Microarray Based Cancer Classification Using Ensemble Learning

by

Nehazee Ahmed
17201074

Mohammad Hasin Bin Sadique
16341007

Timon Protik Mondal
20141038

Rakin Ahmad
16301161

Shawon Alam
16101264

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2020

Declaration

It is hereby declared that

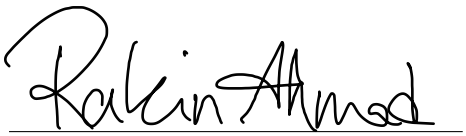
1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

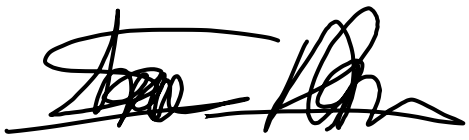
Timon Protik Mondal
20141038

Mohammad Hasin Bin Sadique
16341007

Nehazee Ahmed
17201074



Rakin Ahmad
16301161



Shawon Alam
16101264

Approval

The thesis titled "Microarray Based Cancer Classification Using Ensemble Learning" submitted by

1. Timon Protik Mondal (20141038)
2. Mohammad Hasin Bin Sadique (16341007)
3. Nehazee Ahmed (17201074)
4. Rakin Ahmad (16301161)
5. Shawon Alam (16101264)

Of Summer, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 05, 2020.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Mahbubul Majumdar, PhD
Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Cancer diagnosis and classifications are one of the crucial emerging clinical applications of microarray data. The technology of DNA microarray has allowed us to find the gene expression levels of thousands of genes at the same time giving us a huge opportunity for cancer diagnosis and classification. To determine a computational model, from the given data so that the classes of unknown samples are determined, is the primary objective of microarray data. The majority of researches done in this field dealt with gene expression sequence to differentiate between the cancerous gene expression sequence and healthy gene expression sequence. We are using a microarray dataset which contains the RNA sequence gene expression levels of thousands of genes of hundreds of samples. Our dataset contains information about five types of cancer referred as breast, colon, kidney, prostate and lung cancer. We are using five classification models namely Support Vector Machine(SVM), Random Forest, XGBoost, K Nearest Neighbour(KNN) and Deep Learning for classifying the five different cancers and labeling the groups using the labeled data . Using the ensemble learning technique we combined all the five models in order to make an optimal classification model that can differentiate each sample of the test dataset to its respective cancer class. We used t-SNE (t-distributed stochastic neighbor embedding) model for data visualization and creation of five clusters of each cancer. By training and testing our dataset we were able to determine the accuracy of each model. Using our well-structured model we can easily use any unknown patient's RNA sequence gene expression level to identify if he or she is a patient of any of the five cancers or not.

Keywords: Cancers; Gene expression level; Microarray; Classify; Ensemble Learning; SVM; Feature Selection

Dedication

We would like to dedicate our work to all the cancer patients and medical researchers for whom we have worked so hard so that cancer classification process is done in a faster and easier way. As soon as the cancer type will be identified the appropriate treatment can be started way earlier benefitting the cancer patients in every way possible.

Acknowledgement

Firstly, all praise to the Great Allah due to whom our thesis have been completed without any major interruption. Secondly, to our advisor Md. Golam Rabiul Alam Sir for his kind support and advice in our work. The paper would not be completed without his constant support and involvement. It was all due to his guidance and motivation that pushed us to give our best at this paper. Finally to our parents for their support and prayers which made our research completion possible.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$BRCA$ Breast Cancer Gene

$COAD$ Colon Adenocarcinoma

$KIRC$ Kidney Renal Cell Carcinoma

KNN K-nearest neighbor

$LUAD$ Lung Adenocarcinoma

$PRAD$ Prostate Adenocarcinoma

RBF Radial Basis Function

RF Random Forest

SVM Support Vector Machine

$XGBoost$ Extreme Gradient Boost

Chapter 1

Introduction

One of the crucial research areas in the world of medicine is cancer research. Around the world, millions of people are diagnosed with cancer and close to half of them die due to this[8]. In many countries cancer comes in the second position as the most common cause of death after cardiovascular diseases. Trillions of cells grow and divide on a regular basis in a healthy body as the body needs them to function daily. The specific life cycle of a healthy cell is that it reproduces and dies off in a system which is determined by cell type. As old or damaged cell dies, new cells replace them [32]. This natural process is disrupted by cancer which leads to abnormal growth of cells and it occurs due to the changes or mutations in DNA. In our research we will be focusing on 5 types of cancers which are breast cancer, kidney cancer, colon cancer, lung cancer, and prostate cancer. DNA microarray technology has led to a vast development in oncogenomic analyses and in identification of the complex gene expression patterns of cancer [22]. Microarray is the process used to determine the gene expression of thousands of genes simultaneously. By measuring the binding of mRNA to oligonucleotides attached to the glass slide in microarray experiments, gene expression levels are found. In this chapter, we will discuss the motivation for our work, our contributions towards this research and the thesis orientation which will cover the contents of each chapter in this paper.

1.1 Motivation

Cancer classification types can be of around 200 types and so the specific type of cancer must be detected so that the treatment can be started as early as possible before the tumour starts to spread making the cancer incurable [2]. Mutations in cancerous bodies can cause abnormal situations of cells where unnecessary cells starts to form abnormally creating a disruption in the natural processes of cells. The extra cells which are produced by mutations start to divide without any control or manner which eventually causes the formation of tumor. Variety of health problems are caused by the tumors and the problem varies according to the tumor location in the body parts. Cancer classification is vital to know which type of cancer a particular sample is among various types of cancer data.

As we are working with five different cancer data therefore, training the dataset using various classification methods will result in representing each sample to its corresponding cancer type. Our motive is to classify all the samples into groups of

the same cancer and labeling them according to our labeled data so that any new data of any person's gene expression level can be used to determine if he or she is a patient of any five of the cancer in our dataset. We have chosen a microarray data containing around 801 samples and approximately 16,383 columns of genes for each sample. The dataset contains RNA sequence gene expression levels for each sample for 16,383 genes and by finding the correlation between the values we will determine which samples are of the same group and label them with what cancer group it is. Previous works have been done using gene sequences and microarray data which were used to classify between a healthy and cancerous samples. We are introducing some new methods of classification between different types of cancer which were not widely used.

In summary we will classify the samples of five cancers according to their RNA sequence gene expression level data into five groups using ensemble learning classification method that combines algorithms like Support Vector Machine (SVM), Random Forest, XGBoost, k-Nearest Neighbor (KNN) and Deep Learning then labeling the groups with the names of cancers according to the labeled data from our dataset.

1.2 Major Contribution

Most researchers in the field of cancer classification have worked with mainly gene expressions. The sequences of gene expression were used to classify the cancer from the others as the tumor gene expression sequence is different than a healthy gene expression sequence. Many did work with microarray data using various machine learning approaches such as SVM (support vector machine), Random Forest, K-means, KNN (k –nearest neighbor) and different testing methods to test the data and measure the performance. These methods were used to classify and cluster the data and the accuracy of each model was determined. We are working with a microarray dataset which has five types of cancer data. The data contains the gene expression levels of thousands of genes of hundreds of samples. We have used a unique approach in this research which is the XGBoost and LightGBM algorithm, for feature selection which is not widely used in this type of research, along with SVM, Deep Learning, and Random Forest to cluster and classify the data using ensemble learning.

1.3 Thesis Orientation

The subsequent sections of the thesis have been organized as follows. Chapter 2 is the literature review which contains many of the past research works in this area and existing approaches relevant to our proposed model. Chapter 3 is the background study related to our work and how we are using the resources to get the desired output. Chapter 4 includes we have the proposed model of our works along with relevant graphs and figures. It includes our overall working methodologies and

algorithms, dataset description and the models we used. The predicted results and relevant discussions are showed in chapter 5. Lastly, the summary of the report and conclusion as well as some future work plan is done in chapter 6

Chapter 2

Literature Review

For our research work, we studied different papers that are related to cancer classification and microarray datasets with different methodologies and algorithms.

One of the studies proposed an innovative gene selection method for cancer classification[39]. For that, Gene Selection Programming (GSP) method was proposed in this study. This programming method was used for classifying the cancer in an effective way. It is worked for Gene Expression Programming (GEP) method. GEP consists of two parts including linear chromosomes (Genotype) and Phenotype. Phenotype is tree structure which is called Gene Expression Tree. When the given chromosome is represented, Genotype is then published. Then Genotype can be converted to Phenotype. GEP process includes creating chromosomes for population initialization, evaluating the best individual by identifying a suitable fitness function, achieving the optimal solution by conducting genetic operations and checking the stop conditions whether it is terminated or not. In the first step of GEP method, candidates are constructed from two sets. In between two sets, terminal set represents the attributes to the microarray dataset. Selecting all the attributes into the terminal set will affect the computational efficiency and also create a noise in microarray dataset. So they identified a threshold and selecting n-top (Top 50) ranked genes for reducing the noises from microarray dataset. Inspite of having some disadvantages of this process, they used a different technique named Systematic Selection Approach. This process includes attributes ranking, attributes weight calculation, attribute selection using Roulette Wheel Selection method. By using Information Gain algorithm, they ranked the microarray attributes. Classification process has more influenced by the top ranked attributes. In attribute selection method, based on the weight, there settled all attributes on Roulette Wheel method and top weights attributes has the highest chance to be placed in the terminal element. The process of creating GSP chromosomes includes selecting the terminal set using systematic approach, defining function set, randomly select elements from the candidate selection to create a head of a gene, randomly select elements from terminal set to create the tail of a gene and finally creating phenotypes of all genotypes. Searching the smallest gene subsets was the objective of their proposed method. For that, they defined fitness function that consists accuracy result which is based on the SVM classifier and number of attributes in the dataset. They used 10 microarray datasets for evaluating

the performance of GSP and they follow the criteria consisting classification accuracy, attributes number and CPU time for performance evaluation. After that, they compared results with three well-known Gene Selection algorithms named Particle Swarm Optimization (PSO), Gene Expression Programming and Genetic algorithm. Results on the experiments showed that, GSP algorithm got highest accuracy on the datasets and also took shorter time which was showed in the time results.

In another study of semi-supervised SVM-based Feature Selection for cancer classification using gene expression data[23] aimed to propose a Semi-supervised SVM-based feature selection which exploits the knowledge from unlabeled and labeled data. They first applied the model to microarray data from 181 lung cancer tissue samples with 31 Malignant Pleural Mesothelioma (MPM) and 150 Adenocarcinoma (AD). By SVM-based Recursive Feature Elimination (SVM-RFE) they generated the features ranking using backward feature elimination in where features were arranged according to the weights produced from Supervised SVM and Unsupervised SVM and features with lowest ranked will be removed. In this study, they used SVM-RFE (Semi-supervised SVM based Recursive Feature Elimination) method which integrated the real problem from unsupervised algorithm. In the training process, Supervised SVM trained the labeled data and unsupervised one-class SVM algorithm trained the unlabeled data. In this algorithm, they restricted the training examples to good features indices labeled and unlabeled data, trained the classifier, computed the length dimension, determined the sum of weight, computed the ranking, select the feature, update the ranked list of feature, smallest ranking feature elimination and finally features ranking. Another is SVM-FS (Semi-supervised SVM-based feature selection) method that ordered the list of features according to the ranking scores. Then, by eliminating the lowest ranked feature the dataset will be updated. Their microarray data contained 181 Lung cancer tissue samples. From 10 sets of the data, 9 sets were occupied for training process and the remaining set were occupied for testing purpose. Moreover, 50 percent of the samples from training set were selected randomly and the rest of the samples were trained without class labels. Experimental results on the lung cancer gene expression data showed that, SVM achieved higher accuracy and if it will compare with the supervised method SVM-RFE, it required shorter processing time.

We went through another paper in which they presented a comparative study of seven feature selection methods and evaluated performance by using six different types of classification methods and applied it to four multiclass cancer datasets [13]. In this paper, they identified an approach for the comparison of various feature selection method. Improvement on the ability of the prediction on classification of multiclass cancer was emphasized on this paper. They worked on four benchmark in the comprehensive experiments. The seven Gene Selection method they used includes Correlation based Feature Selection in which they selected the features subset; Chi-Squared (Chi) method in which each gene was evaluated individually by Chi-square measurement; Information Gain algorithm for measuring Entropy; Gain Ratio for computing as information; Symmetrical Uncertainty which is measured by the total classes entropy and entropy of the attributes ;Relief F for calculating relevancy score for each attribute and SVM- RFE for eliminating the gene iteratively. Additionally, for comparing the above seven feature selection algorithms, they applied six state-of-the-art classifiers, namely J48, logistic model trees (LMT),

Bayes Network (Bnet), Naïve Bayes (NB),k-NN, sequential minimal optimization algorithm for training SVM(SMO).Additionally, their data processing includes normalization. In the dataset, there were 7070 probe sets remaining after reducing genes. Then the gene expression values were organized through normalization. Experimental results brought some new insights into the feature selection methods. Small improvements was found in some datasets and accuracy were increased. As the combination of feature selection and classification methods were genuine, they were able to prove the excellent performance and highest prediction accuracy on all the datasets that was around 100 percent.

In another study, they did Gene Selection and Classification of microarray data using random forest through the investigation of the usage of Random Forest for classification in microarray data [7]. They proposed a new method for classification which is based on Random Forest. In this research, firstly they presented an evaluation of random forest because sometimes there might be a problem of performing random forest in classification on microarray datasets. Number of patient's class, number of dimension that are not dependent and genes number of each dimension were simulated on the simulated datasets. They have generated the datasets for two, three and four classes where 20 genes had signal. For that, they selected 4000 genes randomly. Additionally, they generated 4 replicate datasets for each level of number of classes. For taking small sets of genes, they did backward elimination process. They examined the performance of their approach and made the comparison to other related approaches. Then the variable selection approach's predictive performance had been compared. After that, the changes brought in the parameters which gave effects, and then they examined it. They examined all the forests that results from eliminating dropped fraction of the genes which is dropped by 0.2 by default and it helps them to do faster operation with the Aggressive Variable Selection approach. As a result, the method they used returns small sets of genes in comparison with other selection methods. For that, they used KNN, SVM with linear kernel, DLDA concerning 200 gene samples.

Another study was about the comparison of feature selection and classification combination in which they classified cancer datasets through the performance evaluation of over 2000 combinations of feature selection classification algorithms[11]. In this work, they selected 15 sets of genes which obtained 96 percent accuracy for a dataset. In this study, they worked on Colon cancer dataset (2000 genes in 62 samples), Prostate cancer (5845 genes in 102 samples) and Breast cancer (7129 genes in 46 samples) for performance evaluation of various feature selection algorithms. Based on the classification accuracies for these datasets, the toughest classification problem was Colon cancer dataset compared to the other three datasets. Filter methods was used for attribute ranking and by using wrapper method they searched the feature subsets that were optimal. For computing all the feature selection and classification computations, they used a data mining-mining software that is Waikato Environment for Knowledge Analysis (WEKA).In the testing process, features number, a total of 2006 combinations of method of selecting the features and methods of classification were tested and evaluated. In WEKA platform, they implemented all the methods. However, 10-fold cross validation evaluated the classification performance of the methods and after testing the Prostate cancer dataset, accuracy was above 90 percent that was given by the combinations.

In another study, a group of researchers attempted to explore many features and classifiers to evaluate the performances of feature selection methods and machine learning classifiers[55]. Information Gain, Mutual Information and Signal to noise ratio have been used for selecting the features and MLP, k-NN, SVM used for classification. Machine learning was defined for DNA microarray analysis. In this case, they selected discriminative genes that were related to classification, trained the data and then by using learning classifier they classified the new data. When they got gene expression data, they calculated it. Their prediction system divided into two parts including selection of features and classification of stages. For selecting the features, they attempted to get the best feature from large dimensionality. That's why they extracted 25 genes using 7 different methods and cancer predictor were categorized. For classification, they used MLP, k-NN and SVM. In the classification process, they trained the classifier form the samples that were trained and then tested with this samples. In one of their dataset of Colon Cancer consisting 62 samples of colon epithelial cells taken from colon-cancer patients in which each sample contained 2000 gene expression levels. However, in the training data, 31 samples were used and the rest of the samples as testing data.

One of the studies focused on the applications of Support Vector Machine to cancer classification in which they used SVM for cancer classification[6]. The challenges they faced in using gene expression data was that, this kind of data are usually very high dimensional which lasts for over ten thousand. So, to overcome that problem, they went through the process of dimension reduction that was Principal Component Analysis (PCA) in which they transformed the input space by PCs. The input space dimensionality was reduced by PCA analysis. Additionally, by implementing a t-test based gene selection approach, they found the genes that contributed most to the classification. At first, they ranked according to the level of genes with t- score test. From 60 genes they picked, they input 30 genes to the SVM classifier with the basis of rank. Finally, SVM classifier was trained and tested with data. In their dataset of Lymphoma which includes 4026 gene expression data, a part of data was missing. That's why they used k- NN for filling those missing values. The dataset named SBRC included 2308 genes data and among that set, 63 of the samples were trained and 25 of the samples were tested. The last dataset was Leukemia consisting 7129 gene expression data in which the 38 samples were used for training and the other 34 samples were used for testing purpose. For Lymphoma and SBRCT data sets, the data was found normalized. But this normalized data were not found available in Leukemia dataset. So, to overcome this problem, the data of Leukemia should be normalized. This normalization procedure includes reduce the genes with the maximum and minimum of the expression value of a gene and carrying out the logarithmic transformation to all the expression values. As a result, 3571 genes survived after this process. Then they standardized the data in their experiments by reducing the mean.

In another research work, a group of researchers demonstrated the microarray breast cancer classification using Machine Learning approaches[38]. At first, they did not use any feature selection methods and implemented 8 different machine learning algorithms in their data. Then they implemented 2 different feature selection methods. In this work, they used an SVM classifier and took the result of the classification and made the comparison. Initially they divided the data into two parts. By the

training data, they created a learning model and testing data for testing the model. Then they classified data using Machine Learning algorithm. In this work, they used an SVM classifier and take the result of the classification and made the comparison. The first data contained 1919 different proteins from 133 individuals. Among them 122 individuals were breast cancer patients and 11 individuals were healthy. The second data contained 24481 kinds of protein from 97 individuals. Among them, 46 individuals were people that had a disease repetition and 51 individuals were not disease repeated. They used RFE (Recursive Feature Elimination) and RLR (Randomized Logistic Regression) feature elimination methods. In this method, they initialized attribute sets and attribute elimination. They worked on this elimination process until obtaining the remaining desired attributes. 50 features has chosen for stop criterion and thus they ended up the process till 50 attributes remaining. As a result, the accuracy of the dataset has increased. By applying MLP, they examined the effects of neurons number and layers number on the accuracy of a cancer classification process. They used Confusion Matrix as a predictor of correctness or incorrectness of the algorithm. Finally, classification results accuracy was ensured by the k-fold cross validation.

One of the studies focused on the breast cancer diagnosis using k-means Type-2 Fuzzy Neural Network in which it aimed to design a classification by using the k-means clustering algorithm and Type-2 Fuzzy Neural Network[35]. Firstly, they classified the training data into k groups based on its characteristics. After that The Interval Type-2 Fuzzy Neural Network (IT2FNN) will train the k classifiers. In this study, they used k-means algorithm to classify the training data into k sets. In this algorithm, the samples in the similar groups will have more comparative attributes than other groups. After the training process, they had k classifiers with various boundaries. After that, testing data will be determined that in which classifier it belongs to. In K-means algorithm, firstly they initialized the k center point. After that, they calculated the distance between each point to k center point. Then they assigned each point into k cluster. In order to show the effectiveness of the k-means classifier, the Wisconsin breast cancer dataset was used in this study. Finally, the system performance measurements were applied to calculate the accuracy, the sensitivity and the specificity of the Wisconsin breast cancer dataset.

A research on a colon cancer microarray analysis technique proposed a robust microarray analysis technique in which they tried to find the gene interaction involved in colon cancer[31]. They used two different techniques for measuring gene expression which are cDNA arrays and oligonucleotide arrays. At the first stage of an investigation, which is made by R (Robust Statistical Method) and this tool they used for visualizing its libraries and analyzing the gene expression. Then they collected 585 samples of different stages of Colon Cancer patients which they did after reading the data. It also contained information about the samples. They measure gene expression levels and summing up all the features. Then they proceed gene selection method to find the most related genes. Additionally, based on the network correlation and biological functions they characterize gene information. Then they classified genes and finding out the patterns. As a result, they found out the similar and different expression patterns. Finally, for understanding the relationship and behavior inside the pattern, they identified the particular sets of regulated and unregulated genes.

Another research named cancer type prediction and classification based on RNA sequencing data used RNA sequencing data from The Cancer Genome Atlas (TCGA) project and focused on classifying 33 types of cancer patients[34]. For that, they used Decision Tree, k-NN, Linear Support Vector Machine, Polynomial Support Vector Machine and Artificial Neural Network that were conducted to compare the performance of their accuracies and training process. As it is difficult to identify the causes and treatments of different types of cancers, they found the best way to characterize and identified different types of cancers using TCGA (The Cancer Genome Atlas) in which the large data were gathered and analyzed. TCGA launched Pan Cancer analysis to understand the common things and differences of all caner types. Firstly, they collected the raw data which includes 10471 cancer patients samples across 33 types of cancers, a list of 20531 genes based on the RNA expression level. Then, they mapped the information including cancer types to the corresponding cancer type. After that, they preprocessed the data for building classification model. In this step, they checked the raw data whether there were any missing or duplicate values present or not. Rows and columns with null values were removed from the dataset. Several preprocessing methods were considered for model training includes Feature Selection method. This method helped to create best accuracies by reducing the irrelevant genes. In their model, there was a total of 60 testing scenario based on the testing variables and 21 of them were useful experiments that were selected according to the complexity of testing. Then it was conducted to measure their performance. In the dataset, 80 percent data was used on the training set and 20 percent data was used on the testing set. Then the selected 21 useful experiments were further analyzed and compared the accuracy scores.

Another study was about to classification of breast cancer microarray data using Radial Function Network[15]. Microarray breast cancer data was used on this research and they proposed Radial Basis Function Network that classified the normal and cancerous cells. At first, they collected breast cancer microarray dataset from the Center for Computational Intelligence. There were two different classes and 24461 genes containing on the dataset. Then they deal with feature selection method in order to diminish high dimension of microarray data. Thus they achieved accuracy of classification. Then they used Relief-F algorithm as feature selection method. In this method, they estimated the quality of attributes that have threshold value was less than the weight of attributes. As it could not deal with multiclass datasets or large datasets, they used an extended Relief-F algorithm that can handle multiclass datasets. Then they deal with classification process for the classification of normal and cancer cells. Besides, Data classification was applied for classification. Neural Network was embedded on RBF network. In Neural Network, there are hidden units and radial activation was implemented by every hidden units. Then output units produced the weighted sum of the output hidden unit. Gaussian as an activation function worked in pattern classification. Finally, they evaluated the performance of the classifier using ROC (Receiver Operating Characteristics). For that, they analyzed the result of feature selection on 50-high ranked genes and 100-high ranked genes. They also used Information Gain and Chi square and the results showed that, all the classifiers achieved high cross validation accuracy using the selected genes of Relief-F but Information Gain and Chi- square could not achieve this accuracy as this methods ignore feature dependencies. Therefore, Relief-F algorithm was used

in this study.

There was a paper about cancer classification based on microarray gene expression data using Deep Learning in which they used deep learning algorithm based on Multilayer Perceptron to show the better classification performance[29]. The Artificial Neural Network research originated the concept of deep learning. In this deep learning architecture, Multilayer Perceptron is a good example in which there are many hidden layers. Besides, there are several theoretical frameworks for deep learning. In this study, they summarized the feed forward architecture and MLP are the feed forward network with this architecture. Each layer is formed from neurons. In the input layer, neurons received a signal. There was an activation function inside each of the neuron and this function controlled the input signal. This layer included the weighted sum of the previous layers output. Firstly, they tried to find the number of hidden units, activation function type and number of input and output variable by approximating the weight parameters. Then they used training set to estimate the weight parameters. From a set of randomly selected samples, they extracted 60 percent from the dataset, 20 percent were used as the testing set and 20 percent were used as the validation set. They built this operation by training MLP. They trained MLP for learning the multilayer architectures easily. In the input variable, there were 174 parameters, hidden variable has the parameter of (50, 50, 50) and output variable has the parameter of 11 which were used for evaluating the classification performance. As a result, they achieved higher accuracy result in classification by using this parameters.

Another paper was about to the novel Granular Support Vector Machines-Recursive Feature Elimination (GSVM-RFE) for the gene selection[54]. They used this method for eliminating an irrelevant genes on different levels. From 17 genes, GSVM extracted a compact "perfect" gene. Granular computing includes the information of subsets and clustering. Then problems which were targeted being evolved. Original task was also divided into a subtask for huge and complicated problems. Then GSVM combine the granular computing theories principal. In this study, they extended the GSVM framework to a new gene selection algorithm was reliable for gene selection. Firstly they clustered survival genes in every steps by Fuzzy C-Means. They ranked genes in descending order inside the granule. Genes with a higher rank were selected as new survived genes and combined all the new survived genes. In the next step, this genes were clustered again and repeated this process until prespecified threshold is greater than or equals to the number of survived genes. Then GSVM-RFE grouped the genes that had similar expression patters into cluster by Fuzzy C-Means clustering algorithm. As a result, genes with lower rank were removed as irrelevant genes and the higher ranked genes were survived as it was more significant. Finally, they compared GSVM-RFE algorithm with the Signal to Noise algorithm correlation based algorithm and SVM-RFE algorithm. The experimental results showed that, their proposed algorithm showed better accuracy compared to the other algorithms.

We went through another paper that was about to the microarray classification and rule based cancer identification[9]. In this paper, they proposed modified support vector machine (SVM) to classify cancer related five microarray data. They also observed a performance compare to previously used different version of SVM algorithm.

While working on microarray data, they faced challenges on finding out genes that are responsible for a particular cancer. In this case, they used popular rule based machine learning technique C4.5 for finding the genes that are responsible. In this research they worked on five widely used cancer related datasets and analyzed the data inside the dataset to find the causes of cancer. Additionally, they examined their proposed SVM performance with another version of SVM that is SVM Light and showed the efficiency of their proposed algorithm than SVM Light. The five datasets included with the clinical data along with statistical information of five cancers (Breast, Colon, prostate, Leukemia, Lung). Firstly they selected a proper kernel of SVM by observing the nature of these datasets. Then they mapped the distribution of data in Histogram based graphical presentation. Except colon cancer tumor dataset, the other dataset showed closely normal distribution. If Standard deviation is 0.5 in any dataset, they chose radial basis function else they choose polynomial to microarray classification. After that, they compared the classification accuracy by the comparison of their proposed SVM classification with two classifiers including Interesting Rule Group (IRG) and Classification Based on Association (CBA).As a result, accuracy was jumped around 100 percent which was much higher than other classifiers performance but only for leukemia dataset, the accuracy was about same for both proposed SVM and SVM Light. After that, they tried to search the genes that are responsible for a specific cancer. For mining the gene rules from the large gene expression data, they used decision tree entropy method. For identifying cancer type, the two decision tree algorithm parameter, pruning confidence level were tuned for generating the big microarray dataset. Lastly, the classification accuracy measured the rule of performance that were demonstrated all the suitable value. The results on the experiment showed the rule accuracy that was 98.70 percent which means the rule was generated with higher acceptability.

Chapter 3

Background Information

3.1 Gene and Gene Expression

DNA (deoxyribonucleic acid) is the genetic material of all organisms on Earth. Some of the traits or characteristics of the children such as the color of their eyes or the color of their hair can be determined due to the DNA which is passed to the offsprings from parents. DNA has a double helix shape and is comprised of four basic building bases which are adenine (A), cytosine (C), guanine (G) and thymine (T)[27]. Fig 3.1 shows a structure of a DNA[52]. A DNA molecule is divided up into functional units called genes and each gene provides instructions for a functional product which is a molecule needed to perform a job in the cell. The gene is the basic physical unit of inheritance of any living organism. Chromosomes are structures on which genes are placed in a specific consecutive manner. A chromosome contains one, long molecule of DNA, only portion of which is correlated with a single gene.

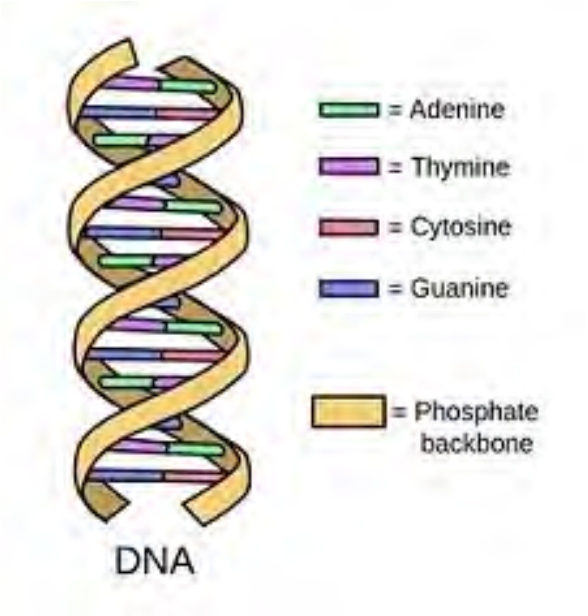


Figure 3.1: Structure of DNA

The process by which the instructions in our DNA are converted into a functional product, such as a protein is gene expression. Gene expression is found when the information stored in our DNA is converted into instructions for making proteins or other molecules. There are two main processes which are associated with gene expression. They are transcription and translation. The process of transcription is responsible for the production of messenger RNA (mRNA) by RNA polymerase which is an enzyme, and the processing of the following molecules of mRNA[20]. As this process allows DNA sequence to be converted into RNA this step is named transcription because it is basically rewriting. The difference of DNA and RNA is that RNA has a single strand of uracil (U) base and DNA consists a base of thymine (T).

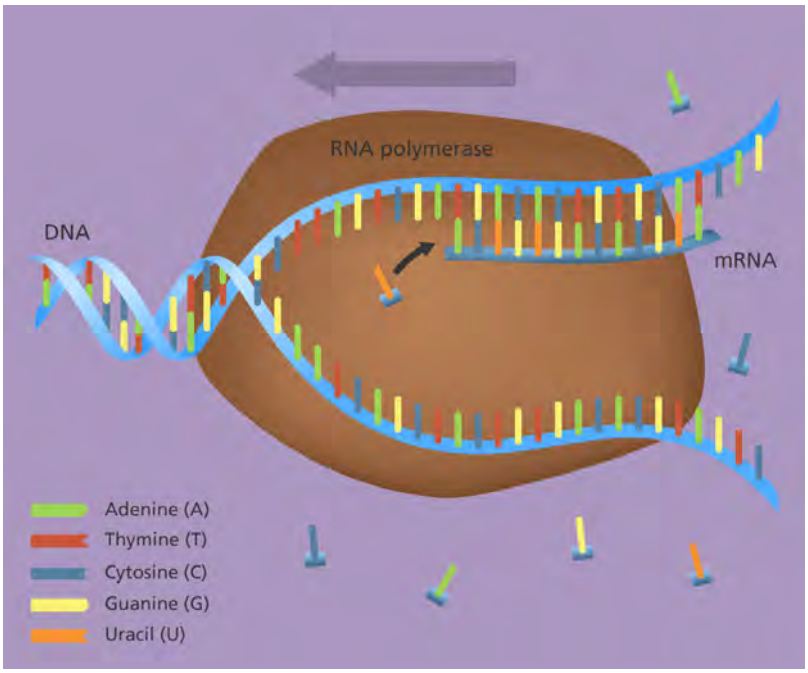


Figure 3.2: Illustration of Transcription

Translation is the process where mRNA is used to conduct and manage the protein synthesis, and it also directs the upcoming processing after translation of the molecules of protein. This happens after the messenger RNA (mRNA) has conveyed the message which was transcribed or passed from the DNA to ribosomes and ribosomes are the producer of proteins in the cells.The message or information carried by the mRNA is read by a carrier molecule which is called transfer RNA (tRNA). The mRNA is read according to the genetic code that relates the DNA sequence to the amino acid sequence in proteins [25]. In mRNA each group has three bases and every group contains a codon. Each codon determines a specific amino acid. To arrange the amino acid chains that form protein in a particular order the mRNA sequence is used as a template.

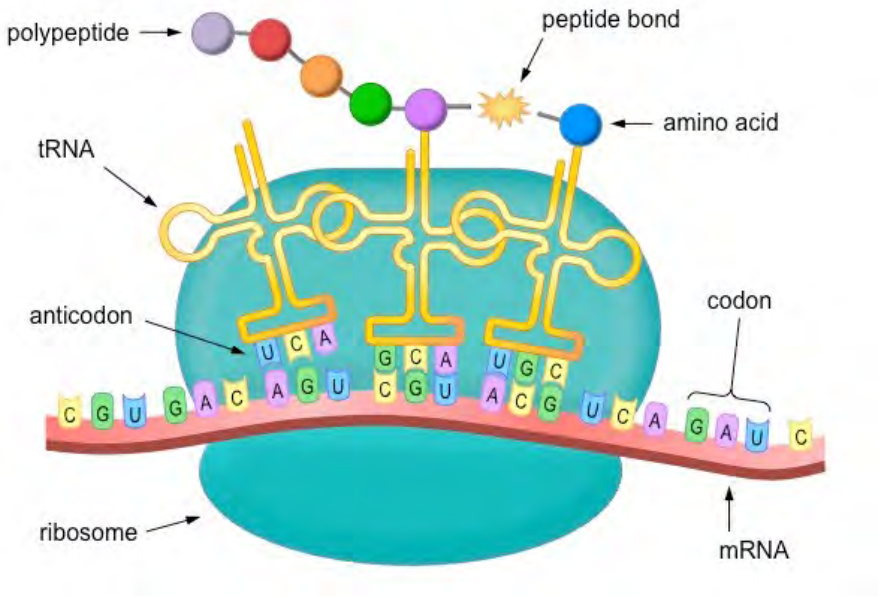


Figure 3.3: Illustration of Translation

Cancer occurs due to mutations or abnormal changes in the genes which contribute in regulation of the growth and development of the cells and are responsible of keeping them in good health. The nucleus of each cell contains genes that is in charge of all the processes happening in each cell [17]. Cell growth follows a well ordered and systematic process where the cells in our bodies replace themselves and old cells die and the new ones take their place. Over time, mutations can initiate any activity in certain genes and also can stop any activity in others. The ability of dividing without any control or order is gained by the changed cell which allows more production of cell resulting in a tumor. A tumor can be malignant which means it has the potential to be harmful for the body or it may be benign which means not harmful or hazardous to health. Tumors that are malignant are only considered to be cancerous as they invade the tissues closer to them and starts to spread to the other body areas.

3.2 Microarray and Cancers

Microarray data play a significant role in identifying and classifying cancer. For the study of genetic phenomena microarrays have been a vital tool for years. DNA microarray technology permits an investigator to rapidly determine which genes are expressed by a cell or tissue. Microarray experiments compare the pattern of gene expression in tumor tissue and normal tissue. When a gene is expressed it is transcribed into mRNA. The mRNAs from the tumor tissue and the normal tissue are isolated and converted into complementary strands of DNA called cDNAs by the enzyme reverse transcriptase. To distinguish between the two pools of cDNAs the molecules are fluorescently labeled red for the tumor derived pool and green for the cDNas direct from the normal tissue. After the RNA is removed DNA microarrays are used to compare the two cDNA samples. DNA microarrays can contain 60000

or more different DNA sequences attached in microscopic spots to a glass slide. The different DNA sequences are oligonucleotides of about 20 bases and lengths. The oligonucleotides represent tiny but unique regions of gene in the genome. The cDNA samples are mixed together and are added to the microarray. cDNAs that are complementary to oligonucleotides on the microarray will bind or hybridize with the DNA and thereby stick to that location on the side. The cDNAs that did not bind are washed away. A scanner detects patterns of hybridization by sensing the fluorescent signals. In the experiments a red spot indicates the expression of the gene is higher in the tumor tissue compared to normal tissue. In contrast a green spot indicates the expression of the gene is higher in normal tissue than in the tumor tissue. Yellow spot indicates that the gene is expressed equally in both tissues. If the gene is not expressed in either tissue the spot will not fluoresce. Each area in microarray contains a known DNA sequence corresponding to a known gene and that is why the identities of the hybridizing cDNAs can be determined. Using these data investigators can establish which genes are expressed differently in the cancerous tissue [50].

The recent advancement of the microarray technology permitted the analysis of enormous amount of genes that have contributed to the cancer advancement. Cancer is known as the result of altered expression of genes in our bodies and the main reason of the alterations of the overall activity of the cell is the turning on of protein which is called gene activation or turning off of protein which is called gene silencing. The genes which are not usually expressed in the cell can be turned on and expressed at higher levels which results in gene mutations [1]. The level determines the estimated amount of gene's RNA copies created in the cell. The value is correlated with the amount of corresponding proteins produced. Gene expression levels determine how many transcripts are generated per locus. Usually by intersecting the genome coordinates of uniquely marked reads with the loci of annotated genes and the transcript length from a given locus, gene expression levels are obtained from RNA sequence based data [28]. The RNA sequence gene expression levels are measured by Illumina HiSeq Platform. Microarray technology and Ilumina HiSeq Platform are commonly used to determine these levels. The level of gene expression do have essential leads that resolve crucial topics that are related to the disease prevention and treatment and drug discovery.

3.3 Illumina HiSeq Platform

The Illumina HiSeq is a deep sequencing tool or machine. It has completely revolutionised genome analysis. With this tool it is possible an entire human genome within just one day. The Illumina sequencing mainly works using the fluorophore detection. The process starts with the fragmentation of DNA testing sample(T) where the whole DNA splits into small parts. Then all the DNA fragments will get ligated or tied up to certain oligonucleotides. It includes two types adapters. Firstly the DNA double stranded fragment is detached as RNA. Now all the detached fragment RNA are now applied on a ship called flow cell containing two types of complementary adapters. The RNA fragments will at this ends hybridize

with the sequencing flow cell as applied to the plate. Here the RNA amplification happens by polymerase chain reaction(PCR). By PCR the complementary sequence to our single-stranded RNA fragment is synthesized. The process is done simultaneously for all RNA fragments on the whole flow cell. After the RNA fragments which are not attached to the flow cells oligonucleotides gets washed away. After that the bridge building procedure happens where the oligonucleotides that are not attached to the flow cells oligonucleotides get attached to the same flow cells oligonucleotides. This whole process repeats for several rounds. After that one of the strands amplified RNA strands is cleaved away for sequencing. Sequencing is now done with specific nucleotides. The nucleotides used during HiSeq are uniquely labeled as compared to 4-channel sequencing instrument:

- "C" nucleotides are labeled with a red fluorophore.
- "T" nucleotides are labeled with a green fluorophore.
- "A" nucleotides are labeled with the yellow color which is the mixture of red and green fluorophore.
- "G" nucleotides are not labeled at all.

These nucleotides are fluorescently labeled such that each nucleotide is attached sequentially with each other such as AT TA GC and CG. During sequencing all nucleotides are pushed into the flow cell and annealing is performed with only one at a time. After applying the annealing method the excess nucleotides are washed away and meanwhile the labeled fluorophore emissions from each cluster are imaged on the illumina hiSeq platform. Lastly for data analysis the informatics tools are used [21]. The whole sequencing generates millions and billions of reads and these reads are overlaid and compared to a reference genome(R) and stored as microarray data. This microarray data is expressed as 2D matrix where each column represents the gene expression and each row represents the expression of a gene across all experiments. Each element is a log ratio defined as $\log_2(T/R)$ where T is the testing sample gene expression level and R is the reference sample gene expression level. Following with the fluorophore process the flow the readable data is visualized with multiple colors green, red, black, grey [53].

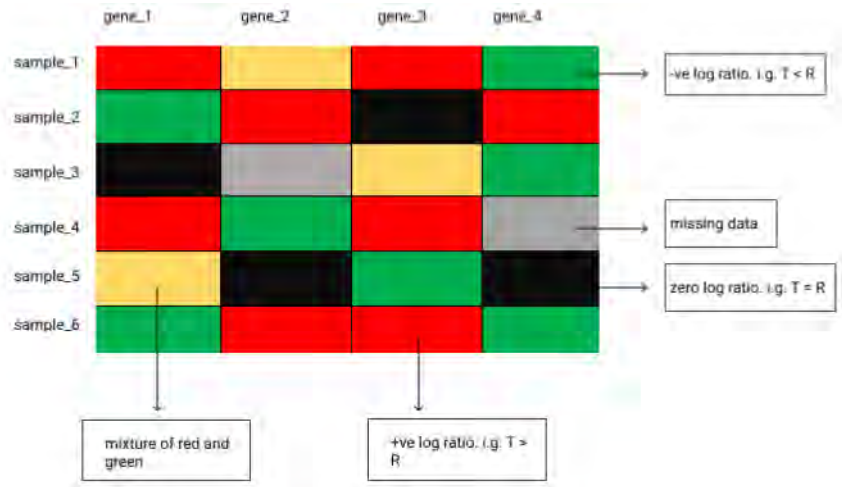


Figure 3.4: Flow Cells Data

Each colors represents the $\log_2$ of ratio.

- Black indicates a log ration of zero means T = R.
- Green indicates negative log ratio means T < R.
- Red indicates positive log ratio means T > R.
- Gray indicates missing data.

After that the log ratio values are stored in the formatted file. The higher value means the tested RNA is more likely different than the referenced RNA.

3.4 Breast Cancer

Breast cancer is one of the most common cancer type in the world especially among women. About 1 in 8 women are diagnosed with breast cancer during their lifetime and there is a huge possibilty of recovery if it is detected at the primary stage. Usually breast cancer begins in the cells of lobules which are the milk producing glands or the ducts. Some symptoms of this cancer are a breast lump or skin thickening that does not feel normal and has a different feel from the surrounding tissue, any change in the breast shape or size or look or changes of the skin in the area such as dimpling, or redness or peeling of skin. Most breast cancer cases which are inherited are found to be associated with mutations in two genes which are BRCA1 (breast cancer gene one) and BRCA2 (breast cancer gene two) [24]. Women with mutations in BRCA1 or BRCA2 genes also have higher possibilities of developing many other cancers such as ovarian, colon, and pancreatic [30].

3.5 Colon Cancer

Colon cancer is the cancer type that begins at the large intestine which is the mainly called the colon which is the end portion of our digestive systems. In the beginning stage of this cancer, polyps are usually found which are tiny noncancerous cell clumps that forms inside the colon and overtime these become tumors. Colorectal cancer and rectal cancers are very common divisions of colon cancer. Rectal cancer originates in the rectum, which is the last few inches of the large intestine that is nearest to the anus [42]. Colon adenocarcinoma (COAD) and rectum carcinoma make up 95 percent of all colorectal cancer cases. Some symptoms of colon cancer or COAD are bowel habit changes, bleeding in rectal, abdominal discomfort or weakness. Genes responsible of COAD are MSH2, MSH6 and MLH1. Mutations in MSH2, MSH6, and MLH1 proteins lead to damaged DNA which results in colon cancer.

3.6 Kidney Cancer

Among the many types of kidney cancer, kidney renal cell carcinoma (KIRC) is the most common one. Almost 9 out of 10 kidney cancer patient are diagnosed with renal cell carcinoma. It is a rapid growing cancer which easily spreads to the lungs and surrounding organs in the body. The malignant cancer cells starts to form in the tubules of the kidney and then spreads in the area nearby. Some causes of this are smoking, being overweight, inherited kidney diseases and hepatitis C. Symptoms may include a lump on side of the belly, blood in urine and weight loss [47]. Genes responsible for KIRC are VHL, PBRM1, BAP1 and SETD2.

3.7 Lung Cancer

Lung adenocarcinoma (LUAD) is a form of non-small cell lung cancer and is the most common type of lung cancer. Lung adenocarcinomas usually originates in the nearby tissues of the outer portion of the lungs and remains there for a long period of time before symptoms starts to appear. Finally when the symptoms starts to appear, the signs are mostly not obvious enough like the other forms of lung cancer, showing with a chronic cough and bloody sputum which are manifested in the more critical stages of the disease. As a result of these less significant signs, some of the more generalized and usual early symptoms which are fatigue, running out of breath, or pain in upper back and chest may be neglected or thought to be of any other causes [19]. Genes responsible for LUAD are KRAS, EGFR, ALK, ERBB2, and BRAF.

3.8 Prostate Cancer

Prostate cancer is the cancer type that occurs in the prostate which is a gland in men that produces the seminal fluid that nourishes and transports sperm. Adenocarcinoma is the name given to the type of cancer that arises in the gland cells. Most cells in the prostate gland are of the glandular type, which means that adenocarcinoma is the most common type of cancer to occur in the prostate. Some symptoms of prostate adenocarcinoma (PRAD) are facing some trouble urinating, decreased force in the stream of urine, blood found in semen, and feeling discomfort mainly in the pelvic area [40]. Genes responsible for this are BRCA1, BRCA2 and HOXB13.

Chapter 4

Proposed Model

According to our research objective we performed several steps for accomplishing better accuracy in classification of the five distinct cancers using our ensemble learning model. Each step is described elaborately in the following sections. A generic flowchart of the entire process is given in the Figure 4.1. In order for the dataset to feed into machine learning algorithms, every feature and label intended for machine learning should be numerical and scaled appropriately. Therefore, after getting the datasets we did some pre-processing of unwanted components from the data before applying the models. Using the label encoder the five distinct cancer gene names were converted into machine readable forms. The classifiers used for classification are Random forest (RF), Support Vector Machine (SVM), XGboost, K Nearest Neighbor(KNN). Furthermore, deep neural network is also used to train in a fully supervised manner using a labeled training set. For feature selection we used XGboost, LightGBM which calculated the feature score and feature importance respectively. Moreover, the effectiveness of each model is decided using the accuracy score and confusion matrix. Finally, our ensemble learning model was generated using the test dataset with its best classification accuracy score and confusion matrix.

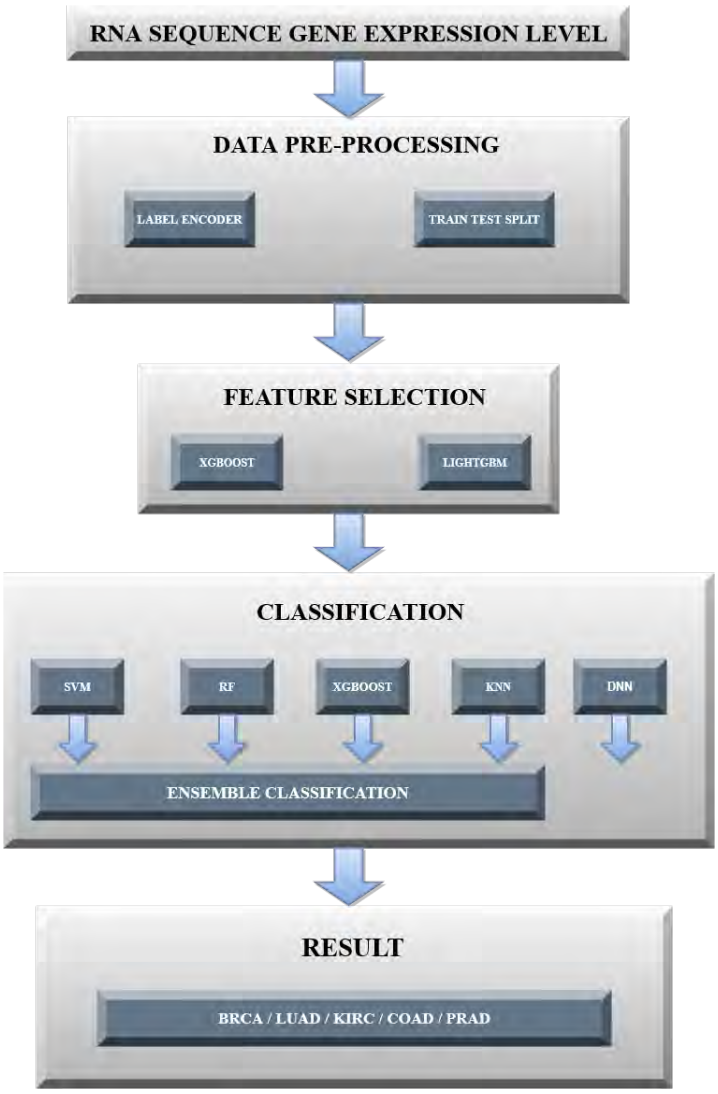


Figure 4.1: Workflow Diagram

4.1 Data Construction

The dataset is collected from (https://archive.ics.uci.edu/ml/datasets.php) UCI machine learning repository. This dataset itself contains two sub datasets named "data" and "label" and both are in csv format. This dataset is open to the public and can be found on the internet. It contains RNA sequence gene expression levels of 801 samples for five different cancers namely Kidney cancer (KIRC), Lung cancer (LUAD), Colon cancer (COAD), Breast cancer (BRCA), Prostate cancer (PRAD). For each individual sample there are more than sixteen thousand RNA sequence gene expression levels.Table (4.1) shows a snippet of our dataset "data" containing few samples in rows and its corresponding features in columns. Some of the columns having null values were removed inorder to avoid problems while training our models. The cancer names were converted into integers using the label encoding technique. For example:["BRCA", "COAD", "KIRC", "LUAD", "PRAD"] were converted into a string array and their index numbers were representing these strings instead.

	gene_1	gene_2	gene_3	gene_4	gene_5	gene_6
sample_0	2.015391	2.017209	3.265527	5.478487	10.432	5.619994
sample_1	2.466601	0.592732	1.588421	7.586157	9.623011	11.05521
sample_2	1.981122	3.511759	4.327199	6.881787	9.87073	8.210248
sample_3	2.874246	3.663618	4.507649	6.659068	10.19618	8.306317
sample_4	2.141204	2.655741	2.821547	6.539454	9.738265	10.14915
sample_5	2.516797	3.467853	3.581918	6.620243	9.706829	6.842765
sample_6	3.023841	1.224966	1.691177	6.572007	9.640511	7.4242
sample_7	2.405856	2.854853	1.750478	7.22672	9.758691	7.373431
sample_8	2.579977	3.992125	2.77273	6.546692	10.48825	10.14762
sample_9	2.296311	3.642494	4.423558	6.849511	9.464466	7.85678
sample_10	3.381962	3.492071	3.553373	7.151707	10.25345	7.965224
sample_11	3.153092	2.941181	2.663276	6.56169	9.376293	9.796182

Table 4.1: Portion of our Dataset

For visualization of the dataset t-distributed stochastic neighbor embedding (t-SNE) is used which produces five different clusters representing each distinct cancer. Figure 5.1 (Left) shows the t-SNE scatter plot visualization. Moreover, we also applied the K-Means clustering algorithm previewed in Figure 5.1 (Right) to see if the clusters made using the K-Means algorithm is same as the t-SNE showed. Subsequently, for training the models we divided the dataset into train and test making a ratio of $8 : 2$.

4.2 Feature Selection

Feature selection is a must when handling comparatively large or high dimensional data, as data that has numerous columns are harder to evaluate. Therefore, it is a necessity to remove or simply ignore non-informative or redundant features that do not contribute to the decision making process. Considering the data we are using to predict our classification model, consists over 801 samples each being featured in over 16,383 gene expression levels, clearly indicating the need for feature selection and importance. We have used two methods for feature selection and to plot the feature score and importance. Therefore these two methods plotted exactly 20 features showing their feature importance and feature score. Among these 20 features, 11 features were found same in both feature importance and feature score. These 11 features were then passed into each of our classification models after splitting the dataset in train and test ratio. Later, Using the selected features our ensemble learning model was generated and the classification of the five cancers were completed. Finally, our model was able to predict the test samples accurately classifying each cancer to its actual class using the 11 features. The two methods used for feature selection are described in the following part of the report.

XGBoosting

The algorithm takes our trained dataset and constructs boosted trees that are formed level wise or horizontally. By invoking the transform() method we are using a threshold to decide what features to select and then evaluating the model on the test dataset reassuring feature importance and what features to select[44]. Finally we calculated 20 features based on their importance and scores to create a subset of 20 important features, where the feature with the highest feature score sits on the top.

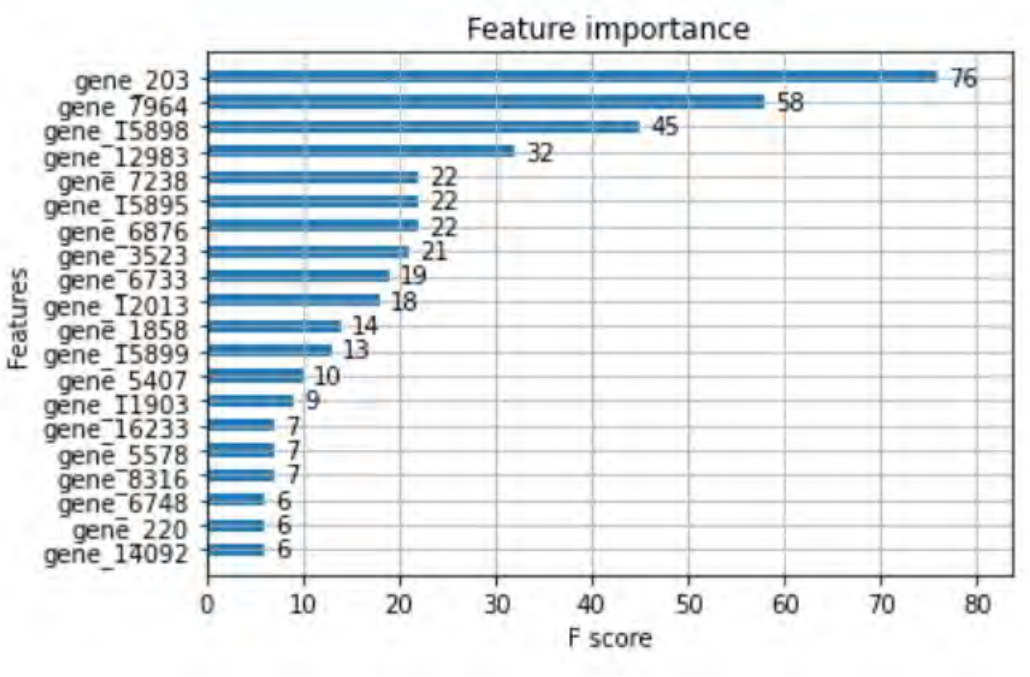


Figure 4.2: F-Score using XGBoost

LightGBM

LightGBM, similar to XGBoost is a tree based learning algorithm but differs from XGBoosting in their way of construction of the boosted trees and processing time. XGBoost constructs it's trees level wise whereas LightGBM constructs it's boosted trees leaf wise, and as the max depth or the maximum number of selected features are being defined LightGBM and can evaluated the test dataset and plot important features faster than XGBoosting. Moreover, LightGBM ensures lower memory usage even for large scale microarray data while maintaining high efficiency.

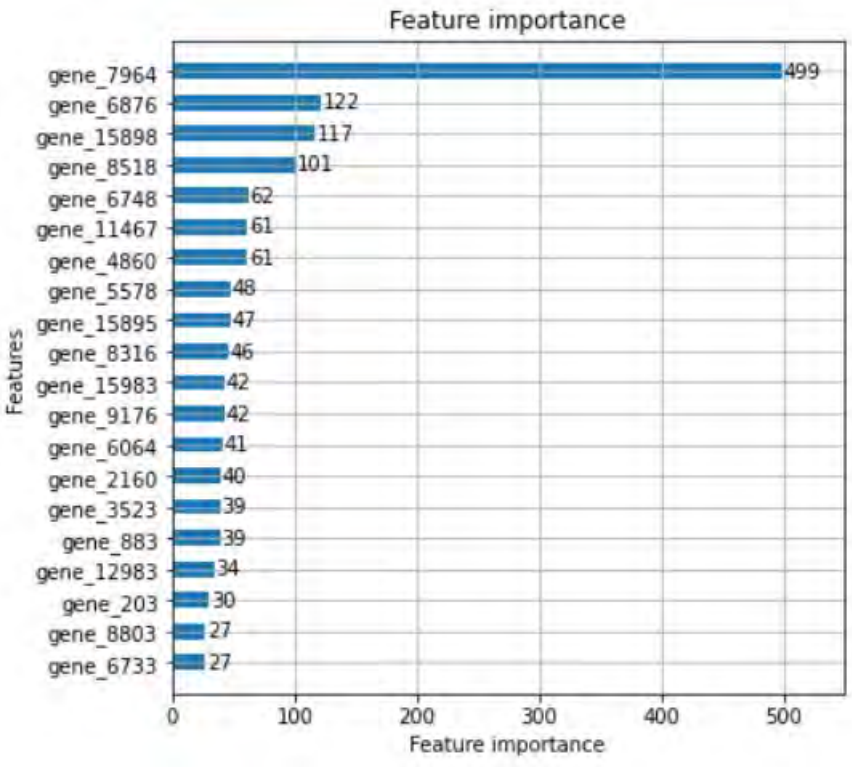


Figure 4.3: Feature Importance using LightGBM

4.3 Model Training

4.3.1 Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm which categorizes or data by drawing a line of separation between the two sets of data. According to the Nature of Statistical Theory, SVM is one of the popular classification methods [18]. In our research SVM is used as a classifier because of its good learning capability, strong generalization ability and most importantly its powerful mathematical background which plays a vital role in various application areas, for instance, computational biology, image segmentation, cancer classification [18], [26]. The SVM algorithm is well known for its effectiveness in high dimensional spaces, says John Shawe-Taylorin in his book "An Introduction to Support Vector Machines and other Kernel-based Learning Methods" [4]. While dealing with high dimensional data the SVM sets hyper-planes in between the data points and divides them into their respective classes. The separation is done by calculating the minimum distance of the data points from the hyper-planes. Figure 4.3 shows SVM classification on multiclass data. Moreover, SVM also uses the kernel trick function that allows pattern analysis by mapping inputs in higher dimensional space to model non-linear decision boundaries. SVM kernels include linear, non-linear, polynomial, radial basis function (RBF), sigmoid. These kernel functions need to be selected depending on the number of observations and the type of problem we are solving. For our research

we focused on polynomial kernel and RBF kernel. The kernel function returns the inner product between two points in suitable feature space.

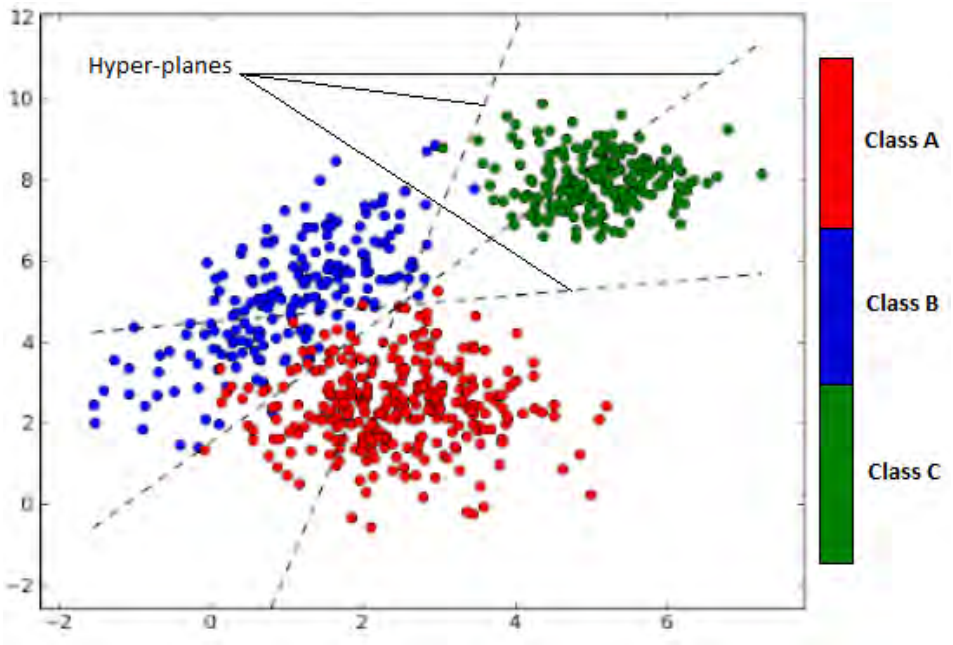


Figure 4.4: SVM Multiclass Classification

Polynomial Kernel

In general, the polynomial kernel is defined as:

$$K\left(X_1, X_2\right) = \left(a + X_1^T X_2\right)^b$$
$$b = \text{ degree of kernel } \& \ a = \text{ constant term.} \tag{4.1}$$

The polynomial kernel only calculates the dot product by increasing the power of the kernel[18]. For example, Let's say originally X space is 2-dimensional such that

$$Xa = (a1, a2),\ Xb = (b1, b2) \tag{4.2}$$

Now, to map our data into higher dimension let's say in Z space which is six-dimensional it may seem like

$$Z_a = \phi\left(X_a\right) = \left(1, a_1, a_2, a_1^2, a_2^2, a_1 * a_2\right)$$
$$Z_b = \phi\left(X_b\right) = \left(1, b_1, b_2, b_1^2, b_2^2, b_1 * b_2\right) \tag{4.3}$$

Therefore, to solve this dual SVM we would require the dot product of (transpose) Z_a^T and Z_b. Solving this by traditional method would be time consuming. So, using the kernel trick method we can simply calculate the dot product by increasing the value of power.

$$Z_a^T Z_b = k\left(X_a, X_b\right) = \left(1 + X_a^T X_b\right)^2$$
$$Z_a^T Z_b = 1 + a_1 b_1 + a_2 b_2 + a_1^2 b_1^2 + a_2^2 b_2^2 + a_1 b_1 a_2 b_2 \tag{4.4}$$

In our case, after splitting of the data we have tuned the hyper-parameters of our model for effective classification. Setting the kernel type to polynomial, the regularization to 0.1 and gamma value to 0.0015 which is the Kernel coefficient for

polynomial kernel in the parameter of our non-linear SVM, we classified the five cancers in five classes. Our train dataset consisted of 640 samples and 16,383 features and it took 9.07 seconds to train our model of SVM as a classifier. Using this SVM classifier we got an accuracy score of 1.0. This means the classifier had successfully classified 100 percent of our five different cancers into five different classes. To test the model efficiency we considered the four criteria such as computational time, memory consumption, debug easability and accuracy . But, this accuracy score was not expected and we were doubtful as we had used all the 16,383 features to train the model. Therefore, we again trained our SVM classification model using those 11 common features that we found same in both feature score and feature importance. Considering those criteria for making our model as efficient as possible we Kept the hyper-parameters same as before and finally we managed to get an accuracy score of 98.76%. This accuracy score shows that our svm model did some misclassifications while trying to classify the 5 different cancers to their respective classes.

4.3.2 Random Forest

Random forest is an ensemble learning method for tasks such as classification and regression. It consists of multiple random decision trees and each of these trees is built using a bootstrap sample of data [10]. At each sample split, the candidate set of variables is a random subset of the variables. Thus, a random forest algorithm uses both the bagging aggregation technique and random variable selection for the tree building [3], [10]. The result of the random forest depends on the majority of the trees voting a particular decision. Figure 4.4 shows an example of trees predicting their results and a voting takes place [49]. The majority of the votes by the trees give the final result of the random forest. Random forest is also one of the most popular classification methods and has excellent performance in this task. A random forest classification algorithm is an ideal choice for its several characteristics which are they can be used for multi-class problems of two or more classes, has good predictive performance, does not overfit and are used to achieve excellent performance there is little need in tuning the hyper-parameters.

Bootstrap Aggregation (Bagging)

The bootstrap is a general tool for assessing statistical accuracy[10]. This creates multiple different models from a single training dataset. Bootstrap Aggregation is a procedure that can be used to reduce the variance for algorithms like decision trees or classification and regression trees (CART) that have high variance[10]. Decision trees are sensitive to their training data. If the training data is changed the predictions of that resulting decision tree can be different. The bagging technique creates random sub-samples with replacement of the training dataset[10][3]. Let's assume we have a dataset of 1000 instances and we are using classification and regression trees (CART) algorithm. Bagging of this algorithm will create many random subsamples of that dataset. When bagging with decision trees, the trees are allowed to grow deep and are not pruned. As a result, these trees will have high variance and low bias.

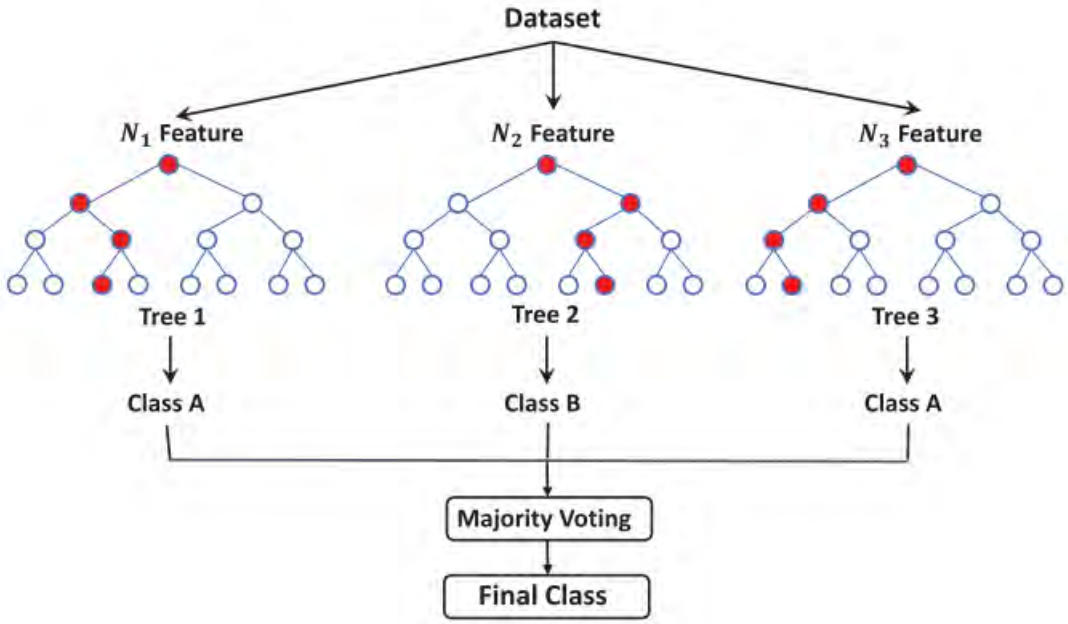


Figure 4.5: RF Model with Majority Voting

In our research, after data pre-processing the hyper-parameters are tuned with necessary parameters for training the random forest classifier. Passing 100 in n estimators as parameters tells the algorithm to make hundred trees for the forest without using the bootstrap technique. As our dataset contains five classes of cancer therefore, setting criterion to entropy for the information gain instead of gini which means gini impurity. Both of these criterions are measures of impurity of a node. After training the model using the default parameters and the tuned hyper-parameters, each tree inside the random forest gives their predictions and a decision is made based on the majority of the decisions by all the 100 trees. An accuracy score of 1.0 is achieved which means 100% classification. To further improve our model, we trained our model for the second time. Using the 11 features found common among the 20 features while in feature selection phase, the model was trained getting an accuracy score of 0.99378. Thus, we were convinced that this algorithm had successfully classified 99.38% percent of the cancers and there were few misclassifications in our effective model. For testing the efficiency of each of our models, both of them were anaylsed based on the computational time, memory consumption and debugging easibility. To sum up, our new model was efficient enough to consider for the ensemble learning model for cancer classification.

4.3.3 XGBoost

In the quotes of Tianqi Chen, the inventor of XGBoost algorithm himself, "The name XGBoost, though, actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms[45]. This is the reason why many people use XGBoost". Extreme Gradient Boosting is a decision-tree based ensemble Machine Learning algorithm that uses a gradient boosting framework[43].

The idea is to resolve the problems of the previous model and create an improved model that minimizes the loss function until no more improvements can be found. This in terms enables us to create an error free and efficient model for evaluation.

XGBoost performs well because:

i. Regularization: getting rid of over-fitting of the model.
ii. Enabled with inbuilt Cross Validation function.
iii. Can handle missing values.
iv. Easily importing functions that can be used to evaluate the performance of the model like: accuracy score, classification report, confusion matrix and metrics.
v. Computation speed: considering the size of our data, it was essential that we needed a classifier to perform fast with high accuracy and XGBoost fulfilled both conditions.

In a XGBoost algorithm, gradient boosting is used for optimizing the trees [5]. Let's assume, the output of a tree be,

$$f(x) = w_q\left(x_i\right) \tag{4.5}$$

Here, x is the input vector and w_q is the score of the corresponding leaf q. Therefore, the output of an ensemble of K trees will be,

$$y_i = \sum_{k=1}^{K} f_k\left(x_i\right) \tag{4.6}$$

Minimizing the following objective function J at step t in XGBoost Algorithm:

$$J(t) = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{t-1} + f_t\left(x_i\right)\right) + \sum_{i=1}^{t} \Omega\left(f_i\right) \tag{4.7}$$

The train loss function L between real class y and output $\hat{y}$ for the n samples in the first term. The second term contains the regularization term, which controls the complexity of the model and helps to avoid over-fitting of the algorithm. The complexity of XGBoost algorithm is defined as,

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4.8}$$

Here, T denotes the number of leaves, y is the pseudo-regularization hyper-parameter and λ is the L2 norm for leaf weights. Now the optimal value of objective function is defined using gradients for second order approximation of the loss function and finding the optimal weights w

$$J(t) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda} + \gamma T \tag{4.9}$$

Here, g_i and h_i are the gradient statistics on the loss function, and I is the set of leaves.

XGBoost parameters can be of three categories: one that takes care of the overall functioning, the second parameter powers the booster parameters and most importantly the third parameter which validates the learning process of the booster[45][46].

• General Parameters:

i. Booster: to introduce tree based model we set the booster parameter to gbtree for efficient classification.

ii. Silent: to help understand the model better and receive run time messages we kept the silent parameter to zero [silent=0].

iii. nthread: This is used for parallel processing and number of cores in the system should be entered, as we wish to run all the cores and let the algorithm detect the numbers automatically, nthread was kept none [nthread=None].

• Booster Parameters:

i. Eta: comparable in certain respects to learning rate in Gradient Boosting Machine, but kept default of 0.3

ii. min child weight: corresponds to minimum sum of weights of all the observations required in a child and effectively preventing over fitting. This parameter was also kept default of one as greater values can lead to under fitting [min child weight=1].

iii. max depth: The default value of maximum number of terminal nodes or leaves in a tree was unchanged, which in terms creates a maximum of 2 to the power 3 leaves.

iv. gamma: specifies the minimum loss reduction required to make a split, this too was kept default [gamma=0].

v. Colsample bytree: indicates the fraction of columns to be randomly samples for each of the trees. Usually ranges from 0.5 to 1, but as we are passing the whole feature column, colsample bytree was updated to 1 learning rate=0.1

• Learning parameters:

i. objective: the objective parameter defines the loss function to be minimized during classification. As our research is to classify a specific sample in one of the five cancer classes, we set the object parameter to multi:softprob. Similar to softmax, it returns a predicted class but also in addition returns the predicted probability of each data point belonging to each of the cancer classes. [objective=' multi:softprob']

ii. Seed was unchanged and kept default [seed=0]. In order to support our classifier to work effectively we preprocessed our data, removed null values and converting them into numpy arrays, encoding them, splitting the data into 80% train and 20% test sets, training our data and applying our XGBoost classifier to the test set.

Although the learning rate being 0.1 reduces the computation speed but minimizes the number of misclassification, imposing better accuracy. To, assess our multi class cancer classification model we printed the accuracy score and confusion matrix, stating 3 out of 161 samples misclassified and the accuracy to be 98.13%.

Our XGBoost model was trained again using the 11 features that was selected in the feature selection process to examine the model based on the criteria for efficiency mentioned in the model training of SVM and Random Forest models. This time using the newly reduced number of features, our XGBoost classifier gave an accuracy score of 0.99378. This clearly shows improvement in the classification process of the cancers. Thus, we came to a conclusion that using the 11 features our classifer classifed 99.38% of the cancers by consuming less memory and taking less time to compute to their actual class.

4.3.4 K-Nearest Neighbor (KNN)

Perhaps the most simple, versatile and easy to use algorithms used in our classification so far, but is fast and efficient in making healthcare predictions and predictions where there are no assumptions for underlying data distribution, thus the name 'non-parametric algorithm'. It has also been described as a 'lazy algorithm' as it does not consider any training data points in model generation, but all the trained data are used while testing, making the training phase faster, despite slightly reducing the pace while testing, all while maintaining great classification accuracy.

For using KNN as a classifier, the parameters of the algorithm were tuned [12], [14]. The following parameters were tuned for implementing this algorithm:

i. **Metric:** considering Manhattan performs better for higher dimensionality data than Minkowski, hamming and Euclidean, we have used 'Manhattan' for generating the best results. The distance equation stands,

$$\text{Manhattan: } \sum_{i=1}^{k} |x_i - y_i| \tag{4.10}$$

ii. **Leaf_size:** was kept default, as changing the leaf_size did not affect the accuracy or the confusion matrix (default=30)

iii. **n_neighbors:** was kept default, similar to the leaf size case, changing the number of neighbours did not affect the prediction or validation, while bearing in mind that the ideal n neighbors is in the range of 3 to 10 in most cases, therefore the n_neighbors was unchanged (default=5)

iv. **p:** the power parameter for Minkowski metric was set to 1 (p=1) as we are using Manhattan distance, the equation states,

$$\text{Minkowski: } \left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{1/q} \tag{4.11}$$

When the value of p=1, the equation is equivalent to the previous Manhattan equation. When the value of p=2, the equation is equivalent to the Euclidean distance equation stated below,

$$\text{Euclidean:} \quad \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{4.12}$$

The weights were also kept uniform. Denoting that, the points in each neighborhood are weighted equally to predict possible values for classification. The K in the algorithm stands for, the number of nearest neighbors, which in terms, is the core deciding factor for the classification process. Suppose 'p' is a point belonging to the kidney cancer class, for which all the labels need to predict. Firstly, we find the k closest point to 'p' and then classify the point 'p' by majority vote of its k neighbors that may be labeled or belong to a certain cancer class. Each object votes for their cancer class and the class with the most votes is taken as the prediction, assigning the point 'p' to that predicted cancer class [36].

The algorithm runs on 3 basics steps, initial step includes fixing a new data point to predict its class and calculating distance to all the closest similar neighboring points in the data. To find the closest similar neighbors, we used the Manhattan distance measuring formula to perform well for high dimension data like ours. In the next step, we assign five nearest neighbors (k = 5) that are fit to vote to predict the cancer class of the data point. Finally, when all of the five neighbors vote for their designated cancer classes, the data point is predicted as the cancer class with the majority of votes voted for it [36].

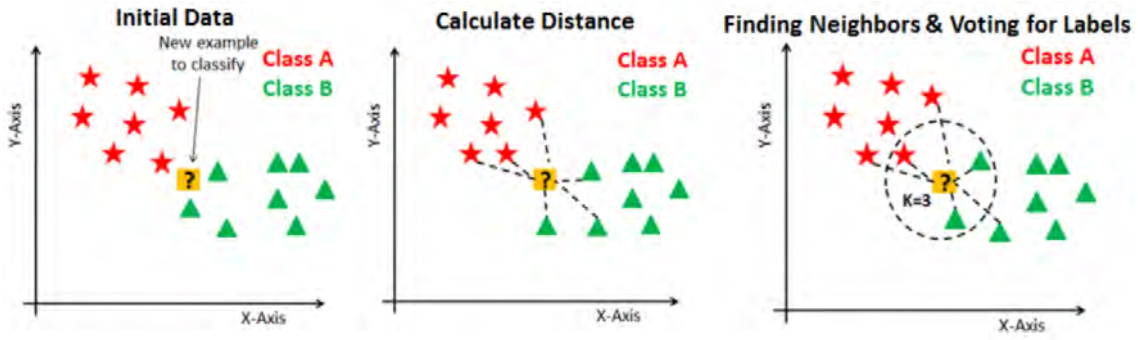


Figure 4.6: KNN Algorithm Steps

Using the 16,383 features our KNN classification model gave an accuracy score of 99.37%. For better classification we trained this model again with the feature reduced to 11 which we found in feature selection process.After training again with these 11 features, we got 100% accuracy score in our new model and this model is fulfilling our criteria to examine the model based on reduced training time, less memory usage and debugging usability . Therefore, this improvement in the classification process of five distinct cancers is preferred.

4.3.5 Deep Learning

When it comes to training a model rigorously to make the prediction state more reliable, one of the option that emerges is deep learning. Deep learning models are constructed using neural networks that take in training data as inputs, processes these data in hidden layers while assigning weights that are adjusted and improved automatically to predict the outcome of test data points[33]. The process can be compared to how neurons in human brain work together to make a decision in certain scenarios[37]. In our research we have applied the sequential model, a deep learning API to classify our test samples, imported from the keras library. After initializing the instance of the sequential class, 4 model layers were created with specified input dimension, activation functions and dense of the layers and added to the model instance[44]. Specifying the dense class and activation function helps us fix the number of nodes in each fully connected hidden layer and what data to pass on from one node in a hidden layer to all the nodes in the next hidden layer. The deep learning diagram portrays a better picture to understand how our system works.

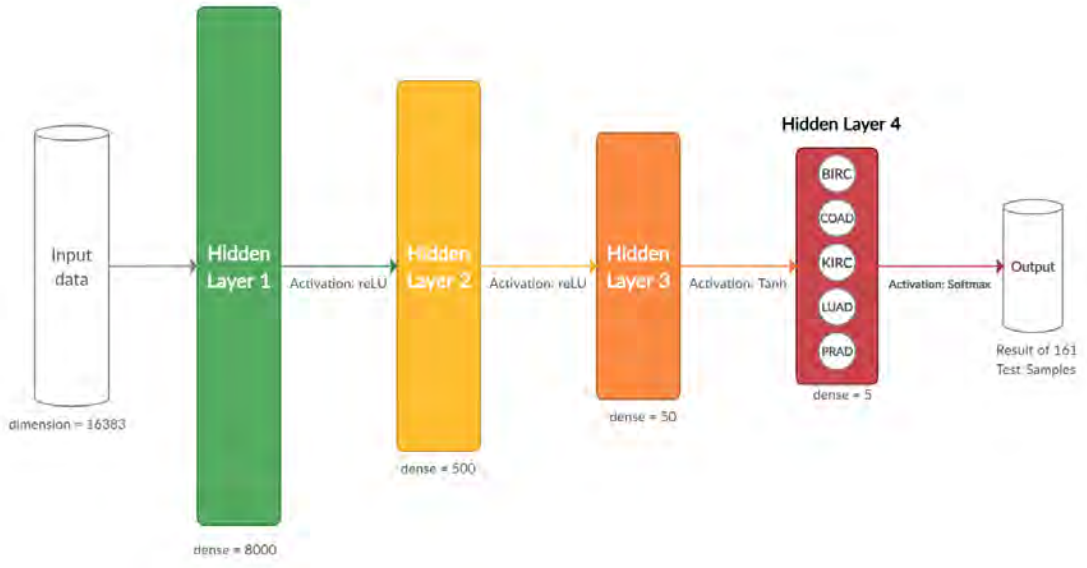


Figure 4.7: Deep Learning Sequential Model

To train our neural network we have used the entire feature column as input dimension (input_dim=16383) so that we can train our model properly. This does increase computational time of the training phase but as each column represents a gene used to test each sample which gives a unique expression level, it is always better to included all the genes having unique expression levels in our training phase. The first hidden layer has 8000 nodes (dense=8000) as it is always ideal to have at least half of the number of input dimension, followed by Rectified Linear Unit (ReLU) as the activation function. The second layer has 500 nodes with the same activation function as the first hidden layer. The third and the fourth layers have 50 and 5 nodes with activation functions Hyperbolic Tangent function (tanh) and softmax respectively. As our prediction model focuses on multiclass classification we have applied the softmax activation function in the last hidden layer to predict

our test sample in one of the five cancer classes. The training data were normalized using the MinMaxScaler() method imported from the sklearn preprocessing library, before fitting the model for the training phase to start.

Certain parameters were tuned before training our model. The loss function was updated to categorical crossentropy as we are focusing on multi class classification and the other loss functions binary crossentropy and mean squared error focuses on binary classification and regression problems respectively [51].

We selected the optimizer to be Adadelta which is a stochastic gradient descent method based on adaptive learning rate optimization, addressing drawbacks like, continuous decay of learning rate during model training and the need of manual selection of global learning rate. Adadelta is more robust than Adagrad and it adapts learning rates based on updates in moving windows while model training.

Some parameters handled while using the Adadelta optimizer from tensorflow keras library was, the learning rate was manually inputted as 0.001, the rho parameter or a floating point value indicating the decay rate was set to be 0.95 and the epsilon parameter to improve the conditioning of gradient updates was set to 1e-07.

Lastly the epoch or in simple terms the number of iterative improvement steps of the model was set to be 500. The model trained well in these 500 steps allowing the learning algorithms to minimize its error (loss) and boasting a training accuracy of 100 percent [41].

4.3.6 Ensemble Learning

According to Robi Polikar, the ensemble learning process consists of multiple models, such as classifiers or experts, which are strategically generated and combined to solve a particular computational intelligence problem [16]. The sole purpose of ensemble learning is to improve the prediction and classification performance of the model and to refer each test sample to its correctly classified cancer class, reducing the probability of sample being predicted in the wrong cancer class. Our Ensemble learning model is a multi-classifier system possessing a combination of diverse classifiers making predictions for each data point. These predictions are considered as votes being carried on to the final voting classifier that takes the majority vote into account and classifies that data point to its correct cancer class [48].

Previously evaluated 5 types of the classification algorithms, the Support Vector Machine, Random Forest, Extreme Gradient Boosting, K-Nearest Neighbor and Deep Learning separately and constructed confusion matrices to understand how they perform and how accurately they can predict data points to corresponding cancer classes. The results of each of the classifiers were satisfactory and almost precise, but to properly diagnose a patient which later may prove to save someone's life in latter stages, suggested that there were room for improvements. We strongly believe that the inclusion of ensemble learning can eliminate the risks of wrong diagnosis, as, instead of one classifier, we are using four classifiers simultaneously that votes for a sample test data point to a cancer class and majority of the votes

is considered as the final prediction.

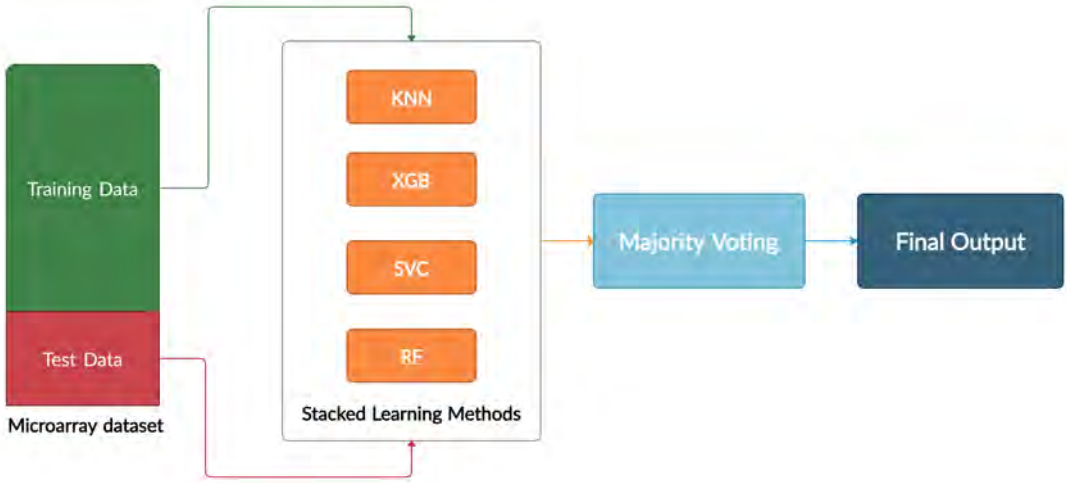


Figure 4.8: Ensemble Learning Model

The ensemble learning process can be divided into 3 steps:

1. **Model Creation:** the process started by importing the voting classifier from the sklearn ensemble learning library and stacking 4 classifiers, KNN, XGB, SVC and RF into it[14]. Each of these classifiers is individual learners and voters, whereas the Voting Classifier performs as a whole to predict the cancer class of one data point.

2. **Feeding data:** in this step, the training data is fed to the voting classifier so that the classifier and all its stacked classifiers can learn and provide effective result in the final step.

3. **Generating results:** after the model is well trained, the test data is passed, each of these classifiers make a prediction or vote for a certain data point in favor of a cancer class each embedded classifier thinks right and the data point is assigned to the majority voted class. For example, if knn predicts a certain data point in favor of class 4 (PRAD), xgb votes for class 0 (BRCA) with SVC and RF both voting for class 2(KIRC), the KIRC class gets an advantage of having majority of the votes, therefore the voting classifier predicts the data point to be of class 2(KIRC).

We also trained our model with a total of 11 selected features with the highest feature scores and importance to evaluate and compare the efficiency (time, space, performance) amongst the two, to propose a best model that can predict test samples.

Chapter 5

Result and Analysis

The data to predict our classification model, needed to be preprocessed and split before introducing algorithms. To visualize our classification results we used t-distributed stochastic neighbor embedding or t-SNE to reduce the dimension of the data used and form a cluster separating the types of cancer classes. We also used k-means clustering to validate our visualization and calculating the values of the centroids of the five clusters, gives us an idea of the span, radius and the mean of each of the clusters of cancer types.

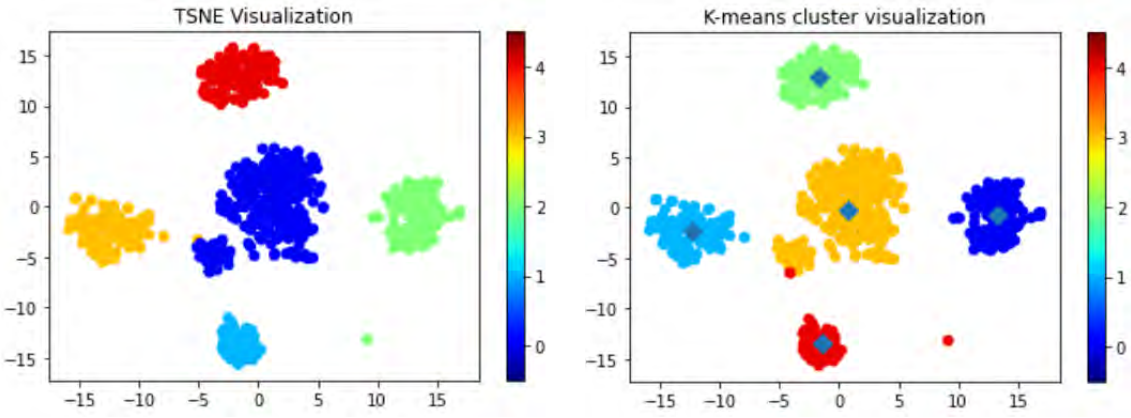


Figure 5.1: (Left): t-SNE Visualization; (Right): K-means Cluster Visualization.

The cluster colors dark blue, light blue, green, yellow and red indicate breast, colon, kidney, lungs and prostate cancer classes respectively. On the other hand unsupervised k-means clustering generates similar cluster patterns with the exception of the highlighted means of each the clusters validating our classification.

In this section we will be evaluating the effectiveness of each model we have used to reach our goal in the research. After training each of our models with the train dataset, we calculated the accuracy score and plotted a confusion matrix using them as a benchmark for our evaluation based on the test data. A confusion matrix is a table that is used to describe the performance of a classification model on the test data for which the predicted values are known. The confusion matrix shows in which way the classification model is confused when it makes predictions. Only calculating

the classification accuracy score can mislead us if the number of observations in each class is unequal or if there are more than two classes in the dataset, according to Jason Brownlee [44]. Therefore, confusion matrix is also accompanied with it. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

5.1 Support Vector Machine(SVM)

After the training of the SVM classifier using polynomial kernel we attempted to evaluate this model. We constructed a confusion matrix using the sklearn library. For cases like ours having five classes i.e. five cancers, the matrix looks like Figure 5.2 (Left) Each row of the matrix corresponds to an actual class and each column of the matrix corresponds to a predicted class. The confusion matrix of our model SVM previews no misclassified values using 16,383 genes or features. But using only 11 features we are getting an accuracy score 98.76%, which is almost same as the previous model but with decreased computational time, reduced memory consumption and enabling debugging easability .The confusion matrix of our new SVM classifier model is shown on Figure 5.2 (Right). Because of the new model trained with 11 important common features, our training time for the SVM classifier decreases. Moreover, less space is required to store the features if we consider 11 genes over 16,383 genes or features. Thus, the accuracy score given by the new SVM classifier is 98.76% correct and we can conclude to the fact that the two misclassifications shown in the confusion matrix are very minor and our ensemble learning model can handle it perfectly.

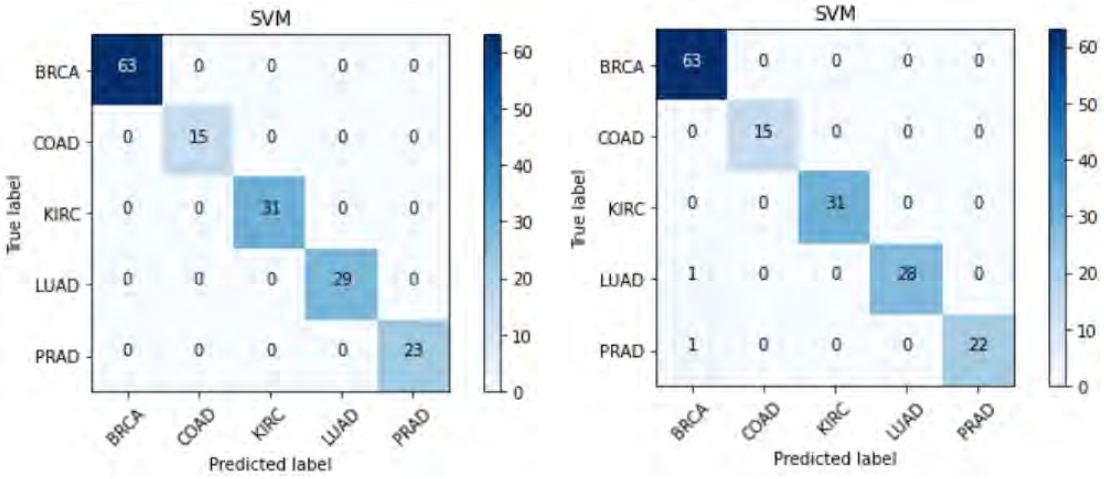


Figure 5.2: (Left): SVM Confusion Matrix Using 16383 Features; (Right): SVM Confusion Matrix Using 11 Features

5.2 Random Forest

The model random forest classifier was evaluated for any misclassifications after the training of the model. The accuracy score alone cannot justify the classification of the cancers using this algorithm. This algorithm may misclassify some of the cancers therefore, a confusion matrix from the sklearn library was generated using the test dataset with 16,383 genes or features. The confusion matrix previewed in Figure 5.3 (Left) tells us about any misclassifications if happened. Misclassification would have occurred if the predicted labels did not match with its correlating true labels. And as the sum of the values are equal to the sum of the trained samples i.e. 161 samples. In our new random forest classification model that we trained with the 11 important genes or features, we got an accuracy score of 99.38%. This new model shows one minor misclassification in the confusion matrix previewed in Figure 5.3 (Right). The new model shows improvement in the duration taken for training our model with 11 genes over 16,383 genes or features. Moreover, space needed for 16,383 features is optimised using only the 11 important features collected from the feature selection phase. Thus, we can say the new model is efficient and close to accurate which leads us to our final statement that our new model of random forest classifier has classified 99.38% of the five cancers in the actual five classes and only one misclassification occurred while classifying the actual colon cancer (COAD). This was misclassified as Lung Cancer (LUAD). And therefore, the new model with 11 features is further used in our ensemble learning model.

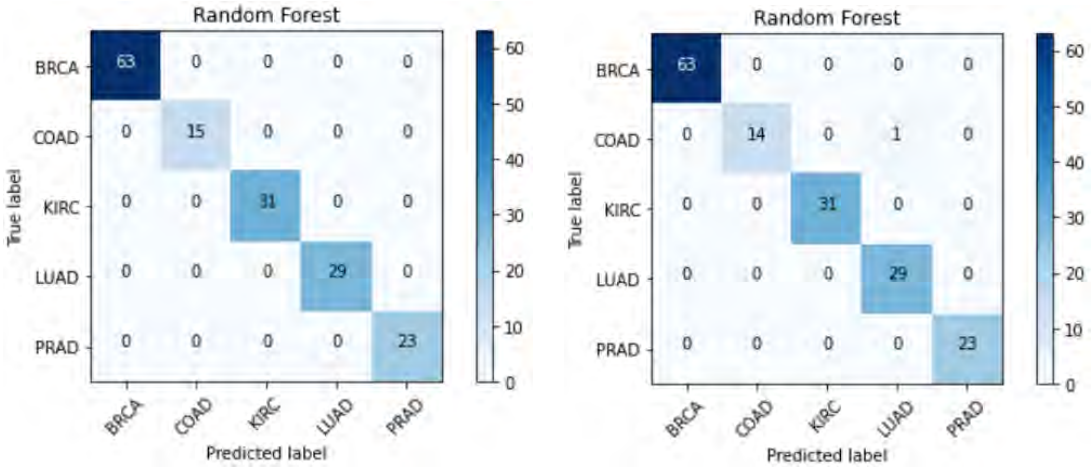


Figure 5.3: (Left): RF Confusion Matrix Using 16383 Features; (Right): RF Confusion Matrix Using 16383 Features

5.3 Extreme Gradient Boosting (XGBoost)

The evaluation process of our extreme gradient boosting (XGBoost) classification algorithms consists of 3 phases. The first is to feed the test data to our already trained model to generate predicted labels for corresponding samples. The next phase uses previously predicted labels and compares them with corresponding true labels of the test data, forming a confusion matrix of the predicted cancer class versus the actual cancer class. Lastly, to evaluate how the classifier performs and how much misclassification it possesses, we printed the accuracy score of the model using the same predicted and true labels. The following Figure 5.4 (Left) shows our XGboost model classifying the test data of 161 samples and 16,383 features/genes.

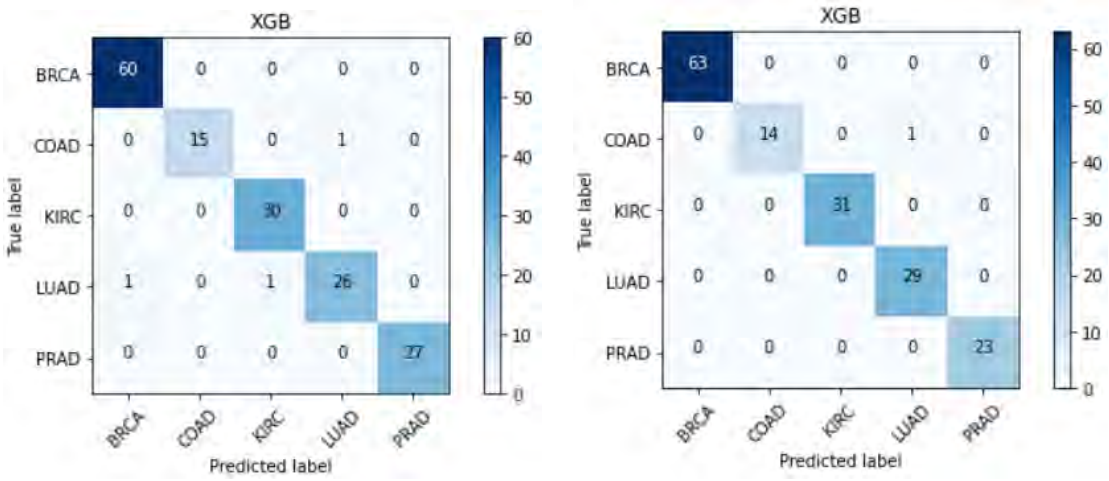


Figure 5.4: (Left): XGBoost Confusion Matrix Using 16383 Features; (Right): XGBoost Confusion Matrix Using 11 Features

Despite the high accuracy of 98.13% using all 16,831 features, the extreme gradient boosting classification algorithm does misclassify 3 sample data. One of them predicted as member of the lung cancer class but actually belongs to the colon cancer class and the rest were predicted as kidney and breast cancers but in reality belonged to the lung cancer class. Given the computational time and the accuracy of the classifier, a misclassification of just 3 samples out of 161 samples, is a healthy tradeoff. To further improve our accuracy score, training time and space complexity we used the 11 important features or genes to again train the XGBoost model. This time the new model gave an accuracy score of 99.38%. After Ploting the confusion matrix previewed in Figure 5.4 (Right), we could easily see the new model has performed better using only 11 features than the previous model with 16,383 features. This new model shows only one misclassification of Colon cancer. To sum up, our new model is efficient in every category i.e accuracy score, time and space complexity by using only 11 important features found at the feature selection phase.

5.4 K-Nearest Neighbors (KNN)

Similar to the previously mentioned extreme gradient boosting algorithm, the K-nearest neighbor algorithm's evaluation consists of the same 3 phases. Where, the first phase involves feeding the unlabeled test data to the KNeighbors Classifier to obtain predicted labels for the corresponding data, which then can be used in forming a confusion matrix with the help of the predicted and true labels. Finally, we presented the accuracy score of the classification model using the predicted and the true labels to evaluate the performance.

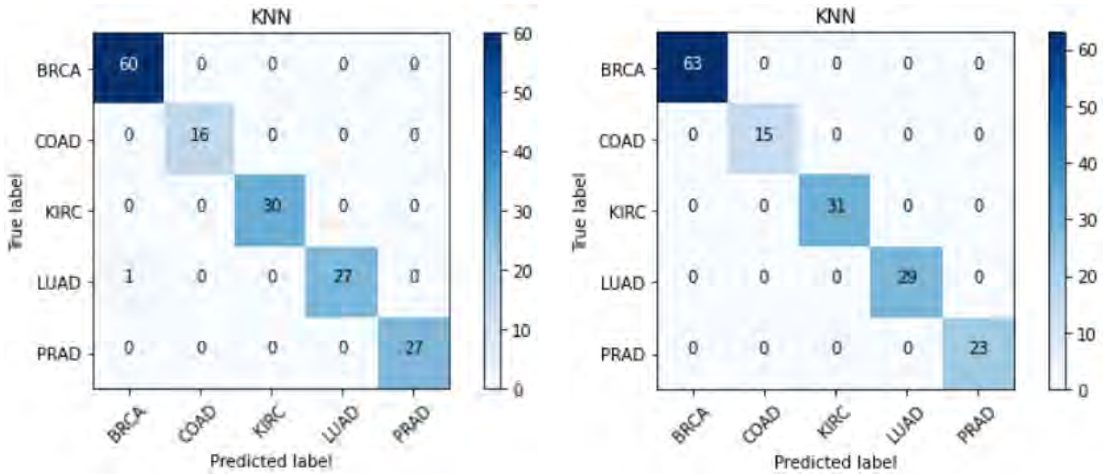


Figure 5.5: (Left): KNN Confusion Matrix Using 16383 Features; (Right): KNN Confusion Matrix Using 11 Features

The K-Nearest Neighbors algorithm took lesser time to train and predict cancer class, considering we only specified the number of neighbors to be 5. If we used more nearest voting neighbors the algorithm would have taken a bit longer. Regardless, lesser numbers of nearest neighbors voting for the sample's label, we managed to conduct the classification with and accuracy over 99.37% (fig 5.5 Left)and only one sample originally from the lung cancer class was misclassified as breast cancer.

When using 11 selected features to reduce the computational time even more, we could obtain a better accuracy despite using the same parameters and number of voting neighbors. Hence, using lesser features or genes while training our model not only increased the proficiency but also reduction in computation time helped us tune the parameters or debug the model for errors whenever needed.

Fig: 5.5(Right) exhibits a fully accurate model that can classify all the test samples into their correct cancer classes, thus proving, selection of a significant number of gene features with highest importance can direct the model into predicting test samples accurately.

In both methods considering 16,383 and 11 features, high accuracy and the least computational time were the standout features of the algorithm in classifying cancer classes.

5.5 Deep Learning

The deep learning sequential model training calculated a loss of 0.0177 which refers that our sequential model, despite taking the longest time to train, proved to be almost 100 percent accurate. When the deep learning model was fully trained we passed the test samples and test labels as validation data to compare the model training and testing accuracy and loss functions. The diagram below shows the graphical comparison of the two phases.

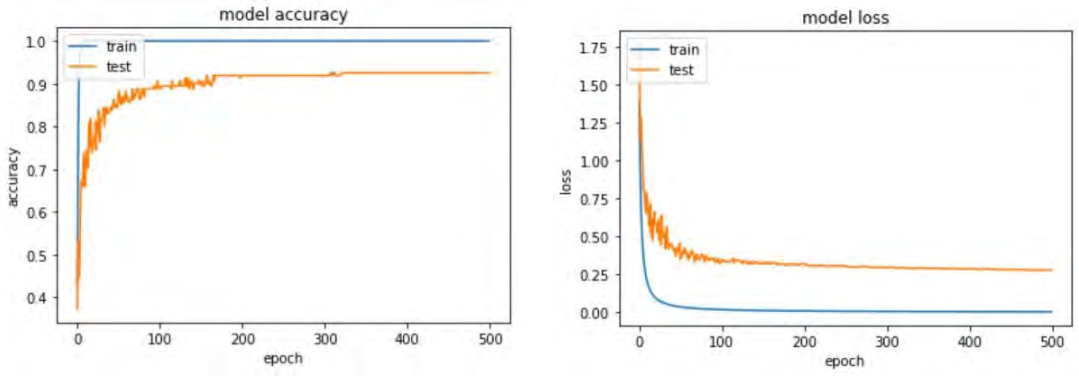


Figure 5.6: (Left): Deep Learning Model Accuracy; (Right): Deep Learning Model Loss

Although the graphs speak highly of our trained model, we inspected the model's performance by evaluating it in a scenario where it matters the most. Therefore, we inserted the testing cancer sample data in our already trained deep learning model and generated the confusion matrix to understand how well the model will perform if random cancer diagnosis data is provided. The confusion matrix displayed accuracy over 97.5% with only 4 misclassifications amongst the total test samples.

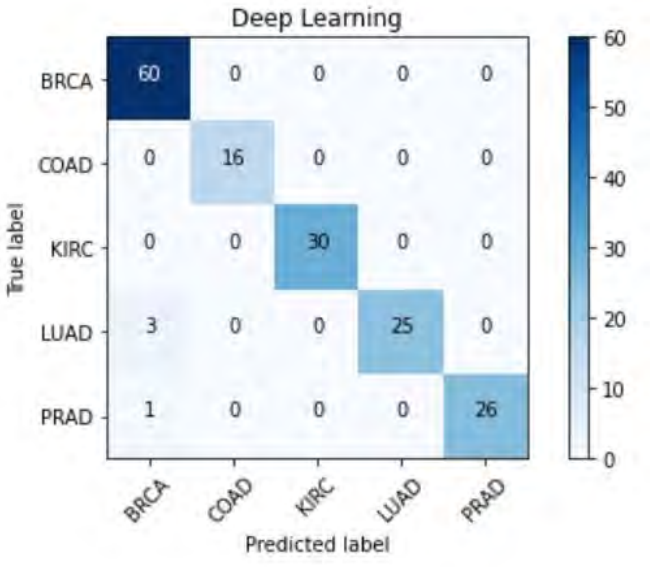


Figure 5.7: Deep Learning Confusion Matrix

To conclude, the deep learning algorithm is an excellent learner as it takes the input,

computes it with weights, carries the computation to all its fully connected nodes, calculates the loss, performs other calculations to minimize the loss and accepts data for prediction with the help of all the knowledge it acquired while training on the data. The amount of time the sequential model needs to acquire knowledge and analyze the data can only suggest how precise the model can be when analyzing patient's diagnosis data. This not only reduces human effort to analyze diagnosis data but also provide clear assumptions on what cancer types can occurs as the deep learning neural network model and its components mimic the human brain.

5.6 Ensemble Learning

In spite of a slight elongation in the computational time while training with all the feature columns, the performance of the model was exemplary, giving an accuracy of over 98.75% with only 2 misclassifications among 161 test samples and both were predicted as BRCA despite being from the LUAD class.

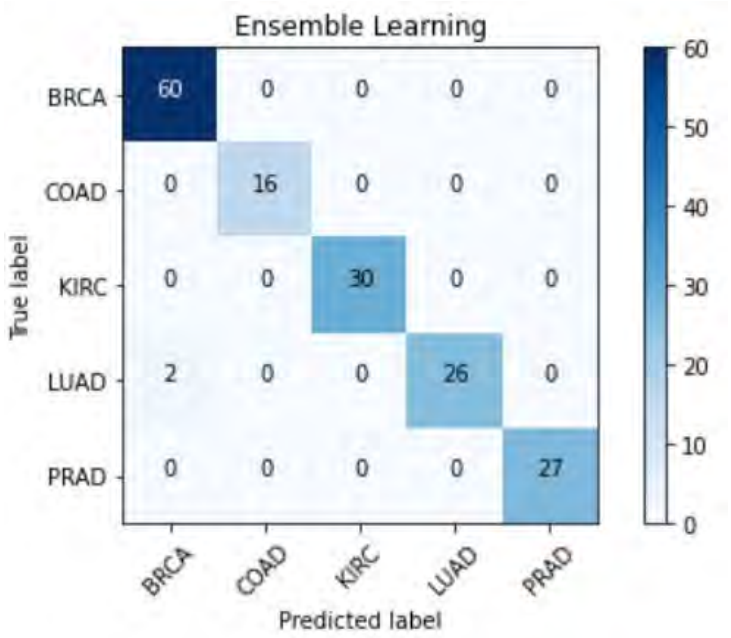


Figure 5.8: Ensemble Learning Confusion Matrix Using 16383 Features

The inclusion of Ensemble learning eliminated the concerns over unique misclassification of each of the classification algorithms used in modeling the voting classifier. To elaborate, suppose the xgb classifier makes a miss classification for a data point which was predicted correctly by all the other or two of the other classifiers, the majority voting model resolves this issue as the majority of the classifiers predicted the data point in favor of the right class. Therefore, only the common misclassifications that were made for a certain data point can be considered as errors. The reliability of the model vastly increases as the probability of making common misclassification decreases. We could visualize each prediction for corresponding test

samples after generating the confusion matrix and calculating the model accuracy score. The prediction result of the first 12 test sample is shown below:

	Predicted Class Number	Predicted Class Name
Test sample 1	4	PRAD
Test sample 2	0	BRCA
Test sample 3	2	KIRC
Test sample 4	2	KIRC
Test sample 5	0	BRCA
Test sample 6	0	BRCA
Test sample 7	2	KIRC
Test sample 8	0	BRCA
Test sample 9	3	LUAD
Test sample 10	2	KIRC
Test sample 11	2	KIRC

Table 5.1: Ensemble Learning Test Sample Results Using 16,383 Features

The term effectiveness, mentioned in previous model evaluation phases hints at the reduction of computational time and storage of the model thus assisting in easy and persistent debugging all while maintaining similar accuracy recommends that we also use our 11 selected features and analyze our ensemble learning results. To evaluate the efficient latter sequential model with the earlier model that considered all 16,383 features, the accuracy score was calculated to be a solid 100% when training with 11 features, suggesting that there were no common misclassification for a particular test sample across a majority of the classifiers in each of the model layers.

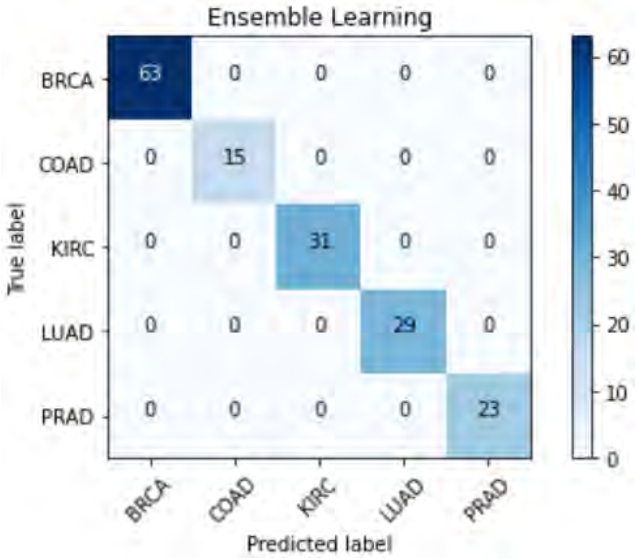


Figure 5.9: Ensemble Learning Confusion Matrix Using 11 Features

Fig 5.9 reflects the efficiency while analyzing 11 selected features with no misclassifications. The model trains and predicts each of the samples into their corresponding classes precisely and in less computational time, proving to be a clear upgrade to the ensemble learning model that considered all the features to train and classify test data points. Moreover, this might also reduce the processing time of microarrays

in laboratories, taking into account, that lesser number of genes as features can be useful in predicting accurate results.

The Ensemble learning model can exponentially boost the performance of our model and can sometimes be the deciding factor in a proper cancer diagnosis[48]. The goal of the entire process and model was to accurately diagnose a patient's gene expression level data and provide accurate predictions within a limited amount of time, and we believe that we have achieved such goal. With the external changes in environment, climate and lifestyle, newer mutations will evolve, inflicting newer challenges in proper diagnosis but what better way than to apply ensemble learning, improve performance of our model and let the majority of our well trained classifiers decide accurate diagnosis to help provide people the time and chance to recover from such miseries.

Chapter 6

Conclusion and Future Work

We have proposed a cancer classification model using ensemble learning technique in this study using feature selection and classification combining different machine learning algorithms. This paper focused on classifying the five distinct cancers successfully with highest accuracy. Feature selection was done using XGBoost and LightGBM which gave us feature score and feature importance respectively. Classifiers such as support vector machine (SVM), Random forest, Deep Learning, XGBoost classifier and K-nearest neighbour (KNN) are combined in an ensemble learning classification model which is then used for classification of the five cancers using the 11 important common features from the feature selection process to efficiently classify without the need for applying all the feature into our model. We managed to achieve more than 98% classification accuracy score of SVM, Random forest, XGboost and Deep Learning whereas KNN showed perfect 100 % accuracy score. Moreover, after combining all our classification models, the ensemble learning model gave us a classification accuracy score of 100% from the dataset of university of California Irvine Machine Learning repository (UCI). This open-source dataset provided information about random extraction of gene expressions levels of patients having different types of tumors. The principle objective of our research is to classify the data into the five cancer groups which are breast, kidney, colon, prostate and lung. Since this research only focused on five cancer types, henceforth we aim to work with bigger datasets of more cancers such as blood, bladder, liver, pancreatic, thyroid cancers etc. and apply our proposed model. Furthermore, in future we also hope to expand our research to classify a cancer in its sub-type for example, kidney cancer can be of renal cell carcinoma, urothelial carcinoma, sarcoma and Wilms tumor. If we can classify the sub-types we can easily identify which patient falls in which category resulting in initiating early treatment of that specific cancer.

Bibliography

[1] B. Alberts, *How genetic switches work*, Jan. 1970. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK26872/.

[2] A. Perez-Diez, *Microarrays for cancer diagnosis and classification*, Jan. 1970. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK6624/.

[3] L. Breiman, "Bagging predictors", *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[4] N. Cristianini, J. Shawe-Taylor, *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *Annals of statistics*, pp. 1189–1232, 2001.

[6] F. Chu and L. Wang, "Applications of support vector machines to cancer classification with microarray data", *International Journal of Neural Systems*, vol. 15, no. 06, pp. 475–484, 2005. DOI: 10.1142/s0129065705000396.

[7] R. Díaz-Uriarte and S. A. D. Andrés, *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006. DOI: 10.1186/1471-2105-7-3.

[8] X. Ma and H. Yu, *Global burden of cancer*, Dec. 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1994799/.

[9] J. Nahar, Y.-P. P. Chen, and A. S. Ali, "Microarray classification and rule based cancer identification", *2007 International Conference on Information and Communication Technology*, 2007. DOI: 10.1109/icict.2007.375339.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[11] V. Vinaya, N. Bulsara, C. J. Gadgil, and M. Gadgil, "Comparison of feature selection and classification combinations for cancer classification using microarray data", *International Journal of Bioinformatics Research and Applications*, vol. 5, no. 4, p. 417, 2009. DOI: 10.1504/ijbra.2009.027515.

[12] S. Sayad, *K nearest neighbors - classification*, 2010. [Online]. Available: https://www.saedsayad.com/k_nearest_neighbors.htm.

[13] X. Li, S. Peng, X. Zhan, J. Zhang, and Y. Xu, "Comparison of feature selection methods for multiclass cancer classification based on microarray data", *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, 2011. DOI: 10.1109/bmei.2011.6098612.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] U. H. Mazlan and P. Saad, "Classification of breast cancer microarray data using radial basis function network", *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)*, 2012. DOI: 10.1109/icssbe.2012.6396523.

[16] R. Polikar, "Ensemble learning", in *Ensemble machine learning*, Springer, 2012, pp. 1–34.

[17] A. V. Devadoss, J. J. Sheeba, and M. A. William, "Finding the peak age of an indian woman victim of breast cancer using cetd matrix", *International Journal of Computer Applications*, vol. 78, no. 10, 2013.

[18] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[19] J. C.-H. Yang, V. Hirsh, M. Schuler, N. Yamamoto, K. J. O'Byrne, T. S. Mok, V. Zazulina, M. Shahidi, J. Lungershausen, D. Massey, *et al.*, "Symptom control and quality of life in lux-lung 3: A phase iii study of afatinib or cisplatin/pemetrexed in patients with advanced lung adenocarcinoma with egfr mutations", *Journal of Clinical Oncology*, vol. 31, no. 27, pp. 3342–3350, 2013.

[20] J. Guo, *Transcription: The epicenter of gene expression*, May 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4076597/.

[21] B. Reese, *Illumina: Nextseq 500*, Aug. 2014. [Online]. Available: https://cgi.uconn.edu/illumina-nextseq-500/.

[22] D. R. Rhodes, J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pander, and A. M. Chinnaiyan, *Oncomine: A cancer microarray database and integrated data-mining platform*, Apr. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1476558604800472.

[23] J. C. Ang, H. Haron, and H. N. A. Hamed, "Semi-supervised svm-based feature selection for cancer classification using microarray gene expression data", *Current Approaches in Applied Artificial Intelligence Lecture Notes in Computer Science*, pp. 468–477, 2015. DOI: 10.1007/978-3-319-19066-2_45.

[24] M. R. Ataollahi, J. Sharifi, M. R. Paknahad, and A. Paknahad, *Breast cancer and associated factors: A review*, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5319297/.

[25] Y. Hou and J. Linhong, "Gene transcription and translation in design", in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, American Society of Mechanical Engineers, vol. 57175, 2015, V007T06A001.

[26] S. Korkmaz, G. Zararsiz, and D. Goksuluk, "Mlvis: A web tool for machine learning-based virtual screening in early-phase of drug discovery and development", *PloS one*, vol. 10, no. 4, e0124600, 2015.

[27] A. Travers and G. Muskhelishvili, *Dna structure and function*, Jun. 2015. [Online]. Available: https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.13307.

[28] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, *et al.*, "A survey of best practices for rna-seq data analysis", *Genome biology*, vol. 17, no. 1, p. 13, 2016.

[29] P. Guillen and J. Ebalunode, "Cancer classification based on microarray gene expression data using deep learning", *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2016. DOI: 10.1109/csci.2016.0270.

[30] A. Mehrgou and M. Akouchekian, *The importance of brca1 and brca2 genes mutations in breast cancer development*, May 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4972064/.

[31] I.-O. Petre and C. Buiu, "A colon cancer microarray analysis technique", *2017 E-Health and Bioengineering Conference (EHB)*, 2017. DOI: 10.1109/ehb.2017.7995412.

[32] t. H. E. Team, *Cancer types - carcinoma, sarcoma, leukemia, lymphoma*, Oct. 2017. [Online]. Available: https://www.healthline.com/health/cancer.

[33] E. Allibhai, *Building a deep learning model using keras*, Nov. 2018. [Online]. Available: https://towardsdatascience.com/building-a-deep-learning-model-using-keras-1548ca149d37.

[34] Y.-H. Hsu and D. Si, "Cancer type prediction and classification based on rna-sequencing data", *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018. DOI: 10.1109/embc.2018.8513521.

[35] T.-L. Le, T.-T. Huynh, C.-M. Lin, and F. Chao, "Breast cancer diagnosis using k-means type-2 fuzzy neural network", *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018. DOI: 10.1109/smc.2018.00703.

[36] A. Navlani, *Knn classification using scikit-learn*, Aug. 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn?utm_source=adwords_ppc.

[37] Playment, *What is deep learning and why you need it?*, Aug. 2018. [Online]. Available: https://becominghuman.ai/what-is-deep-learning-and-why-you-need-it-9e2fc0f0e61b.

[38] S. Turgut, M. Dagtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods", *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018. DOI: 10.1109/ebbt.2018.8391468.

[39] R. Alanni, J. Hou, H. Azzawi, and Y. Xiang, "A novel gene selection algorithm for cancer classification using microarray datasets", *BMC medical genomics*, vol. 12, no. 1, p. 10, 2019.

[40] Y. Braizer, *Prostate cancer: Symptoms, treatment, and causes*, Aug. 2019. [Online]. Available: https://www.medicalnewstoday.com/articles/150086.

[41] J. Brownlee, *Difference between a batch and an epoch in a neural network*, Oct. 2019. [Online]. Available: https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/.

[42] P. Crosta, *Colon cancer: Symptoms, treatment, and causes*, Aug. 2019. [Online]. Available: https://www.medicalnewstoday.com/articles/150496.

[43] V. Morde, *Xgboost algorithm: Long may she reign!*, Apr. 2019. [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.

[44] J. Brownlee, *What is a confusion matrix in machine learning*, Aug. 2020. [Online]. Available: https://machinelearningmastery.com/confusion-matrix-machine-learning/.

[45] R. Dwivedi, *Introduction to xgboost algorithm for classification and regression*, Jul. 2020. [Online]. Available: https://analyticssteps.com/blogs/introduction-xgboost-algorithm-classification-and-regression.

[46] A. Jain, *Xgboost parameters: Xgboost parameter tuning*, Jul. 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/.

[47] M. Khatri, *Renal cell carcinoma: Symptoms, diagnosis, and treatment*, Jan. 2020. [Online]. Available: https://www.webmd.com/cancer/renal-cell-carcinoma.

[48] A. Singh, *Ensemble learning: Ensemble techniques*, May 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/.

[49] P. Tahmasebi, S. Kamrava, T. Bai, and M. Sahimi, "Machine learning in geo- and environmental sciences: From small to large scale", *Advances in Water Resources*, p. 103 619, 2020.

[50] M. M. Babu, *Chapter 11 an introduction to microarray data analysis*. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.7755.

[51] L. M. Kitchell, *Loss functions and optimizers*. [Online]. Available: https://kitchell.github.io/DeepLearningTutorial/7lossfunctionsoptimizers.html.

[52] L. Learning, *Biology for non-majors i*. [Online]. Available: https://courses.lumenlearning.com/wmopen-nmbiology1/chapter/storing-genetic-information/.

[53] Y. Qui, *Microarray data analysis*. [Online]. Available: http://compbio.uthsc.edu/microarray/lecture1.htm.2006.

[54] Y. Tang, Y.-Q. Zhang, Z. Huang, and X. Hu, "Granular svm-rfe gene selection algorithm for reliable prostate cancer classification on microarray expression data", *Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)*, DOI: 10.1109/bibe.2005.34.

[55] H. won and S. Cho, *Machine learning in dna microarray analysis for cancer classification.* [Online]. Available: https://www.researchgate.net/publication/221118082_Machine_Learning_in_DNA_Microarray_Analysis_for_Cancer_Classification.