National College of Ireland

Breast Cancer Classification on Gene Expression Data Using Ensemble Learning

MSc Research Project
Data Analytics

Siddharth Sharma
Student ID: x16148371

School of Computing
National College of Ireland

Supervisor: Mr. Vikas Tomer

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Siddharth Sharma
Student ID:	x16148371
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Supervisor:	Mr. Vikas Tomer
Submission Due Date:	20/12/2018
Project Title:	Breast Cancer Classification on Gene Expression Data Using Ensemble Learning
Word Count:	7153
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	27th January 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Breast Cancer Classification on Gene Expression Data Using Ensemble Learning

Siddharth Sharma
x16148371

Abstract

Breast cancer is the most common cancer diagnosed in women and the leading cause of death for them; if the correct tumor is diagnosed at an early stage, the disease is treatable. This research classifies breast cancer with the help of ensemble learning using RNA-Seq gene expression levels data (New technologies are being developed to extract information from human DNA and RNA strands.), extracted by illumina Hi-Seq system consisting of 14,000 genes for a mere 800 samples with 5 tumors: BRCA, PRAD, KIRC, COAD and LUAD. Evaluating the classification accuracy of the models: Stacking, Bagging, Support Vector Machine-Radial Basis Function (SVM-RBF), k-Nearest Neighbors (kNN) and decision trees (C5.0) resulted in an optimal outcome for this research. The most significant features were selected using random forest and high correlation matrix. Stacking's performance with an acme accuracy of 98.75% and a kappa value of 0.9835. This research will help in finding the most important genes responsible for Breast cancer using the RNA-Seq Data with the help of ensemble learning which has proven to be quite remarkably efficient for classification problems.

1 Introduction

World is full of diseases and epidemics such as diabetes, Ebola, Zika virus, AIDS and Cancer. However, cancer is the one that has been troubling people in every part of the world, be it United States, Ireland, India and China, it is the growth of abnormal cells which can spread over the whole body or some specific parts of the body. There are various kinds of cancer such as skin cancer, breast cancer, leukemia, Lung cancer, kidney cancer and prostate cancer; almost all types of cancer get their name from the location of their origin (Vanitha et al.; 2015). People have been dying in large numbers on account of cancer for the past few decades especially women due to Breast cancer as it is a pretty common occurrence among women of all ages. Governments, Non-Government organisations, various trusts and charities have been trying hard towards the awareness of breast cancer for a long time now.

A considerable amount of research has been accomplished with the help of supervised and unsupervised machine learning techniques such as decision trees, linear classifiers, random forest, support vector machine, clustering analysis, artificial neural networks, deep learning, Bayesian networks, statistical learning, ensemble learning and regularisation algorithms. The DNA microarray datasets have too much information, i.e. thousands of features extracted from a small number of samples which is known as the "Large N

– small N " paradigm (Zhang et al.; 2009). The authors (Hsieh et al.; 2012) have used ensemble machine learning models for diagnosis of breast cancer with the help of feature selection using information gain. Models used for ensemble machine learning are Neural Fuzzy (NF), k-Nearest Neighbors (k-NN), and Quadratic Classifier. In 2016, (Salem et al.; 2017) carried out research to classify different types of cancer using Genetic programming with the help of gene expression profiles, i.e. microarray datasets by using Information Gain (IG) & Genetic Algorithm (GA) for feature selection on seven different cancer datasets. The authors Vanitha et al. (2015), have done gene data classification using support vector machine because according to the author other statistical methods are not fully capable of classifying unlabeled gene expression data; the research was carried out on Colon cancer and Lymphoma datasets using Mutual Information Technique (MI) and Support Vector Machines (SVM).

In this study gene expression levels of RNA sequencing have been used for the classification of five tumors BRCA (Breast invasive carcinoma), KIRC (Kidney Renal clear cell Carcinoma), COAD (Colon Adenocarcinoma), LUAD (Lung Adenocarcinoma) and PRAD (Prostate Adenocarcinoma) [1]. Outline for the research is as follows: the study carried out is depicted by six sections starting with section 1 that covers the introduction and motivation for the research, section 2 consists of the related work done in this domain. Section 3 will explain the methodology used in the due process followed by the models used for implementation in section 4. Section 5 consists of implementation which then leads to section 6, i.e. the evaluation and results found in this study and finally section 7 with the conclusion and future work.

1.1 Motivation

Cancer is such a disease which comes to one's notice when it is diagnosed, so we have to be prepared for the worse. With the help of this classification of gene expression data, we can help people to be aware of a disease they did not know existed. Cancer has many forms and types; with an efficient classification system, becoming aware of which gene to target and prevent a disastrous end to any human's life. There exists some type of breast cancer which may or may not be diagnosed quickly. Over the years there has been a decline in the number of deaths due to extensive research going on all over the world. The whole month of October is dedicated to a worldwide campaign towards breast cancer awareness, which in turn helps a common man to be on a lookout for the symptoms and get treatment to lead a healthy life instead of an untimely death. To save millions of lives by detecting breast cancer in any way possible is very important, it can be done in two ways, first by observing the physical changes that happen gradually with the course of time and second choice would be exploring the vast pool of genes that cause breast cancer.

1.2 Research Question

RQ: *How can gene expression levels of RNA-Seq data be classified with the help of Ensemble Learning (Bagging and Stacking) in the classification of Breast Cancer by selecting the most significant biomarkers using Random Forest?*

[1]Genomic Data Abbreviations: `https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations`

1.3 Purpose

Purpose of this research is to classify tumors using gene expression data for breast cancer. Gene expression analysis is a very vast field to explore, and with a time-bound project, one can explore very few areas of interest without the help of certified professionals concerning this particular domain (Cancer). Gene expression analysis is not an easy task to accomplish; it needs special equipment to extract essential features from a vast pool of attributes, sensitive and private data. Skilled professionals are required for handling genetic data. There are still some types of cancer that cannot be detected using gene information or otherwise. The medical community has joined hands with data scientists to figure out a way to find ways to foresee breast cancer as early as possible because, if spotted at an early stage breast cancer is treatable.

2 Related Work

2.1 Ensemble Learning Techniques

Ensemble machine learning techniques have contributed to achieving high accuracy for classification of breast cancer diagnosis and prognosis. It is a known fact that accuracy for classification varies with the models implemented on the same dataset as can be seen in(Khuriwal and Mishra; 2018) and (Hsieh et al.; 2012). Wisconsin Breast cancer dataset has been utilised for implementation of the machine learning models by (Khuriwal and Mishra; 2018). Artificial Neural Network and logistic regression were tested individually, and then Adaptive voting ensemble learning was applied to both the algorithms; which resulted in an accuracy of 98.50%. Sixteen features were selected with the help of Recursive feature elimination algorithm. As it can be seen in the research done by (Hsieh et al.; 2012) in 2012, information gain is the means for selecting the significant features and then with the help of Neural fuzzy, KNN and quadratic classifier an ensemble learning model was developed for the breast cancer classification along with the distinct implementation of each model. This resulted in producing and accuracy of 97.14% which is more than the accuracy provided by the individual algorithms on the data grabbed from the UCI machine learning repository.

(Tarek et al.; 2017) describes the importance of predicting and diagnosing cancer for the appropriate treatment can be given to the patients as early as possible with the help of ensemble learning techniques on three DNA microarray datasets, i.e. Leukemia, Colon & Breast cancer. Author has used a combination of majority voting, weighted voting, stacking and naïve Bayes algorithms to overcome the drawbacks, i.e. result accuracy, over-fitting, application of ensemble models on various types of cancer. Most essential genes have been selected with the help of Backward Elimination Hilbert-Schmidt Independence Criterion [BAHSIC], Extreme value Distribution based gene selection (EVD) and Singular Value Decomposition Entropy Gene Selection (SVDE). KNN was used to classify with five base classifiers; each having different features at the time of implementation; evaluation was carried out using Bolstered Resubstituting Error (BRE). BAHSIC model generated poor accuracy. Ensemble learning model incorporating the five classifiers by majority voting demonstrates significant performance improvements for the Colon & Leukemia data; whereas it was almost the same for breast cancer data.

Authors (Zhang et al.; 2009) have made a comparison between the ensemble learning techniques such as AdaBoost, LogitBoost & random forest and classification techniques such as logistic regression and support vector machine. For implementation two different set of datasets were used i.e. Van De Vuver (Van De Vijver et al.; 2002) & Wang dataset (Wang et al.; 2005). Subnetwork markers were created using gene expression and protein-protein interaction network so that the common genes in both the datasets can be recognised Inner dataset comparison & cross-dataset comparison between the van De Vuver and Wang dataset were carried out for evaluation of the ensemble and classification techniques. After implementation accuracy provided by random forest seems to outperform other algorithms, i.e. SVM & logistic regression by 4.8% & 22.4% whereas for AUC (Area under the curve) LogitBoost outperforms logistic regression and SVM by a margin of 29.3% & 5.1%. In the cross-dataset comparison evaluation of the techniques suggest that ensemble learning performs way better than the classification techniques for classifying breast cancer metastasis accurately which also inspired to take up this study.

(Jaffar et al.; 2009) did "An Intelligent Ensemble Based Systems for Breast Cancer Diagnosis". SVM-RBF has been used for classification under the single classifier system. The Wisconsin data set has been used to complete the implementation through MATLAB. When compared to other algorithms such as C4.5, naïve Bayes, Logitboost, the proposed technique had the highest accuracy of 99.78% for diagnosing breast cancer.

2.2 Machine Learning Techniques

2.2.1 Microarrary Data

Authors (Gharibi et al.; 2015) researched "Identification of Gene Signatures for Classifying of Breast Cancer Subtypes using Protein interaction Database and Support Vector Machines". Most significant genes were selected with the help of wrapper & embedded methods, and integrated with the protein-protein interaction network to which SVM-RFE was applied and contrasted with decision tree algorithm. The selected microarray datasets produce an accuracy of 91.2% for SVM-RFE and 78.6% for decision tree.

(Vanitha et al.; 2015) did "Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection". Two different set of microarray datasets were used, i.e. Colon and Lymphoma cancer. Mutual Information was used for the selection of genes which were then used for classification using SVM (linear, RBF, Quadratic, polynomial), KNN and ANN; highest accuracy being 97.77% for linear SVM.

In 2017, (Salem et al.; 2017) authors showed interest in gene expression profiles for the classification of cancer with the help of genetic programming. Predefined thresholds played a significant role in selecting significant genes from a set of seven distinct datasets (Leukemia, Colon tumor, Lung & Prostate cancer) using information gain (IG); which were then dimensionally reduced using a genetic algorithm (GA). Results showed that the proposed algorithm is capable of diagnosing cancer at an early stage as the classification accuracy came out to be pretty high in the case of Lung cancer and Prostate cancer along with Leukemia & Colon Tumor datasets.

Author (Gypas et al.; 2011) has used support vector machine, recursive feature elimination based on linear neuron weights (RFE-LNW), least absolute shrinkage and selection operator (LASSO) and multilayer perceptron to identify significant biomarkers and for distinguishing the irrelevant genes in the microarray dataset of breast cancer. The results from this study have been compared with a disease-agnostic tool called the Genotator to check the acquired gene signatures. The study showed that a gene named

ERRFI1 as a biomarker of breast cancer. Data has been procured from the Netherlands Cancer Institute which was used by (Van't Veer et al.; 2002) in previous research. Three gene signatures were found during implementation; SVM has been used for the sole purpose of examining the accuracy for classification of genes, i.e. 82-83%, 74% and 80-85% respectively with variation in the number of genes for each model, i.e. 190, 82 & 152 respectively. The obtained gene signatures are put through the biological evaluation with the help of Genotator the disease-agnostic tool; which helped in identifying the most vital and known genes from the three gene signatures acquired earlier.

Authors from Spain (González-Navarro and Belanche-Muñoz; 2011) have used many classifiers, i.e. kNN, Naïve Bayes, LDC, QDC, Logistic Regression, SVM with linear, quadratic & radial kernels. Bootstrap resampling technique was applied on the two microarray datasets for the procurement of best gene subset (BGS). Gene selection was made using entropy filter methods which results in higher predictive accuracy and a greater number of significant genes.

Authors (Zeng et al.; 2009) in China during the year 2009 lead research for microarray-based cancer classification utilising biological pathways. Many significant features were discovered using the pathway than the usual biomarkers. Four different datasets have been integrated to find the genetic, and there are predefined gene lists for the pathway data extracted from EuGene software gathered from KEGG. Fisher's exact test has been used on the gene expression data, SVM is used for classification which shows that biological pathway can have a good effect on the cancer diagnoses.

Authors (Chiu et al.; 2008) carried out research, "Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer". Support Vector Regression model (SVR) has been used on the features selected using a wrapper, and ten-fold cross-validation methods on a microarray dataset with 25,000 cDNA. 44 genes were chosen after a greedy search algorithm was applied. The genes selected were evaluated using many techniques such as Self-Organizing maps for regression, features selected based on correlation, IBk algorithm and random selection; which leads to inclusion of two genes named CXCL14 and ER (concerning breast cancer) into the dataset used for regression modelling. When compared to other researches carried out in the same domain it was found that the SVM regression model helps in better prediction of clinical metastases time for cancer patients which is more important than the prognosis prediction according to the author.

Microarray datasets are used for almost every research concerning the classification of cancer as they are readily available, but for this study, the authors (Castillo et al.; 2017) have used a novel integration technique on three distinct datasets with RNA-Seq & microarray data to classify breast cancer. With the integration of these data sets potential bio makers were found which could help in breast cancer diagnosis, six to be exact & 98 differentially expressed genes (DEGs). For classification SVM, RF, k-NN have been used to apply features selected from minimum-Redundancy Maximum-relevance (mRMR) algorithm. Accuracy for SVM and RF came out to be 97% using the DEGs.

2.2.2 Gene Expression Data

Gene expression data is beneficial in the classification of cancer as it provides a detailed set of attributes to work with, i.e. more information from RNA strands. In 2017 authors (Danaee et al.; 2017), used deep learning techniques for diagnosing cancer and finding the significant genes responsible for the same. Stacked Denoising Autoencoder

(SDAE) has been implemented to analyse the relevant genes that are procured from a high dimensionality dataset. RNA-Seq gene expression data of breast cancer has been considered for the research; taken from The Cancer Genome Atlas (TCGA) database; which consists of 1097 samples. Artificial neural network (ANN), support vector machine (SVM), principal component analysis (PCA) and kernel principal component analysis (KPCA) have been used for evaluation of features & their dimensionality reduction; then the deeply connected genes (DCGs) are analyzed for Gene Ontology (GO) with the help of Panther pathway analysis. SVM-RBF (Radial basis function) had the highest accuracy and high sensitivity, when genes from the SDAE were applied to it, i.e., 98.26% & 98.73% respectively. When DCGs are applied to machine learning algorithms, the highest accuracy comes out to be 94.78% for SVM-RBF, 98.13% as the highest sensitivity for ANN. Limitation of the study was the size of the dataset but regardless of that SDAE succeeded in the classification of breast cancer with the help of deeply connected genes.

In 2015 (Firoozbakht et al.; 2015) did "A Novel Approach for Finding Informative Genes in Ten Subtypes of Breast Cancer". A bottom-up hierarchical classification process has been implemented with the help of machine learning algorithms such as random forest; Decision trees & SVM; which have been used to find the most crucial genes as well as assess the similarity levels among the ten subtypes of breast cancer. Following the bottom-up approach; Ward's method, average, complete and single linkage were utilised to determine the distance between the two clusters; Euclidean distance being the basis of the metric system put into use. Dataset has been grabbed from the European Genome-Phenome archive consisting gene expression data of 2000 samples of the breast tumor and copy numbers obtained by Illumina HT & Affymetrix SNP arrays respectively. Performance of random forest was the best as it had sizeable AUC and out of the four models applied Ward's method produced the most balanced tree where subtypes evenly distributed.

In 2014 authors (List et al.; 2014) approach towards the classification of subtypes of cancer was different from what we have seen till now. The use of DNA methylation data in addition to gene expression data leads to a better understanding of genes relevant for causing breast cancer at an epigenetic level. Data was procured from TCGA containing 547 samples from patients who have breast cancer. For classification RF and bootstrapping was put to use; the variables were grabbed with the help of varSeIRF. After implementation of the models proposed, it was seen that gene expression data is more useful for classification of subtypes of breast cancer than Methylation data as very few features were contributing towards the successful classification of breast cancer.

In 2002, authors (De Jong et al.; 2002) explored many genes seeking polymorphisms related to breast cancer by conducting a pooled analysis by computing the odds ratio, 95% confidence levels, Population attributable risk (PAR) and sample size for all the features selected to find genes other than BRCA1 & BRCA2. There were a few genes that would have an impact on causing breast cancer at a much higher risk such as HRAS1, GSTM1, GSTP1, CYP1B1, CYP2D6, CYP19, VDR and there are some polymorphisms have less tendency of causing breast cancer in women namely intron3 & intron 6, XbaI, PROGIN in Tp53, ER & PR genes respectively. After the implementation 12 polymorphisms in 9 genes were found, such as NAT1 NAT2, GSTT1, CYP1A1, CYP17, AR, COMT AND UGT1A1; to be entirely impertinent to breast cancer.

A different approach has been followed by the author (Otoom et al.; 2015) by using image shape-based features along with microarray data for diagnoses of breast cancer. It is a novel study evaluated with ten-fold cross validation along with classifiers such as SVM,

MLP, NB, j48 & RF decision tree, bagging, radial basis function neural networks (RBF-NN), MPPCA, a mixture of normalised linear transformations and Gaussian Mixture Model (GMM). Microarray data from the University of Stanford as well as Wisconsin Diagnostic breast cancer dataset has been used for the study. Highest accuracy was achieved by bagging algorithm with 98.1% for the Wisconsin dataset, whereas SVM and RBF-NN achieved overall accuracy as high as 97.6%.

The main focus of every research regarding cancer diagnosis and prognosis has been feature selection because there are thousands of genes which may or may not be significant. Authors (Wang and Gotoh; 2010) did "A Robust Gene Selection Method for Microarray-based Cancer Classification" which lead them to a comparative study of various feature selection methods such as Chi-Square method, depended degree, IG, Relief-F and symmetric uncertainty. The research suggests that the proposed method is excellent to the other feature selection methods for classification of biomarkers in a gene expression dataset. Classifiers used for evaluation are Naïve Bayes, Decision Tree, SVM and k-NN. For the implementation of the algorithms, eight distinct datasets regarding cancer were used. The α-depended degree when subjected to fine-tuning of the value α performs way better than the normal depended degree for feature selection. The performance of the classifiers varied with the features selected from different methods. This research was an inspiration for using random forest as a basis of feature selection and brings novelty to the research.

3 Methodology

This section comprises a detailed description of the methodology followed in this research. *'CRISP-MED-DM'* – Cross Industry Standard Process for medical data mining; as depicted in figure 1 it is a hierarchical approach that has been undertaken for this study. Because for medical domain there is no predefined framework or methodology (Niaksu; 2015).

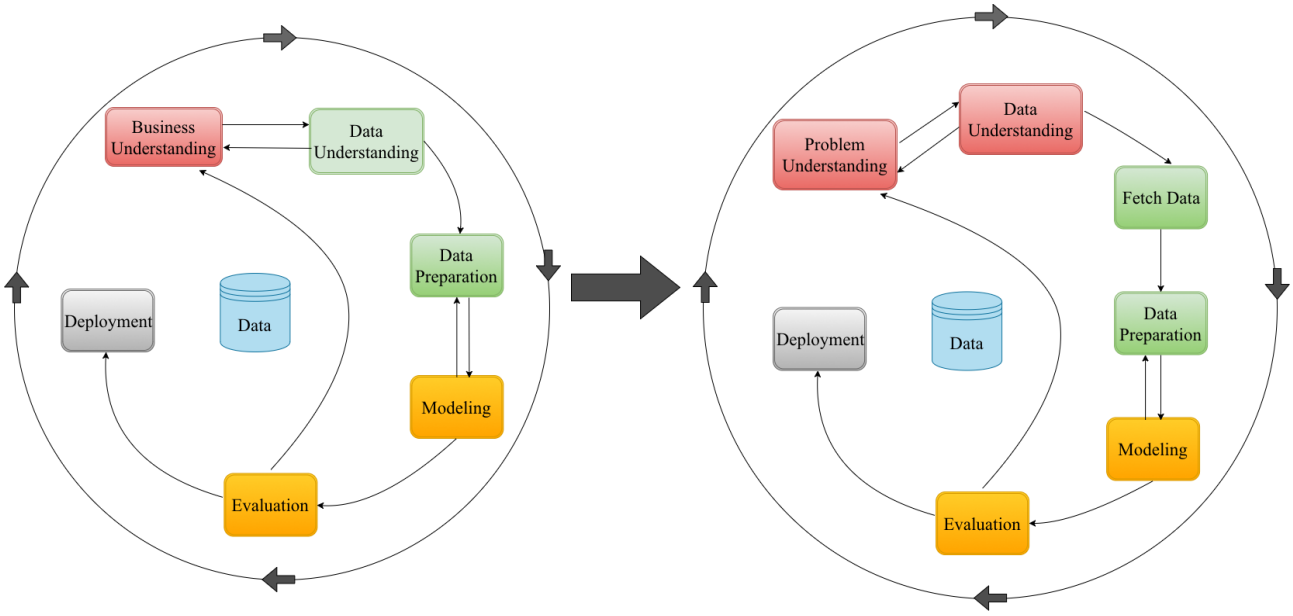

Figure 1: CRISP-MED-DM (Niaksu; 2015), (Chapman et al.; 2000)

3.1 Problem Understanding

Breast cancer is a topic that can be explored to vast lengths. The research carried out until now has been with some help from medical organisations or medical professionals having an in-depth knowledge of this domain, which has been a contributing factor towards successful research. Limitation of this research is that no help or advice was taken from any medical staff or organisation. Many areas can be explored such as breast cancer prognosis, breast cancer classification (two types: metastasis and biomarkers), breast cancer diagnosis, risk factors causing breast cancer and Breast cancer cell lines. Complex nature of this topic is that every area mentioned has tremendous potential for exploration and experimenting in itself, with a lot of pros & cons that are very hard to handle. The goal of this study was to deal with the classification of breast cancer. This can be done using different types of data such as RNA-Sequencing data, gene expression microarray data and clinical data which helped in drawing up a plan of data understanding.

3.2 Data Understanding

The first and foremost challenge that comes up involving medical data is accessible to the required dataset as it involves information regarding the patient which gives rise to privacy and ethical issues against usage of such data. The sensitive and complex nature of patient's data causes one of the major problems, i.e. storage of data in a clean and understandable state. Data related to cancer or genes is not easily available as it contains privileged information about the patients. Data available for cancer research is quite complex as it involves the usage of DNA and genes. Data is available in the form of gene expression microarray, RNA-Seq and clinical data. The tcga_pancancer dataset has been considered for this study, which is available for the public at "Synapse [2] which is an open-source research platform provided by Sage Bionetworks.

The chosen dataset has certain restrictions as the gene expression levels for RNA-Seq (RNA-Sequencing) were calculated by the dataset provider beforehand using an Illumina HiSeq Sequencing system, which was unavailable as it is a sophisticated machine with particular specifications as depicted in figure 2 [3]. HiSeq Sequencing system is required for the simplification of the data extracted from DNA strands which are highly complex and unstructured. RNA-Seq helps in exploring the RNA family which include miRNA, tRNA, small RNA and the changes in the gene expression concerning time. With advancement in technology, there has been significant progress in studying gene expression data with the help of next-gen sequencing of cDNA (complementary DNA) [4]

3.3 Fetch Data

As discussed in the previous stage getting data is not an easy task. There are three types of datasets available, i.e. microarray data, gene expression data and clinical data. Initially, the motive was to work on breast cancer prognosis, but due to a few shortcomings in the clinical dataset, the focus was shifted to breast cancer classification. The dataset had been earlier used by (Van't Veer et al.; 2002), which comprised of gene expression data of 25,000 genes corresponding to 117 patients. However, the data was structured

[2]Synapse; URLhttps://www.synapse.org/#! Synapse:syn300013/wiki/70804

[3] illumina HiSeq 2500 System; https://goo.gl/EZ2f9D

[4]RNA-Sequencing: https://en.wikipedia.org/wiki/RNA-Seq#Gene_expression

Figure 2: illumina HiSeq 2500 System[5]

in such a way that it could not be used for classification of breast cancer prognosis; therefore it proved to be useless. Clinical data is hard to get as it requires permissions from the supplier and is a lengthy process. A relevant dataset from European Genome-Phenome Archive (Metabric breast cancer samples, targeted sequencing of 173 genes) was requested and access was granted, but there were some limitations to the dataset that prevented its use first was the small size of the dataset which would not have led to a fruitful outcome at all. Second, the features required for implementation of machine learning models were not present. After much research, data was found on Synapse [5] i.e. RNA-Seq data containing gene expression data of five tumors, i.e. BRCA, PRAD, KIRC, LUAD and COAD for 800 patients & 14,000 genes.

3.4 Data Preparation

This stage of methodology involves cleaning of data and feature selection. All the missing values (NULL, 0 and special characters), irrelevant attributes are dealt with because models should perform at their full potential for a successful/unsuccessful result. This is the most important stage as feature selection takes place which molds the course for rest of the research. Feature selection was carried by applying correlation matrix and random forest (Gini plot) which eliminate the redundant variables from the dataset. The variables having high correlation are excluded as they have a negative impact on the performance of classification models as shown in figure 3. There were almost 14,000 genes it was reduced to a handful of 51 variables were selected for implementation of various algorithms using feature selection algorithms such as random forest and high correlation matrix. Dataset was checked for class imbalance, but there was none. Data scaling has been done for the dataset as the value of variables is not between a given range, i.e. 0 to 1. Therefore scaling provides the right balance between a variable with high value and a variable with a low value. Data cleaning has been done in R and Microsoft Excel so that model can perform efficiently.

[5]Synapse; URLhttps://www.synapse.org/#!Synapse:syn300013/wiki/70804

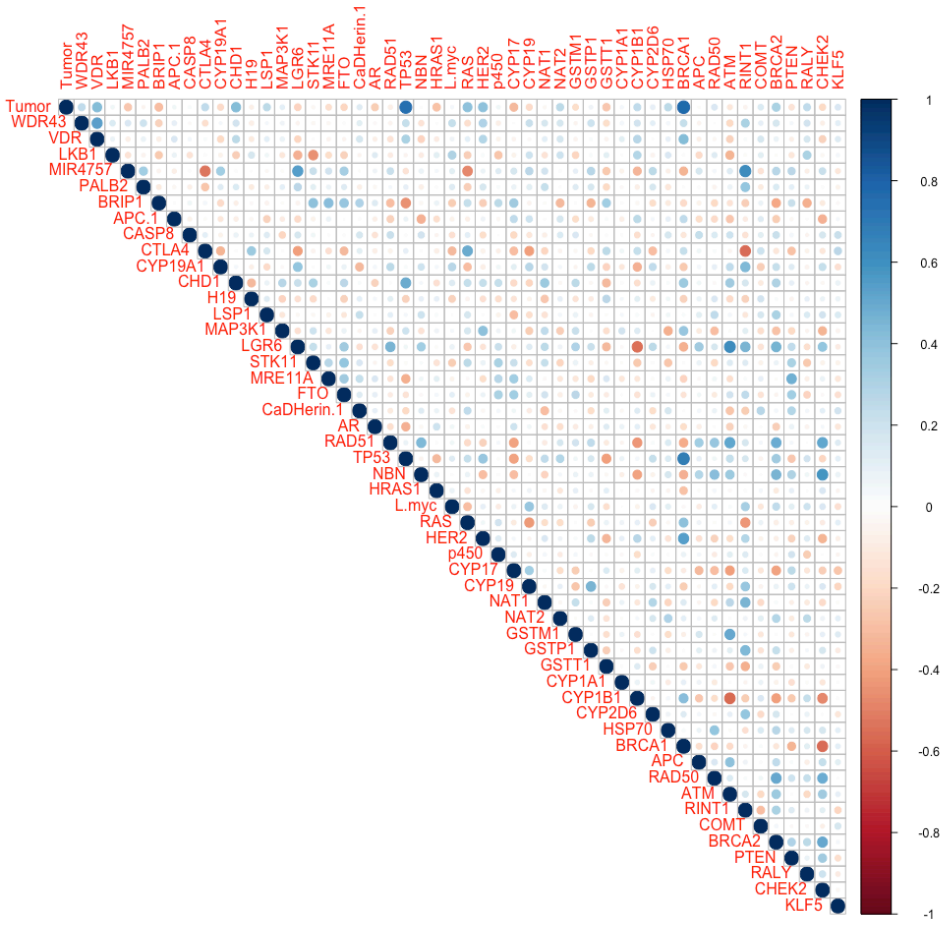

Figure 3: Correlation Matrix

3.5 Modelling

This study uses ensemble machine learning techniques after a thorough literature review for breast cancer classification. Great performance has been achieved by the preceding researchers using ensemble learning as well as other machine learning methods (SVM, NB, SDAE). In case of ensemble learning (Khuriwal and Mishra; 2018) used adoptive voting ensemble method and generated an accuracy of 98.5% whereas in 2009 (Jaffar et al.; 2009) produced an accuracy of 99.78% using ensemble learning consisting C4.5, NB & Logitboost inspired to pursue the current research with ensemble models. In case of support vector machines (Gharibi et al.; 2015) got 91.2% accuracy with SVM-RFE on a microarray dataset. (Vanitha et al.; 2015) attained 97.77% accuracy for linear SVM. Recurring algorithms found throughout the literature review are SVM, RF & kNN; these three models have been put into practice for the current research too because of their significance towards classification problems. Breast cancer classification has been performed for five tumors BRCA, PRAD, KIRC, COAD and LUAD with the main focus on genes causing breast cancer; only significant genes have been considered after conducting feature selection algorithms. A detailed description of the execution of these models will be explained in the sections to come next.

3.6 Evaluation

Depending on the method of classification or prediction, assessment of the model can vary with respect to the subject of research. In this study for the classification of breast cancer confusion matrix is utilised which provides the overall statistics of the model. For this study focus is on classification accuracy, kappa value, precision, recall and F-measure.

- **Classification Accuracy**

  It depicts how correctly has the classification taken place on a particular set of data. Performance measures are: Positive cases (P), Negative cases (N), True Positive:- correctly diagnosed positive cases (TP), True negative:- correctly diagnosed negative cases (TN), False Positive:- Negative cases diagnosed as positive (FP): Type I error, False Negative:- Positive cases diagnosed as negative (FN): type II error. Equation one shows how accuracy is calculated (Salem et al.; 2017).

$$"ACC = \frac{TP + TN}{TP + TN + FP + FN}" \tag{1}$$

- **Kappa**

  Cohen's kappa coefficient's upper limit is +1, and the lower limit is between 0 & -1. Higher the value of kappa better the classification for the predicted values by the model but it is not entirely reliable judge when dealing with multi-class classification problems (Cohen; 1960)

- **Recall**

  It calculates the actual positive cases correctly diagnosed. It is also known as true positive rate [6]. Expressed as:(Salem et al.; 2017)

$$"Recall = \frac{TP}{TP + FN}" \tag{2}$$

- **Precision** It is known as the positive predictive value which diagnoses the actual positives from the predicted positives cases [7].

$$"Precision = \frac{TP}{TP + FP}" \tag{3}$$

- **F-measure** It calculates the harmonic mean of precision and recall when combined together (Khuriwal and Mishra; 2018).

$$"F = \frac{precision * recall}{precision + recall} * 2" \tag{4}$$

3.7 Deployment

There is no deployment for this study as there are certain limitations that need attention such as small dataset, fewer variables in play, lack of proper medical equipment, a professional medical officer for guidance so that a medical point of view can also be integrated into the research. In real time there are plenty of variables such as physical effects, daily lifestyle and cell lines. This is a pretty sensitive domain and needs to be dealt with utmost precaution. The analytical approach for this study is purely for research-based and not for practical application.

[6]Wikipedia: `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`
[7]Wikipedia: `https://en.wikipedia.org/wiki/Precision_and_recall`

4 Models and Techniques Used in Reseacrh

4.1 Random Forest

It is an ensemble of decision trees used for classification & feature selection. For classification, every tree in the forest gets an input vector allotted by the algorithm. After a successful compilation of algorithm, a class with the greatest number of votes is selected from all the trees (Castillo et al.; 2017). The hierarchical structure of trees is more beneficial for the random forest as the predictive power exceeds over classification of a single tree and the voting technique helps in improving the accuracy and stability of the model (Elshazly et al.; 2013). Random forest shows low bias and high variance as an effect of overfitting of the training datasets, averaging of various decision trees finally leads to the reduction of variance [8].

4.2 KNN

The algorithm k-nearest neighbours are used for classification as well as regression while performing data mining. It also follows the voting method, the most number of votes select the target variable to be classified among the neighbours who are then allocated to the most common class in the k-nearest neighbours utilised for classification [9].

4.3 C5.0

C5.0 is an algorithm that has a tree-like structure where the output is the terminating node, and the decision is made when nodes are split into a leaf node forming conjunction. Decision trees are used extensively for classification purposes; it consists of various parts, i.e. parent node, branches, leaf node and end node; each playing a role in the outcome of the model. A decision tree is generally used for training and testing of the dataset. C5.0 is an upgraded version of C4.5 in R.

4.4 SVM

It is a binary classifier that separates the classes by locating the hyperplane between them, which then provides a high-dimensional classification in a very efficient manner (Xu et al.; 2012). An optimal hyperplane is detected by the SVM in the feature space with the help of dot product function's called kernels, the combination of the input points known as support vector points towards the optimal hyperplane. There are four types of kernels present in the feature space, i.e. linear, polynomial, quadratic and radial basis function (Vanitha et al.; 2015). SVM is used because it achieves high accuracy and deals with high dimensionality data such as gene expression data (Gypas et al.; 2011).

4.5 Ensemble Learning

When a set of classifiers are bundled up using voting, stacking or weighted average, it makes an ensemble method. As various algorithms have been bundled result in producing better accuracy and lower variance compared to the performance of any single algorithm

[8] Random Forest: `https://en.wikipedia.org/wiki/Random_forest`

[9] k-NN: `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm`

(Giarratana et al.; 2009). According to (Elshazly et al.; 2013) ensemble methods help in better understanding of high dimensional data.

- **Bagging:** A method that creates numerous versions of a classifier so that they can be used in an aggregated form. For a numeric result, the aggregated form averages over the versions for prediction voting takes place. In the case of classification, the classifier can be made optimal using bagging (Breiman; 1996).
- **Stacking:** A meta-classifier is used for the aggregation of various classification models. The first step involves the usage of training dataset for preparing the base level models. In the second step outcome from step one is used as an attribute for meta-classifier models [10].
- **Gradient Boosting Method:** GBM is quite the opposite to what is done in bagging or stacking, here the weak classifiers are targeted, and their performance is improved as it a sequential process which will keep going until it hits the threshold. It can't be applied to complex classifiers with high variance (Neural networks) [11].

5 Implementation

The implementation of this study was carried out using an R programming language in RStudio. This can be seen in the process flow diagram in figure 4.

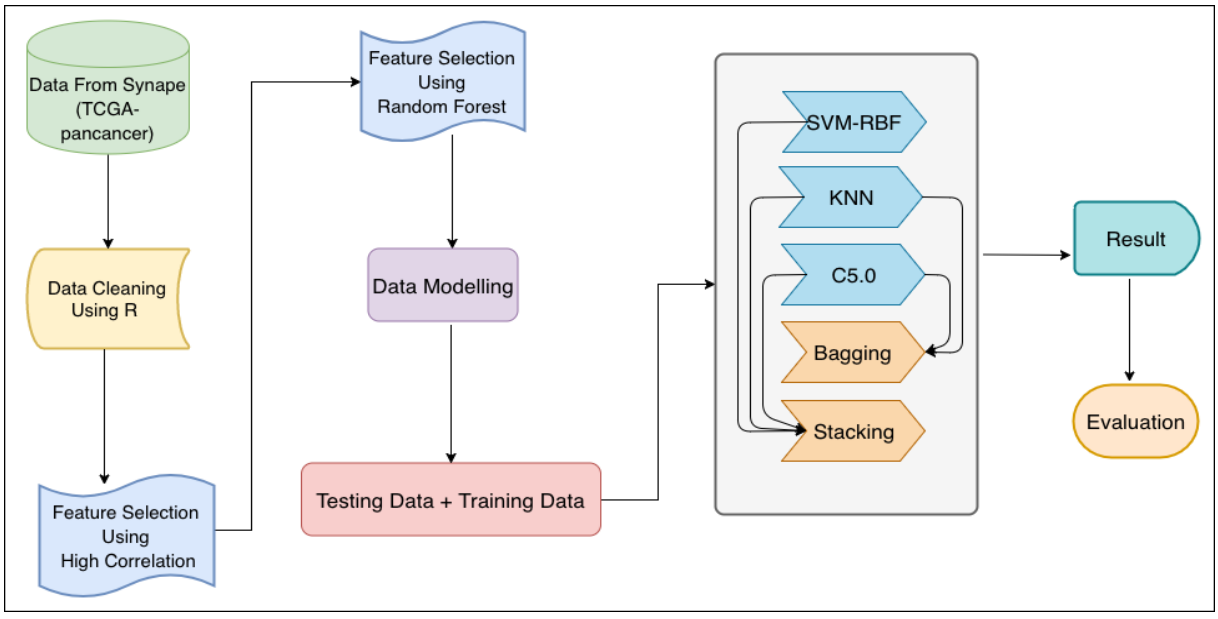

Figure 4: Architecture of Research

First and foremost a working directory is set for the research project. For the analysis to start dataset was loaded into R using 'read.csv' function followed by what can be said as the most important step which is data cleaning. All the redundancies were eliminated from the dataset as explained in Data Preparation in section 3 (Methodology). And as a last measure, NA's are checked for all the attributes in the dataset.

[10]Stacking: `https://blog.statsbot.co/ensemble-learning-d1dcd548e936`
[11]GBM: `https://datascienceplus.com/gradient-boosting-in-r/`

A tumor probability table was produced with the help of 'prop.table' function, according to the probability of which, the vast pool of genes classify BRCA as 37.45%, which is one of the causes for Breast Cancer. The second step was to load and install the packages that were necessary for the execution of algorithms. The Package 'caret' is used for classification as well as regression models but in this case, for radial-SVM and various functions from the package 'caret' were used in this study such as confusion matrix, recall, train and varImp; which led to the betterment of all corresponding results from the models.

Subsequent to data cleaning high correlation matrix was used to filter the superfluous genes after which feature selection was made using 'randomForest' function, which uses many parameters. First, the number of trees (n) is 100 because the dataset is small in size and it takes less time for computation. Second, 'importance' parameter is set to true, so that mean decrease gini plot & mean decrease accuracy plot are produced which are an integral part of cleaning the dataset. Because these plots provide the significant variables which in addition to the high correlation matrix, where 'corrpot' helped in plotting correlation matrix as seen in figure 3. Random forest and high correlation matrix helped to reduce the number of genes from fourteen thousand to fifty-one. These variables will be used as an input in the rest of the models used for classification. After selecting the most important variables, data is split into training and testing data, i.e. a ratio of 80% and 20% respectively. As the variables having more significant value in the dataset tend to dominate other variables, therefore, scaling of data is essential by using 'prop.table' & 'plot' functions.

Individually models were trained and tuned to enhance their performance. Support Vector Machine has been implemented with the help of 'train' function. The parameter passed in the function is radial SVM because we have a multiclass problem at hand. A multiclass classification problem cannot be plotted in 2-D, it needs a 3-D hyperplane to distinguish the classes. SVM has kernels that help solve such problems i.e. linear, quadratic, polynomial and radial [12]. As radial SVM has high variance and overfitting, 10-fold cross-validation has been applied for tuning of the model. After implementation of SVM, k-nearest neighbours has been applied using the package 'class'. All the significant variables have been used for training of the model. Parameter tuning leads to improved performance of the model. One advantage of tuning is that it deals with the issue of overfitting. Decision trees have been implemented using algorithm C5.0 from the package 'C50' for breast cancer classification. Decision trees are susceptible to overfitting, 10-fold cross-validation overcomes this problem.

Ensemble learning uses SVM, k-NN and C5.0 models for implementation of bagging and stacking techniques. Bagging was implemented with the help of averaging. Significant Features from RF are uniquely selected for each model by selecting odd and even variables. The function "apply" assisted in calculating the average of applied models. While in stacking individually trained models act as base models using the significant variables. Their prediction probabilities have been calculated in the form of a data frame which is then passed on to the top layer in the model where gradient boosting machine is used as the meta-classifier with the help of "gbm" package. At the top layer, the weak classifiers are trained iteratively resulting in overall better performance of the model.

[12]Support Vector Machine: `https://goo.gl/Gwh3VA`

6 Evaluation and Results

For the evaluation of this research, a confusion matrix was used to analyse the performance of the five models, i.e. SVM-RBF, kNN, C5.0, Bagging and Stacking. The overall statistics and statistics by class of all the confusion matrix's created in the due process can be seen in table 1 and table 2. Precision, recall and f1 score were calculated for the five breast cancer tumors (BRCA, PRAD, KIRC, COAD and LUAD) using the formulas as mentioned earlier in section 3. As seen in table 1 comparison of the evaluation measures against each model has been depicted. For ensemble learning techniques, it is evident that stacking has outperformed the other models by classifying the actual positives of each variable correctly with a precision of 0.98. All the other models that have been implemented individually, SVM-RBF has the best results against kNN & C5.0. Stacking is the best overall.

Table 1: Evaluation of Models

Model Name	Class->	BRCA	KIRC	COAD	LUAD	PRAD
SVM-RBF	Precision	0.9524	1.0000	0.9412	1.0000	1.0000
	Recall	1.0000	0.9655	1.0000	0.8929	1.0000
	F1-score	0.9756	0.9825	0.9697	0.9434	1.0000
kNN						
	Precision	0.9831	1.0000	1.00000	0.9286	0.9000
	Recall	0.9667	0.9655	0.93750	0.9286	1.0000
	F1-score	0.9748	0.9825	0.96774	0.9286	0.9474
C5.0						
	Precision	0.9365	1.0000	1.0000	0.9600	0.9643
	Recall	0.9833	1.0000	0.93750	0.8571	1.0000
	F1-score	0.9593	1.0000	0.96774	0.9057	0.9818
Bagging						
	Precision	0.9833	0.931	1.0000	0.8214	1.0000
	Recall	0.8939	1.0000	1.0000	0.9583	1.0000
	F1-score	0.9365	0.9643	1.0000	0.8846	1.0000
Stacking						
	Precision	0.9833	1.0000	1.0000	0.9643	1.0000
	Recall	1.0000	1.0000	0.9412	0.9643	1.0000
	F1-score	0.9916	1.0000	0.9697	0.9643	1.0000

Breast cancer classification has been done using five models that have performed to their full potential with the available dataset. In table 2 we can see that comparison of various models has been done to check their performance with the help of classification accuracy and kappa value. The best outcome for accuracy and kappa value was given by Stacking i.e. 98.75% as seen in table 2. As observed in previous research work not a lot of work has been done using RNA-Seq data, for microarray datasets highest accuracy achieved is 99.78% in case of (Jaffar et al.; 2009) using ensemble learning, Although, going by the inferences gathered from (Salem et al.; 2017) accuracy was staggering 100% for particular datasets and not the overall result; which was obtained with Information Gain (IG) and Standard Genetic Algorithm (SGA). From section 2 it is clear that the accuracy of the models implemented majorly depends on the type of dataset used but as mentioned in section 2, there are instances when same dataset has been used with different models, different feature selection methods resulting in different results. The data varies

from one research to other as the data is gathered from patients and it depends on their vitals, health and numerous other variables in play as gene expression data is involved in the research. Therefore it depends on the models being implemented and dataset under consideration. There are various instances in the previous research work where it can be seen that accuracy has gone up to 98.7%, 98.5% and 97% with microarray data. These findings infer that this research has obtained quite impressive results for ensemble learning using RNA-Seq data.

Table 2: Breast Cancer Evaluation		
Model Name	Accuracy (%)	Kappa
RF	94.38	0.9251
SVM-RBF	97.5	0.9669
kNN	96.25	0.9506
C5.0	96.25	0.9502
Bagging	95	0.9333
Stacking	98.75	0.9835

Random forest used for feature selection has generated an accuracy of 94.38%, which shows that the features used for classification of breast cancer were chosen correctly and were significant enough to produce such good results. The Genes: BRAC1, BRAC2, ATM, PTEN, SKT11, BRIP1, LKB1, CDH1, RAD51, MRE11A, CHEk2, LGR6, NBN, CYP17, RAD50 and TP53 have contributed towards the classification of breast cancer as they were present in the significant genes selected by RF is shown in figure 5.

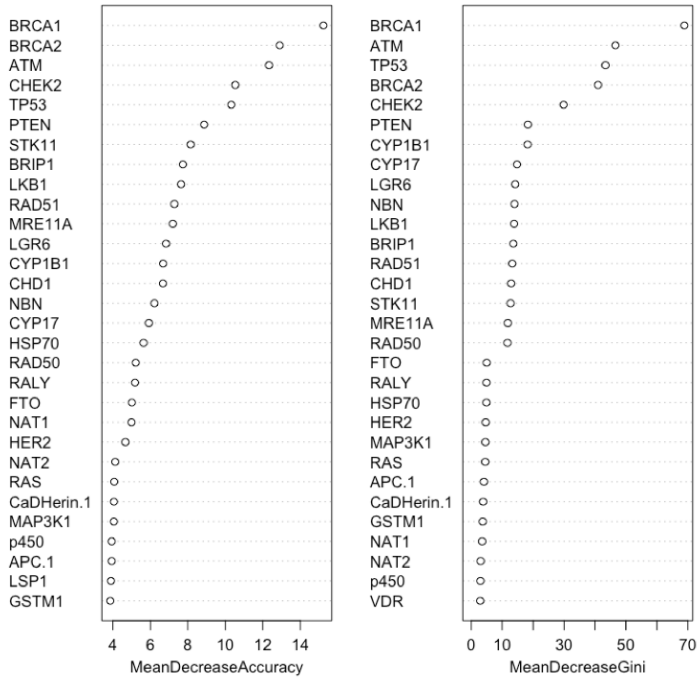

Figure 5: Significant Features

Figure 6 shows the plot for C5.0 (Decision Tree) of accuracy against the boosting iterations done by C5.0. The decision tree has been trained on two types of models, i.e. tree & rule-based as shown in figure 6. Model-based on rules generally disintegrates the original tree into rules that are mutually exclusive. The comparison has been made

between winnowing and no winnowing. Independent variables are omitted, and the performance of the model is calculated in the winnowing section, which leads to a decrease in the performance of the model as the accuracy dips down to almost 94%. The dip in the winnowing section occurs as the features responsible for hiked error rate are flagged & then the unflagged variables are used to train the model; significant variables are evaluated beforehand. On the other hand, the no winnowing section is the one where all the significant variables are considered, produces an accuracy of 96.25%.

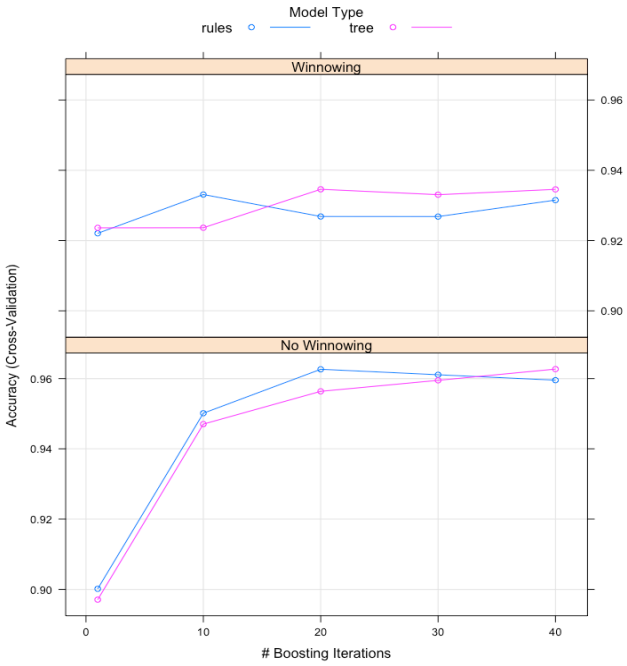

Figure 6: Decision Tree Plot

Figure 7 shows the plot for Gradient Boosting Machine which depicts the relative influence of the variables responsible for prediction. When the GBM model is boosted with optimisation, it shows the importance score of relative features. In this research gradient boosting machine helped in the process of stacking (ensemble learning) were SVM comes out to be the model with most influence among other models , which helped in achieving highest accuracy for stacking.

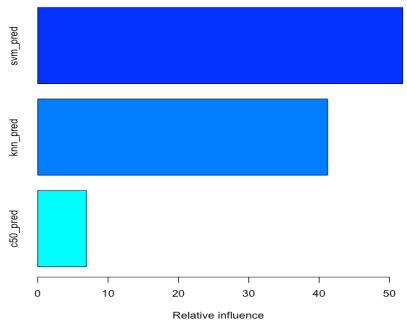

Figure 7: Gradiant Boosting Machine Plot

7 Conclusion and Future Work

The primary objective of this study was to classify the breast cancer tumors with the help of genes provided by RNA-Seq gene expression levels using ensemble learning. After analysis of the RNA-Seq data, conclusion inferred from the results is that ensemble learning is an efficient technique for classification of breast cancer. It is evident from the results that with an accuracy of 98.75% and a high kappa value of 0.9835, stacking is the best ensemble learning algorithm among other algorithms implemented for classifying RNA-Seq data. The most essential genes found with the help of random forest (mean decrease accuracy plot and mean decrease gini plot) for classification of five tumors at hand were BRCA1, BRCA2, CYP17, CHEK2, ATM, TP53 and PTEN. As seen in previous research work these are the genes that are responsible for triggering abnormal growth of the cells i.e. BRCA tumor, thus leading to breast cancer. Therefore in case of breast cancer these are the genes to look out for while dealing with different set of tumors.

Keeping in mind the limitations of the dataset in this study, the scope for research in the domain of breast cancer is vast. Already much effort has been put up to tackle the classification of breast cancer using different machine learning algorithms such as SVM, RF, LR, RVM, KNN, ANN, NB and QC but deep learning techniques (DBN, SDAE) are yet to be explored to their full potential. Hinderance in the successful execution of a deep learning technique is the limitation of data availability. Once this issue is dealt with it could prove to be vital as the number of genes that can be extracted from human DNA and RNA is in surplus. Use of microarray datasets has been overtaken by the new RNA-Seq data with gene expression levels to move to a more advanced level of research in the near future.

Acknowledgment

First, I would like to express my sincere gratitude to my supervisor Mr Vikas Tomer for the continuous support of my research which would have been impossible to complete without his patience, immense knowledge and motivation throughout the semester. I want to thank the open platform Synapse for providing the data for breast cancer patients. I am profoundly grateful to National College of Ireland library for providing the necessary resources for the completion of my research. Finally, I would like to thank my parents and friends for the constant support and motivation.

References

Breiman, L. (1996). Bagging predictors, *Machine learning* **24**(2): 123–140.

Castillo, D., Gálvez, J. M., Herrera, L. J., San Román, B., Rojas, F. and Rojas, I. (2017). Integration of rna-seq data with heterogeneous microarray data for breast cancer profiling, *BMC bioinformatics* **18**(1): 506.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Chiu, S.-H., Chen, C.-C. and Lin, T.-H. (2008). Using support vector regression to model the correlation between the clinical metastases time and gene expression profile for breast cancer, *Artificial intelligence in medicine* **44**(3): 221–231.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and psychological measurement* **20**(1): 37–46.

Danaee, P., Ghaeini, R. and Hendrix, D. A. (2017). A deep learning approach for cancer detection and relevant gene identification, *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, World Scientific, pp. 219–229.

De Jong, M., Nolte, I., Te Meerman, G., Van der Graaf, W., Oosterwijk, J., Kleibeuker, J., Schaapveld, M. and De Vries, E. (2002). Genes other than brca1 and brca2 involved in breast cancer susceptibility, *Journal of medical genetics* **39**(4): 225–242.

Elshazly, H. I., Elkorany, A. M., Hassanien, A. E. and Azar, A. T. (2013). Ensemble classifiers for biomedical data: performance evaluation, *Computer Engineering & Systems (ICCES), 2013 8th International Conference on*, IEEE, pp. 184–189.

Firoozbakht, F., Rezaeian, I., Ngom, A., Rueda, L. and Porter, L. (2015). A novel approach for finding informative genes in ten subtypes of breast cancer, *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, IEEE, pp. 1–6.

Gharibi, A., Sehhati, M. R., Vard, A. and Mohebian, M. R. (2015). Identification of gene signatures for classifying of breast cancer subtypes using protein interaction database and support vector machines, *Computer and Knowledge Engineering (ICCKE), 2015 5th International Conference on*, IEEE, pp. 195–200.

Giarratana, G., Pizzera, M., Masseroli, M., Medico, E. and Lanzi, P. L. (2009). Data mining techniques for the identification of genes with expression levels related to breast cancer prognosis, *Bioinformatics and BioEngineering, 2009. BIBE'09. Ninth IEEE International Conference on*, IEEE, pp. 295–300.

González-Navarro, F. F. and Belanche-Muñoz, L. A. (2011). Parsimonious selection of useful genes in microarray gene expression data, *Software Tools and Algorithms for Biological Systems*, Springer, pp. 45–55.

Gypas, F., Bei, E. S., Zervakis, M. and Sfakianakis, S. (2011). A disease annotation study of gene signatures in a breast cancer microarray dataset, *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, IEEE, pp. 5551–5554.

Hsieh, S.-L., Hsieh, S.-H., Cheng, P.-H., Chen, C.-H., Hsu, K.-P., Lee, I.-S., Wang, Z. and Lai, F. (2012). Design ensemble machine learning model for breast cancer diagnosis, *Journal of medical systems* **36**(5): 2841–2847.

Jaffar, M. A., Hayder, Z., Hussain, A. and Mirza, A. M. (2009). An intelligent ensemble based systems for breast cancer diagnosis, *Proceedings of the 2009 International Conference on Computer Engineering and Applications, Manila, Philippine*, pp. 6–8.

Khuriwal, N. and Mishra, N. (2018). Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm, *2018 IEEMA Engineer Infinite Conference (eTechNxT)*, IEEE, pp. 1–5.

List, M., Hauschild, A.-C., Tan, Q., Kruse, T. A., Baumbach, J. and Batra, R. (2014). Classification of breast cancer subtypes by combining gene expression and dna methylation data, *Journal of integrative bioinformatics* **11**(2): 1–14.

Niaksu, O. (2015). Crisp data mining methodology extension for medical domain, *Baltic Journal of Modern Computing* **3**(2): 92.

Otoom, A. F., Abdallah, E. E. and Hammad, M. (2015). Breast cancer classification: Comparative performance analysis of image shape-based features and microarray gene expression data, *International Journal of Bio-Science & Bio-Technology* **7**(2): 37–46.

Salem, H., Attiya, G. and El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles, *Applied Soft Computing* **50**: 124–134.

Tarek, S., Elwahab, R. A. and Shoman, M. (2017). Gene expression based cancer classification, *Egyptian Informatics Journal* **18**(3): 151–159.

Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer, *New England Journal of Medicine* **347**(25): 1999–2009.

Vanitha, C. D. A., Devaraj, D. and Venkatesulu, M. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection, *procedia computer science* **47**: 13–21.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *nature* **415**(6871): 530.

Wang, X. and Gotoh, O. (2010). A robust gene selection method for microarray-based cancer classification, *Cancer informatics* **9**: CIN–S3794.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J. et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, *The Lancet* **365**(9460): 671–679.

Xu, X., Zhang, Y., Zou, L., Wang, M. and Li, A. (2012). A gene signature for breast cancer prognosis using support vector machine, *Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on*, IEEE, pp. 928–931.

Zeng, T., Luo, F. and Liu, J. (2009). Biological pathway conducting microarray-based cancer classification, *Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on*, IEEE, pp. 1–5.

Zhang, W., Zeng, F., Wu, X., Zhang, X. and Jiang, R. (2009). A comparative study of ensemble learning approaches in the classification of breast cancer metastasis, *Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS'09. International Joint Conference on*, IEEE, pp. 242–245.