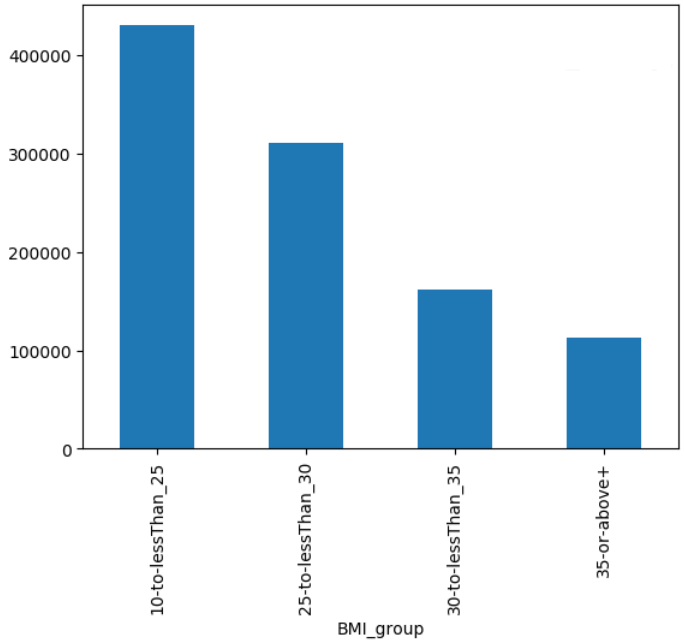

Figure 2.3. Bar graph of BMI group for BCSC risk factors data.

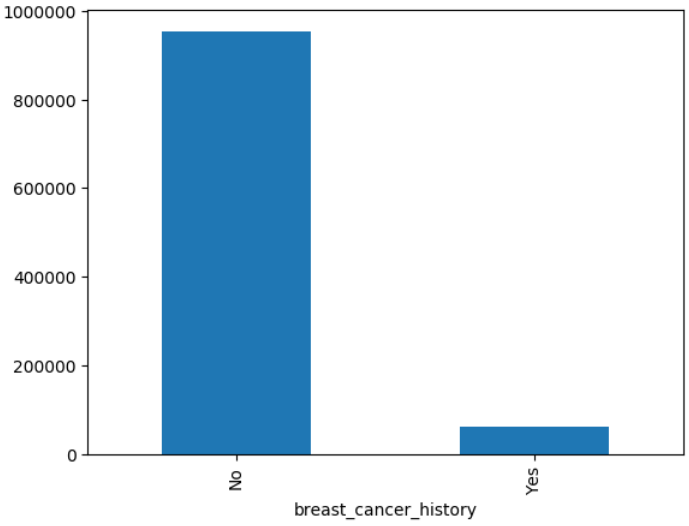

Figure 2.4. Bar graph of prior breast cancer for BCSC risk factors data.

2.2.3. Conversion of Data Set into Transaction-like Database

For association and class rule mining, the data set has been converted into transactions. For instance, for feature such as race or ethnicity there were a total of six values namely non-Hispanic white, non-Hispanic black, Asian, native American, Hispanic, and mixed/other; for that six columns have been created accordingly with values Yes or No. For example, if an individual is a Native American, then Yes or 1 would be in the corresponding column and the remainder would be No or 0. This way, a total of 46 columns have been created. So, in total there were 1015583 records and 46 items or columns.

2.2.4. Problem Statement

Let, $P = \{p_1,p_2,p_3,...,p_n\}$ be the set of n patients and $D = \{d_1,d_2,d_3,...,d_m\}$ be the characteristics of patients, where m is the number of attributes of the patients. We define, $C = \{c_1,c_2\}$ be the class information or the breast cancer history (yes or no) of patients. In this research, we are interested in finding the relationships among breast cancer risk factors. More specifically, we are interested to find the characteristics or rules in terms of risk factors of both the breast cancer and non-breast cancer individuals (i.e. $\{d_1, d_3, d_6\} \Rightarrow c_1$ and $\{d_2, d_5, d_7\} \Rightarrow c_2$).

2.3. Analytical Workflow

In this section, we provide an overview of our framework. First, we used the logit model on the Breast Cancer Surveillance Consortium (BCSC) data set to identify appropriate factors that may affect the likelihood of breast cancer. After that we applied association rule mining and class association rule mining on these risk factors to find significant rules of both non-breast cancer and breast cancer patients.

2.3.1. Logit Model

In the current study, the dependent attribute of breast cancer (Yes or 1) or no breast cancer (No or 0) is dichotomous and thus represented as a binary variable. The binary logit model is extensively used in breast cancer investigations where the response variable is binary [39]. The model takes the natural logarithm of the likelihood ratio such that the dependent variable is 1 (breast cancer) as opposed to 0 (no breast cancer). Let, p_1 and p_0 represents the probabilities of

the response to variable categories breast cancer and no breast cancer, respectively. The binary logit model is given as:

$$Y = log\left[\frac{p_0}{p_1}\right] = \alpha + \beta_i X_i \tag{2.1}$$

where Y is the Binary response or class variable; α is the intercept to be calculated; β_i is the estimated vector of parameters, and X_i is the vector of independent variables.

In Equation (1), the maximum likelihood estimation technique is used to estimate the parameters. The unit increase in the independent variables X_i, while keeping all the remaining factors constant, will result in the increase of the likelihood ratio by $exp(\beta_i)$. This states that the relative magnitude by which the response outcome (breast cancer) will increase or decrease, while considering a one-unit increase in the explanatory variable. The probability of breast cancer (p_1) is given by:

$$p_1 = \frac{exp(\alpha + \beta_i X_i)}{1 + exp(\alpha + \beta_i X_i))} \tag{2.2}$$

Similarly, the probability of no breast cancer (p_0) is given by:

$$p_0 = \frac{1}{1 + exp(\alpha + \beta_i X_i)} \tag{2.3}$$

We used the logit model to identify and select appropriate factors that may affect the likelihood of breast cancer.

2.3.2. Association Rule Mining

Association Rule Mining (ARM) is one of the key techniques to discover and extract useful information from a large data set. Mining association rules [7] can formally be defined as: Let $I = \{i_1, i_2, i_3, ..., i_n\}$, be a set of n binary attributes called items, and Let, $D = \{t_1, t_2, t_3, ..., t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$. The sets of items or item sets X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule, respectively. Often rules are restricted to only a single item in the consequent.

Association rules are rules which surpass a user-specified minimum support and minimum confidence threshold. The support $supp(X)$ of an item set X is defined as the proportion of transactions in the data set, which contain the item set and confidence of a rule as defined as:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{2.4}$$

Therefore, an association rule $X \rightarrow Y$ will satisfy $supp(X \cup Y) \geq \phi$ and $conf(X \rightarrow Y) \geq \delta$, which are the minimum support and minimum confidence, respectively. Minimum confidence can be interpreted as the threshold on the estimated conditional probability, the probability of finding the RHS of the rule in the transactions under the condition that these transactions also contain the LHS. Another popular measure for association rules used throughout this research is lift [40]. The lift of a rule is defined as:

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)} \tag{2.5}$$

It can be interpreted as the deviation of the support of the whole rule from the support expected under independence given the support of both sides of the rule. Greater lift values ($>> 1$) indicate stronger associations. Measures like support, confidence, and lift are generally called interest measures because they help with focusing on potentially more interesting rules. For example, consider a rule such as $\{milk, sugar\} \Rightarrow \{bread\}$ with support of 0.1, confidence of 0.9, and lift of 2. Now, we know that 10% of all transactions contain all three items together, thus the estimated conditional probability of seeing bread in a transaction under the condition that the transaction also contains milk and sugar is 0.9; and we see the items together in transactions at double the rate we would expect under independence between the item sets milk, sugar and bread [41].

Rules can be generated from data sets having specified classes as their consequences under the name of class association rule mining. These rules have the form $\{A_1, A_2, A_3, ..., A_n \Rightarrow class\}$. The objective here is to focus on using exhaustive search techniques to find all rules with the specified classes as their consequences that satisfy support and confidence [37]. Appropriate values of support and confidence is the key for generating rules since keeping a very low support value will generate large rules and if the support value is too high, we may lose rare but important

rules. In this research, we generated rules from the data set having specified classes such as rules or characteristics of patients who have prior breast cancer. We also generated or mined rules for non-breast cancer individuals. Our goal is to find rules or characteristics rules for these two groups.

2.4. Experiments and Results

Results of the logit model and association rule mining are discussed in this section. Association rule mining and class association rule mining has been applied on the data set. By selecting the optimum value of support and confidence, we mined strong rules for both breast cancer, and non-breast cancer patients. In this section, we also interpret few strong rules for both groups.

2.4.1. Output of Logit Model

The binary logit regression model was used to estimate the coefficients of significant explanatory variables in the final model. The software package SAS was used for the model development. For the model, all attributes were used as input for the likelihood of breast cancer. Interestingly, all explanatory variables turned out to be statistically insignificant ($p < 0.0001$). Table 2.11 shows the predictor variables which are significant at the corresponding significance levels in the binary logit model, which can contribute to the likelihood of breast cancer.

Table 2.11. Predictor variables with corresponding p values.

Parameter	**DF**	**Estimate**	**Standard Error**	**ChiSq**	**Pr >ChiSq**
Intercept	1	-9.1986	0.0544	28589	<.0001
Age_group	1	0.223	0.00228	9580	<.0001
Race_eth	1	0.0376	0.00463	66	<.0001
First_degree _relative	1	0.1068	0.0109	95	<.0001
Age_menarche	1	0.0259	0.00651	16	<.0001
Age_first_birth	1	0.0729	0.00375	377	<.0001
BIRADS_breast _density	1	-0.1035	0. 00682	230	<.0001
HRT	1	-1.9993	0.0238	7052	<.0001
Menopaus	1	0.4206	0.0132	1009	<.0001
BMI group	1	-0.0164	0.00512	10	0.0014
biopsy	1	5.511	0.0386	20417	<.0001

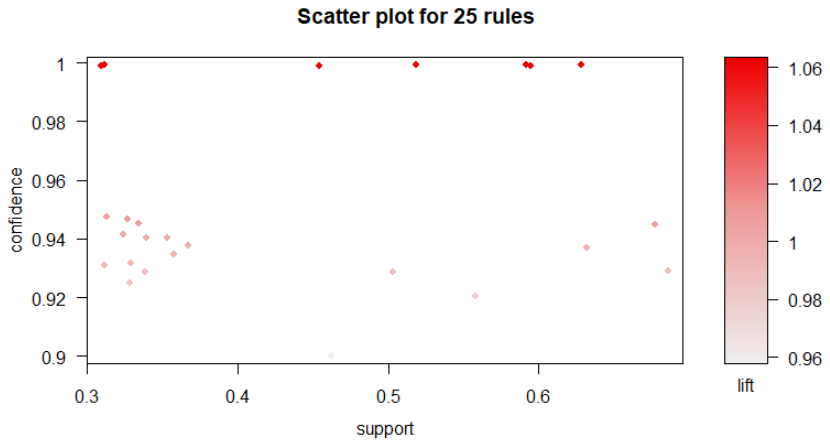

Figure 2.5. Scatter plot of 25 rules with minimum support, and confidence of 30% and 80%, respectively.

Positive values of coefficients express that the probability of breast cancer will increase by a certain amount for the specific predictor variables. Interestingly, all explanatory variables are significant at $p < .0001$ except the BMI group which is significant at .0014. From the table it can be referred that age group, race, first degree relatives, age menarche, age first birth, menopause, and biopsy has a positive relationship with previous breast cancer history. However, BIRADS breast density, HRT, and BMI group have negative relationship with breast cancer history.

2.4.2. Rules Generation from BCSC Risk Factors Data Set

Our goal is to extract characteristics of patients who have prior breast cancer and who do not have breast cancer. For that, we generated rules using the association rule technique with the specified support and confidence. We defined the consequent of a rule so that we can get our target rules that represent the characteristics of the patients who have breast cancer ($Breast\_cancer\_history = Yes$) or who do not have breast cancer ($Breast\_cancer\_history = No$). Support and confidence play an important role in rule generation. Initially, we set the minimum values of support and confidence to 30% and 80%, respectively. Also, we set the minimum length to 3, which means that the generated rules should have at least three items including the consequent. With these specified parameters the algorithm generated 37 rules and after pruning redundant rules we got 25 rules. The scatter plot of these 25 rules are shown in Fig. 2.5. From these 25 rules, 11 rules whose lift values are greater than or equal to one are shown in Table 2.12 sorted by higher lift value with corresponding support, and confidence. The software R was used for the experiments.

Table 2.12. Rules generated using the association rule technique with minimum support, and confidence value 30% and 80%, respectively.

SL	Rules	Supp. (%)	Conf. (%)	Lift
1	{Race=Non-Hispanic-White, First_degree_relative =No, biopsy=No} =>{breast_cancer_history= No}	52	99	1.06
2	{Age_menarche=Age_12_13, biopsy=No} =>{breast_cancer_history=No}	31	99	1.06
3	{First_degree_relative=No, biopsy=No} =>{breast_cancer_history=No}	59	99	1.06
4	{Race=Non-Hispanic-White, biopsy=No} =>{breast_cancer_history=No}	63	99	1.06
5	{HRT=No, biopsy=No} =>{breast_cancer_history=No}	60	99	1.06
6	{BIRADS_breast_density= scattered_fibroglandular_densities, biopsy=No} =>{breast_cancer_history=No}	31	99	1.06
7	{Menopaus=post_menopausal, biopsy=No} =>{breast_cancer_history=No}	45	99	1.06
8	{First_degree_relative=No, BIRADS_breast_density = Heterogeneously_dense} =>{breast_cancer_history=No}	31	95	1.01
9	{First_degree_relative=No, BMI_group= 10-to-lessThan_25} =>{breast_cancer_history=No}	33	95	1.01
10	{First_degree_relative=No, Age_menarche= Age_12_13} =>{breast_cancer_history=No}	33	95	1.01
11	{Race=Non-Hispanic-White, First_degree_relative =No} =>{breast_cancer_history=No}	68	95	1.01

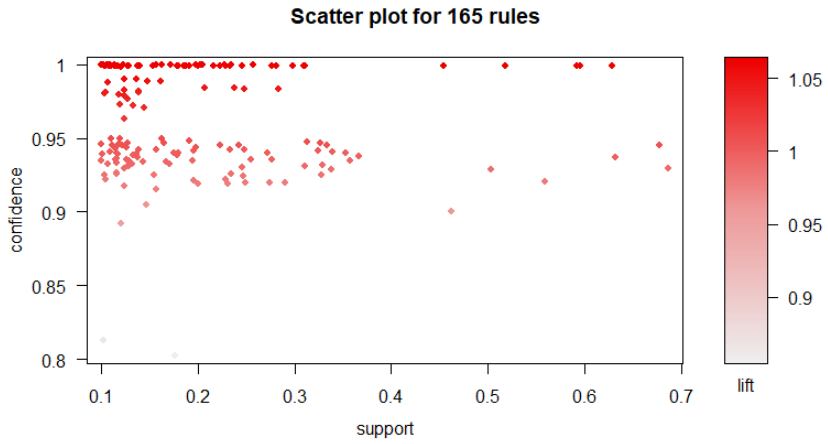

Figure 2.6. Scatter plot of 165 rules with minimum support and confidence of 10% and 80%, respectively.

It is worth to mention that we did not obtain any rules of patients who have prior breast cancer ($Breast\_cancer\_history = Yes$) for the specified support and confidence. This is due to the given values of support, and confidence; also a very small number of instances in which patients have breast cancer compared to their counterpart (ratio is about 1:16).

To obtain the rules of patients having breast cancer we set support to 10% and keep the confidence the same as before (80%). After pruning the redundant rules, we have 165 rules. The scatter plot of these rules is shown in Fig. 2.6. We still did not obtain any rules having the consequent equals to Yes, which means rules of breast cancer patients.

After several experiments, we assigned the value of support to 0.001% but a high confidence value of 90%, and obtained 67 rules. Here, we set the consequent or class value to Yes ($breast\_cancer\_history = Yes$) so that we can get the rules of breast cancer patients only. The scatter plot of these 67 rules is shown in Fig. 2.7. And from these 67 rules, the top 10 rules sorted by lift are shown in Table 2.13.

2.4.3. Generating Strong Rules

We obtained many rules using our methods described earlier. Here, we show a few rules for both breast cancer and non-breast cancer patients that are strong or important as they have higher confidence and lift values. Strong rules of both non-breast cancer patients and breast cancer patients are shown in Table 2.14 and Table 2.15, respectively.

Table 2.13. Rules generated using association rule technique with minimum support and confidence of 0.001% and 90%, respectively and consequent fixed for breast cancer patients only.

Rules	**Supp.** (%)	**Conf.** (%)	**Lift**
{Age_group=age_greater_equal_85, Race=Hispanic, Age_first_birth=Age_less_20, BIRADS_breast_density = Almost_entirely_fat, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	100	16.7
{Age_group=age_75_79, Race=Non-Hispanic-Black, Age_first_birth =Age_20_24, BMI_group=35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Age_group=age_greater_equal_85, Race=Non-Hispanic-White, First_degree_relative=No, Age_first_birth=Nulliparous, BMI_group=35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Age_group=age_75_79, First_degree_relative=Yes, Age_first_birth=Nulliparous, BIRADS_breast_density= Almost_entirely_fat,BMI_group=25-to-lessThan_30, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Age_group=age_75_79, Race=Asian_or_Pacific Islander, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, BIRADS_breast_density=Heterogeneously_dense, HRT=No, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Race=Asian_or_Pacific Islander, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, Age_first_birth= Age_less_20, BIRADS_breast_density= scattered_fibroglandular_densities, HRT=No, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Race=Asian_or_Pacific Islander, First_degree_relative=Yes, Age_first_birth=Age_less_20, BIRADS_breast_density= scattered_fibroglandular_densities, HRT=No, BMI_group= 25-to-lessThan_30, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	99	16.7
{Race=Hispanic, First_degree_relative=Yes, Age_menarche= Age_greaterEqual_14, Age_first_birth=Nulliparous, BIRADS_breast_density= Heterogeneously_dense, HRT=No, Menopaus=post_menopausal, BMI_group=10-to-lessThan_25, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	95	16.7
{Age_group=age_80_84, First_degree_relative=Yes, Age_first_birth=Age_25_29, BIRADS_breast_density= scattered_fibroglandular_densities, BMI_group=35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.002	95	15.66
{Age_group=age_80_84, First_degree_relative=Yes, Age_menarche=Age_greaterEqual_14, Age_first_birth= Age_less_20, BIRADS_breast_density= scattered_fibroglandular_densities, BMI_group=25-to-lessThan_30, biopsy=Yes} =>{breast_cancer_history=Yes	0.002	95	15.66

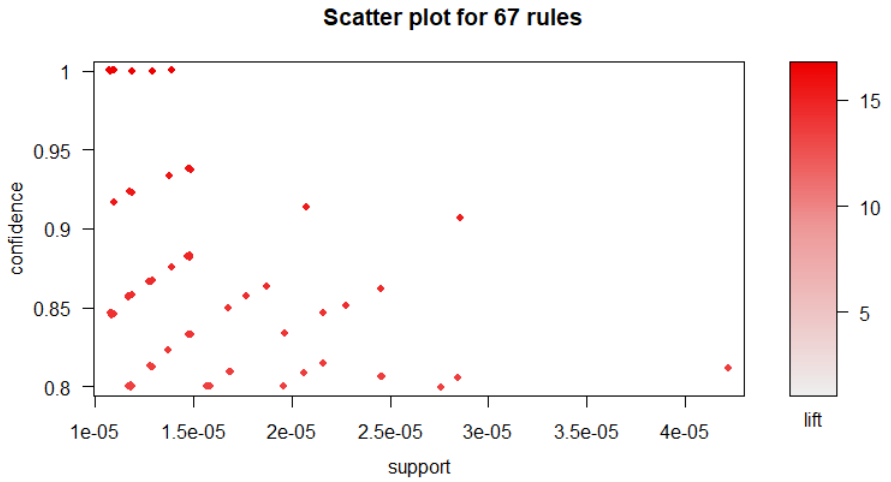

Figure 2.7. Scatter plot of 67 rules with minimum support (0.001%) and confidence (90%) when the consequent is fixed for breast cancer patients only ($breast\_cancer\_history = Yes$).

Table 2.14. Strong rules for non-breast cancer patients with corresponding support, confidence, and lift values.

SL	Rules	Supp. (%)	Conf. (%)	Lift
1	{Race=Non-Hispanic-White, First_degree_relative =No, biopsy=No} =>{breast_cancer_history=No}	52	99	1.062
2	{Race=Non-Hispanic-White, First_degree_relative =No} =>{breast_cancer_history=No}	68	95	1.005
3	{Age_menarche=Age_12_13, biopsy=No} =>{breast_cancer_history=No}	31	99	1.063
4	{First_degree_relative=No, BMI_group=10-to-lessThan_25} =>{breast_cancer_history=No}	33	95	1.007

2.4.4. Interpreting Strong Rules

Rule 1 of Table 2.14 can be interpreted as "If a person is a non-Hispanic white with no breast cancer of first degree relatives, and has not had a previous breast biopsy then the individual is a non-breast cancer patient". Rule 4 can be interpreted as "If a person's first-degree relatives do not have breast cancer, and a person's BMI range is between 10 and 25 then the individual is a non-breast cancer patient".

Table 2.15. Strong rules for breast cancer patients with corresponding support, confidence, and lift values.

Rules	**Supp. (%)**	**Conf. (%)**	**Lift**
{Age_group=age_greater_equal__85, Race=Hispanic, Age_first_birth=Age_less_20, BIRADS_breast_density =Almost_entirely_fat, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	100	16.70
{Age_group=age_75_79, Race=Non-Hispanic-Black, Age_first_birth=Age_20_24, BMI_group= 35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	100	16.70
{Age_group=age_greater_equal__85, Race =Non-Hispanic-White, First_degree_relative=No, Age_first_birth=Nulliparous, BMI_group =35-or-above+, biopsy=Yes} =>{breast_cancer_history=Yes}	0.001	100	16.70
{Age_group=age_75_79, First_degree_relative=Yes, Age_first_birth=Nulliparous, BIRADS_breast_density = Almost_entirely_fat, BMI_group=25-to-lessThan_30, biopsy=Yes}=>{breast_cancer_history=Yes}	0.002	100	16.70
{Race=Asian_or_Pacific Islander, First_degree_relative =Yes, Age_menarche=Age_greaterEqual_14, Age_first_birth= Age_less_20, BIRADS_breast_density= scattered_fibrogland ular_densities, HRT=No, biopsy=Yes} =>{breast_cancer_history=Yes}	0.002	100	16.70

We can interpret Rule 1 of Table 2.15 as "If a patient's race is a Hispanic with age greater or equal to 85 and having had the first birth less than 20 years ago, with a BIRADS breast density

being almost entirely fat, and had a previous breast biopsy then the person is a breast cancer patient". Likewise, Rule 2 can be interpreted as "If a person is a non-Hispanic black with an age between 75 and 79 years, the first birth age range between 20 to 24 years, BMI value 35 or above, and had a previous breast biopsy then the individual is a breast cancer patient".

2.4.5. Interpreting Rules based on Support, Confidence, and Lift

If we consider the rules of both breast cancer and non-breast cancer individuals we can see the significant differences. For both non-breast cancer and breast cancer individuals, its observed confidence, which indicates how often the rule has been found to be true in the data set, is very high (close to 100 %). In case of support, which demonstrates how frequently the item set or factors appear in the data set, it is high (more than 30%) for non-breast cancer patients. However, for breast cancer patients support value is very low (about 0.001%).

For both groups, if we look at the lift value that measures the degree of dependence between the antecedent and the consequent value, we can see the differences. For non-breast cancer individual, lift value is just above 1.0 that means the relationship between factors of these rules (antecedent part) and consequent (non-breast cancer patients) are very low. On the other hand, for the breast cancer patients' lift value is very high (more than 16.0) that indicates a greater association between factors in the antecedent and the consequent (breast cancer patients).

2.5. Discussion

Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as this has direct influence in their daily practice and clinical service. A reliable prediction will help oncologists and other clinicians in their decision-making process and allow clinicians in choosing the most reliable and evidence-based treatment and prevention strategies for their patients. Although, recent research has looked into various data mining techniques to aid clinicians in the diagnosis of breast cancer, however, there still remain gaps in suggesting an accurate prediction model. Our research explores association rules for breast cancer and non-breast cancer patients by data mining of the BCSC risk factors data set. Our findings suggest association rules that could be used to predict breast cancer risks among the target population. The data-driven approach that we used in this research can guide the efficient process of clinical data set to discover behavioral risk factor patterns and reveal hidden information for early detection and initiate prevention efforts as

well as treatment strategies of at risk breast cancer patients. However, any prediction should be combined with clinical judgment and individual patient circumstances.

There are several limitations of the current research. First, we used the BCSC data set which is robust, however, we did not have any control of the overall quality of the data collected. Second, in our data set there are a small number of instances in which patients have breast cancer compared to non-breast cancer patients. In our approach, we specified different support values for both target populations; for breast cancer patients we set a very low support value. In literature[42], we found that researchers used multiple support value for rare item problems and by using a low support value we attained rules of breast cancer patients that are rare in our cases. Although we used a low support value for breast cancer patients, however we set a high confidence value that represents the predictive strength of the rules.

2.6. Summary

Extracting useful rules has been generated from a breast cancer risk factor data set using association rule mining. Before applying association rule mining, we used the logit model to check the statistical significance of all predictors. We mined rules for both breast cancer and non-breast cancer patients with specified support and confidence. The experimental results showed that the generated rules hold the highest confidence level for both groups. However, in case of breast cancer patients we have to set a very low support value due to the imbalance of the data (small number of instances of patients having breast cancer compared to non-breast cancer individuals). We also mined strong rules from a huge set of generated rules and interpreted those rules accordingly. This research is an important step in improving risk prediction for people with potential risks for breast cancer.

We intend to extend this research by considering more risk factors to extract more useful and significant rules not only for breast cancer but also other cancer types using the association rule mining algorithm. Furthermore, we plan to build a predictive model using machine learning techniques for the breast cancer data set.

3. CLASSIFICATION OF BREAST CANCER RISK FACTORS USING SEVERAL RE-SAMPLING APPROACHES

Breast cancer is the most common cancer in women worldwide and the second most common cancer overall. Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as it has direct influence in daily practice and clinical service. Classification is one of the supervised learning models that is applied in medical domains. Achieving better performance on real data that contains imbalance characteristics is a very challenging task. Machine learning researchers have been using various techniques to obtain higher accuracy, generally by correctly identifying majority class samples while ignoring the instances of the minority class. However, in most of the cases the concept of the minority class instances usually is of higher interest than the majority class. In this research, we applied three different classification techniques on a real world breast cancer risk factors data set. First, we applied specified classification techniques on breast cancer data without applying any re-sampling technique. Second, since the data is imbalanced meaning data has an unequal distribution between the classes, we applied several re-sampling methods to get better performance before applying the classifiers. The experimental results show significant improvement on using a re-sampling method as compared to applying no re-sampling technique, particularly for the minority class.

The remainder of the chapter is organized as follows. Related work is discussed in Section 3.1. Methodology of the proposed classification model to handle imbalance data is discussed in Section 3.2. Section 3.3 shows the experimental results; the proposed techniques were evaluated using breast cancer risk factors data and their results are presented. Section 3.4 is the summary section; we conclude the chapter and suggest possible future research directions.

3.1. Related Work

Researchers have developed different models for breast cancer risk prediction, and association between risk factors [32]–[35]. In [32], the authors applied statistical methods to show a positive association between Hormone Replacement Therapy (HRT) and breast cancer risk, although this relationship varies according to race/ethnicity, BMI (Body Mass Index), and breast

density. The Gali model is used to estimate the number of expected breast cancers for white females who are examined annually [33]. In [34], the authors used commonly identified risk factors such as race/ethnicity, breast density, BMI, and the use of hormone therapy, type of menopause, and previous mammographic results to improve the model using logistic regression. In [35], the Breast cancer risk score is determined using k-nearest-neighbor (KNN) to improve readability for physician and patients.

Machine Learning (ML) or Data Mining (DM) algorithms are applied in the medical domain in order to assist with the decision-making process, for example, for the prediction of cancer risk. ML and DM algorithms [2], [7], [32] can be classified into supervised or unsupervised learning depending on the goal of the data mining task. Classification is a supervised learning techniques and the goal of the classification model is to predict qualitative or categorical outputs which assume values in a finite set of classes (e.g. Yes/No or Benign-cancer/Malignant-cancer, etc.) without an explicit order [34]. The primary objective of traditional classifiers is to get higher accuracy by reducing the overall classification error [35]. However, the overall classification error is biased towards the majority class for imbalanced data problems.

The problem of class imbalance is common that affects ML or classification models due to having a disproportionate number of different class instances in practice [43]. There are many approaches that deal with this problem such as cost function based, and sampling based solutions. In this research, we focused on sampling based approaches that can be classified into three major categories - random under-sampling, random over-sampling, and hybrid of over-sampling and under-sampling.

Sampling methods modify the data set to balance the class distribution before using the data set to train the classifier. Random under-sampling is the process of removing some of instances of the majority class whereas over-sampling is the process of adding more samples of the minority class so it has a larger effect on the ML algorithm. Although the methods are simple, however, both of these techniques have some shortcomings. The random under-sampling technique has the potential to lose information as it removes instances from the major class. On the other hand, over-sampling generates instances from the minority class that creates the potential risk of over-fitting. The hybrid method is a mix of the oversampling and under-sampling technique.

The Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic minority instances to balance the class distribution [44] and has been widely used. SMOTE produces synthetic minority instances by linear interpolation between neighbors in the input space. ENN (Edited Nearest Neighbor) is a technique of under-sampling of the majority class. It removes points or instances whose class labels differ from the majority of its k nearest neighbors [44]. Tomek Link [45] is a method of under-sampling which is used as a method of guided under-sampling where the observations from the majority class are removed. The combinations of these techniques are also applied in the literature to achieve better performance.

In this research, we applied three different classification algorithms on breast cancer risk factors data, and calculated the predicted performance on a test set. Since the data is imbalanced, we also applied various resampling techniques on the training data and applied classifiers on the 'modified' training data. Performance comparisons on the test data based on the all classification models were also conducted.

3.2. Methodology

3.2.1. Classification Phase

We used three different classifiers namely Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) to train the breast cancer data set of imbalanced data (original data) as well as modified training data obtained by using different resampling methods. These trained models were used to predict the target class for the test data set. The three classifiers that are used in this research are briefly described below.

3.2.1.1. Decision Tree (DT)

DT is a supervised learning approach that learns from class-labeled instances. It works very well with different types of data and results are easy to interpret. In addition, building a model using decision tree is comparatively easy, and data can be represented in a visualizing form. The decision tree model generation is however sensitive to overfitting and may get stuck in local minima. When the number of dimensions gets too high, the decision tree model generation may fail. The decision tree classifier has been widely applied to solving many real world problems including in areas of healthcare, medicine, business, education, and so on [6] [21], [46]. A standard decision tree algorithm for classification problems is the C4.5 decision tree algorithm that was initially developed

by Ross Quinlan [21]. The C4.5 algorithm extends Quinlan's earlier ID3 tree algorithm to address certain practical issues such as over-fitting and the type of variables accepted in the input.

3.2.1.2. Random Forest (RF)

RF is a powerful classification and regression tool that generates a forest of classification trees, rather than a single classification tree [47]. RF creates decision trees on randomly selected data samples, obtains the prediction from each tree and selects the best solution by means of voting. There are two stages in the RF algorithm, the first one is RF building, and the second stage is to make a prediction from the RF classifier created during the first stage. RF is considered as a highly accurate and robust method because of the number of decision trees participating in the process. In addition, if there are more trees in the forest, the RF classifier will avoid the over-fitting problem. RF is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then voting is performed on all predictions. This whole process is time-consuming. However, RF is widely used to various problems for its good performance and it does not overfit. RF has been used extisively in the area of medical and bioinformatics [29].

3.2.1.3. Extreme Gradient Boosting (XGBoost)

XGBoost [29] is an implementation of gradient boosted decision trees designed for speed and performance. XGBoost provides a wrapper class to allow models to be treated like a classifier or a regressor in the scikit-learn framework. The XGBoost model for classification is called XGB-Classifier. XGBoost is a scalable and accurate implementation of the gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit of memory and hardware resources available for the tree boosting algorithm.

Boosting is an ensemble method that aims to create a strong classifier based on several weak classifiers. By adding models on top of each other iteratively, the errors of the previous model are corrected by the next predictor, until the training data is accurately predicted by the overall model. Gradient boosting also comprises an ensemble method that sequentially adds predictors and collects previous models. However, instead of assigning different weights to the classifiers after every iteration, this method fits the new model to new residuals of the previous prediction and

then minimizes the loss when adding the latest prediction. Thus, in the final model it actually uses gradient descent and hence the name gradient boosting. XGBoost specifically, implements this algorithm for decision tree boosting with an additional custom regularization term in the objective function.

XGBoost has been widely used in a number of machine learning and data mining challenges. For example, in Kaggle, which is a ML competetion site; among the 29 challenge winning solutions published on the Kaggle site during 2015, 17 solutions used XGBoost. The second most popular method was deep neural network and was used in 11 solutions [48]. Examples of the problems in these winning solutions include: store sales prediction; web text classification; customer behavior prediction; motion detection; ad click through rate prediction; malware classification; hazard risk prediction; massive online course dropout rate prediction, and so on. The most significant factor behind the success of XGBoost is its scalability in all scenarios. The system runs more than ten times faster than existing popular solutions on a single machine and scales billions of examples in distributed or memory-limited settings. More importantly, XGBoost exploits out-of-core computation and enables data scientists to process hundred millions of instances on a desktop.

3.2.2. Resampling Phase

The data set that we used in this research is imbalanced data, meaning there are significantly more samples for one category than the other. For that reason, different resampling techniques were applied to the training data set (imbalanced) and thus the training data is modified accordingly. The resampling techniques that were used in this work are briefly discussed below.

3.2.2.1. Random under-sampling (RUS) of majority class

is a form of data sampling that randomly picks majority class instances and removes them from the dataset until the desired class distribution is achieved [49]. This means that for a dataset containing 100 positive and 500 negative instances, RUS removes 400 negative instances in order to achieve a 50:50 post-sampling positive:negative class ratio.

3.2.2.2. Random over-sampling (ROS) of minority class

is a form of data sampling that randomly picks minority class instances with replacement until the desired class distribution is achieved [49]. This means that for a dataset containing 100 positive and 500 negative instances, ROS adds 400 positive instances in order to achieve a 50:50 post-sampling positive:negative class ratio.

3.2.2.3. SMOTE

works by creating synthetic observations based upon the existing minority instances [44],[50]. For each minority instance, SMOTE calculates the k nearest neighbors. Depending upon the amount of oversampling needed, one or more of the k-nearest neighbors are selected to create synthetic examples.

3.2.2.4. Edited Nearest Neighbor (ENN)

is the technique of under-sampling of the majority class [44]. It removes points or instances whose class label differs from a majority of its k-nearest neighbors.

3.2.2.5. SMOTE + ENN

combines the over-sampling and under-sampling techniques [44]. It performs over-sampling using SMOTE and under-sampling or cleaning using ENN. Thus, instead of removing only the majority class examples, instances from both classes are removed. ENN tends to remove more instances than Tomek links do, so it is expected that it will provide more in-depth data cleaning.

3.2.2.6. SMOTE + Tomek Link

also combines over-sampling and under-sampling techniques. It performs over-sampling using SMOTE and under-sampling or cleaning using Tomek links [44],[45]. Thus, instead of removing only the majority class examples, instances from both classes are removed. Tomek links remove less instances compared to ENN.

3.2.3. Proposed Approach

To obtain a better classification performance, we used specified classifiers to train the model using the original training data. We also used various types of resampling methods on the training data to train the model using specified classifiers with the modified training data. We then used all the trained models to obtain class information on the test data. The diagram of our proposed approach is shown in Fig. 3.1 consisting of the following main steps:

3.2.3.1. Step 1

This step includes obtaining the classification model data and test data for classification; we constructed the classification model data, or training data, and sample, or test data, for classification. The training set contains 80% of the data while the test set contains the remaining 20%. The stratified shuffle split technique was used since it preserves the percentage of samples for each

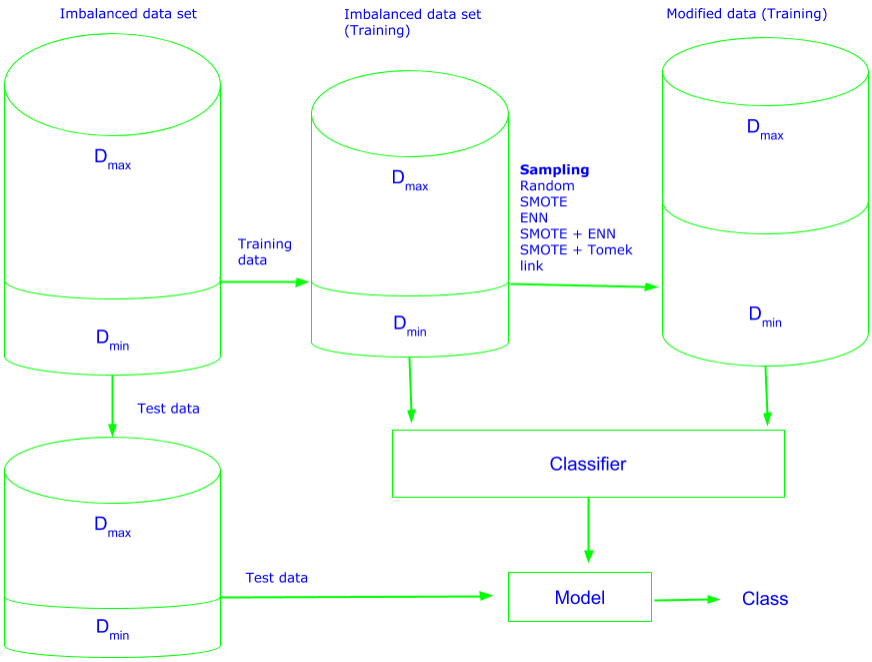

Figure 3.1. Proposed model to handle imbalanced data.

class which is important for imbalanced data. The Stratified shuffle split technique available in scikit-learn (sklearn), a machine learning library for the Python programming language, was used since it preserves the percentage of samples for each class which is important for imbalance data. *Dmax* is the number of instances belonging to the negative class, or majority class, while *Dmin* is the number of instances of the positive, or minority class.

3.2.3.2. Step 2

This step resamples the training data. Several resampling techniques were used on training data that changed the number of instances of the training data. Based on the techniques of the resampling methods, the instances of the majority class were removed and/or instances of the minority class were added. The test data was kept unchanged.

3.2.3.3. Step 3

In this step, the classification model data was trained with the specified classifiers. First, we used the original training data without using any sampling methods, and built models using the specified classifiers. Second, for the training we used the modified training data obtained by applying the different sampling techniques. Each of these training data sets were used to train all three classifiers. All of the above models were saved for the prediction on the test data.

3.2.3.4. Step 4

The last step was to apply test data on the saved models obtained in Step 3 to generate predictions on the test data.

3.3. Experiments and Results

Detailed data description and pre-processing is discussed in this section. This section also presents the experimental results and performance evaluation of the different models.

3.3.1. Data Description and Pre-processing

The dataset includes information from 6,318,638 mammography examinations obtained from the Breast Cancer Surveillance Consortium (BCSC) database collected from January 2000 to December 2009 [38]. Data for this study was obtained from the BCSC Data Resource and more information is available at `http://www.bcsc-research.org`.

The data is aggregated such that the total number of instances or records is 1,144,565, with 13 attributes or columns. The dataset also contains missing or unknown values denoted by 9. To build a reliable model, we discarded the records containing at least one missing or unknown value. We also removed the attribute year that represents the calendar year of the observation. After discarding these records and one attribute, there are 219,524 available records with 12 attributes. In the dataset, there is an attribute named count, representing the number of records that have the combination of variable-values shown in the row. For instance, the value of the count column for the particular row is 12. It indicates that there were 12 similar records, the same as that particular row in the original data. For that reason, we created the number of rows or records the same as the count value in the original dataset, and discarded the count column after that. Finally, there are a total of 1,015,583 records with 11 attributes for building the model. Among 1,015,583 records, 60,800 individuals have prior breast cancer, and 954,783 are non-breast cancer individuals. Among the 11 attributes, "prior breast cancer" values yes or no is considered as the response or dependent variable and the remaining 10 attributes are considered as explanatory or predictors or independent variables.

The summary of the BCSC data along with train/test split are shown in Table 3.1.

We used different resampling methods on the training data. The distribution of the training data after applying different resampling techniques is shown in Table 3.2.

Table 3.1. Summary of BCSC data with train/test spilt.

Types	**Class = yes**	**Class = no**	**Total**
BCSC data	60,800	954,783	1,015,583
Training (80 %)	48,640	763,826	812,466
Test (20%)	12,160	190,957	203,117

Table 3.2. Distribution of modified training data after applying different resampling methods.

Resampling type	**Class = yes**	**Class = no**	**Total**
Random under-sample	48,640	48,640	97,280
Random over-sample	763,826	763,826	1,527,652
SMOTE	763,826	763,826	1,527,652
ENN	48,640	685,963	734,603
SMOTE + ENN	437,256	658,167	1,095,423
SMOTE + Tomek link	763,825	763,825	1,527,650

3.3.2. Evaluation Measures

To measure the performance of our model, several evaluation measures were used such as accuracy, recall, precision, area under the Receiver Operating Characteristic curve (ROC) or AUC, and F-measure [51]. These were derived from the confusion matrix, and applied to the classifier evaluation.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{3.1}$$

$$Recall = TP/(TP + FN) \tag{3.2}$$

$$Precision = TP/(TP + FP) \tag{3.3}$$

where Y is the binary response or class variable; α is the intercept to be calculated; β_i is the estimated vector of parameters, and X_i is the vector of independent variables. Here, TP denotes the number of positive examples correctly classified, TN denotes the number of negative samples correctly classified, FN represents the number of positive observations incorrectly classified, and FP indicates the number of negative samples incorrectly classified by the estimator.

The ROC curve is a representation of the best decision boundaries for the cost between the True Positive Rate (TPR), and the False Positive Rate (FPR). The ROC curve plots TPR against FPR. TPR, and FPR are defined as follows.

$$TPR = TP/(TP + FN) \tag{3.4}$$

$$FPR = FP/(FP + TN) \tag{3.5}$$

The area below the ROC curve is called AUC and is widely utilized for weighing classifier performance. The value of AUC ranges from 0.0 to 1.0, where a value of AUC equals 1.0 means perfect prediction, a value of 0.5 means random prediction, and a value less than 0.5 is considered as a poor prediction.

If only the performance of the positive class in this case the minority class is considered, two measures namely recall, and precision are important. Recall or true positive rate denoting the percentage of retrieved objects that are relevant, while precision or positive predictive value denoting the percentage of relevant objects that are identified for retrieval. The F-measure or F1 score is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test, which is defined as follows:

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \tag{3.6}$$

The harmonic mean of two numbers tends to be closer to the smaller of the two. Hence, a high F-measure value ensures that both precision and recall are reasonably high.

It is to be noted that for balanced class F1 score can effectively be ignored, the accuracy is key. For the imbalance class, if the class distribution is highly skewed, then the classifier can have a higher accuracy simply by choosing the majority class. In such a situation, the classifier that gets a high F1 score on both classes, as well as high accuracy should be selected. However, if a particular class generally the minority class is more important than the other then it is more important to correctly classify instances for the minority or important class as opposed to the majority class. In this case, the classifier that has a good F1 score only on the important class should be considered.

3.3.3. Results

In this research, we applied three different classifier models on the original training data (imbalanced), and the modified training data sets. We compared the performance of the trained models on the test data. The overall performance of the classification models (built based on the original training data) on test data are shown in Table 3.3 whereas the performance of the minority class is shown in Table 3.4, respectively.

Table 3.3. Overall performance of specified classifiers on test data (trained with the original training data).

Methods	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
DT	0.92	0.94	0.92	0.9406	0.9272
RF	0.92	0.94	0.92	0.9399	0.9164
XGBoost	0.92	0.94	0.91	0.9404	0.9287

Table 3.4. Performance of minority class on test data.

Methods	**Precision**	**Recall**	**F1-score**
DT without sampling	0.54	0.05	0.10
RF without sampling	0.49	0.11	0.18
XGBoost without sampling	0.54	0.03	0.05

Although, the performance of these classifiers seem very good (according to Table 3.3) when no resampling techniques were used, however, the performance of classifying the instances of the minority class was very low. For the minority class, maximum recall, and F1-score were reported as 0.11 and 0.18, respectively for the RF classifier.

Different resampling methods on the training data were used to modify the training data accordingly. The modified training data sets were used for the training of the specified classifiers. Results were obtained from the models applied to the test data. Table 3.5 shows the overall performance of the DT classification models (built based on the modified training data) on test

data for all the different training data sets, whereas Table 3.6 shows the performance of the minority class for the DT classifier, respectively.

Table 3.5. Overall performance of DT classifier (model built on modified training data) on test data.

Methods	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
DT with RUS	0.95	0.82	0.86	0.8171	0.9255
DT with ROS	0.95	0.82	0.86	0.8157	0.9263
DT with SMOTE	0.95	0.82	0.87	0.8244	0.9266
DT with ENN	0.93	0.91	0.92	0.9069	0.9249
DT with SMOTE + ENN	0.94	0.87	0.90	0.8722	0.9207
DT with SMOTE + Tomek link	0.95	0.82	0.86	0.8208	0.9270

Table 3.6. Performance of minority class on test data based on DT classifier.

Methods	**Precision**	**Recall**	**F1-score**
DT with RUS	0.24	0.96	0.39
DT with ROS	0.24	0.97	0.39
DT with SMOTE	0.25	0.95	0.39
DT with ENN	0.33	0.56	0.42
DT with SMOTE + ENN	0.29	0.80	0.43
DT with SMOTE + Tomek link	0.24	0.96	0.39

For DT, the best accuracy obtained was 90.69% when sampling method ENN was applied, but the AUC value was little (0.0021) less than the highest AUC value of 0.9270. For the minority class, The best recall (0.80) and the best F1-score (0.43) values were obtained when the resampling technique SMOTE and ENN were applied.

Table 3.7 shows the overall performance of the RF classification models (built based on the modified training data) on test data for all the different training data sets whereas Table 3.8 shows the performance of the minority class, respectively.

Table 3.7. Overall performance of RF classifier on test data.

Methods	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
RF with RUS	0.95	0.82	0.87	0.8219	0.9180
RF with ROS	0.95	0.84	0.87	0.8356	0.9145
RF with SMOTE	0.95	0.85	0.89	0.8540	0.9140
RF with ENN	0.93	0.88	0.90	0.8820	0.9039
RF with SMOTE + ENN	0.94	0.88	0.91	0.8855	0.8606
RF with SMOTE + Tomek link	0.95	0.85	0.89	0.8532	0.9135

Table 3.8. Performance of minority class on test data based on RF classifier.

Methods	**Precision**	**Recall**	**F1-score**
RF with RUS	0.24	0.94	0.39
RF with ROS	0.26	0.91	0.40
RF with SMOTE	0.27	0.85	0.41
RF with ENN	0.28	0.63	0.39
RF with SMOTE + ENN	0.31	0.74	0.44
RF with SMOTE + Tomek link	0.27	0.85	0.41

For RF, the best accuracy obtained was 88.55% when sampling method SMOTE followed by ENN was applied. But in case of SMOTE followed by ENN, the AUC value (0.8606) was the lowest among all other sampling methods. The maximum AUC (0.9180) for RF was reported when

RUS used. For the minority class, the best recall (0.94) was found when RUS was applied and the highest F1-score (0.44) was obtained when resampling technique SMOTE and ENN were applied.

Table 3.9 shows the overall performance of XGBoost classification models (built based on the modified training data) on test data for all the different training data sets whereas Table 3.10 shows the performance of the minority class, respectively.

Table 3.9. Overall performance of XGBOOST classifier on test data.

Methods	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
XGBoost with RUS	0.95	0.81	0.86	0.8118	0.9287
XGBoost with ROS	0.95	0.81	0.86	0.8128	0.9288
XGBoost with SMOTE	0.95	0.82	0.87	0.8218	0.9284
XGBoost with ENN	0.93	0.91	0.92	0.9149	0.9281
XGBoost with SMOTE + ENN	0.95	0.86	0.89	0.8626	0.9270
XGBoost with SMOTE + Tomek link	0.95	0.82	0.86	0.8210	0.9282

Table 3.10. Performance of minority class on test data based on XGBOOST classifier.

Methods	**Precision**	**Recall**	**F1-score**
XGBoost with RUS	0.24	0.97	0.38
XGBoost with ROS	0.24	0.97	0.38
XGBoost with SMOTE	0.25	0.96	0.39
XGBoost with ENN	0.35	0.52	0.42
XGBoost with SMOTE + ENN	0.29	0.87	0.43
XGBoost with SMOTE + Tomek	0.25	0.96	0.39

For XGBoost, the best accuracy obtained was 91.49% when the sampling method ENN was used. Surprisingly, the AUC value (close to 0.93) remained almost same for all the resampling

techniques. For the minority class, the best recall (0.97) was found when both RUS and ROS were applied, and the highest F1-score (0.43) was obtained when the resampling technique SMOTE and ENN were applied.

3.3.4. Performance Comparison

Although we obtained the best overall performance for all the classifiers when no resampling methods were used for the training phase, however, for minority class performance was very low. The accuracy for all three classifiers were about 94% when no resampling methods were applied which is about 3% more than the best accuracy obtained when the resampling techniques were used.

However, for the minority class, the performance was not better when no resampling methods were used. For instance, the best recall and F1 score for the minority class for RF were reported as 0.11 and 0.18, respectively when no resampling was used on the training data. Yet, the best recall and F1 score for the minority class were reported as 0.87 and 0.43, respectively for the XGBoost classifier when the resampling method SMOTE and ENN was used. It is also worth to mention that the overall performance for the same combination was also good (not best). For example, the accuracy and AUC score for this combination were reported as 86.26% and 92.70%, respectively. The performance for the minority class was far better when applying all the specified resampling methods as compared to not applying any resampling method. Thus, it is important to consider all the factors when dealing with imbalanced data such as if both classes are important or only the minority class is significant. Therefore, the appropriate model should be selected based on the objective.

3.4. Summary

Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as this has direct influence in their daily practice as well as their clinical service. In this research, we explored breast cancer risk factors data and applied different resampling techniques before applying machine learning methods. The data that we used in this research was severely imbalanced (60,800 versus 954,7834). Our main objective was to improve the classification performance of the standard machine learning algorithms towards the prediction of the important or minority class. We compared the impact of using several resampling techniques on the training data before using the specified classifiers in terms of the overall performance and the performance

of the minority class. Experimental results show that the performance improves particularly for the minority class when the resampling techniques were used as compared to applying the classification techniques without using any resampling techniques.

We intend to extend this research by considering more risk factors not only for breast cancer but also for other cancer types. Furthermore, we plan to build more accurate predictive models that could provide better performance for both the minority and the majority class.

4. ENHANCING THE PERFORMANCE OF CLASSIFICATION USING SUPER LEARNING

Classification is one of the supervised learning models, and enhancing the performance of a classification model has been a challenging research problem in the fields of Machine Learning (ML) and data mining. The goal of ML is to produce or build a model that can be used to perform classification. It is important to achieve superior performance of the classification model. Obtaining a better performance is important for almost all fields including healthcare. Researchers have been using different ML techniques to obtain better performance of their models; ensemble techniques are also used to combine multiple base learner models. The ML technique called super learning or stacked-ensemble is an ensemble method that finds the optimal weighted average of diverse learning models. In this chapter, we presented two different forms of super learner or stacked ensemble. First one uses two base learners namely Gradient Boosting Machine (GBM) and Random Forest (RF), and the second one uses three base learners namely GBM, RF and Deep Neural Network (DNN); and for both cases a meta-learner called Generalized Linear Model (GLM) is used [26], [47]. We used four well-known benchmark data sets related to the healthcare area and compare the performance of both super learners with the individual base learners, baseline ensemble and the state-of-the-art classifiers. Our evaluations confirm that the super learner method has the ability to perform better compared to individual base learners, baseline ensemble approach, and some of the state-of-the-art techniques on four benchmark data sets.

The rest of the chapter is organized as follows. Section 4.1 describes state-of-the-art techniques; Section 4.2 presents methodology where proposed solution is discussed. Section 4.3 shows the experimental results; the proposed techniques are evaluated using four benchmark data sets and their results are presented. Section 4.4 is the summary section; we conclude the chapter and suggest possible future research directions.

4.1. Related Work

In this research, a ML technique called super learning or stacked ensemble [52], [53], [54] has been used to improve the performance of four benchmark data sets related to healthcare.

Stacked generalization in the context of neural net ensembles used leave-one-out Cross-Validation (CV) to generate level-one data [55], which is the cross-validated predicted values generated from cross-validating base learners on the training data. The authors extended the previous stacking framework [55] to regression problems [30] and proposed to use k-fold CV to generate level-one data. In this work, the authors also suggested non-negativity constraints for the meta-learner. It was proposed combining estimates in regression and classification that provided a general framework for stacking and compared CV-generated level-one data to bootstrapped level-one data [56]. Ensemble or combining learners in various methods showed better performance over a single candidate learner, but there is a concern that these methods may over-fit the data and may not be the optimal way to combine the candidate learners [52]. Researchers suggest a solution to this problem in the form of a new learner and named it super learner. In the context of prediction, a super learner is itself a prediction algorithm, which applies a set of candidate learners to observed or training data, and chooses the optimal learner for a given prediction problem based on the cross-validated risk. Theoretical results show that the super learner will perform asymptotically as well as or better than any other candidate learners [52], [57].

Using super learning for dynamic accuracy prediction in various domains is becoming popular. Researchers have used a super learning model to enhance anomaly detection in cellular networks [58]. It was also used in predicting violence among inmates from the 2005 census of state and federal adult correctional facilities [59]. Researchers investigated different ensemble learning methods including super learning for network security and anomaly detection. In their research, they showed that the super learner provides better results than any of the single models like Naïve Bayes (NB), Decision Tree (DT), Neural Network (NN), Support Vector Machine (SVM), K-nearest Neighbors (KNN) and RF [60].

Different ML and DM techniques have been developed and used in various data sets in healthcare. Researcher used ensemble-based techniques with 10 fold cross-validation on Messidor data for enhancing the performance [61]. Classifier methods like multi-layer perceptron (MLP), and NB have been used to assess the performance of the Wisconsin breast cancer (WBC) data sets [62]. Sequential minimal optimization (SMO) technique, which is an optimization algorithm widely used for training SVM, has also been used to assess the performance of the WBC data set [62]. In addition, bagging and boosting methods have been used to compare the performance of the

WBC data set [3]. The NB classifier has been used on the Pima Indian Diabetes Dataset (PIDD). In order to get superior performance over the NB classifier, researchers used a Genetic Algorithm (GA) approach for attribute or feature selection [63]. For the Indian Liver Patient Dataset (ILPD) data set, authors used an ensemble classifier with 5 fold cross-validation and obtained acceptable results [64]. Researchers showed the comparative analysis of diverse ML algorithms like NB, SVM, MLP, random forest (RF) for various data sets including ILPD with the best accuracy for ILPD using SVM [65].

In this research, we used the super learner or stacked ensemble approach that is discussed in the following section applied to the four benchmark data sets.

4.2. Methodology

Super learning or stacked ensemble is a ML method that uses two or more learning algorithms. It is a loss-based supervised learning method that finds the optimal combination of a collection of prediction algorithms. It is a cross-validation-based approach for combining machine learning algorithms that produce predictions that are at least as good as those of the best input algorithm [52], [60].

4.2.1. Super Learning or Stacking

Stacking is a broad class of algorithms that involves training a second-level meta-learner of an ensemble. Super learning or stacking [52] is a procedure for ensemble learning in which a meta-learner is trained on the output of a collection of base learners. The output from the base learners, also called the level-one data, can be generated using cross-validation. Construction of level-one data is discussed in the following section. The original training data set is often referred to as the level-zero data. The pseudo-code of the super learning or stacking is shown in Algorithm 1 [53], [54], and the concept diagram of the super learning method is illustrated in Fig. 4.1.

4.2.1.1. Constructing level-one data

The super learner theory requires cross-validation to generate the level-one data. Assume that the training set is comprised of n independent, and identical distributed observations, $\{O_1, O_2, O_3\}$ where $O_i = (X_i, Y_i)$ here X_i is the feature value, and Y_i is the outcome or class value [53] [54]. Consider an ensemble comprised of a set of L base learning algorithms, $\{B_1, B_2, ..., B_L\}$ each of which is indexed by an algorithm class, and a specific set of model parameters. Then, the

Algorithm 1 Super learning algorithm

1: Input: data set D with set of X examples, and response column Y.
2: Output: ensemble-model.

3: Set up the ensemble

- Specify a list of L base algorithms (with a specific set of model parameters).
- Specify a meta-learning algorithm.

4: Train the ensemble.

- Train each of the L base algorithms on the training set.
- Perform k-fold cross-validation on each of the L learners, and collect the cross-validated predicted values from k-fold CV that was performed on each of the L base learners.
- The N cross-validated predicted values from each of the L algorithms can be combined to form a new matrix, $Z(NXL)$. This matrix Z, along with the original response vector is called the "level-one" data (N = number of instances in the training set).
- Train the meta-learning algorithm on the level-one data (Z,Y). The ensemble model consists of the L base learning models, and the meta-learning model, which can then be used to generate predictions on a test set.

5: Predict new data.

- To generate ensemble predictions, first generate predictions from the base learners.
- Feed those predictions into the meta-learner to generate the ensemble predictions.

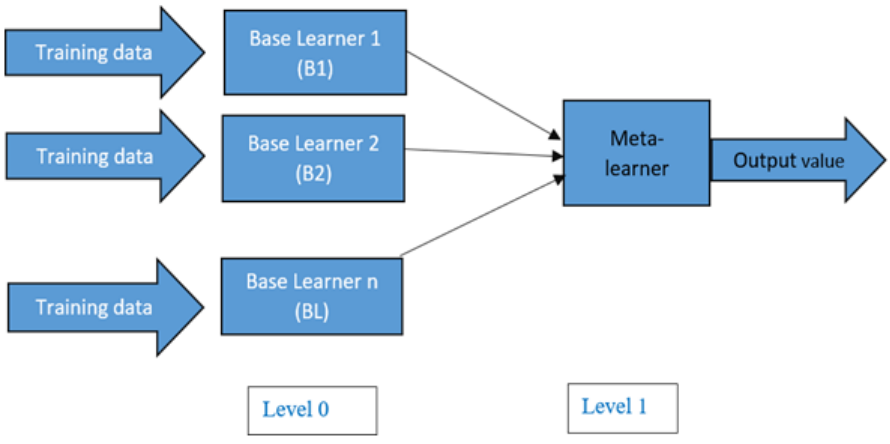

Figure 4.1. Concept diagram of super learner

process of constructing the level-one data will involve generating a $n \times L$ matrix, referred to as Z of the k-fold cross-validated predicted values as follows:

1. The original training set X is divided at random into $k = V$ roughly-equal pieces $X(1), X(2), ..., X(V)$.

2. For each base learner in the ensemble, B_L V-fold cross-validation is used to generate n cross-validated predicted values associated with the l^{th} learner. These n-dimensional vectors of cross-validated predicted values become the L columns of Z.

The level-one data set Z, along with the original outcome vector $\{Y_1, Y_2, ..., Y_n\}$, is used to train the meta-learning algorithm. Finally, each of the L base learners are fitted to the full training set and these fits are saved. The final ensemble fit is comprised of the L base learner fits, along with the meta-learner fit. To generate a prediction for new data using the ensemble, the algorithm first generates the predicted values from each of the L base learner fits, and then passes those predicted values as input to the meta-learner fit, which returns the final predicted value for the ensemble.

4.2.1.2. Base Learners

It is recommended that the base learners should include a diverse set of learners, for example, linear model, SVM, RF, GBM, Neural Net, etc., however, the super learner theory does not require any specific level of diversity among the set of base learners [53], [54]. It is also allowable to include the same algorithm multiple times as a base learner by different sets of parameters. For example, the user could specify multiple Distributed Random Forest (DRF) method, each with a different splitting criterion, tree depth, number of folds, or number of trees. Typically, in stacking-based ensemble methods, the prediction functions are fit by training each base learning algorithm on the whole training data set and then combining these fits using a meta-learning algorithm. In this research, we first used two base learners namely Gradient Boosting Machine (GBM) and Distributed Random Forest (RF). In addition, we used another base learner called Deep Neural Network (DNN) with GBM and RF that are briefly discussed below.

Gradient Boosting Machine (GBM) [47] produces a prediction model in the form of an ensemble of weak prediction models. It builds the model in a stage-wise fashion and is generalized by allowing an arbitrary differentiable loss function. It is one of the most powerful methods available today. GBM for regression and classification is a forward learning ensemble

method. The guiding heuristic is that good predictive results can be obtained through increasingly refined approximations. H2O is an open source, in-memory, ML, and predictive analytics platform [47] which is used in this work. GBM is available in H2O, which is built upon the distributed, open source, Java-based machine learning platform for big data [47]. H2O's GBM sequentially builds regression trees on all the features of the data set in a fully distributed way - each tree is built in parallel. Additional features have been incorporated into the new version of H2O like the per-row observation weights, per-row offsets, N-fold cross-validation, and support for more distribution functions (such as Gamma, Poisson, and Tweedie).

Distributed Random Forest (DRF) [47] is a powerful classification and regression tool. When given a set of data, Random Forest (RF) generates a forest of classification (or regression) trees, rather than a single classification (or regression) tree. Each of these trees is a weak learner built on a subset of rows and columns. More trees will reduce the variance. Both classification and regression take the average prediction over all of their trees to make a final prediction, whether predicting a class or numeric value. For a categorical response column, DRF maps factors (e.g. 'dog', 'cat', 'mouse') in lexicographic order to a name lookup array with integer indices (e.g. 'cat' - 0, 'dog' - 1, 'mouse' - 2).

Deep Neural Network (DNN) [26] is an architecture of deep learning based on an Artificial Neural Network (ANN) that is inspired by biological neural networks. A DNN has basically many connected units arranged in layers of varying sizes with information being fed forward through the network. DNNs have been successfully applied to fields such as computer vision and natural language processing system and achieved better or similar accuracy rates compared to humans in classification tasks[66]. H2O's deep learning is based on a multi-layer feedforward ANN that is trained with stochastic gradient descent using back-propagation. The network can contain a large number of hidden layers consisting of neurons with activation functions such as tanh, rectifier, and maxout. Advanced features such as dropout, L1 or L2 regularization, grid search, etc. enable high predictive accuracy.

4.2.1.3. Meta-learning algorithm

The meta-learner is used to find the optimal combination of the L base learners. The Z matrix of cross-validated predicted values, described previously, is used as the input for the meta-learning algorithm along with the original outcome from level-zero training data $\{Y_1, Y_2, ..., Y_n\}$. In

the super learning algorithm, the meta-learning method is specified as the minimizer of the cross-validated risk of a loss function of interest, such as squared error loss or rank loss. Historically, in stacking implementations, the meta-learning algorithm is often some sort of regularized linear model, however, a variety of parametric and non-parametric methods can be used as a meta-learner to combine the output from the base fits [53] [54]. For this research, we used Generalized Linear Models (GLM) as the meta-learner, which is described briefly as follows.

Generalized Linear Models (GLMs) are an extension of traditional linear models. They have gained popularity in statistical data analysis due to the following three characteristics [26]. Firstly, the flexibility of the model structure unifying the typical regression methods (such as linear regression, and logistic regression for binary classification). Secondly, the recent availability of model-fitting software, and finally, the ability to scale well with large data sets.

GLM provides flexible generalization of ordinary linear regression for response variables with error distribution models other than a Gaussian (normal) distribution. GLM's estimate regression models for outcomes follow exponential distributions. In addition to the Gaussian (i.e. normal) distribution, these include Poisson, binomial, and gamma distributions. Each serves a different purpose, and depending on the distribution and link function choice, either can be used for prediction or classification [47].

4.2.2. Proposed Approach

To obtain better performance, we selected three base learners from H2O namely Gradient Boosting Machine (GBM), Random Forest (RF), and Deep Neural Network (DNN) [47]. For the meta-learner, we used Generalized Linear Model (GLM) [26], [47]. It is a particular implementation of the Super Learner, using a probability-based weighting function to combine the outputs of the first level learners. In a nutshell, we used the probabilities of success of each class to build exponentially decayed weighting functions, adding a control variable to reduce the overall influence of low accuracy models in the final prediction.

Our proposed method has the following main steps:

1. Classification Model Data and Sample Data for Classification

    - We construct the classification model data and sample data for classification whereby for the training data set the class information is known whereas the class information is

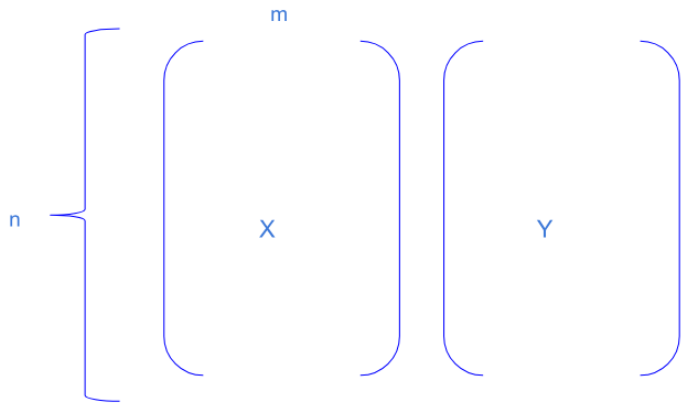

Figure 4.2. Level-0 data

unknown for the testing data set. The data sets are referred to as level-0 data, which is shown in Fig. 4.2 where X is the training data set with n rows, and m columns; the class value column is separated from the training data, which is referred to as Y.

2. Classifiers and Model Selection

- To set up the stacked ensemble or super learner, we need to specify the base learners and a meta-learner algorithm. For this research, we first selected two base learners namely GBM and RF. We also selected another base learner DNN with the previous two base learners and for the meta-learner we specified GLM.

  For the model selection process, we used the cartesian grid search and specified a set of values for particular parameters to search over each base learner. The parameters that underwent a model selection phase in the grid search are shown in Table 4.1 with the corresponding range of values. If a hyper-parameter of a learner is not listed in the Table 4.1, default values of the implementation of the algorithm were used. Best parameters for all three base learners of the four specified data sets using grid search were listed in Table 4.2. For the meta learner algorithm, we used the default parameters available in H2O. The training of the ensemble has the following two steps:

  (a) Base learners

  – We trained GBM, RF, and DNN individually on the training data set with the specific parameters obtained using the grid search. Here, 10-fold cross-

Table 4.1. Classifiers with the corresponding hyper-parameter values in grid search.

Classification algorithm	**Hyper parameters in grid search with corresponding range of values**
GBM	learn_rate: [0.01, 0.03, 0.05, 0.1] sample_rate: [0.6, 0.7, 0.8,0.9] col_sample_rate_per_tree:[.7, .8, .9] ntrees:[50, 80, 100,120] max_depth:[6,8,10,12,15]
RF	sample_rate: [0.6, 0.7, 0.8,0.9] col_sample_rate_per_tree:[.7, .8, .9] ntrees:[50, 80, 100,120] max_depth:[6,8,10,12,15]
DNN	activation: [tanh, rectifier, maxout] hidden_layers: [[30], [50],[30,30],[50,50]] epochs:[10, 15, 20, 30] l1:[0,1e-3, 1e-5] l2:[0,1e-3, 1e-5]

validation is performed on each of these learners and we kept the cross-validation prediction parameter specified as True. For all three base learners, the Bernoulli distribution was specified since the response column is of type categorical with two classes. In addition, for the base learners the fold-assignment modulo was selected which is a simple deterministic way to evenly split the data set into the folds. It is important to note that in our experiments we first used two base learners (GBM and RF) and then three base learners (GBM, RF, and DNN).

The N cross-validated predicted values of the three base learners GBM, RF, and DNN are defined as P1, P2, and P3 respectively. For the ensemble consisting of two base learners (GBM and RF), the predicted values P1 and P2 are combined to form a $n \times 2$ matrix. This matrix along with the class value (Y) of the training data is called the level-1 data for the ensemble having two base learners, which is shown in Fig. 4.3.

For the stacked ensemble consisting of three base learners, level-1 data is constructed similarly. However, instead of using the cross-validated predicted values P1 and P2, we used P1, P2, and P3, which are combined to form a $n \times 3$ matrix. This matrix along with the class value (Y) of the training data is called the

Table 4.2. Classifiers with the corresponding hyper-parameter best values from grid search for the specified data sets.

Classifiers / Data set	**GBM**	**RF**	**DNN**
Messidor	learn_rate: 0.03 sample_rate: 0.6 col_sample_rate_per_tree: 0.8 ntrees: 120 max_depth: 6	sample_rate: 0.7 col_sample_rate_per_tree: 0.9 ntrees: 80 max_depth: 15	activation: rectifier hidden_layer: [30] epochs: 30 l1: 0.001 l2: 0.0
WBC	learn_rate: 0.03 sample_rate: 0.6 col_sample_rate_per_tree: 0.7 ntrees: 120 max_depth: 8	sample_rate: 0.8 col_sample_rate_per_tree: 0.8 ntrees: 50 max_depth: 6	activation: rectifier hidden_layer: [50,50] epochs: 20 l1: 1e-5 l2: 0.0
PIDD	learn_rate: 0.03 sample_rate: 0.7 col_sample_rate_per_tree: 0.9 ntrees: 100 max_depth: 6	sample_rate: 0.8 col_sample_rate_per_tree: 0.8 ntrees: 80 max_depth: 6	activation: tanh hidden_layer: [50] epochs: 20 l1: 1e-5 l2: 0.001
ILPD	learn_rate: 0.03 sample_rate: 0.7 col_sample_rate_per_tree: 0.8 ntrees: 50 max_depth: 8	sample_rate: 0.6 col_sample_rate_per_tree: 0.9 ntrees: 120 max_depth: 8	activation: rectifier hidden_layer: [50] epochs: 30 l1: 0.0 l2: 0.01

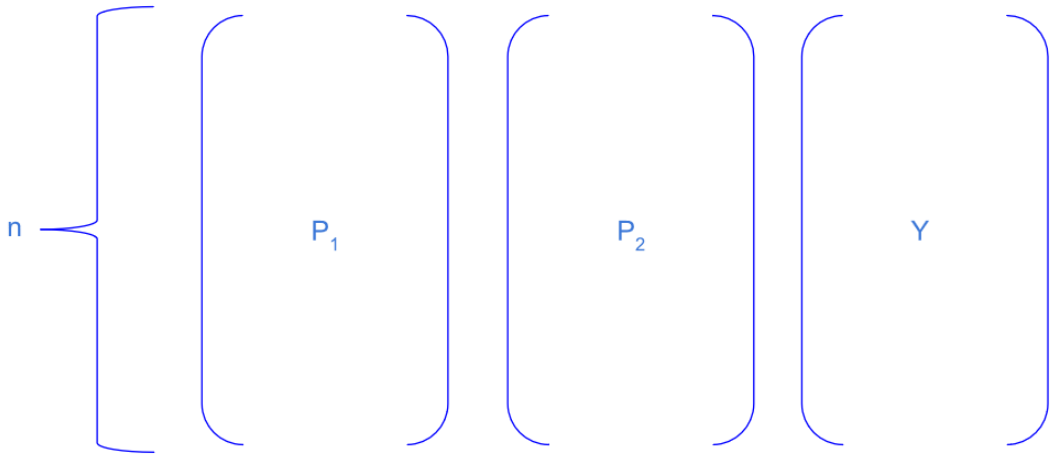

Figure 4.3. Level-1 data for two base learners (GBM and RF).

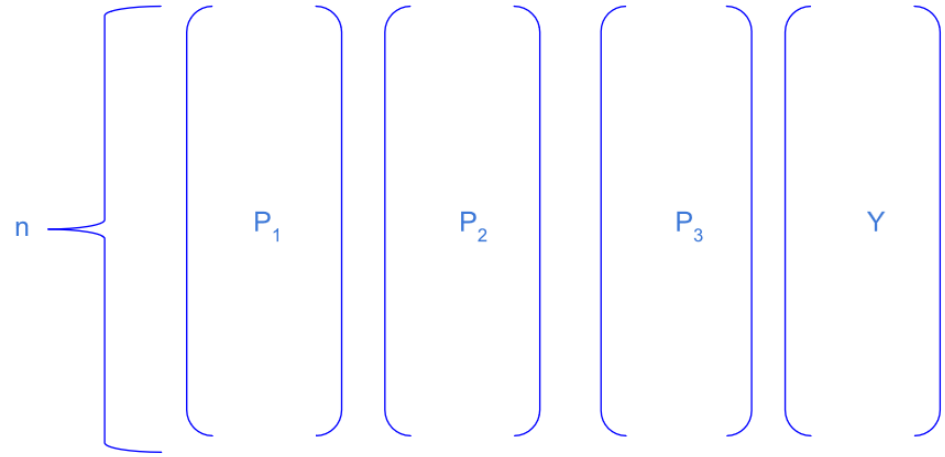

Figure 4.4. Level-1 data for three base learners (GBM, RF, and DNN).

level-1 data for the ensemble having three base learners, which is shown in Fig. 4.4.

Please note that for each of the base learners the best model was selected based on the mean squared error (MSE) which is the average squared difference between the estimated values and the actual values. This was done once the grid search on the training data was complete, and then we queried the grid object and sorted the models by the performance metric MSE. Finally, for each base learner the model having the minimum MSE was selected.

(b) Meta-learner

– In the stacked ensemble, for the meta-learner we have used GLM available in H2O. We trained the level-1 data using GLM with default parameters to get the prediction values for the training data set. Firstly, for the stacked ensemble having two base learners GBM and RF were used for the parameter named base model with the other specified default parameters discussed in *Step (a)*. Secondly, for the stacked ensemble having three base learners GBM, RF and DNN were used for the parameter named as base model.

It is important for the stacked ensemble that all base models must have been cross-validated and they all must use the same cross-validation folds. Also, a parameter named 'keep cross-validation prediction' was set to True. In our case, we considered that by using 10 fold cross-validation and setting the 'keep cross-validation prediction' parameter as True for all the base learners.

3. Output generation / results stage

- The last part of our approach was to use the super learner or ensemble-model to generate predictions on the test data.

4.3. Experiments and Results

This section presents the experimental results and performance evaluation of our model. For our experiment we used H2O. We chose Python as the programming language for the implementation using H2O.

4.3.1. Benchmark Data Sets

To evaluate the performance of our model, we used four benchmark data sets related to healthcare. The data sets were chosen from the UCI Machine Learning repository [61], [67]. The first data set named Diabetic Retinopathy Debrecen data, also called Messidor data set, contains features extracted from the Messidor image set to predict whether an image contains signs of Diabetic Retinopathy (DR) or not. It has a total of 1151 instances, 19 attributes, and a class label with binary outcome 1 or 0, where 1 represents 'sign of DR' and 0 represents 'no sign of DR'. The second data set that we used is the original Wisconsin Breast Cancer (WBC) data set. The goal of this data set is to predict breast cancer. There are 699 records in this database. Each record in the database has nine attributes. In this database, there are a total of 699 instances, among them

241 (65.5%) records are malignant and 458 (34.5%) records are benign. We also used the Pima Indian Diabetes Database (PIDD) and the objective of this data set is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the data set. Various constraints were placed on the selection of these instances from a large database. For example, all patients should include female patients who are at least 21 years old and of Pima Indian heritage. There is a total of 768 records with 268 (34.9%) diabetes patients and 500 (65.1%) non-diabetes patients. The final data set that we used in our evaluation process is the Indian Liver Patient data set (ILPD) that contains 10 variables and a binary variable as output (liver patients or not). The data set contains 441 male and 142 female patient records. There are a total of 583 records with 416 (71.4%) liver patients and 167 (28.6%) non-liver patients. The summary of these four data sets is shown in Table 4.3.

Table 4.3. Data sets description.

Name	**Number of instances**	**Number of attributes**	**Class label**
Messidor	1151	9	Class 0: no sign of DR (540) Class 1: contains sign of DR (611)
WBC	699	19	Class 2: benign (458) Class 4: malignant (241)
PIDD	768	8	Class 0: non-diabetes patients (500) Class 1: diabetes patients (268)
ILPD	583	10	Class 1: liver patients (416) Class 2: non-liver patients (167)

We constructed the training data and the test data for all the data sets that we used in this research. The training set contains 80% of the data while the test set contains the remaining 20%. The Stratified shuffle split technique available in scikit-learn (sklearn), a machine learning library for the Python programming language, was used since it preserves the percentage of samples for each class.

4.3.2. Evaluation Measures

To measure the performance of our model, several evaluation measures were used such as Sensitivity, Specificity, and Accuracy [51]. These were derived from the confusion matrix, and applied to the classifier evaluation, and are shown in Equation (1) through (3).

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{4.1}$$

$$Sensitivity = TP/(TP + FN) \tag{4.2}$$

$$Specificity = TN/(TN + FP) \tag{4.3}$$

where TP is the number of positive examples correctly classified; TN is the number of negative samples correctly classified; FN is the number of positive observations incorrectly classified and FP is the number of negative samples incorrectly classified.

In addition, the Area under the Receiver Operating Characteristic curve (ROC) were also measured [51]. This is because almost all data sets used in this research can be considered as imbalanced data sets. This metric has been widely used as the standard measure for comparison of the performance. The ROC curve is a representation of the best decision boundaries for the cost between the True Positive Rate (TPR), and the False Positive Rate (FPR) that are defined in Equation (4) and (5). The ROC curve plots TPR against FPR.

$$TPR = TP/(TP + FN) \tag{4.4}$$

$$FPR = FP/(FP + TN) \tag{4.5}$$

The area below the ROC curve is called AUC and is widely utilized for weighing classifier performance. The value of AUC ranges from 0.0 to 1.0, where a value of AUC equals 1.0 means perfect prediction, a value of 0.5 means random prediction, and a value less than 0.5 is considered as a poor prediction.

4.3.3. Results

In this research, we compared the performance of our method with the individual base learners used in this research, baseline ensemble, and best results available so far in the literature. We applied the stacked ensemble or super learner (SL) methods on the training data. For the evaluation of the model, we used the test data set. Table 4.4 shows the performance (different evaluation metrics) of the proposed technique (SL having two base learners - GBM and RF) on the test data for the different data sets while Table 4.5 shows the performance of SL having three base learners namely GBM, RF, and DNN on test data for all the data sets used in this research.

Table 4.4. Performance of the proposed techniques on test data (SL consisting of two base learners - GBM and RF).

Data sets	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**AUC**
Messidor	90.24	45.37	69.26	0.806
WBC	100.00	97.83	98.57	0.997
PIDD	90.74	76.00	81.17	0.882
ILPD	94.12	51.81	64.10	0.733

Table 4.5. Performance of the proposed techniques on test data (SL consisting of three base learners - GBM, RF and DNN).

Data sets	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**AUC**
Messidor	78.86	79.63	79.22	0.847
WBC	100.00	98.91	99.29	0.998
PIDD	96.30	73.00	81.17	0.886
ILPD	70.59	72.29	71.80	0.730

Comparing Table 4.4 and Table 4.5, for all the data sets used in this research, best results (based on test data) were obtained using the super learner methods (either SL consisting of two base learners or SL consisting of three base learners). For the Messidor data, best AUC, specificity,

and accuracy were obtained when SL consisting of three base learners applied on the test data and for sensitivity best results were reported using SL with two base learners. Interestingly, for WBC the best performance was obtained when SL consisting of three base learners applied on the test data for all the performance measures considered in this research. For the PIDD data set, best AUC, sensitivity and accuracy were obtained when SL consisting of three base learners applied on the test data and for specificity best results were reported SL with two base learners. For the ILPD, best AUC and sensitivity were obtained with SL consisting of three base learners and for accuracy and specificity best results were achieved SL consisting of two base learners.

In addition, the accuracy comparison using single base learners, baseline ensemble, the SL consisting of two base learners (GBM and RF), and the SL having three base learners (GBM, RF, and DNN) on test data are presented in Table 4.6. We also compare AUC using single base learners, baseline ensemble, SL consisting of two base learners (GBM and RF), and the SL that consist of three base learners (GBM, RF, and DNN) on the test data set are shown in Table 4.7.

Table 4.6. Accuracy comparison using single base learners, baseline ensemble, and super learner consisting of two base learners (BLs) and three BLs on test data (**bold** indicates the best value).

Data set	Accuracy (%) (GBM)	Accuracy (%) (RF)	Accuracy (%) (DNN)	Accuracy (%) (ensemble)	Accuracy (%) (SL - 2 BLs)	Accuracy (%) (SL - 3 BLs)
Messidor	71.86	67.53	77.92	69.86	69.26	**79.22**
WBC	**99.29**	98.57	**99.29**	97.90	98.57	**99.29**
PIDD	79.22	**81.17**	74.68	75.33	**81.17**	**81.17**
ILPD	63.32	64.10	65.81	70.16	64.10	**71.80**

Table 4.7. AUC comparison using single base learners, baseline ensemble, and super learner of having two base learners (BLs) and three base learners (**bold** indicates the best value).

Data set	AUC (GBM)	AUC (RF)	AUC (DNN)	AUC (ensemble)	AUC (SL - 2 BLs)	AUC (SL - 3 BLs)
Messidor	0.815	0.765	0.838	0.740	0.806	**0.847**
WBC	0.997	0.997	**0.998**	0.996	**0.998**	**0.998**
PIDD	0.876	0.882	0.872	0.808	0.882	**0.886**
ILPD	0.718	0.727	0.733	0.730	0.727	**0.734**

From the Table 4.6, it is explicit that our proposed method SL having three base learners performs slightly better (or equal in few cases) than other methods for all the data sets used in this research. For the Messidor data set, SL with three base learners has the best accuracy (79.22%) followed by the individual learner DNN (77.92%). For PIDD, the best accuracy (81.17%) was obtained with both SL methods (having two and three base learners) and with an individual learner named RF. For ILPD, the best accuracy (71.80%) was obtained when the SL method with three base learners was applied on the test data followed by the baseline ensemble (70.16%).

Similar trends are also observed in Table 4.7, the best AUC value was obtained using the super learner having three base learners for all the data sets used in this research. For the Messidor data set, the best AUC value (0.847) was reported with SL consisting of three base learners followed by individual base learner DNN (0.838). For WBC, the best AUC score (0.998) was reported with both SL methods (using two and three base learners) and an individual learner named DNN. For PIDD, the best AUC (0.886) was attained with the SL method consisting of three base learners followed by SL with two base learners and an individual learner RF (0.882). For ILPD, the best AUC (0.734) was obtained with the SL method having three base learners followed by an individual learner named DNN (0.733).

We also present the ROC analysis for all data sets that have been used in this research using all the base learners and the super learner. The ROC plots using the base learners namely GBM, RF, and DNN for all data sets (test) are shown in Figure 4.5, Figure 4.6, and Figure 4.7, whereas the ROC plots using the super learner or stacked-ensemble for the data sets are shown in Figure 4.8.

4.3.4. Performance comparison of four benchmark data sets with other methods

Several ML techniques have been used for the four benchmark data sets that we used for the evaluation of the performance. Authors in [61] used an ensemble-based technique on the Messidor data set with 10-fold cross validation; they obtained 90% sensitivity, 91% specificity, 90% accuracy, and 0.989 AUC. Authors in [62] showed the comparison of five different classifiers based on 10-fold cross validation on the WBC data sets. Among these classifiers, the best accuracy (about 97%) was obtained by SMO. The authors also used feature selection method named Principal Component Analysis (PCA) on the WBC data set with the J48, an open source java implementation of the C4.5 decision tree algorithm and MLP classifiers, and the best accuracy achieved was 97.57%.

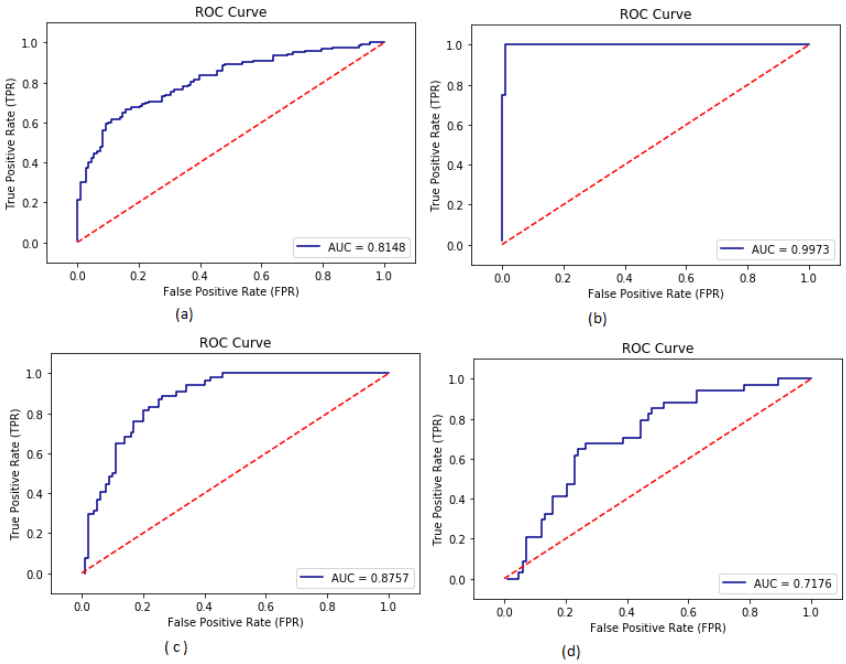

Figure 4.5. ROC analysis using GBM for different data sets used: (a) messidor / diabetic retinopathy (DR), (b) wisconsin breast cancer diagnostics, (c) pima indian diabetes, and (d) ILPD (indian liver patient data set).

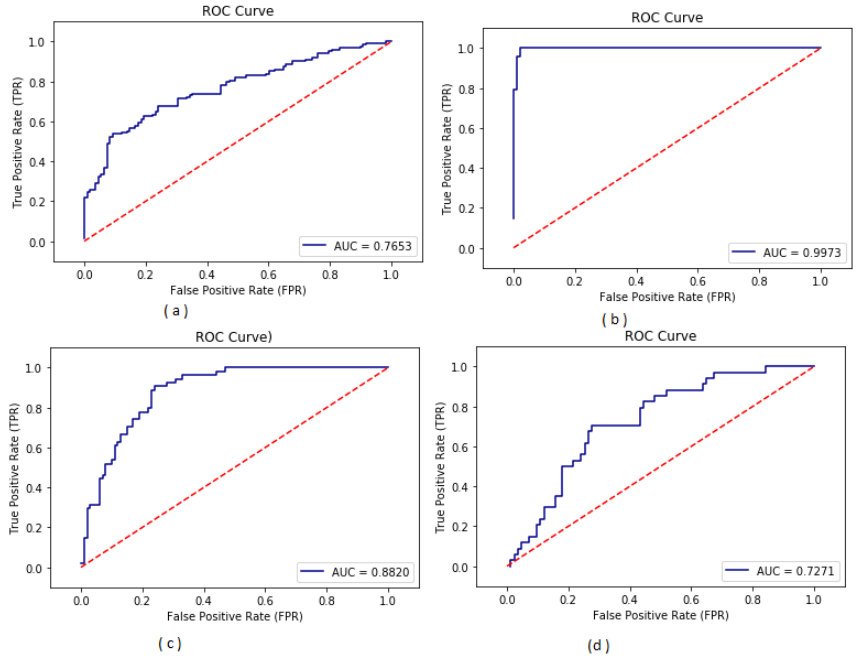

Figure 4.6. ROC analysis using RF for different data sets used: (a) messidor / diabetic retinopathy (DR), (b) wisconsin breast cancer diagnostics, (c) pima indian diabetes, and (d) ILPD (indian liver patient data set).

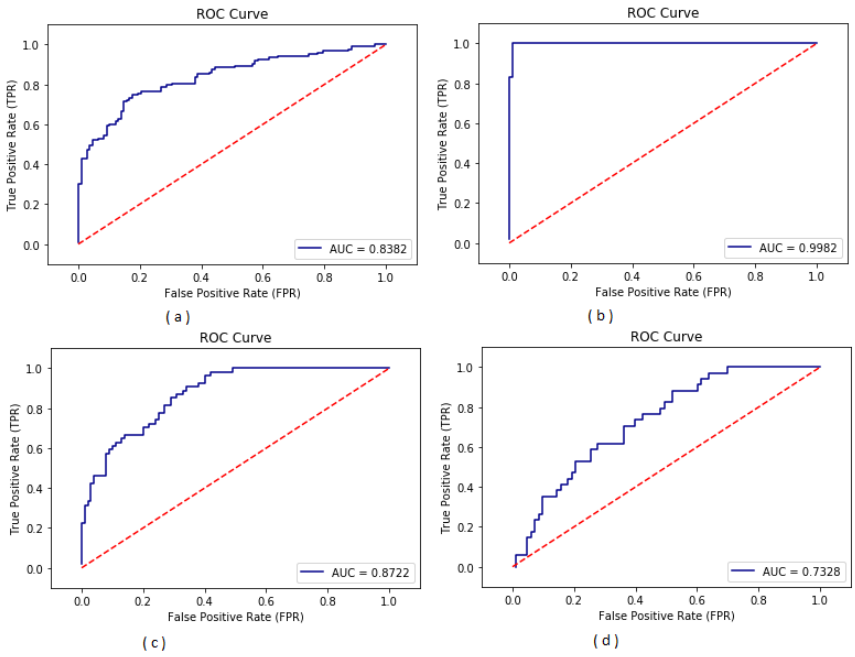

Figure 4.7. ROC analysis using DNN for different data sets used: (a) messidor / diabetic retinopathy (DR), (b) wisconsin breast cancer diagnostics, (c) pima indian diabetes, and (d) ILPD (indian liver patient data set).

In [3], authors compared the performance in terms of accuracy of bagging, and boosting with a hybrid approach of a Hierarchical and Progressive Combination of Classifiers (HPCC). They found a 83.34% accuracy for HPCC, and 82.39% for bagging with GLM. The authors did not explicitly mention the number of cross-validation they used in their experiments. In [63], the authors used GA for attribute or feature selection methods, and a NB classifier has been used for classification on PIDD. For PIDD, the authors partitioned the data set with a split of 70% / 30% for training and testing, respectively. They obtained an accuracy of 77.3%, and 76.95% for training and testing, and an AUC of 0.816 and 0.846, respectively. For the ILPD data set, the best accuracy (79.38%) was found using an ensemble classifier with 5 fold cross-validation [64]. In [65], the authors provided a comparative analysis of different ML algorithms for the diagnosis of different data sets. For ILPD, the best accuracy (79.66%) was obtained by SVM.

We summarized and compared the results that we obtained using the SL methods with the state-of-the-art (SA) best results based on the four benchmark data sets outlined in Table 4.8. From the table, for the SL methods all the values were obtained using three base learners except the sensitivity for the Messidor data (indicated in *italics*), which were achieved using two base

learners. It is important to note that in our experiments, we used 80% for training and 20% for testing for all data sets used and the results were evaluated on the test data.

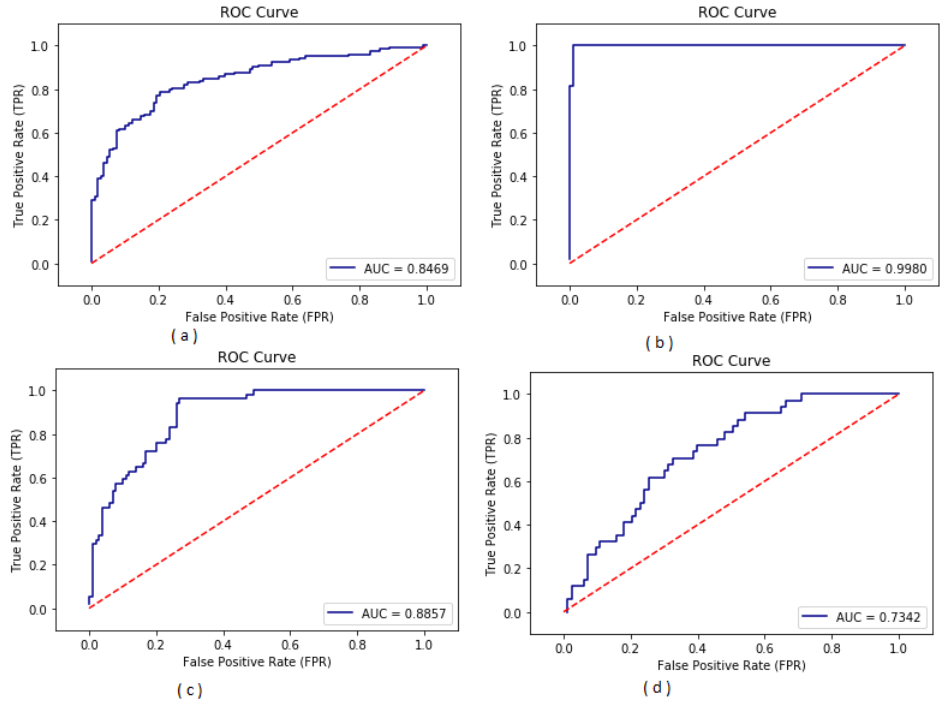

Figure 4.8. ROC analysis using the super learner (using three base learners) for different data sets used: (a) messidor / diabetic retinopathy (DR), (b) wisconsin breast cancer diagnostics, (c) pima indian diabetes, and (d) ILPD data sets.

Table 4.8. Comparison of super learner (SL) methods, and state-of-the-art (SA) best results for the four benchmark data sets (*italics* indicates that the result is obtained using the SL method consisting of two base learners).

Data set	**Sensitivity (%)**		**Specificity (%)**		**Accuracy (%)**		**AUC**	
	SA	**SL**	**SA**	**SL**	**SA**	**SL**	**SA**	**SL**
Messidor	90.00	*90.24*	91.00	79.63	90.00	79.22	0.990	0.847
WBC	-	100.00	-	98.91	97.57	99.29	-	0.998
PIDD	-	96.30	-	76.00	76.95	81.17	0.846	0.886
ILPD	-	94.12	-	72.29	79.66	71.80	-	0.733

4.4. Summary

Classification is one of the important tasks of machine learning that predicts the target class for each example in the data. To achieve good performance on the available data sets,

researchers are using appropriate single classifiers. However, selecting the best data mining or machine learning model for a specific problem is complex. Due to this researchers are using multiple different models for a particular problem to obtain good performance. In this chapter, we focused on the improvement of the classification performance in terms of sensitivity, specificity, accuracy, and AUC for four benchmark data sets related to healthcare. To do so, we used the super learning or stacked-ensemble method that finds the optimal weighted average of diverse learning models. For the base learners we first used GBM and RF and then used another base learner DNN along with the previous two - GBM and RF. To find the optimal combination of the base learner models used in this research, Generalized Linear Models (GLM) was used as the meta-learner.

From our experimental results, we showed that super learning has a better performance compared to individual base learners, baseline ensemble approach, and some of the state-of-the-art techniques for these four benchmark data sets.Using the stacked ensemble or super learner methods (using two base learners or three base learners), we achieved better or equal performance compared to the individual base learners and the baseline ensemble for all the evaluation metrics considered in this research.

In our future work, we plan to apply this technique to other health related big data problems. In addition, we will investigate research problems by including more diverse base learners and other meta-learner. Finally, this technique could be applied to other real world problem domains such as cyber security, Geographic Information System, transportation, and agriculture.

5. CLASSIFICATION MODELS AND SURVIVAL ANALYSIS FOR PROSTATE CANCER USING RNA SEQUENCING AND CLINICAL DATA

Early detection of cancer can significantly increase the chance of successful treatment. This research performs a study on early cancer detection for prostate cancer patients from whom cancer tissue was analyzed with Illumina Hi-Seq ribonucleic acid (RNA) Sequencing (RNA-Seq). Cancer relevant genes with the most significant correlations with the clinical outcome of the sample type (cancer / non-cancer) and the overall survival (OS) were assessed. Traditional cancer diagnosis primarily depends on physicians' experience to identify morphological abnormalities. Gene expression level data can assist physicians in detecting cancer cases at a much earlier stage and thus can significantly improve the potential of patient treatment. In this research, for the classification task, we applied machine learning and data mining approaches to detect cancer versus non-cancer based on gene expression data. Our goal was to detect cancer at the earliest stage. Besides, for the regression task, survival outcomes in prostate cancer patients were performed. Regression trees were built using cancer-sensitive genes along with clinical attribute 'Gleason score' as predictors, and the clinical variable 'overall survival' as the target variable. Knowledge in the form of rules is one of the vital tasks in data mining as it provides concise statements of easily understandable and potentially valuable information. For the classification model, we derived rules from a decision tree and interpreted these rules for cancer and non-cancer patients. For the regression or survival model, we generated rules for predicting or estimating the survival time of cancer patients. In this study, cancer-relevant genes were analyzed as predictors, although various genes may interact with genes currently known to contribute to cancer. These findings have implications for assessing gene-gene interactions and gene-environment interactions of prostate cancer as well as for other types of cancer.

In this chapter, for detecting cancer different classification models were built and for the prediction or estimation of survival time several regression models were presented. Gene expression of prostate data of cancer-relevant genes along with the clinical variable 'Gleason score' were used

as predictors. For the classification task, sample type (cancer / non-cancer) was used as the target variable while for regression or survival prediction the overall survival (OS) was used as the target variable.

The rest of the chapter is organized as follows. Section 5.1 describes state-of-the-art techniques; Section 5.2 presents methodology including data characteristics, feature selection techniques for both classification and survival analysis. Also, model building along with brief descriptions of the algorithms are provided. Section 5.3 illustrates the experimental results where the results obtained from various feature selection models are provided and discussed. Also, the rule generation from both decision tree and regression tree are shown in this section. Section 5.4 is our discussion section. Section 5.5 is the summary section; we conclude the chapter and suggest possible future research directions.

5.1. Related Work

In machine learning or data mining, classification is an example of supervised learning techniques. The goal of classification training a classification model is to predict qualitative or categorical outputs which assume values in a finite set of classes without an explicit order [16]. Regression models are used to predict one variable from one or more variables. Regression learns a function that maps a data item to a real-valued prediction variable. Many regression methods exist in mathematics, such as linear, non-linear, logistic, and multi-linear regression. Regression models provide the data miner with a powerful tool, allowing predictions about past, present, or future events to be made with information about past or present events [16]. Data mining is often referred to as knowledge discovery in databases and describes the process of nontrivial extraction of implicit, previously unknown and potentially valuable information from a large amount of data[7]. The mined information is referred to as knowledge provided in the form of rules, constraints, and regularities. In data mining, rule mining is one of the vital tasks since rules provide concise statements of potentially valuable information that can be easily understood by end users[10].

Researchers have developed different statistical, data mining, and machine learning models for various cancers detection and estimation of survival time. In most of the cases, the researchers used clinical or patient data. However, gene expression abnormalities always appear before morphological changes can be observed. Therefore, in this research we build the model by investigating gene expression data along with the clinical data.

Next-generation sequencing has revolutionized the field by not only increasing the sequencing depths and accuracy but also reducing the time and cost to an affordable level for individual cancer patients. Therefore, gene expression profiling has become a feasible cancer diagnosis and prognosis. Researchers have developed various models with promising results. Authors in [68] investigated six different machine learning techniques on publicly available datasets of predicting cancer outcome. Besides, the authors also used different feature selection approaches of identifying relevant genes for maximizing predictive information. In [69], the early diagnosis of breast cancer is done using genetic algorithms (GA) along with artificial neural networks (ANN). The authors used GA for feature extraction and parameter optimization of the ANN. Rule generation is one of the vital tasks since rules provide concise statements of potentially relevant information that can be easily understood by end users[10]. The authors in [70], discovered useful rules of breast cancer and non-breast cancer patients from risk factors data using association rule mining techniques.

In this research, we used the gene expression of prostate data of cancer-relevant genes along with a clinical variable 'Gleason score' as predictors. For the classification task, sample type (cancer / non-cancer) was used as the target variable, while for regression or survival prediction the overall survival (OS) was used as the target variable. Furthermore, knowledge in the form of rules was generated from both the classification and regression models. These rules can be useful for physicians or biologists to investigate i) the relationship between the overall survival and specific gene expression levels, and ii) the association between sample type and specific gene expression levels in prostate cancer.

5.2. Methods

5.2.1. Data Characteristics

RNA-seq and clinical variables available from the National Cancer Institute Genomic Data Commons (GDC) were investigated in this research. These variables were integrated in order to detect cancer cases and survival predictions based on the level of individual variables as well as the interaction of these variables, including RNA-Seq and clinical predictors. Illumina Hi-Seq RNA sequencing $log2(x+1)$ normalized data was merged with clinical variables accessible from the GDC.

There were a total of 550 instances in the prostate cancer data set. Among them, 497 were primary tumor samples (cancer patients), and 52 were solid tissue standard samples (non-cancer individuals). There was only one sample named as metastatic tumor, which has been considered

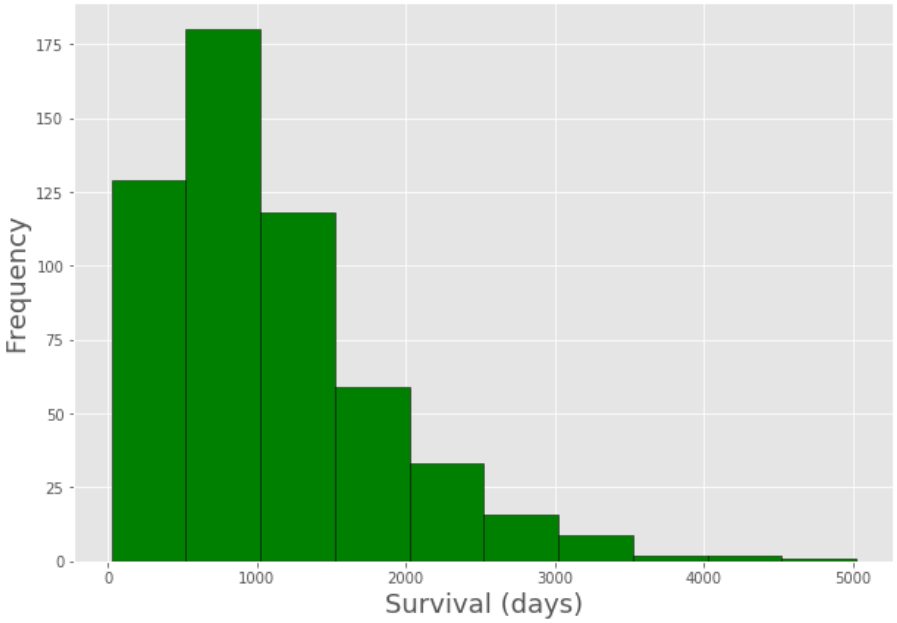

Figure 5.1. Distributions of clinical variable overall survival (OS).

as a primary tumor. So, total primary tumor or cancer samples counted were 498, and normal or non-cancer samples were 52. There were more than twenty thousand (20,000) genes, however, in this research we only consider 36 common genes that are associated with cancer (according to the National Cancer Institute Genomic Data Commons [71]).

Thirty-six (36) cancer-relevant genes (AR, BRCA1, BRCA2, CD82, CDH1, CHEK2, EHBP1, ELAC2, EP300, EPHB2, EZH2, FGFR2, FGFR4, GNMT, HNF1B, HOXB13, IGF2, ITGA6, KLF6, LRP2, MAD1L1, MED12, MSBM, MSR1, MXI1, NBN, PCNT, PLXNB1, PTEN, RNASEL, SRD5A2, STAT3, TGFBR1, WRN, WT1, and ZFHX3) and clinical variable 'Gleason score' (an index of cancer stage) of prostate cancer were assessed as predictors of tissue type (cancer or non-cancer).

Besides, these cancer-sensitive genes, along with clinical variable 'Gleason score' of prostate cancer were assessed as predictors in survival analysis to predict overall survival (OS). The goal of the survival analysis is to increase the ability to predict survival time based on the expression level of predictors genes and the clinical variable 'Gleason score'. The distribution of overall survival is shown in Fig. 5.1.

5.2.2. Feature Selection Approaches

A multivariate correlation analysis was performed to observe the correlation of predictor variables with the target variable. Predictors were filtered and sorted with the absolute correlation coefficient value. A value closer to 0 implies a weaker relationship, and a value closer to 1 means a stronger correlation with the target. For the classification model, a multivariate correlation analysis was performed to observe the association of predictor variables with the target variable named as sample type (cancer or non-cancer). For survival prediction, the correlation was done with the target variable named as overall survival (OS). Predictors or genes were filtered and sorted with the absolute correlation coefficient value with cancer/sample type and then with OS, respectively.

The area of feature selection in machine learning has become quite robust. There are numerous feature selection algorithms which identify the features from given data that contributes the most to the target variable [72]. An extra-trees classifier and select-K-best approaches were investigated to obtain relevant or essential features for building the classification models. An extra-tree or extremely randomized trees classifier [73] implements a meta-estimator that fits several randomized decision trees named as extra-trees on various sub-samples of the data set. It is very similar to a Random Forest Classifier and only differs in the way the construction of the decision trees is done using the forest. In the feature selection process, the Gini index is used in the creation of the forest. Each feature is ordered in descending order according to the Gini importance of each feature, and the user can select the top K features accordingly. The Select-K-best algorithm extracts features according to the highest scores. It calculates a chi-square statistic between each feature and the target variable. The implementation of these algorithms was performed using the Scikit-learn python package [74].

The Cox (proportional hazards or PH) model is the most commonly used multivariate approach for analyzing survival time data in medical research [75]. The Cox regression model extends survival analysis methods to assess the effect of several risk factors on survival time simultaneously. The model is used to identify the impact of predictors on the survival of cancer patients. This model makes it possible to isolate variables that have little effect on survival. Furthermore, the model allows estimating the risk or danger of death for an individual based on the prognostic (good for survival) variables. The output of the Cox (ph) regression model, along with the hazard ratio,

was investigated to select a good predictor (good prognostic factor) for survival. The Hazard Ratio (hr) assesses the overall survival or the risk of death by the predictors. Generally, the value of hazard ratio less than 1.0 is considered a good predictor (good prognostic factor) for survival, while the value of hr greater than 1.0 is considered not good for survival (bad prognostic factor).

5.2.3. Classification and Regression Techniques

The classification techniques that we investigated in this paper are decision tree (DT), random forest (RF), and multi-layered neural network (MLP or NN). Besides, for the survival analysis, the decision tree regressor was investigated.

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label [76], [77]. It works very well with different types of data, and results are easy to understand. The process of model building is comparatively easy compared to other algorithms, and data can be represented in a visual form (tree-like form). From the tree, we can generate or form rules that can be used to classify unknown values. The decision tree classifier has been widely applied to solve many real-world problems in different fields [78], [79].

Random forest is a robust classification and regression technique that generates a forest of classification trees, rather than a single classification tree. RF creates trees on randomly selected data instances and obtains the prediction from each tree to choose the best solution through voting. RF is considered as a highly accurate and robust technique because it generates many trees in the process [78], [79].

A neural network is a set of connected input/output units in which each connection has a weight associated with it [16]. During the learning phase, the network learns by adjusting the weights to be able to predict the correct class label of the input tuples. Neural networks involve long training times and are therefore more suitable for applications where this is feasible. The most popular neural network algorithm is back-propagation – Multilayer feed-forward networks. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.

A regression tree is similar to a classification tree, except that the target variable is continuous, and a regression model is fitted to each node to return the prediction value of target variable

[80]. Here, the tree is used to predict the value for unknown cases. For regression, the prediction error is typically measured by the squared difference between the observed and predicted values.

5.2.4. Building Models

For cancer detection, we built classifier models with 36 cancer-sensitive genes and the clinical variable 'Gleason score'. The target variable is tissue or sample type (cancer or non-cancer). A decision tree, a random forest, and multi-layer neural networks were selected as classifiers. The default parameter values were used for the random forest algorithm. For the decision tree algorithm, the maximum depth of the tree was specified as six, and for the multi-layered neural network two hidden layers with 25 and 12 nodes were used. The same procedure was followed with predictors or genes that were considered or selected using the feature selection approaches.

For building the classification models and the prediction (survival) model, we split the data into 70 % training set, and 30 % test set with stratified train test split.

5.2.5. Rule Generation from Tree

From the built trees, we generated knowledge in the form of rules. For the classification model, a decision tree was built, and from the tree, rules were generated for both cancer and non-cancer patients. For the regression model, we created rules for the estimation of survival time. To obtain a rule, we need to follow the tree down from the root to the leaf nodes.

5.3. Experiments and Results

Results of the feature extraction for both the classification models and survival prediction are discussed in this section. Moreover, the performance measure of the classifiers and both classification (decision) tree and regression tree are shown here. Finally, knowledge discovery in the form of rules from both decision tree (cancer detection) and regression tree (survival prediction) are shown and elaborated.

5.3.1. Output of Feature Selection for Classification Model

Genes correlated with sample or cancer type were determined. Correlations of selected cancer-relevant genes with a clinical variable named as sample type were represented in heat maps and genes in the order of those with the highest absolute value of association with cancer type are EZH2, HOXB13, RNASEL, FGFR2, SRD5A2, CD82, MXI1, MAD1L1, IGF2, ITGA6, PTEN.

The important genes with clinical variable sample type that were obtained using the extra tree classifier are shown as a bar graph in Fig. 5.2. The genes are given in the order of the

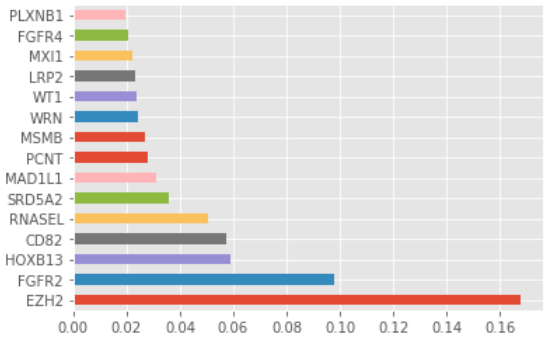

Figure 5.2. Important features that were obtained using extra tree classifier.

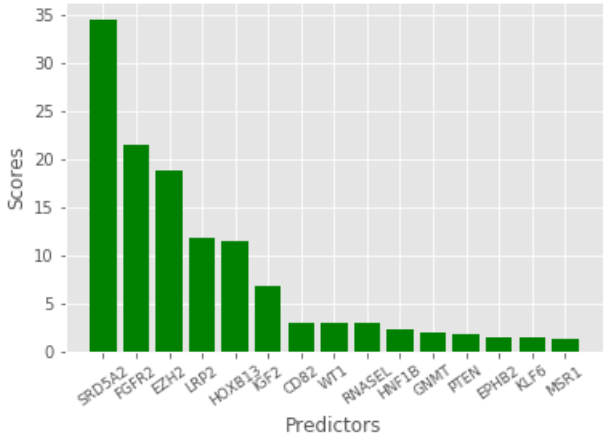

Figure 5.3. Feature selection using K best features.

importance, which are EZH2, FGFR2, HOXB13, CD82, RNASEL, SRD5A2, MAD1L1, PCNT, MSMB, WRN, WT1, LRP2, MXI1, FGFR4, PLXNB1.

The SelectKBest technique selects K best features according to the highest scores. Fifteen ($K = 15$) predictors or genes according to the highest score are: SRD5A2, FGFR2, EZH2, LRP2, HOXB13, IGF2, CD82, WT1, RNASEL, HNF1B, GNMT, PTEN, EPHB2, KLF6, MSR1 are shown in Fig. 5.3.

5.3.2. Selected Predictors for Classification Model

The three aforementioned feature extraction approaches were applied. Most essential predictors that were common in all three techniques are EZH2, HOXB13, RNASEL, FGFR2, SRD5A2, and CD82. We also selected more features (MXI1, MAD1L1, IGF2, PTEN, WT1, and LRP2) that were common in any two techniques.

5.3.3. Performance Measure of Classifiers

To evaluate the performance, several measures were used such as accuracy, recall, precision, area under the Receiver Operating Characteristic curve (ROC) or AUC, and F-measure [78], [79], [51]. These were derived from the confusion matrix and applied to the classifier evaluation.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{5.1}$$

$$recall = TP/(TP + FN) \tag{5.2}$$

$$precision = TP/(TP + FP) \tag{5.3}$$

Here, TP denotes the number of positive examples correctly classified, TN denotes the number of negative samples correctly classified, FN represents the number of positive observations incorrectly classified, and FP indicates the number of negative samples incorrectly classified by the estimator. The ROC curve is a representation of the best decision boundaries for the cost between the True Positive Rate (TPR) and the False Positive Rate (FPR). The ROC curve plots TPR against FPR. TPR and FPR are defined as follows:

$$TPR = TP/(TP + FN) \tag{5.4}$$

$$FPR = FP/(FP + TN) \tag{5.5}$$

The F-measure or F1 score is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test, which is defined as follows:

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \tag{5.6}$$

Detailed information about these measures can be found in [78], [79].

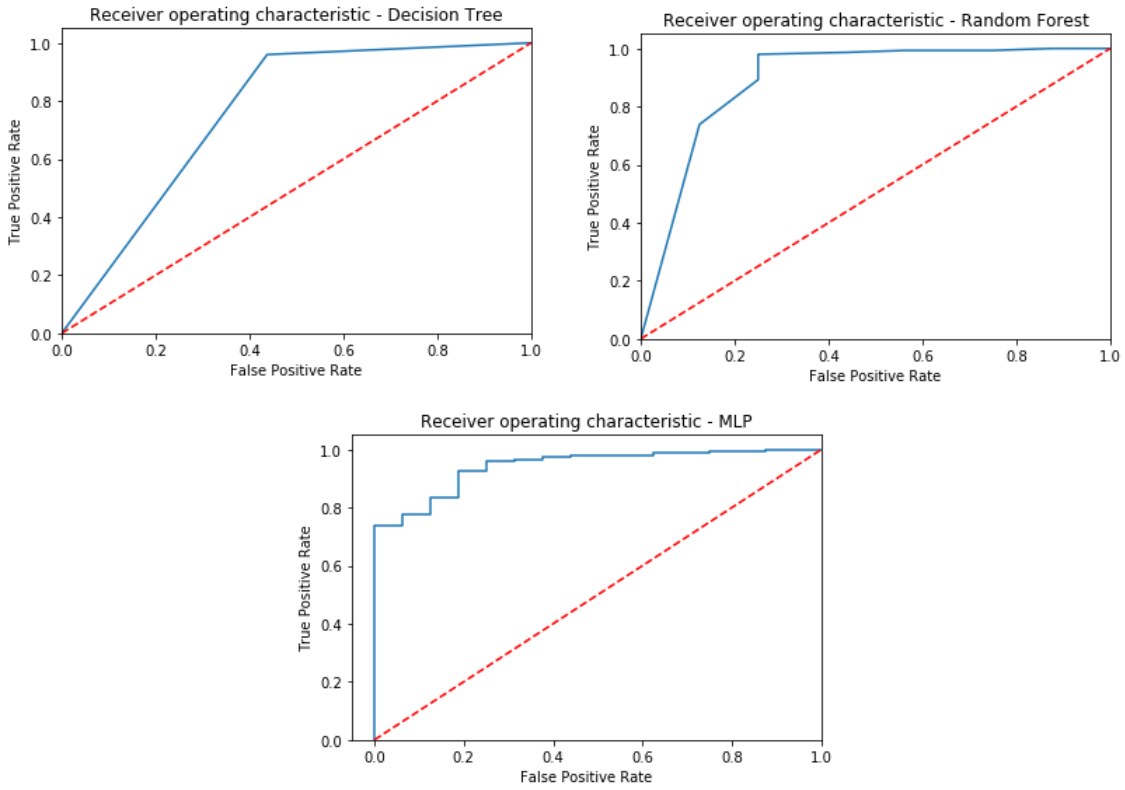

Figure 5.4. ROC curve for three specified classifiers.

5.3.4. Results of Classifiers

In this paper, we applied three different classifier models on the training data and compared the performance of the trained models on the test data. The overall performance of the classification models is shown in Table 5.1. Results were evaluated on the test data.

Table 5.1. Overall performance based on test data.

Methods	**Precision**	**Recall**	**F1-score**	**Accuracy (%)**	**AUC**
DT	0.92	0.92	0.92	92.1212	0.7611
RF	0.96	0.96	0.96	95.7576	0.8920
MLP	0.94	0.94	0.94	93.9393	0.9421

Area under the Receiver Operating Characteristic curve (ROC) or AUC for these three classifiers are shown in Fig. 5.4.

In the second step of our classification technique, we trained data that were obtained using feature selection approaches (discussed in Section 5.3.2). The overall performance of classifiers are shown in Table 5.2.

Table 5.2. Overall performance based on test data (classifiers trained with selected features).

Methods	**Precision**	**Recall**	**F1-score**	**Accuracy (%)**	**AUC**
DT	0.91	0.91	0.91	90.9090	0.89052
RF	0.96	0.93	0.93	93.3333	0.8744
MLP	0.93	0.93	0.93	93.9393	0.90772

Comparing both tables, we can see that in general multi-layered neural networks (MLP) performs better when we trained the model without the feature selection approach. If we look at the F1 measure, which is the weighted harmonic mean of the precision and recall, the classifiers trained with all features (without feature selection) perform well compared to the trained models with the selected predictors (important features). The reason for this is that all the features contribute to the detection of prostate cancer rather than using fewer predictors.

5.3.5. Generated Rules from Decision Tree

In Fig. 5.5, a tree was shown that was built by applying the decision tree classifier with all 36 genes and the Gleason score as predictors. The root node, with the most information gain indicates the significant gene in determining cancer or non-cancer for prostate data, which is EZH2. The impurity is the measure as given at the top by the Gini score. Samples show the number of instances available to classify, and the value indicates how many samples are in class 0 (non-cancer) and how many samples are in class 1 (cancer).

If we follow the tree down from the root to the leaf nodes, we can find a rule. From the tree, we generated some rules for both cancer (Rules 1 through 5) and non-cancer (rules 6 through 8) patients that are shown as follows:

Rule 1: If the gene expression level of EZH2 is less than or equal 5.494, and the gene expression level of CD82 is less or equal 9.51, then there is a chance that individual will be a cancer patient.

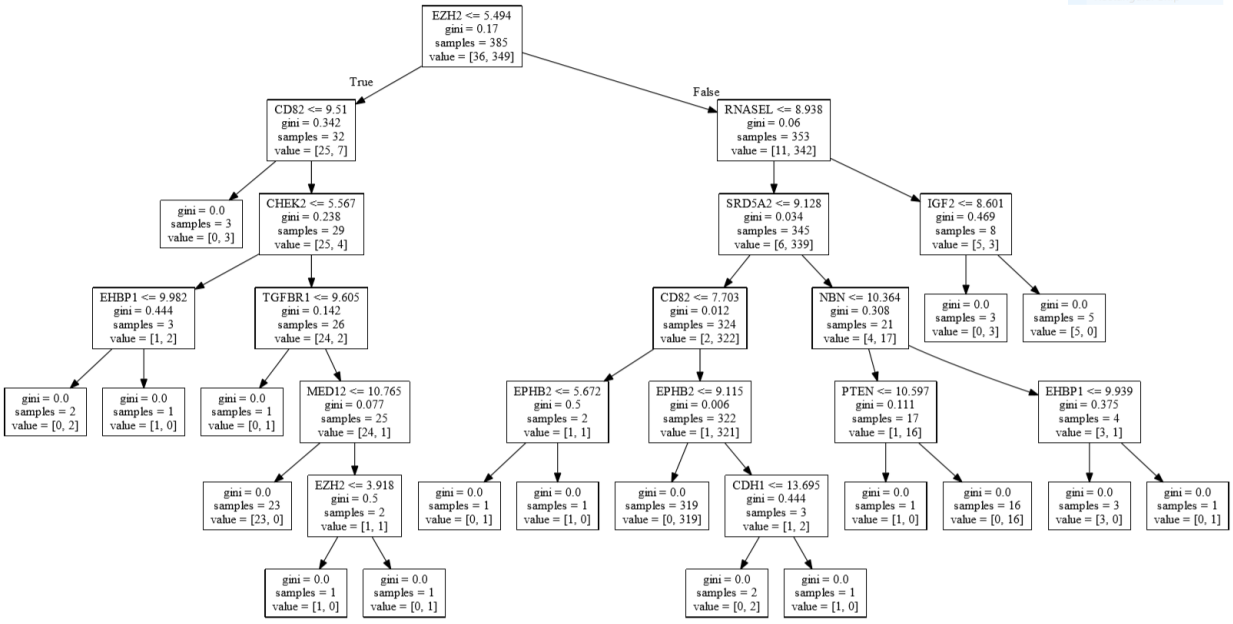

Figure 5.5. A decision tree that was built by using cancer-sensitive genes (without feature selection).

Rule 2: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51, and the gene expression value for CHEK2 is less or equal 5.567, and EHBP1 is less or equal 9.982 then there is a chance that individual will be a cancer patient.

Rule 3: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51, and the gene expression value for CHEK2 is greater than 5.567, and TGFBR1 is less or equal 9.605 then there is a chance that individual will be a cancer patient.

Rule 4: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51, and the gene expression value for CHEK2 is more than 5.567, and TGFBR1 is more than 9.605, and MED12 is less or equal 10.765 then there is a high chance that individual will be a cancer patient.

Rule 5: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51 and the gene expression value for CHEK2 is more than 5.567 and TGFBR1 is more than 9.605 and MED12 is higher than 10.765, and the gene expression level of EZH2 is less than 3.918 then there is a chance that individual will be a cancer patient.

Rule 6: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51, and the gene expression value for CHEK2 is less or equal to 5.567, and EHBP1 is more than 9.982 then an individual will not be a cancer patient.

Rule 7: If the gene expression level of EZH2 is less than or equal 5.494 and the gene expression level of CD82 is higher than 9.51 and the gene expression value for CHEK2 is greater than 5.567 and TGFBR1 is more than 9.605 and MED12 is higher than 10.765, and the gene expression level of EZH2 is more than 3.918 then there is a chance that individual will be a non-cancer patient.

Rule 8: If the gene expression level of EZH2 is higher than 5.494 and the gene expression level of RNASEL is more than 8.938, and the gene expression value for IGF2 is less than 8.601 then there is a chance that individual will be a non-cancer patient.

5.3.6. Results of Feature Selection for Survival Prediction

Genes correlated with a clinical variable overall survival (OS) were determined. Correlations of selected genes with clinical variable overall survival are represented in a heat map that is shown in Fig. 5.6. Genes are given in the order of those with the greatest absolute value of correlation with overall survival (OS): AR, BRCA2, CD82 , CDH1, EPHB2, FGFR2, FGFR4, IGF2, ITGA6, LRP2, MAD1L1, MED12, MSMB, MSR1, PLXNB1, RNASEL, ZFHX3.

In the Cox (ph) regression model, the p-value for all three tests – likelihood ratio test (p = 0.008), Wald test (p= 0.02), and Score (log-rank) test (p= 0.02) are significant, indicating that the model is significant. Also, in the multivariate Cox analysis, the covariates BRCA1, EZH2, and MED12 remain significant. However, other covariates fail to be significant. The output of the Cox (ph) regression model along with the hazard ratio are shown in Fig. 5.7. The hazard ratio (HR) assesses the overall survival or the risk of death by predictors. Good predictors or good prognostic factors that were obtained by applying multivariate Cox (proportional hazards) regression based on hazard ration are BRCA1, CHEK2, EHBP1, EP300, EPHB2, GNMT, HNF1B, IGF2, ITGA6, MAD1L1, MSR1, MXI1, NBN, PCNT, PLXNB1, SRD5A2, WRN, gleason_score.

5.3.7. Decision Tree Regressor for Survival Predictions

We build three regression models by applying the decision tree regressor. In the first model, all cancer sensitive genes along with the clinical variable gleason_score were used as the predictor for predicting the survival time (overall survival - OS). In the second model, variables that had a

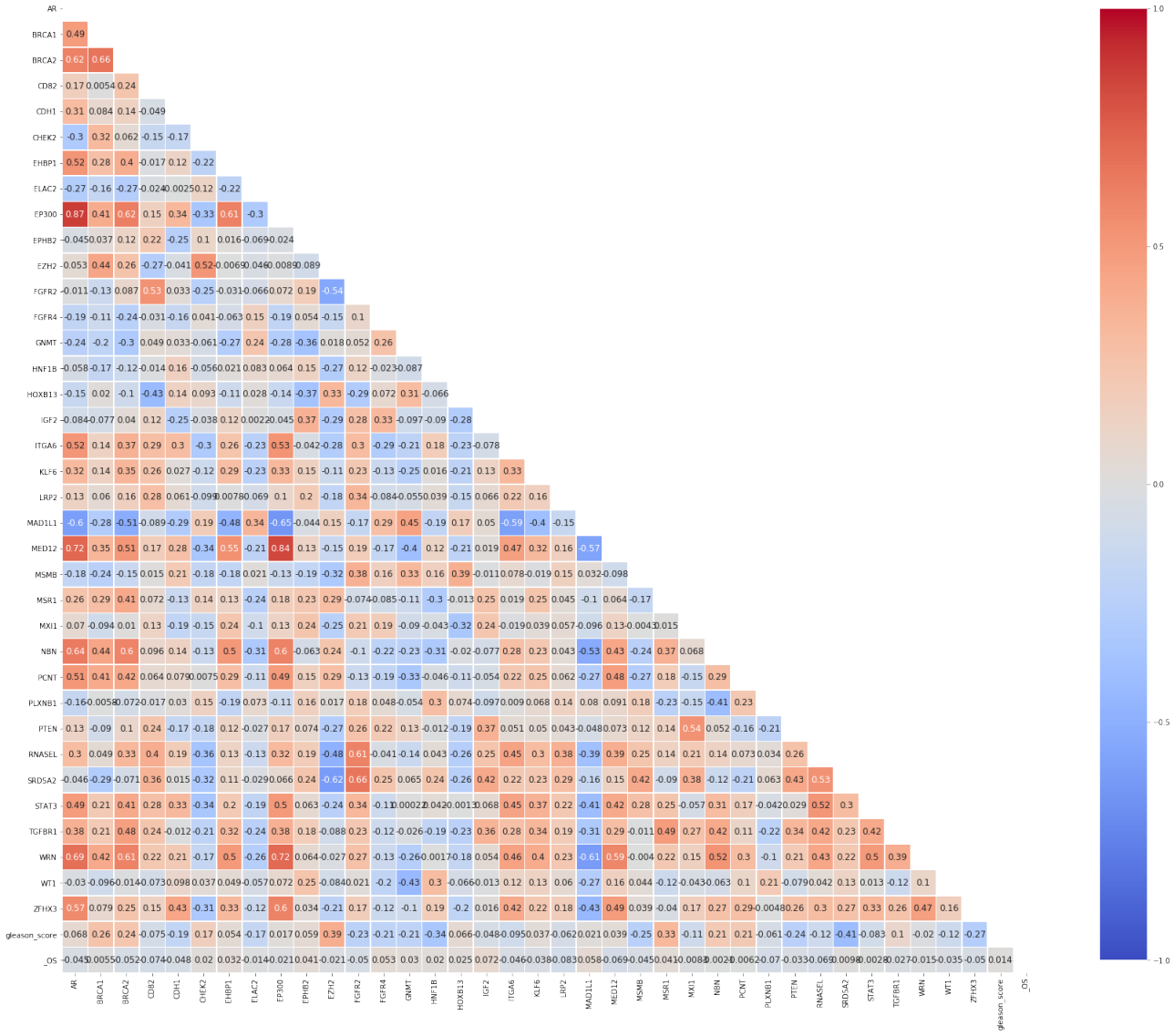

Figure 5.6. Heat map of correlations of cancer-relevant genes with a clinical variable overall survival (OS).

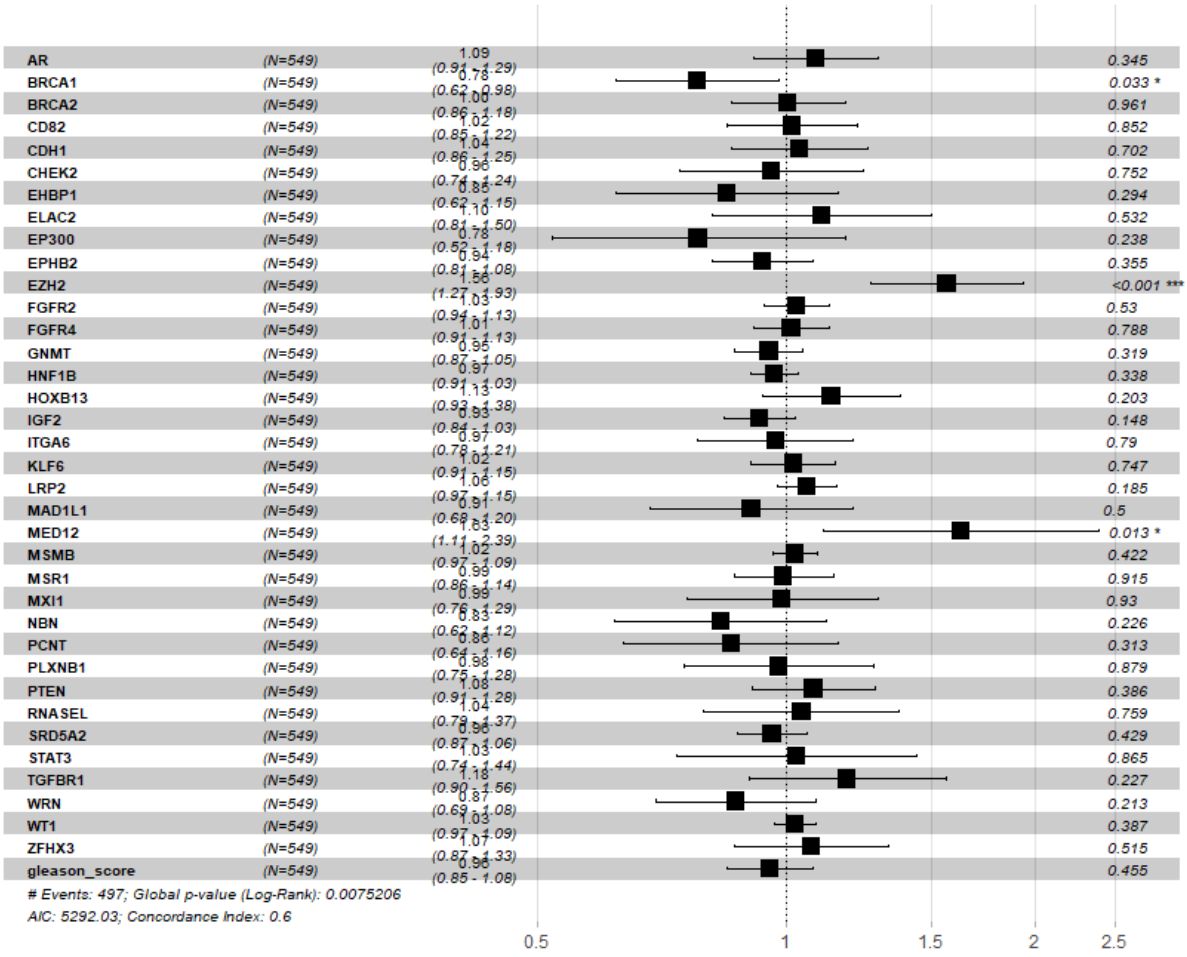

Figure 5.7. The output of cox (ph) regression model along with hazard ratio.

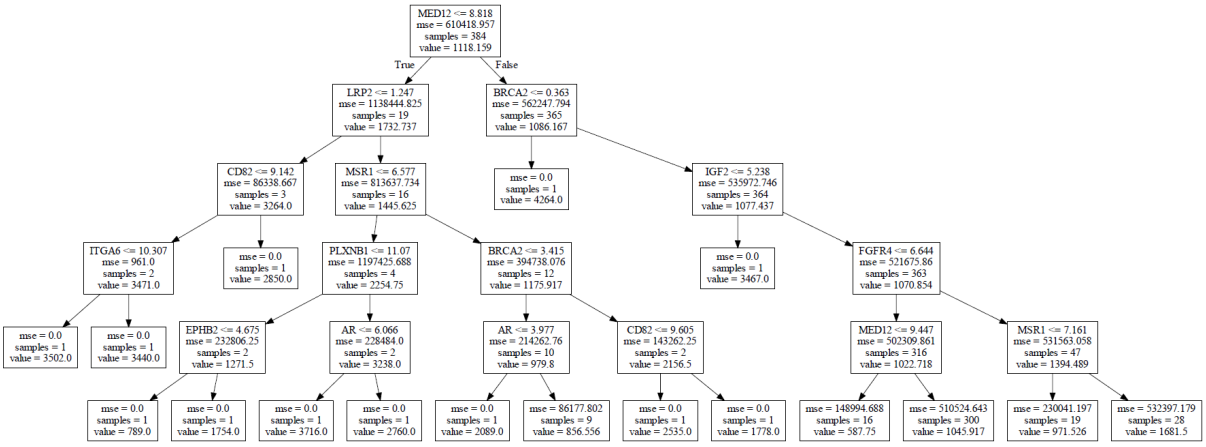

Figure 5.8. A regressor tree that was built by using higher correlation genes with overall survival (OS).

higher correlation with the overall survival were considered. In the final model, variables that were obtained from the Cox (ph) regression model based on the hazard ratio were used for predictors. For the performance evaluation, mean square error (MSE) was considered for the test data. Among these three models, the second model was selected for further study as it has a lower MSE value than the other models.

In Fig. 5.8, a tree was shown that was built by applying the decision tree regressor on higher correlation genes with overall survival (OS). The root node can be considered as the most informative feature or gene for survival prediction. In our cases, MED12 is the most informative gene and then LRP2 or BRCA2 based on the expression value of MED12.

5.3.7.1. Predictions of Survival Time from Decision Tree Regressor

The root node MED12 can be considered as the most important gene for overall survival prediction. If we visit from the root node to a particular leaf node, we can find a rule for survival time prediction. From the regression tree, we can generate the number of rules or knowledge that will be helpful to predict patients' survival time. Some of the rules generated from the regression tree are shown as follows:

Rule 1: If the gene expression level of MED12 is less than or equal 8.818 and the gene expression level of LRP2 is less or equal 1.247 and expression level of CD82 is less or equal to 9.142, and the gene expression level of ITGA6 is less or equal to 10.307 then there is a chance that the patient will survive about 3502 days.

Rule 2: If the gene expression level of MED12 is less than or equal 8.818 and the gene expression level of LRP2 is less or equal 1.247 and expression level of CD82 is less or equal to 9.142, and the gene expression level of ITGA6 is higher than 10.307 then there is a chance that the patient will survive about 3440 days.

Rule 3: If the gene expression level of MED12 is less than or equal 8.818 and the gene expression level of LRP2 is less or equal 1.247 and expression level of CD82 is higher than 9.142 then there is a likelihood that the patient will survive about 2850 days.

Rule 4: If the gene expression level of MED12 is higher than 8.818, and the gene expression level of BRCA2 is less or equal 0.363, then there is a chance that the patient will survive about 4264 days.

Rule 5: If the gene expression level of MED12 is higher than 8.818 and the gene expression level of BRCA2 is more than 0.363 and IGF2 is less or equal 5.238 then there is a possibility that the patient will survive about 3467 days.

Rule 6: If the gene expression level of MED12 is higher than 8.818, and the gene expression level of BRCA2 is more than 0.363, and IGF2 is greater than 5.238, and the gene expression level of FGFR4 is more abundant than 6.644 and MSR1 is less or equal to 7.161 then there is a possibility that the patient will survive about 971 days.

Rule 7: If the gene expression level of MED12 is higher than 8.818 and the gene expression level of BRCA2 is greater than 0.363 and IGF2 is larger than 5.238 and the gene expression level of FGFR4 is greater than 6.644 and MSR1 is higher than 7.161 then there is a chance that the patient will survive about 1682 days.

From the regressor tree, we can generate rules as discussed above and can estimate or predict the survival time or overall survival for a particular patient.

5.4. Discussion

The National Institutes of Health Genomic Data Commons may be utilized to determine which clinical variables and RNA-Seq expression levels detect clinical outcomes, such as sample types and overall survival. In this research, in order to get a clear understanding of RNA-Sequencing and clinical data, we investigated 36 cancer-sensitive genes and few clinical variables. Based on the classification models for cancer detection, we see that the model performs better for unseen cases when we applied all 36 genes and the clinical variable 'Gleason score' as predictors; instead

of applying only a few predictors (obtained by using feature selection approaches). This has implications that for predicting cancer cases, almost all features contribute rather than the selected features. It also implies that for building classification models in cancer detection, all genes (about twenty-thousand) along with other clinical variables should be investigated further.

Furthermore, in survival prediction or estimation, we see the model that uses higher correlated features with overall survival (OS) performs better than the other models. Overall, the correlation of features with overall survival (OS) was very low, which also implies that all genes contribute to the overall survival. This means that for our further studies in survival prediction we should use all the predictors.

In this research, we also generated rules from the decision tree and the regression tree. By looking at the rules, we can see that the level of gene expression plays a vital role in determining if an individual could be a cancer patient or non-cancer patient. For instance, have a look at the rule (Fig. 5.5 - part of the right subtree), if the level of expression of the gene EZH2 is greater than 5.494, and the expression level of gene RNASEL is larger than 8.938 then the gene expression level of IGF2 plays a key function in determining cancer or non-cancer for a particular patient. If the gene expression level IGF2 is less or equal to 8.601, then there is a possibility that an individual will not have cancer; otherwise, there is a high chance that the patient will have cancer. These types of relationships among various genes with corresponding expression levels and clinical variables can be further investigated for personalized medicine research. These type of associations can be found from the regression tree as well.

5.5. Summary

RNA-seq and clinical variables available on the National Cancer Institute Genomic Data Commons (GDC) were investigated in this research. For detecting clinical variable cancer type, we built three different classification models based on decision tree (DT), random forest (RF), and multi-layered neural networks (MLP) using gene expression data. Different feature selection techniques were also investigated to find the most predictive genes, and we developed models using the three aforementioned classification methods on these selected genes. The results showed that MLP performs better on test data when we built the model without applying any feature selection approach.

Also, the prediction of the clinical variable 'overall survival' in prostate cancer was performed by applying i) all 36 genes and the clinical variable 'Gleason score' as predictors, and ii) genes obtained from the feature selection approach. Furthermore, rule generation was performed from a selected decision tree classifier for both cancer and non-cancer patients. Rules discovery was also performed from a selected regression tree for estimating survival outcome.

In this research, we utilized 36 cancer-sensitive genes along with few clinical variables. Future studies will assess all genes (about twenty-thousand) along with more clinical variables.

6. CONCLUSION AND FUTURE WORK

Machine Learning (ML) and data mining (DM) have become an integrative part of contemporary scientific methodology, providing insights about data and offering prediction based on historical observations. The use of DM and ML techniques requires a reasonable understanding of their mechanisms, properties, and constraints in order to understand them better and interpret their outcomes.

In this dissertation, we investigated DM methods on healthcare and medical data to find important information in the form of rules. The aim was to utilize these rules for improving public awareness of different cancer symptoms that could also initiate prevention strategies. Furthermore, we designed and implemented different ML techniques and applied these models in healthcare, medical, and RNA-Sequencing data having imbalanced and high-dimensional characteristics. The intention was to enhance the performance of the model for the unseen data.

In this dissertation, we first investigated class association rule mining, a variant of association rule mining, on healthcare data. These rules can be used to promote public awareness of different cancer symptoms and could also be useful to initiate prevention strategies.

Secondly, ML techniques have been applied in healthcare or medical data with imbalanced characteristics to build a predictive model. Three different classification techniques have been examined. Various resampling approaches have been employed before applying the classifiers. We showed that there was a significant improvement in performance when applying a resampling technique as compared to applying no resampling technique.

Thirdly, super learning techniques that utilize multiple base learners have been studied to boost the performance of classification models. We applied two different forms of super learners - the first one used two base learners, while the second one used three base learners. For evaluating the models, we used four well-known benchmark data sets related to the healthcare domain. The results confirmed that the SL model performs better than the individual classifier and the baseline ensemble.

Finally, we assessed cancer-relevant genes of prostate cancer with the most significant correlations with the clinical outcome of the sample type and the overall survival. For detecting cancer,

we built different classification models, including a decision tree model, and for the estimation of survival time, we constructed a decision tree regressor model. Finally, we generated rules from both the decision tree and the regressor tree models.

Our future work aims to apply these techniques to other real-world healthcare problems. Moreover, the techniques that we applied in this research can be utilized in different domains, such as cybersecurity, business intelligence, and geographic information systems.

One of our immediate future work is to continue analyzing biomedical data to automate the early detection of prostate cancer based on gene expression levels. In our previous research [81], we considered a few number of genes along with few clinical variables. Future studies will assess all genes (about twenty-thousand) along with more clinical variables. The biggest challenge in this research is the imbalanced and high dimensionality of the gene expression data. To handle the imbalance issue, in addition to data-level approaches like the resampling technique that we applied in our previous research [78], we also plan to investigate cost-sensitive methods. To overcome the curse of dimensionality, we plan to use different deep neural network (DNN) architectures to build a model for early detection of prostate cancer patients. Applying DNN in gene expression profiling is worthwhile as it perfectly fits the need for high dimensional data processing and capturing gene-gene interactions. Moreover, we plan to use feature selection or dimensionality reduction techniques to examine if it can further improve the model.

REFERENCES

[1] Louppe, Gilles, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. "Understanding variable importances in forests of randomized trees." In Advances in neural information processing systems, pp. 431-439. 2013.

[2] Ferlay, Jacques, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012." International journal of cancer 136, no. 5 (2015): E359-E386.

[3] Kaur, Harnoor, and Shalini Batra. "HPCC: An ensembled framework for the prediction of the onset of diabetes." In 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC), pp. 216-222. IEEE, 2017.

[4] Gibbons, Chris, Suzanne Richards, Jose Maria Valderas, and John Campbell. "Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy." Journal of medical Internet research 19, no. 3 (2017): e65.

[5] Silwattananusarn, Tipawan, Wanida Kanarkard, and Kulthida Tuamsuk. "Enhanced classification accuracy for cardiotocogram data with ensemble feature selection and classifier ensemble." Journal of Computer and Communications 4, no. 4 (2016): 20-35.

[6] Yao, Zheng, Peng Liu, Lei Lei, and Junjie Yin. "R-C4. 5 Decision tree model and its applications to health care dataset." In Proceedings of ICSSSM'05. 2005 International Conference on Services Systems and Services Management, 2005., vol. 2, pp. 1099-1103. IEEE, 2005.

[7] Han, Jiawei, and Micheline Kamber. "Data mining concept and technology." Publishing House of Mechanism Industry (2001): 70-72.

[8] Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. "Machine learning applications in cancer prognosis and prediction." Computational and structural biotechnology journal 13 (2015): 8-17.

[9] Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pp. 207-216. 1993.

[10] Rahman, SMMonzurur, MdFaisal Kabir, and F. A. Siddiky. "Rules mining from multi-layered neural networks." International Journal of Computational Systems Engineering 1, no. 1 (2012): 13-24.

[11] Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. "Association rule mining to detect factors which contribute to heart disease in males and females." Expert Systems with Applications 40, no. 4 (2013): 1086-1093.

[12] Khalilian, Majid, and Seyedeh Talayeh Tabibi. "Breast mass association rules extraction to detect cancerous masses." In 2015 International Congress on Technology, Communication and Knowledge (ICTCK), pp. 337-341. IEEE, 2015.

[13] Ordonez, Carlos, Cesar A. Santana and Levien de Braal. "Discovering Interesting Association Rules in Medical Data." ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2000).

[14] Stilou, S., Panagiotis D. Bamidis, Nicos Maglaveras, and Constantinos Pappas. "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare." Studies in health technology and informatics 2 (2001): 1399-1403.

[15] Agrawal, Rakesh, Sakti Ghosh, Tomasz Imielinski, Bala Iyer, and Arun Swami. "An interval classi er for database mining applications." In Proc. of the VLDB Conference, pp. 560-573. 1992.

[16] Rahman, S M Monzurur and Md Faisal Kabir, and Muhammad Mushfiqur Rahman. "Integrated Data Mining and Business Intelligence." In Encyclopedia of Business Analytics and Optimization. edited by John Wang, 1234-1253. Hershey, PA: IGI Global, 2014. http://doi:10.4018/978-1-4666-5202-6.ch114

[17] Aggarwal, Charu C. "Data Classification: Algorithms and Applications." Chapman & Hall/CRC publisher, 1st edition (2014).

[18] Murthy, Sreerama K. "Automatic construction of decision trees from data: A multi-disciplinary survey." Data mining and knowledge discovery 2, no. 4 (1998): 345-389.

[19] Breiman, Leo, Joseph H Friedman, R. A. Olshen and C. J. Stone. "Classification and Regression Trees." (1983).

[20] Quinlan, J. Ross. "Induction of decision trees." Journal of Machine Learning, volume no. 1 (1986): pages 81-106.

[21] Salzberg, Steven L. "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993." (1994): 235-240.

[22] Murphy, Kevin P. "Machine learning: a probabilistic perspective." MIT press, Cambridge, MA, 2012.

[23] Kabir, Md Faisal, Chowdhury Mofizur Rahman, Alamgir Hossain, and Keshav Dahal. "Enhanced Classification Accuracy on Naive Bayes Data Mining Models." International Journal of Computer Applications 28, no. 3 (2011): 9-16.

[24] Aha, David W. "In Lazy learning." pp. 7-10. Kluwer Academic Publishers (1997).

[25] Steinwart, Ingo, and Andreas Christmann. "Support vector machines." Springer Publishing Company, Incorporated, ISBN:978-0-387-77241-7 (2008).

[26] Aiello, Spencer, et al. "Machine Learning with Python and H20." Edited by Lanford, J., Published by H 20 (2016): 2016.

[27] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32. DOI:https://doi.org/10.1023/A:1010933404324

[28] Breiman, Leo. "Bagging predictors." Machine learning 24, no. 2 (1996): 123-140.

[29] Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." BMC bioinformatics 7.1 (2006): 3.

[30] Wolpert, David H. "Stacked generalization." Neural networks 5.2 (1992): 241-259.

[31] Bennett, Casey C. "Utilizing RxNorm to support practical computing applications: capturing medication history in live electronic health records." Journal of biomedical informatics 45.4 (2012): 634-641.

[32] Hou, Ningqi, Susan Hong, Wenli Wang, Olufunmilayo I. Olopade, James J. Dignam, and Dezheng Huo. "Hormone replacement therapy and breast cancer: heterogeneous risks by race, weight, and breast density." Journal of the National Cancer Institute 105, no. 18 (2013): 1365-1372.

[33] Gail, Mitchell H., Louise A. Brinton, David P. Byar, Donald K. Corle, Sylvan B. Green, Catherine Schairer, and John J. Mulvihill. "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually." JNCI: Journal of the National Cancer Institute 81, no. 24 (1989): 1879-1886.

[34] Barlow, William E., Emily White, Rachel Ballard-Barbash, Pamela M. Vacek, Linda Titus-Ernstoff, Patricia A. Carney, Jeffrey A. Tice et al. "Prospective breast cancer risk prediction model for women undergoing screening mammography." Journal of the National Cancer Institute 98, no. 17 (2006): 1204-1214.

[35] Gauthier, Emilien, Laurent Brisson, Philippe Lenca, and Stéphane Ragusa. "Breast cancer risk score: a data mining approach to improve readability." In The International Conference on Data Mining, pp. 15-21. CSREA Press, 2011.

[36] Li, Wenmin, Jiawei Han, and Jian Pei. "CMAR: Accurate and efficient classification based on multiple class-association rules." Proceedings 2001 IEEE international conference on data mining. IEEE, 2001.

[37] Paul, Razan, Tudor Groza, Jane Hunter, and Andreas Zankl. "Inferring characteristic phenotypes via class association rule mining in the bone dysplasia domain." Journal of biomedical informatics 48 (2014): 73-83.

[38] Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C).A list of the BCSC investigators

and procedures for requesting BCSC data for research purposes, last retrieved July 2018 from http://www.bcsc-research.org.

[39] Seddik, Ahmed F., and Doaa M. Shawky. "Logistic regression model for breast cancer automatic diagnosis." In 2015 SAI Intelligent Systems Conference (IntelliSys), pp. 150-154. IEEE, 2015.

[40] Brin, Sergey, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. "Dynamic itemset counting and implication rules for market basket data." In Proceedings of the 1997 ACM SIGMOD international conference on Management of data, pp. 255-264. 1997.

[41] Hahsler, Michael, Bettina Grün, and Kurt Hornik. "Introduction to arules–mining association rules and frequent item sets." SIGKDD Explor 2, no. 4 (2007): 1-28.

[42] Liu, Bing, Wynne Hsu, and Yiming Ma. "Mining association rules with multiple minimum supports." In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 337-341. 1999.

[43] Mathew, Josey, Chee Khiang Pang, Ming Luo, and Weng Hoe Leong. "Classification of imbalanced data by oversampling in kernel space of support vector machines." IEEE transactions on neural networks and learning systems 29, no. 9 (2017): 4065-4076.

[44] More, Ajinkya. "Survey of resampling techniques for improving classification performance in unbalanced datasets." arXiv preprint arXiv:1608.06048 (2016).

[45] Elhassan, T., and M. Aljurf. "Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method." (2016).

[46] Kabir, Md Faisal, Salahuddin Aziz, Suman Ahmmed, and Chowdhury Mofizur Rahman. "Information theoretic SOP expression minimization technique." In 2007 10th international conference on computer and information technology, pp. 1-6. IEEE, 2007.

[47] Vanerio, Juan, and Pedro Casas. "Ensemble-learning approaches for network security and anomaly detection." Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks. 2017.

[48] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. 2016.

[49] Batista, Gustavo EAPA, Ronaldo C. Prati, and Maria Carolina Monard. "A study of the behavior of several methods for balancing machine learning training data." ACM SIGKDD explorations newsletter 6, no. 1 (2004): 20-29.

[50] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[51] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

[52] Van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard. "Super learner." Statistical applications in genetics and molecular biology 6.1 (2007). doi:10.2202/1544-6115.1309.

[53] Nykodym, Tomas, et al. "Generalized linear modeling with h2o." Published by H2O. ai Inc (2016).

[54] LeDell, Erin. "Scalable Super Learning." Handbook of Big Data 339 (2016).

[55] LeDell, Erin. Scalable Ensemble Learning and Computationally Efficient Variance Estimation. Diss. UC Berkeley, 2015.

[56] Breiman, Leo. "Stacked regressions." Machine learning 24.1 (1996): 49-64.

[57] LeBlanc, Michael, and Robert Tibshirani. "Combining estimates in regression and classification." Journal of the American Statistical Association 91.436 (1996): 1641-1650.

[58] van der Laan, Mark J., Sandrine Dudoit, and Aad W. van der Vaart. "The cross-validated adaptive epsilon-net estimator." (2004).

[59] Casas, Pedro, and Juan Vanerio. "Super learning for anomaly detection in cellular networks." 2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). IEEE, 2017.

[60] Van der Laan, Mark J., and Sherri Rose. Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media, 2011.

[61] Baćak, Valerio, and Edward H. Kennedy. "Principled machine learning using the super learner: an application to predicting prison violence." Sociological Methods & Research 48.3 (2019): 698-721.

[62] Antal, Bálint, and András Hajdu. "An ensemble-based system for automatic screening of diabetic retinopathy." Knowledge-based systems 60 (2014): 20-27.

[63] Salama, Gouda I., M. Abdelhalim, and Magdy Abd-elghany Zeid. "Breast cancer diagnosis on three different datasets using multi-classifiers." Breast Cancer (WDBC) 32.569 (2012): 2.

[64] Choubey, Dilip Kumar, et al. "Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection." Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016). 2017.

[65] Abdar, Moloud, et al. "Performance analysis of classification algorithms on early detection of liver disease." Expert Systems with Applications 67 (2017): 239-251.

[66] Ciregan, Dan, Ueli Meier, and Jürgen Schmidhuber. "Multi-column deep neural networks for image classification." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.

[67] Fatima, Meherwar, and Maruf Pasha. "Survey of machine learning algorithms for disease diagnostic." Journal of Intelligent Learning Systems and Applications 9.01 (2017): 1.

[68] Luque-Baena, Rafael Marcos, et al. "Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data." Theoretical Biology and Medical Modelling 11.1 (2014): S7.

[69] Ahmad, Fadzil, et al. "Intelligent breast cancer diagnosis using hybrid GA-ANN." 2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks. IEEE, 2013.

[70] Kabir, Md Faisal, Simone A. Ludwig, and Abu Saleh Abdullah. "Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

[71] "GDC." GDC, 2018, portal.gdc.cancer.gov/, accessed on January, 2019.

[72] Hemphill, Edward, et al. "Feature selection and classifier performance on diverse bio-logical datasets." BMC bioinformatics. Vol. 15. No. 13. BioMed Central, 2014.

[73] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63.1 (2006): 3-42.

[74] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830.

[75] Husain, Hartina, et al. "The Application of Extended Cox Proportional Hazard Method for Estimating Survival Time of Breast Cancer." Journal of Physics: Conference Series. Vol. 979. No. 1. IOP Publishing, 2018.

[76] Quinlan, J. Ross. "C 4.5: Programs for machine learning." The Morgan Kaufmann Series in Machine Learning, San Mateo, CA: Morgan Kaufmann,— c1993 (1993).

[77] Yao, Zheng, Peng Liu, Lei Lei, and Junjie Yin. "R-C4. 5 Decision tree model and its applications to health care dataset." In Proceedings of ICSSSM'05. 2005 International Conference on Services Systems and Services Management, 2005., vol. 2, pp. 1099-1103. IEEE, 2005.

[78] Kabir, Md Faisal, and Simone A. Ludwig. "Classification of Breast Cancer Risk Factors Using Several Resampling Approaches." 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018.

[79] Kabir, Md Faisal, and Simone A. Ludwig. "Enhancing the Performance of Classification Using Super Learning." Data-Enabled Discovery and Applications 3.1 (2019): 5.

[80] Loh, Wei-Yin. "Classification and regression trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1.1 (2011): 14-23.

[81] Kabir, Md Faisal, and Simone A. Ludwig. "Classification Models and Survival Analysis for Prostate Cancer Using RNA Sequencing and Clinical Data." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.