Lung Cancer Prediction Using Gene Expression Data with Statistical Approach for Gene Selection

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD
OF DEGREE
OF

MASTER OF TECHNOLOGY IN
COMPUTER SCIENCE AND ENGINEERING

Submitted By:

MANISH ANAND

2K18/CSE/23

Under the supervision of

NIPUN BANSAL

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of Engineering)

Bawana Road, Delhi-110042
JULY, 2020

DECLARATION

I, Manish Anand, Roll No. 2K18/CSE/23 student of M.Tech (Computer Science & Engineering), hereby declare that the Project Dissertation titled "**Lung Cancer Prediction Using Gene Expression Data with Statistical Approach for Gene Selection**" which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi, in partial fulfillment for the requirement of the award of the degree of Master of Technology is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Manish Anand

Place: DTU, Delhi Manish Anand

Date: (2K18/CSE/23)

CERTIFICATE

I hereby certify that the Project Dissertation titled “**Lung Cancer Prediction Using Gene Expression Data with Statistical Approach for Gene Selection**” which is submitted by Manish Anand, Roll No. 2K18/CSE/23, Department of computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment for the requirement of the award of degree of Master of Technology (Computer Science and Engineering) is a record of a project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Nipun

Place: Delhi **(Mr. Nipun Bansal)**

Date: **SUPERVISOR**

ABSTRACT

Cancer is caused by abnormal cell growth or cell division in the body. This uncontrolled and undesired growth is a reflection of genetic variation causing abnormal functioning of genes, causing a change in gene expression. This change in gene expression is brought understudy for cancer prediction, diagnosis, and treatment. Machine Learning techniques when applied to gene expressions can predict one's susceptibility towards cancer. The tough task is to determine those genes that possess a stronger capability or show greater variation in expression when in an abnormal state than the normal state. This paper proposes gene selection techniques for selecting an optimal subset of genes that are highly important for accurate prediction. The lung cancer gene expression data has been taken from Kent Ridge Biomedical Dataset Repository. The main focus of the project is to select the optimal subset of genes to have a high value of AUC_ROC and F-measure to make a correct assessment of the model dealing with an imbalance dataset.

ACKNOWLEDGEMENT

The success of a project requires help and contribution from numerous individuals and organizations. Writing the report of this project work gives me an opportunity to express my gratitude to everyone who has helped in shaping up the outcome of this project.

I express my gratitude to my major project guide **Mr. Nipun Bansal,** Asst. Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to him for the extensive support and encouragement he provided.

I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.

I am also thankful to my family for constantly encouraging me and giving me unconditional support while pursuing this research.

Manish Anand
Roll No. 2K18/CSE/23
M.Tech.(CSE)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST of SYMBOLS, ABBREVIATIONS And NOMENCLATURE

$\sum$	Summation
μ	Mean
Acc	Accuracy
Adj	Adjective
Adv	Adverb
CV	Cross Validation
DNA	Deoxyribonucleic acid
Diff	Difference
F1	F-measure
GNV	Gaussian Naive Bayes
MBA	Mean Based Approach
MIFS	Mutual Information based Feature Selection
ML	Machine Learning
NB	Naive Bayesian
Norm	Normal
P	Precision
R	Recall
RF	Random Forest
SVM	Support Vector Machine
Tum	Tumor
WHO	World Health Organization

CHAPTER 1 INTRODUCTION

1.1. OVERVIEW

Gene microarray or DNA chip technique is based on the principle of depositing thousands of multiple DNA sequences on a small surface which usually happens to be a glass slide, also called chip. The technique helps in calculating and analyzing the gene expression of large numbers of samples. Golub et al. first analyzed gene expression data for cancer classification [1]. Gene expression data is a numerical value that shows the magnitude of the genes at some point [2]. Analysis of this data can help in predicting cancer and also the type of it by comparing it to the normal gene expression value. The data generated by the microarray technique is extensive and it’s a complex task to analyze them. All the genes do not contribute the same to class. So there is a need for gene selection to obtain an optimal subset of genes such that informative and significant genes are selected and noisy and redundant ones are excluded. In literature, many techniques talk about gene selection.

Cancer is the abnormal growth of cells in the body. It is one of the major causes of deaths in humans. Cancer is often attributed to be a fatal disease. This myth can be overcome with early detection and timely treatment. In many cases, people are asymptotic to symptoms at early stages of their cancer. Thus, need of the hour is a technique that can detect cancer even when asymptotic or in early stage. Genes analysis is an effective measure to predict cancer. The challenging part is to identify the most informative genes, the genes that can show high variation in expression when in diseased condition to when the cells are disease free.

In this project, we propose a statistic based gene selection method that selects the most significant genes and filters out the noisy genes and applies machine learning algorithms that analyze the gene expression data for lung cancer classification. The data has been taken from Kent Ridge Biomedical Dataset Repository [8]. Here, we have 96 samples out of which 86 are tumorous and 10 are normal. We have selected 72 genes out of 7129 genes using the proposed method and sampling has been done to deal with the imbalanced dataset. Then machine learning algorithms have been used to classify and calculate f-score and AUC-ROC.

1.2. RESEARCH OBJECTIVE

The idea presented here is to predict Lung cancer by analyzing gene expression data and to propose genes that are observed to be major contributors in causing lung cancer. The focus has been on selecting the optimal subset of genes that are highly important for accurate prediction. We propose a statistics based gene selection method that selects the most significant genes and filters out the noisy ones. Machine learning algorithms, that analyze the gene expression data, have been used for lung cancer classification.

1.3. ORGANISATION OF THESIS

The organization of the report is as follows. The project report has been divided into five chapters. Each chapter deals with one component related to this thesis. Chapter 1 being an introduction to this thesis, gives us the brief overview of the project, the literature on gene selection and cancer prediction is reviewed in chapter 2, the proposed methodology of gene selection and application of machine learning model is described in chapter 3, the implementation and results are discussed in chapter 4 and the overall conclusions and future scope are summarized in chapter 5.

CHAPTER 2 LITERATURE SURVEY

2.1. RELATED WORK

Gene selection refers to the technique of selecting the most significant and informative genes and removing the noisy genes which are not important for analysis. There are several gene selection techniques which talk about predicting cancer by selecting the optimal number of genes. Lung cancer prediction by analyzing gene expression data and considering environmental factors affecting some of the genes has been discussed in [3]. The paper takes the subset of total genes and applies machine learning algorithms for the classification. In [4] Liu, Wei, et al. identified a subset of genes that are significantly deregulated in LUAD (Lung adenocarcinoma) by analyzing independent microarrays. G. Russo et al. presented the advantages and limitations of microarray technology in human cancer [5].

Gene selection becomes very important when dealing with a large number of genes as not all genes make the same contribution to the class. In [6], Ahmad et al. has talked about the impact of noise and sampling on feature ranking of a biological dataset. Yijun et al. [7] used local-learning-based feature selection for high dimensional data analysis. D. Pavithra and B. Lakshmanan [9] have presented a classification of different cancer data. Battiti et al [10] proposed Mutual-Information based Feature Selection (MIFS) that expresses the amount of data shared between two random variables. It is often used to show relationships among features and class.

Along with feature selection, the class imbalance is a very common challenge faced in many of the biological datasets. Our approach not only addresses the issue of feature selection but also handles the problem of an imbalanced dataset. Sampling is a common technique to solve the class imbalance problem, which is done by modifying the training dataset to enhance its balance.

2.2. BACKGROUND CONCEPTS

2.2.1. About Dataset

The data set for this project is an authenticated one. It has been taken from the Kent Ridge Biomedical Dataset Repository. The data set is named ‘Lung cancer’. To be precise, there are 96 records and a total of 7129 features(genes).

Out of 96 records, 86 are labelled to be cancerous and remaining 10 are non-cancerous in nature.

2.2.2. Gene Expression

Genes encode proteins, which are responsible for cell functioning. So we can say that the cumulative genes present in the cell determine what the cell can do. Biological processes can be assessed effectively by analyzing the variation in gene expression because variation is a sensitive indicator of biological activity and a changing gene expression pattern is reflected in a change of biological process. Gene expression analysis typically involves the isolation or capture of transcribed RNA within a sample, followed by amplification and subsequent detection and quantitation.[22].

The value of expression of the genes is determined and a data set is constructed. By observing the value of each of the genes across different samples, we can find biomarkers, the genes that are capable of showing observable variations when in diseased conditions as compared to when in a normal state. The challenge is to determine the optimal set for the target disease and also eliminate outliers. The variation in genes may be subject to conditions other than the disease itself, thus leading to false negatives and false positives. Some variations and genes may be highly interdependent. Identifying such anomalies and their elimination can lead to highly precise and accurate predictions and classifications.

2.2.3. Stratified K- Cross Validation

Model evaluation is a challenging task, involving the use of available dataset to train a model and then use it to estimate the performance of prediction on unseen data. A model can simply repeat the labels and become perfect in classification if it is made to test the same data or a part of the same data that it has used to learn the labels from. To avoid this situation of overfitting in classification models, we have used cross validation. By using cross-validation, we make sure a part of the data is set out as test data or unseen data, so that we can estimate the skill of the machine learning model, by ensuring a less biased estimate.

Since our data set is visibly unbalanced, we have used stratified K-folds cross validation to ensure that each fold is representative of the complete dataset. This helps to overcome the bias present in supervised learning algorithms used. In the k-fold cross-validation technique, the dataset is divided into k-folds of approximately equal size, and the model is trained on one fold, remaining k-1 folds are used for testing. The model is evaluated k times so that each fold gets a chance to evaluate the model exactly once. In the case of stratified k-folds cross-validation, the folds are stratified so that they contain approximately the same proportions of labels as the original dataset [13]. I have used 5-folds stratified cross-validation as shown in figure 1. Stratified K folds prove better in terms of bias as well as variance.

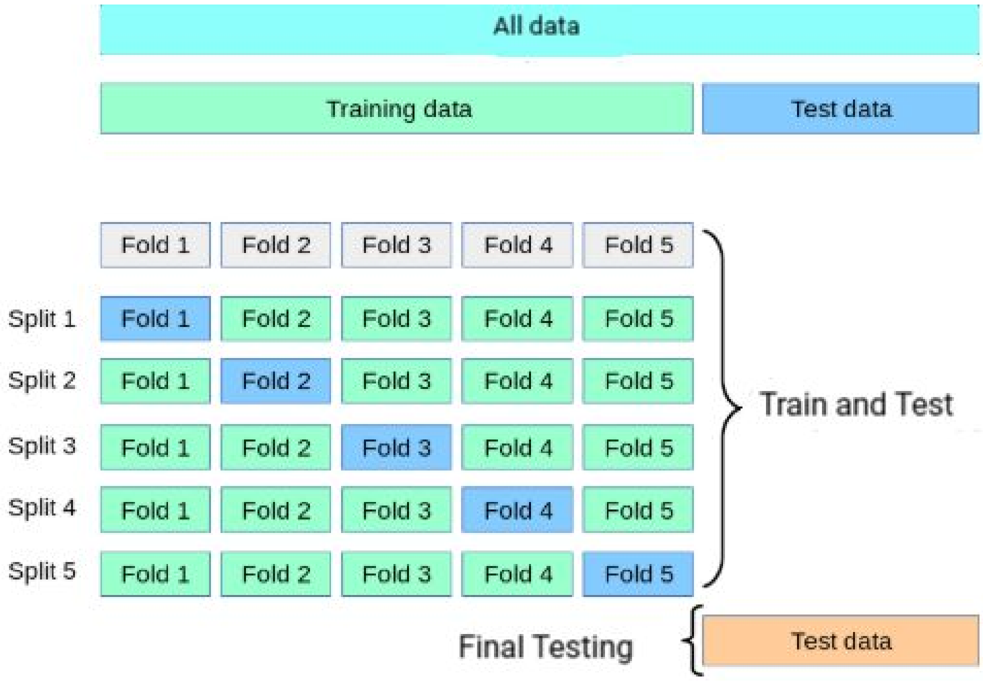


Fig. 1. 5-Folds Cross-Validation

2.2.4. Supervised Machine Learning

Machine Learning is a data analysis method. Machine learns with inputs given to it and in turn predicts or classifies the unseen data. If the machine is made to learn by an already classified or labelled data, we term it as supervised machine learning. In case of absence of learning examples, machine learned by analysis of patterns in data, termed as unsupervised learning. In this project, we use supervised learning models.

Our training dataset consists of a pair of input objects and the desired output, namely the gene expression behaving as input object and the label of the expression i.e. being cancerous or non-cancerous in nature is the desired output. Our models will use this training dataset and classify the unseen gene expressions, based on whether the value is close to its normal expected one or is showing considerable variation due to abnormality present.

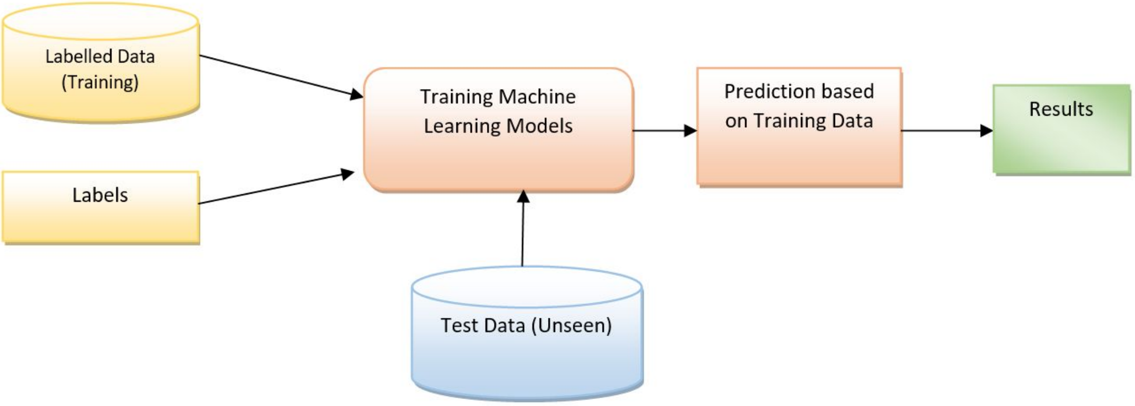

Labelled Data (Training)
Labels
Training Machine Learning Models
Prediction based on Training Data
Results
Test Data (Unseen)

Fig. 2. Supervised Machine Learning

2.2.5. Gaussian Naive Bayes

Gaussian Naive Bayes is a variant of the Naive Bayes algorithm which follows Gaussian distribution. It is one of the most commonly used supervised machine learning algorithms and is a powerful one for classification. It follows the Naive Bayes' assumption that each feature is independent and contributes equally to the outcome, but the algorithm seems to perform well even when this assumption is not completely applicable to the dataset. Each feature is supposed to be distributed over the Gaussian curve. Robustification of Naïve Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis has been discussed in [14]. It is useful for binary classification, as in the case of this project where we are to classify a record either as cancerous or as non-cancerous, as well as for multinomial classes. The advantage of using this algorithm is its fast learning nature and highly accurate prediction.

2.2.6. Multilayer Perceptron

Multilayer Perceptrons (MLPs) are feedforward networks with one or more layers of units between the input and output layers, such layers are called hidden layers. As the name suggests, it has a number of perceptrons arranged into several layers. MLP can be used for the classification of linearly inseparable patterns and for function approximation [15]. It is a supervised learning algorithm. A back propagation algorithm is used to train MLP. Each layer in MLP consists of neurons (nodes). Each neuron in the hidden layer transforms the values from the previous layer using edge weight and activation function.

The advantages of MLP include its ability to learn the nonlinear models and can also learn in real-time. This model tends to provide approximate solutions even for highly complex problems and find a place in a large number of real time applications. In many cases it beats the conventional statistical techniques for prediction and classification in various fields of applications [16].

2.2.7. Support Vector Machine

Support Vector Machine (SVM) [17] is a supervised machine learning algorithm that offers solutions for both classification and regression problems. It introduces one of the most robust prediction methods. It is a non-probabilistic, linear, binary classifier. The aim of SVM is to find a hyperplane in N-dimensional space that can distinctly classify the data. When given classes cannot be linearly separated in the original input space, the Support Vector machine first (non -linearly) transforms the original input space into a higher dimensional feature space. This transformation can be achieved by using various nonlinear mappings [18]. Once this transformation is done the SV machine finds the linear optimal separating hyperplane in this feature space. SVM does not need a large number of labelled instances for classification problems. It is highly effective in high dimensional space and is also memory efficient.

2.2.8. XGBoost

XGBoost stands for eXtreme Gradient Boosting [19]. It is a model that was first proposed by Tianqi Chen and Carlos Guestrin in 2011 and has been continuously optimized and improved in the follow-up study of many scientists [20]. It is moreover a library that is flexible to use and easy to implement. It is an implementation of a gradient boosting technique. It is efficient in terms of parallel computing of tree structures, training large models and handling large data sets. Its best features include, providing a boost to an already constructed model to handle new input data and also, to handle missing values in the data set automatically. XGBoost produces results fast as compared to the implementation of other gradient boosting.

2.2.9. Performance Metrics

Selection of performance metrics is an essential task in classification problems. If the right metrics are not used, correctness of the model cannot be assured. It influences the comparisons we make and conclusions that we drew from the experiment. Selection of the metrics depends on various factors, like the nature of the dataset, the kind of problem we are dealing with, the machine learning models we use etc. The performance metrics that we have used in this project are discussed below.

- **Accuracy**

  Accuracy can be defined as the ratio of correct predictions by the model to the total number of predictions by the model. Accuracy is easy to understand and calculate. Accuracy, however, does not prove to be a very efficient metric, especially for highly unbalanced data, thus I have used other metrics along to get a clearer picture of model performance. I have reported accuracy because it is the metric that is generally used to summarize the performance of a machine learning model.

Accuracy = (TP + TN) / (TP + FP + FN + TN)

Where:

TP - True Positive

TN- True Negatives

FP - False Positive

FN - False Negatives

- **F1 measure**

  Precision and Recall are 2 metrics that are widely computed for their ability to provide a better estimation of model performance by taking into considerations, the false positives and false negatives given by the model. Precision gives estimates of correct positive predictions thus focusing on the minority class. It focuses on minimizing false positives. Recall is one such measure that tends to minimize false negatives. In order to take advantage of both precision and recall, we use F measure.

  F-Measure = (2 * Precision * Recall) / (Precision + Recall)

  Where,

  Precision = TP / (TP + FP)

  Recall = TP /(TP + FN)

  F score proves to be a better performance estimator than accuracy in cases with highly unbalanced data.

- **AUC_ROC curve**

  ROC curve stands for receiver operating characteristic curve. This is basically a graph that shows the model's performance at various thresholds. ROC curve is plotted with True Positive Rate on the Y axis and False Positive Rate on the X-axis. It is highly used with classification problems.

AUC is the area under the ROC plot. It basically denotes how efficient the model is to distinguish between the classes in the data set. If the value of area is 0, it signifies that the model is predicting positives to be negatives and negatives to be positives. Value = 1 signifies correct prediction under all thresholds and is ideal to be achieved.

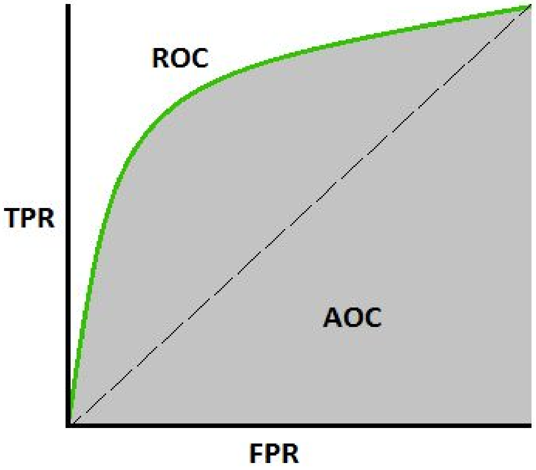


Fig. 3. Sample ROC Plot

CHAPTER 3 PROPOSED METHOD

3.1. OVERALL WORKFLOW

The overall flow of the experiment done has been described in this section.

The first step is the input data. The entire data set downloaded is taken as input data. The next step is the selection of the optimal set of genes that are capable to show variation when in diseased condition from normal state. The higher the capacity of differentiation , the more reliable the gene is. I have used a mean and standard deviation approach as described in 3.2 for gene selection. The top 72 genes obtained in these methods are used, rather than the entire set of 7000 genes present in the dataset. Selection of optimal genes is the most important step as we eliminate the genes that do have a significant contribution towards classification. The genes that do not show much variation may tend to influence the results, thus they are eliminated from further process of classification.

After the genes have been selected, the data is divided into a 70:30 ratio, where 70 5 of the complete data is used for training the model and the rest 30% is kept aside for testing purposes. The 30% kept aside is unseen to the model, thus can prove to be a good test for its performance. The 70% data is subjected to sampling because of high imbalance in the data. There are just 10 non-cancerous samples as compared to 86 positive samples. To nullify this imbalance, random oversampling technique, to oversample minority class has been adopted. To get rid of the overfitting problem , we use 5 folds stratified cross validation over the test data.

Gaussian Naive Bayes, Multilayer perceptron, SVM and XGBoost machine learning models trained with training dataset prepared. The testing of model's performance is done by testing the 30% dataset that was kept aside. Accuracy, F1 score and acu-roc measures are reported.

The flow diagram below gives an overall view of the methodology followed. The results of the experiment are discussed in the next chapter. Graphs and plots have been provided for better understanding and interpretation.

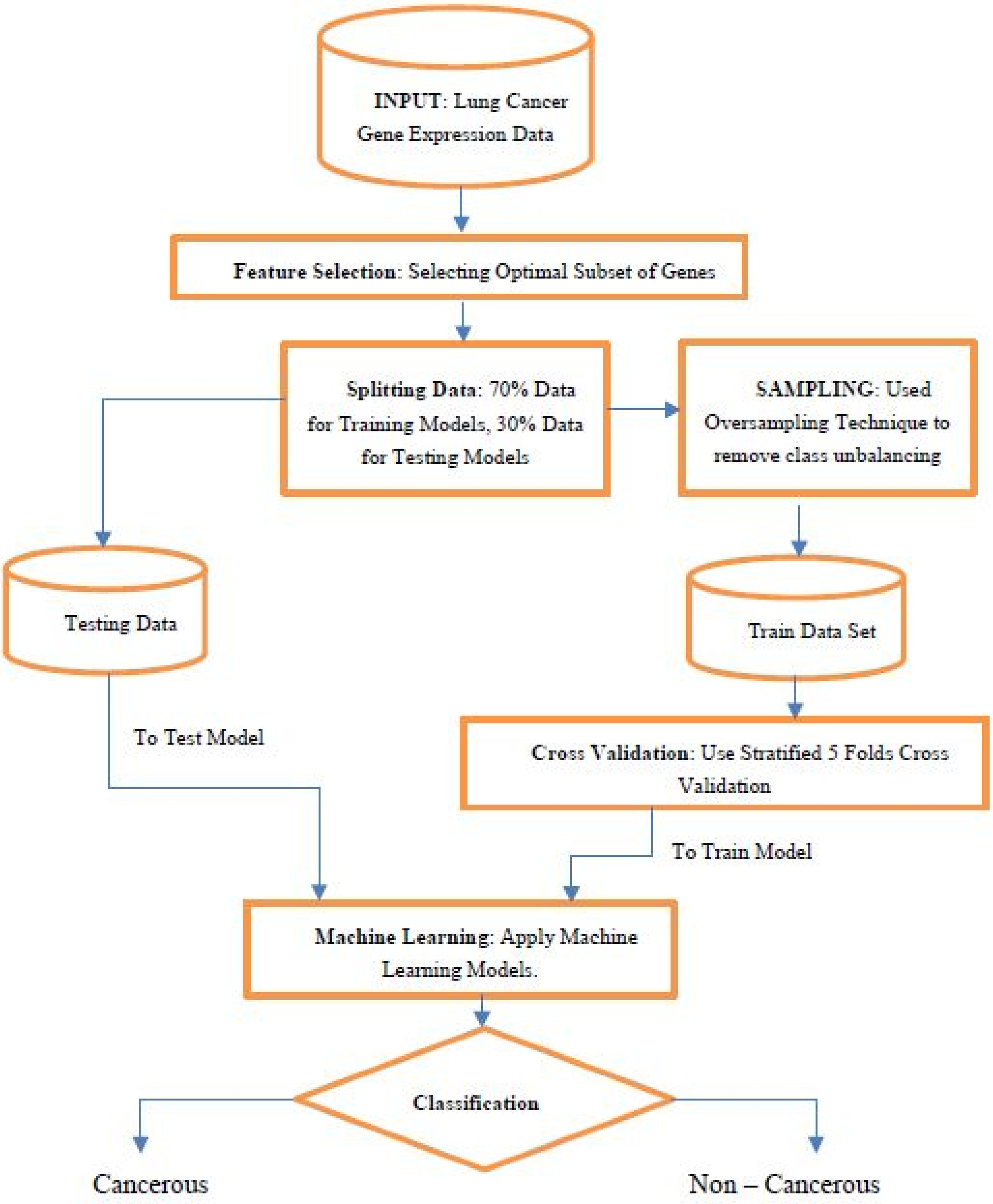

INPUT: Lung Cancer Gene Expression Data
Feature Selection: Selecting Optimal Subset of Genes
Splitting Data: 70% Data for Training Models, 30% Data for Testing Models
SAMPLING: Used Oversampling Technique to remove class unbalancing
Testing Data
Train Data Set
To Test Model
Cross Validation: Use Stratified 5 Folds Cross Validation
To Train Model
Machine Learning: Apply Machine Learning Models.
Classification
Cancerous
Non – Cancerous

Fig. 4. Flow Diagram for Computational Modeling of Gene Expression Data

3.2 GENE SELECTION

The value of the gene expression is an indicator of the sample being cancerous or non-cancerous. The data set consists of 7129 gene expressions per sample. It becomes important to identify the genes that show a significant increase or decrease in value and become accurate indicators for either presence of carcinoma or its absence, to reduce false positives and false negative prediction. This paper presents two statistical methods for selecting genes. The first method is to consider the mean value of expression for every gene and the second method takes into consideration the standard deviation in the expression value.

3.3. SAMPLING

High Dimensional Data with fewer numbers of samples can lead to biased model performance when the problem of class imbalance is not checked [11]. Sampling is one of the most used techniques to counter the class imbalance dataset. It is done by modifying the training dataset to enhance its balance. In this paper, we have used oversampling of minority class [12]. The dataset before and after sampling is shown in Fig. 3 and Fig. 4 respectively.

3.4. STATISTICS BASED METHODS

3.4.1 Mean Base Approach

The first approach is to calculate the mean value of gene expression of all the 7129 genes and tabulate them for cancerous as well as non-cancerous samples. Then the difference in mean values for cancerous and non-cancerous samples for every gene is calculated. This gives an estimate of the difference between expressions of positive and negative samples. The higher the difference, the more variation the gene can show when in diseased state as compared to normal state. The genes are sorted in decreasing order of the calculated difference. Top 72 genes are selected, their scores are shown in Table 1.

Difference in expression value = |Mean (Tumor)-Mean (Normal)|

These 72 genes have been used to classify the records in two ways:

- On the basis of threshold which is calculated by taking the average of Mean(Tumor) and Mean(Normal), and
- Using the supervised machine learning algorithms.

Results and observations of both the classifications are given in chapter 4.

3.4.2 Standard Deviation Based Approach

The second approach is to calculate standard deviation value for each gene in the data set. Standard Deviation was taken into account because it is a better measure to judge the dispersion, thus providing a complete overview of the entire dataset. The difference in mean value calculated in the previous approach is divided by the sum of standard deviation for cancerous and non-cancerous records.Next, genes are sorted in decreasing order based on this value, mathematically

sorting_function = | mean_diff | / (sd_tum + sd_norm)

Top 72 genes are selected with this approach; their scores have been shown in Table 3.

With this approach as well the records have been classified in two ways:

- On the basis of threshold which is calculated as we have calculated for the previous method, and
- Using supervised machine learning models.

CHAPTER 4 IMPLEMENTATION AND RESULTS

This chapter describes the implementations and results of the entire experiment that have been carried out. Plots and graphs provide an efficient means of comparison of results.

4.1 CLASSIFICATION USING MEAN BASED APPROACH

4.1.1. Using Threshold Value

Top 72 genes are selected using the mean based approach discussed in chapter 3 section 3.4.1 out of a total 7129 genes. Following table gives the details of the top 72 genes and their threshold value.

Table 1 Genes selected using Mean Based Approach

Genes	Threshold	Score
M25079_s_at	16001.84756	27609.86488
J03890_rna1_at	10310.3457	18082.2286
M68519_rna1_at	14637.63302	17009.09395
M30838_at	11471.45081	12334.37837
M87789_s_at	23597.79965	11072.1193
HG3925-HT4195_s_at	7687.482674	10423.57465
HG1428-HT1428_s_at	7840.684302	10277.7314
M13686_s_at	7833.737326	10139.86535
X53331_at	10274.00674	8459.706512

Z19554_s_at	14272.23384	8226.872326
AFFX-HUMGAPDH/M33197_5_at	9133.664186	7350.748372
X57809_s_at	13794.73605	7339.272093
Y09267_at	4581.148256	7129.503488
M11313_s_at	6961.035	7080.47
M17733_at	17635.09581	7072.968372
AFFX-HUMGAPDH/M33197_M_at	7799.892674	7004.405349
S71043_rna1_s_at	23803.09349	6894.326977
HG2815-HT4023_s_at	16206.70174	6402.396512
X01677_f_at	6839.90686	6340.953721
U01102_at	4075.697326	6265.785349
X00274_at	16128.90837	6087.063256
AFFX-HUMGAPDH/M33197_3_at	9901.182907	6084.445814
V00594_s_at	16972.11198	6058.996047
M63438_s_at	21220.02547	5916.19093
M55998_s_at	9634.886628	5875.673256
M17885_at	17121.40535	5340.070698
M12963_s_at	3484.116047	4875.187907
X04470_s_at	5343.696628	4869.946744

L19686_rna1_at	5501.145	4832.81
X98482_r_at	5817.106047	4780.332093
Z84721_cds2_at	3525.503023	4528.533953
M18728_at	5296.803837	4115.287674
HG2815-HT2931_at	7831.002791	4069.054419
J00105_s_at	16797.4007	3992.238605
M14328_s_at	5497.415349	3837.950698
X65614_at	2052.264419	3641.168837
hum_alu_at	24091.54023	3616.259535
Z18951_at	2408.948721	3552.162558
X17206_at	13929.68419	3473.308372
X02152_at	5719.854186	3445.868372
U60115_at	2180.62	3403.84
HG3431-HT3616_s_at	3215.057907	3388.204186
AFFX-CreX-5_at	5906.666744	3370.273488
U21931_at	4319.209535	3307.86093
X16832_at	6074.769302	3284.201395
V00599_s_at	6957.733023	3267.506047
X86693_at	2534.212326	3187.415349

Z48501_s_at	4333.215	3150.77
X57351_s_at	16426.70663	3088.666744
U89336_cds3_at	1879.633837	3054.172326
S73591_at	8571.74314	3046.853721
X00351_f_at	13637.29256	2994.245116
M80563_at	3663.736395	2982.947209
HG2809-HT2920_s_at	3007.02686	2961.086279
M94250_at	1897.135116	2922.290233
M57710_at	9956.838023	2885.303953
AFFX-HSAC07/X00351_M_at	17310.21965	2882.779302
X16064_at	13865.65605	2842.867907
D45370_at	2291.289535	2791.42093
M24461_at	5569.847791	2758.644419
M95787_at	4081.977674	2607.964651
X68277_at	4819.899651	2606.180698
M17886_at	14350.43407	2593.84814
L20688_at	7326.88593	2538.48814
Z23090_at	5883.60814	2534.516279
X12876_s_at	2568.30314	2515.926279

AF001548_rna1_at	2003.519767	2514.360465
HG2279-HT2375_at	4536.635349	2503.110698
J04988_at	7240.314302	2493.208605
J03909_at	4282.634535	2451.21093
M60854_at	10107.16686	2435.333721
M24485_s_at	5569.008372	2399.236744

The records are classified as tumorous or normal by each gene based on its expression value and the threshold value for that particular gene.

The graph below in fig 5 shows the F score for each of the top 72 genes selected based on their mean value. These are obtained by considering each gene individually for classification of the 96 samples available in the dataset.

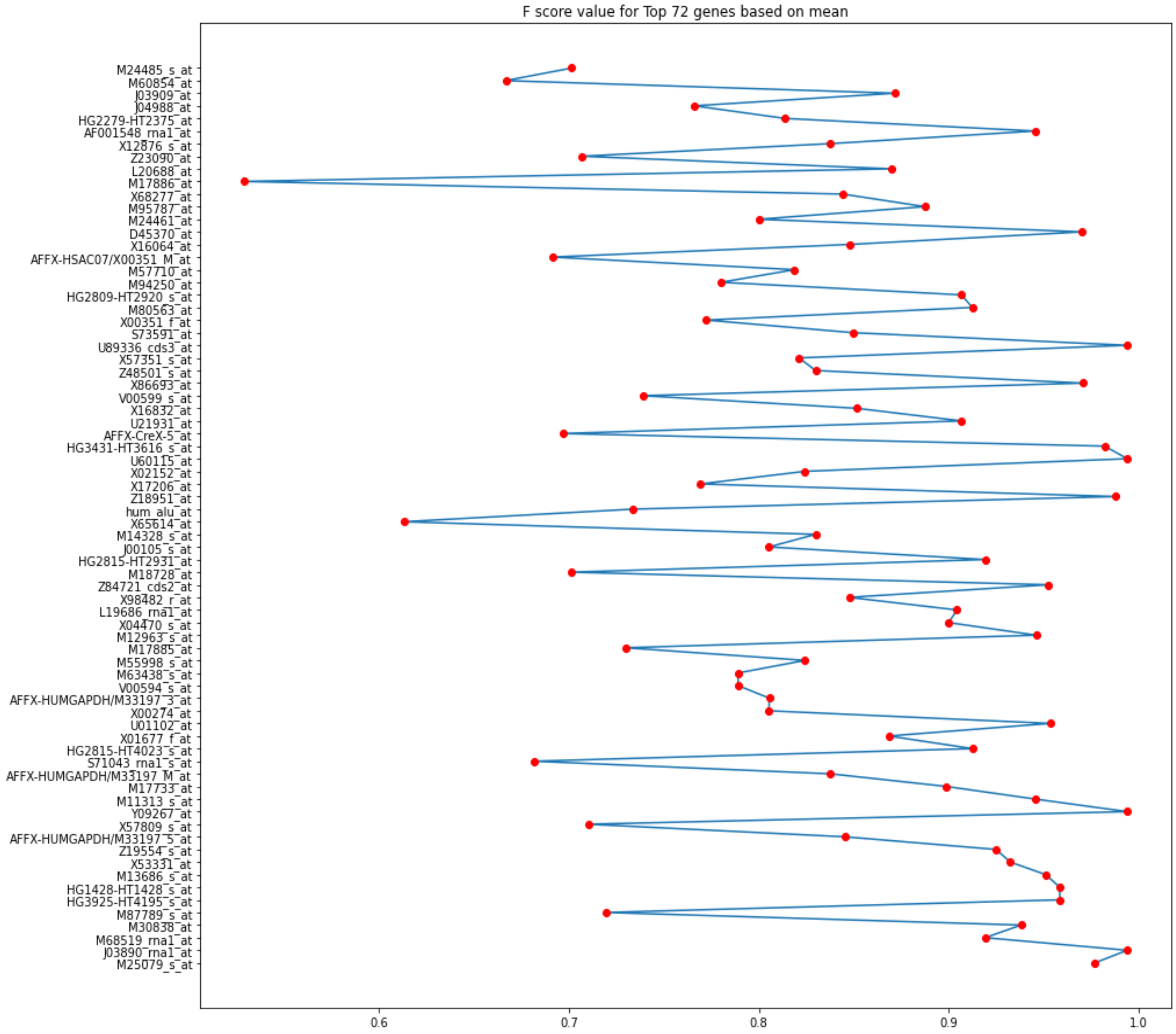


Fig 5. F score value for Top 72 genes using mean based approach

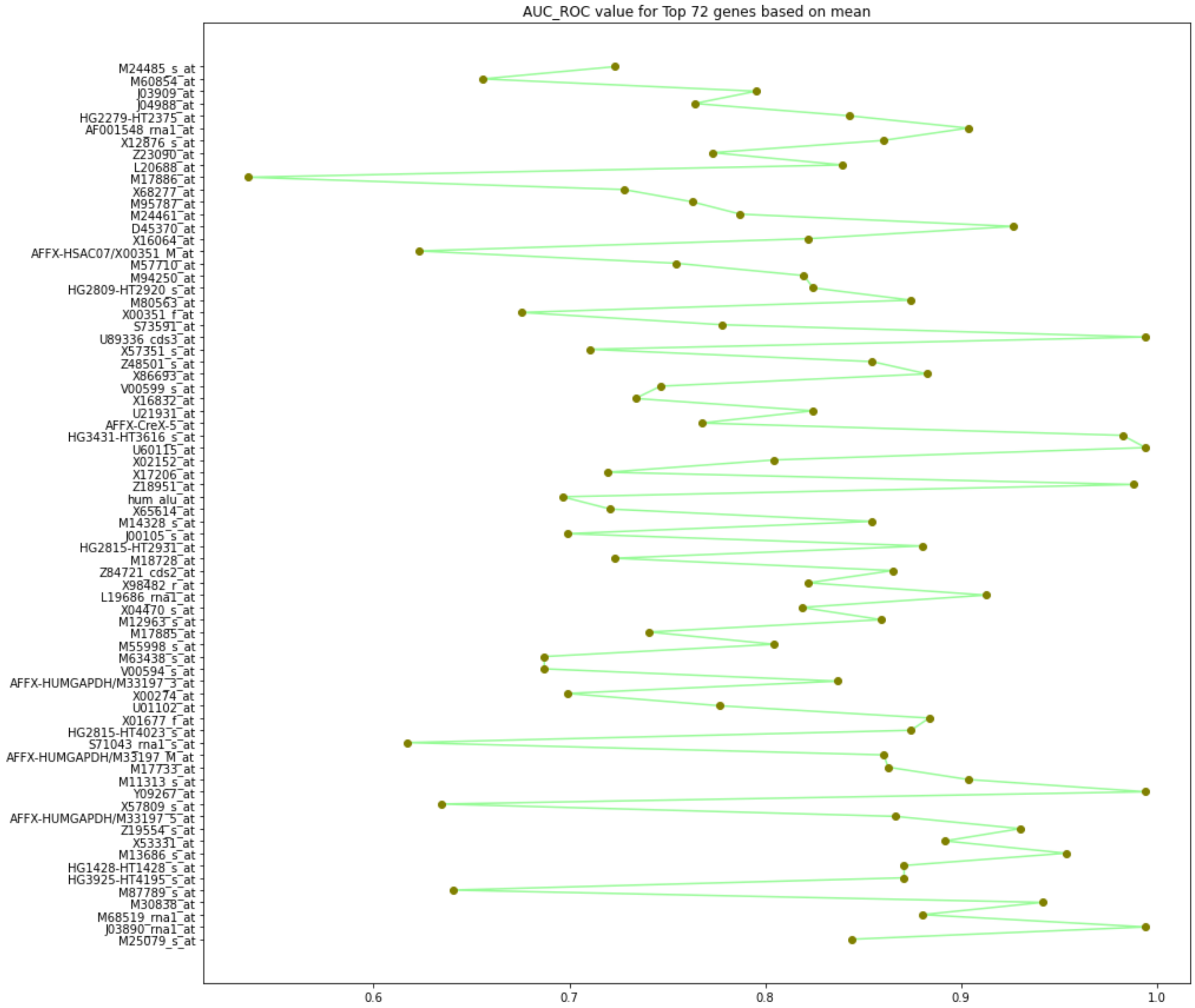


Fig 6. AUC_ROC Value for Top 72 Genes Using Mean Based Approach

Fig 6, shows a graph of AUC_ROC for each of the top 72 genes selected based on their mean value. These ROC scores are obtained by considering each gene individually for classification of the 96 samples available in the dataset.

4.1.2. Using Machine Learning Models

- **The Unbalanced Data Set and Sampling**

Count plot [21] is used to determine if the data is balanced or imbalanced. The count plot is shown in fig 7. Class 1 represents the cancerous samples and 0 represents normal non-cancerous samples. The original dataset has been shown in fig. 7. It can be seen we have in total 96 records and 86 of these records are tumorous while 10 are normal. It can be observed from the count plot fig. 7, the data set is highly imbalanced. This problem is resolved by using the Oversampling method which creates synthetic observations for the minority class using the available samples for the minority class. The balanced data set is the result of the RandomOverSampler function as shown in fig. 8.

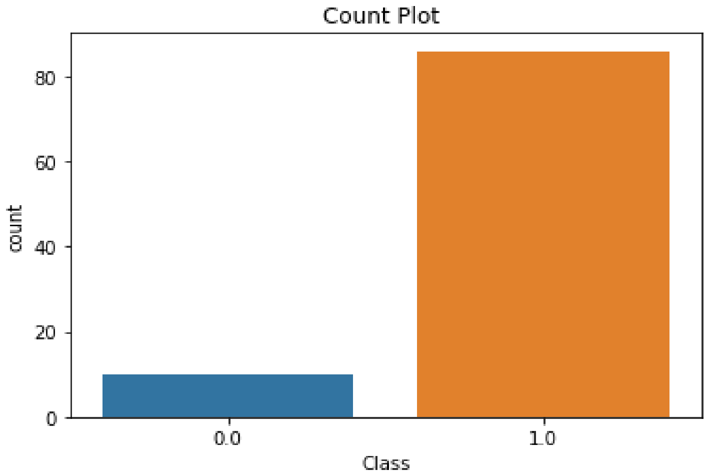


Fig 7 Count Plot for Data Set

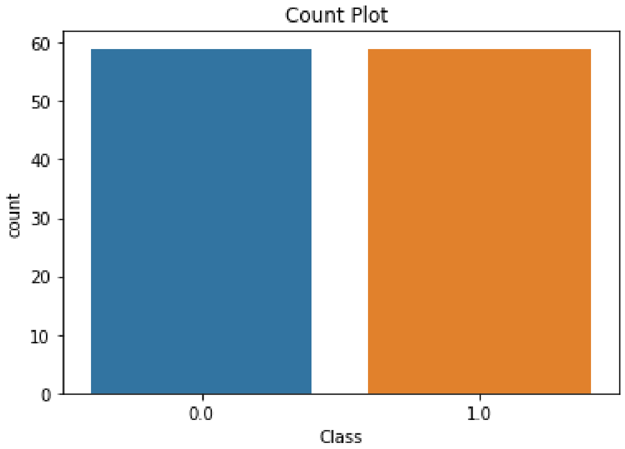


Fig 8 Training Dataset after Sampling

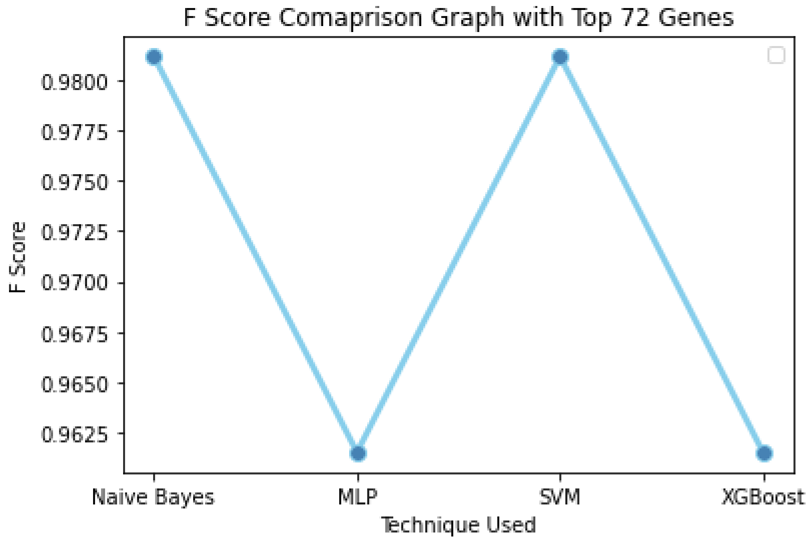


Fig 9 F Score Comparison Mean Based Techniques

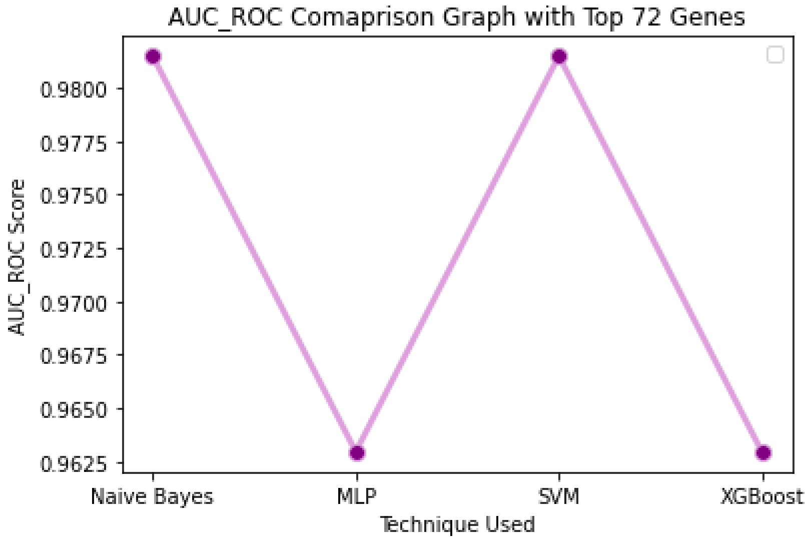
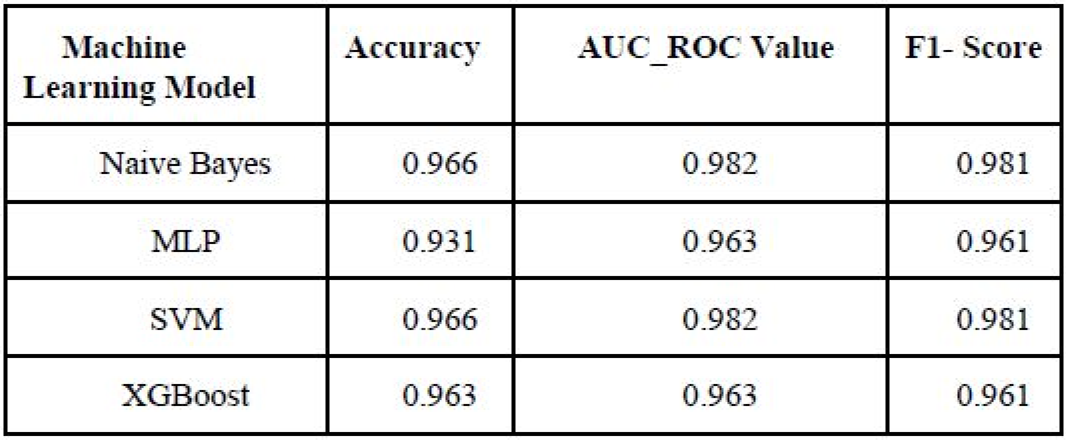


Fig 10 AUC_ROC Score Comparison Mean Based Techniques

From the above graphs, it can be observed that SVM and Naive Bayes perform better than MLP and XGBoost algorithms for gene selection using mean based approach.

The top 72 genes are used for prediction using various Machine Learning models. The results are tabulated in Table

Table 2 Test Results on genes selected by mean based approach



Machine Learning Model	Accuracy	AUC_ROC Value	F1- Score
Naive Bayes	0.966	0.982	0.981
MLP	0.931	0.963	0.961
SVM	0.966	0.982	0.981
XGBoost	0.963	0.963	0.961

4.2 CLASSIFICATION USING STANDARD DEVIATION APPROACH

The standard deviation based approach is the second statistical method used for gene selection. The score is calculated by dividing the difference in mean value by the sum of the standard deviation for cancerous and non-cancerous records

4.2.1. Using Threshold Value

Top 72 genes are selected using the standard deviation based approach discussed in chapter 3 section 3.4.2 out of a total 7129 genes. Following table gives the details of the top 72 genes and their threshold value.

Table 3 Genes selected using Standard Deviation Based Approach

Genes	Threshold	Score
X64559_at	946.7986047	2.36709624
U60115_at	2180.62	2.16363437
Z11793_at	1326.078837	2.153383053

X03350_at	1089.678256	2.062262626
Y09267_at	4581.148256	2.04338031
Z18951_at	2408.948721	2.025579805
J03890_rna1_at	10310.3457	2.004587238
U60060_at	214.1345349	1.978041916
M37984_rna1_at	680.9205814	1.935003584
U89336_cds3_at	1879.633837	1.915521563
M83186_at	896.4310465	1.873168746
L34657_at	895.7677907	1.850156288
D13628_at	426.3927907	1.842176911
J02874_at	718.4337209	1.823787135
M61906_at	271.8384884	1.797192798
U97105_at	1685.135	1.747871952
U76764_s_at	1033.339419	1.717242308
HG3431-HT3616_s_at	3215.057907	1.704994383
X86693_at	2534.212326	1.677663594
U39447_at	585.4111628	1.677016071
L27479_at	234.6773256	1.61659385
X62466_at	1126.765814	1.609783469

L15388_at	637.6872093	1.590883708
X85116_rna1_s_at	2579.647442	1.568286859
U48959_at	1010.765465	1.556857608
J02871_s_at	912.5082558	1.554101858
M25322_at	246.1662791	1.552189608
X05130_s_at	2799.409535	1.541329053
U03090_at	95.68709302	1.518361918
U13219_at	211.2360465	1.51009011
D13626_at	144.8395349	1.507745109
X72889_at	674.4923256	1.503573872
M11313_s_at	6961.035	1.502192473
L08895_at	314.3088372	1.490660197
X61118_rna1_at	262.3405814	1.489247331
HG2175-HT2245_s_at	372.8972093	1.478952929
U52100_at	1949.500814	1.465959088
D25304_at	382.7515116	1.46546604
M98833_at	109.4486047	1.454379736
M30838_at	11471.45081	1.450992721
L76380_at	367.9067442	1.445893251

M36284_s_at	409.8543023	1.443280638
U29171_at	937.7786047	1.440934021
M84526_at	1682.214535	1.437744031
U45973_at	543.4468605	1.436369828
S74017_at	1012.365814	1.426113102
U24488_s_at	234.4012791	1.424353992
M94856_at	1228.365581	1.420997945
M31210_at	114.7694186	1.416553564
D50683_at	2078.860465	1.408602307
U76421_at	273.5025581	1.406726998
M60721_at	372.1160465	1.404803197
U19247_rna1_s_at	870.6225581	1.401129355
D13168_at	102.8334884	1.399264427
M10321_s_at	814.0325581	1.398988608
L36531_at	636.0567442	1.392456765
M13686_s_at	7833.737326	1.390005018
Z37987_s_at	1014.120116	1.37979064
Z19554_s_at	14272.23384	1.370145396
U31384_at	534.4244186	1.369342819

M19722_at	797.5853488	1.363643993
M14338_at	391.3189535	1.361969309
Z84721_cds2_at	3525.503023	1.357441648
U25956_at	650.529186	1.355302331
U40038_at	198.8346512	1.350798135
HG1428-HT1428_s_at	7840.684302	1.348030255
D79990_at	513.7326744	1.346434865
M96803_at	838.9132558	1.334985851
U52101_at	1675.301395	1.334718636
AF001294_at	577.3125581	1.331914418
J03600_at	633.1619767	1.317902295
X54131_at	190.4876744	1.310509353

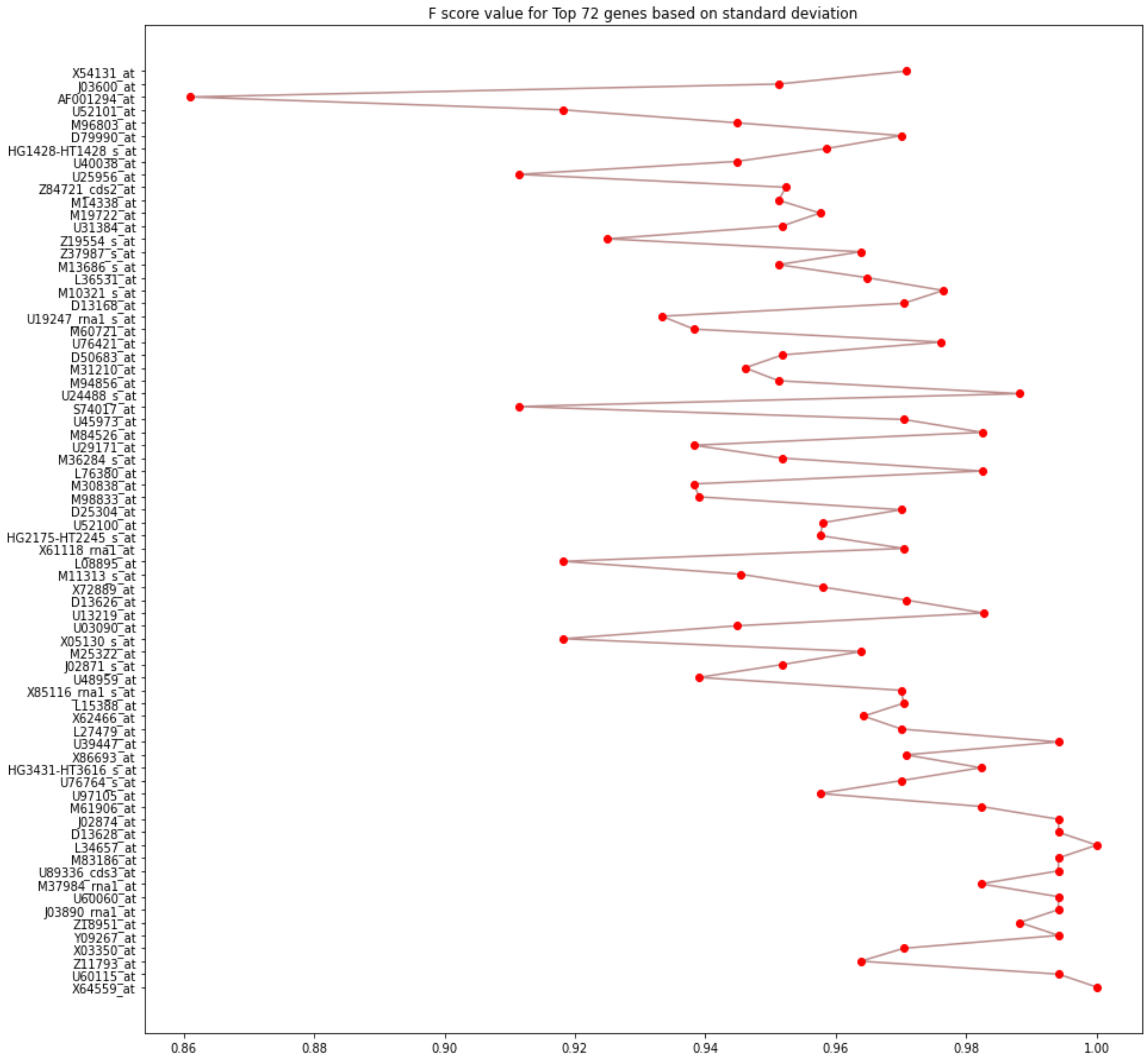


Fig 11. F score value for Top 72 genes using SD based approach

Fig 11, shows a graph of F score for each of the top 72 genes selected based on their standard deviation value. These scores are obtained by considering each gene individually for classification of the 96 samples available in the dataset.

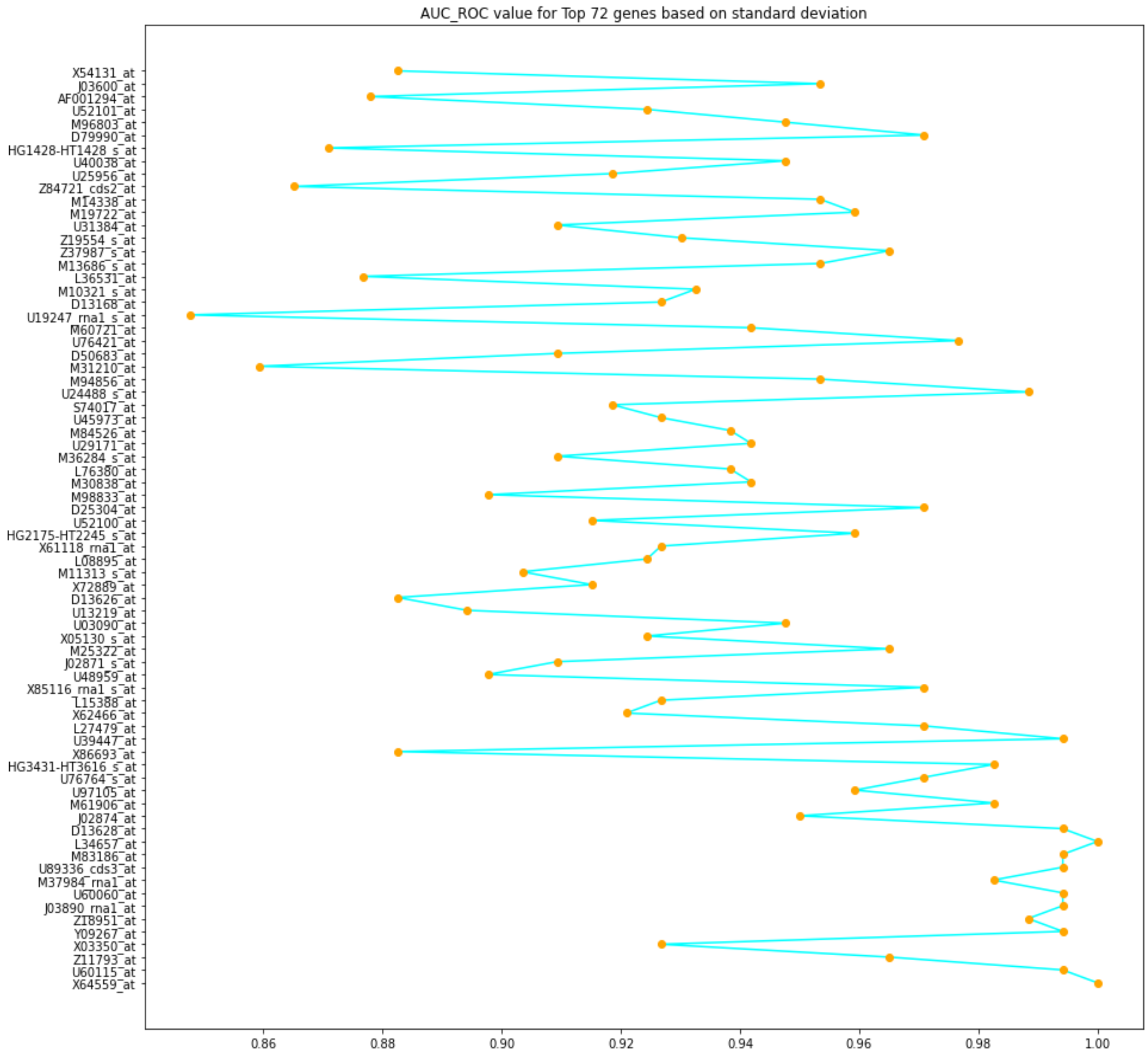


Fig 12. auc-roc value for Top 72 genes using SD based approach

Fig 12, shows a graph of auc_roc score for each of the top 72 genes selected based on their standard deviation value. These scores are obtained by considering each gene individually for classification of the 96 samples available in the dataset.

4.2.2. Using Machine Learning Models

Supervised Machine learning models such as GNB, Multilayer perceptron , SVM and XGBoost have been used for the classification using top 72 genes.

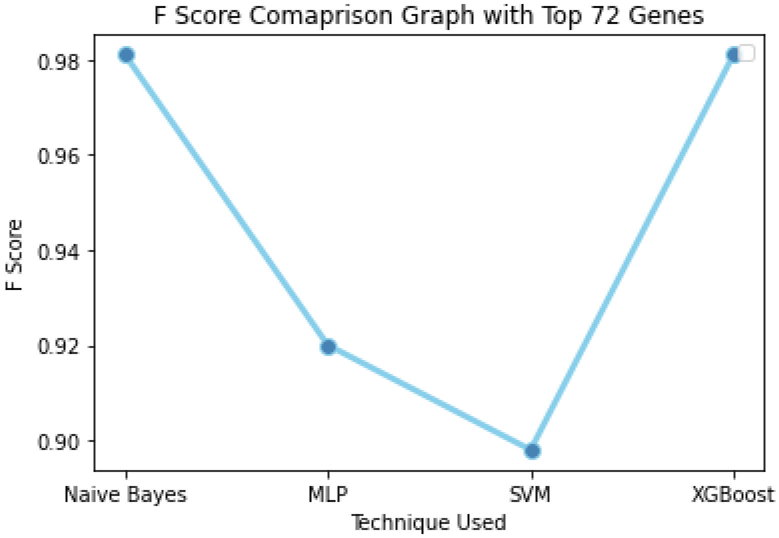


Fig 13 F score Comparison for Standard Deviation Based Approach

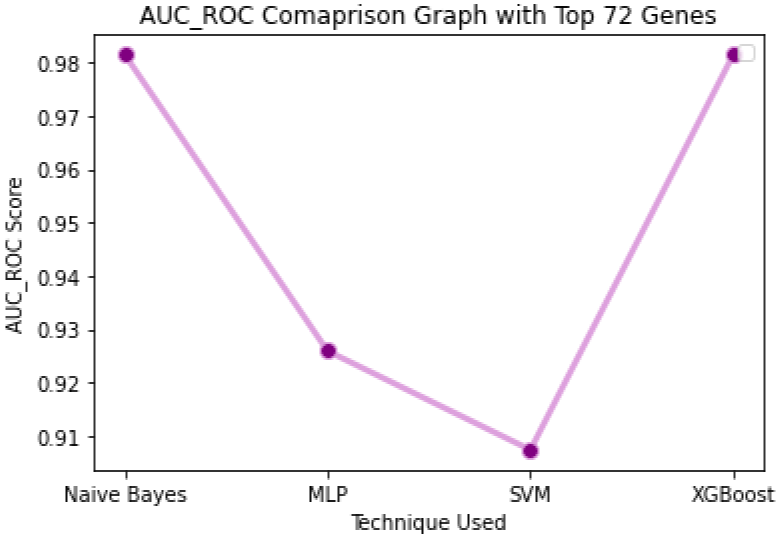
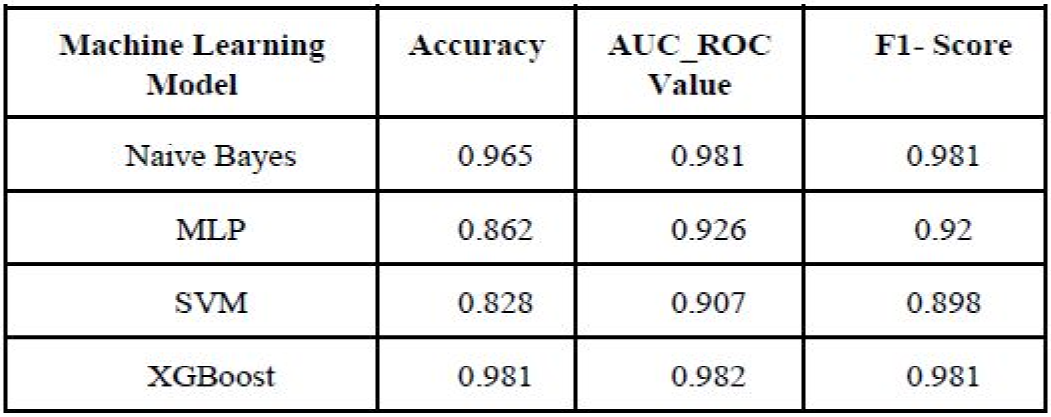


Fig 14 AUC_ROC score Comparison for Standard Deviation Based Approach

The top 72 genes are used for prediction using various Machine Learning models. The results are tabulated in Table

Table 4 Test Results on genes selected by SD based approach



Machine Learning Model	Accuracy	AUC_ROC Value	F1- Score
Naive Bayes	0.965	0.981	0.981
MLP	0.862	0.926	0.92
SVM	0.828	0.907	0.898
XGBoost	0.981	0.982	0.981

CHAPTER 5 CONCLUSION AND FUTURE SCOPE

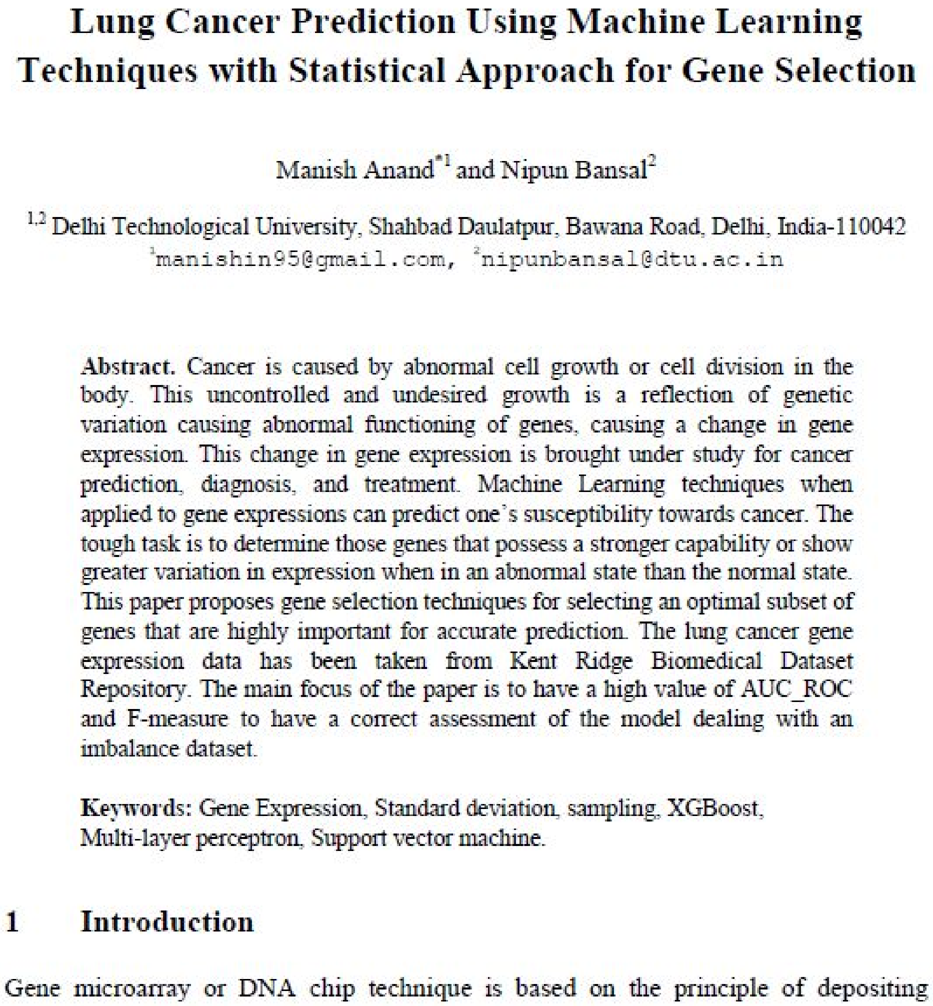
Cancer is a leading cause of death worldwide. According to WHO (World Health Organization) data, the most common type of cancer is lung cancer (2.09 million cases in 2018). The early detection of anomalies in the gene expression of an individual can be very effective in fighting such types of disease. This paper presents the analysis of gene expression data using supervised machine learning algorithms for lung cancer prediction.

The focus has been given on the gene selection process so that most informative genes are selected and noisy genes are excluded. Basic techniques of mean and standard deviation were used to determine the top 72 genes. There were several challenges like class imbalance in the dataset, which were overcome with techniques like sampling and using metrics that are better capable of assessing the models in this scenario. Seeing the results, it can be concluded that the object of the paper is met.

In future work, this approach can be taken forward to deep learning as well. We found that Y09267_at, J03890_rnal_at, Z18951_at, U60115_at, HG3431-HT3616_s_at, M11313_s_at, and M30838_at were the genes that came up as the most informative genes by our techniques. These genes appeared from computation, so the next step can be to check their biological relevance. Environmental factors affecting these genes and causing a change in the expression value can also be studied. Outliers in the data tend to affect statistical methods. Determining those outliers and their removal from the dataset can refine the results further.

APPENDICES

APPENDIX 1: LIST OF PUBLICATIONS (ACCEPTED & REGISTERED)



Lung Cancer Prediction Using Machine Learning Techniques with Statistical Approach for Gene Selection

Manish Anand*[1] and Nipun Bansal[2]

[1,2] Delhi Technological University, Shahbad Daulatpur, Bawana Road, Delhi, India-110042
[1]manishin95@gmail.com, [2]nipunbansal@dtu.ac.in

Abstract. Cancer is caused by abnormal cell growth or cell division in the body. This uncontrolled and undesired growth is a reflection of genetic variation causing abnormal functioning of genes, causing a change in gene expression. This change in gene expression is brought under study for cancer prediction, diagnosis, and treatment. Machine Learning techniques when applied to gene expressions can predict one's susceptibility towards cancer. The tough task is to determine those genes that possess a stronger capability or show greater variation in expression when in an abnormal state than the normal state. This paper proposes gene selection techniques for selecting an optimal subset of genes that are highly important for accurate prediction. The lung cancer gene expression data has been taken from Kent Ridge Biomedical Dataset Repository. The main focus of the paper is to have a high value of AUC_ROC and F-measure to have a correct assessment of the model dealing with an imbalance dataset.

Keywords: Gene Expression, Standard deviation, sampling, XGBoost, Multi-layer perceptron, Support vector machine.

1 Introduction

Gene microarray or DNA chip technique is based on the principle of depositing

REFERENCES

[1] Golub T. R., Slonim D. K., Tamayo P., et al. Molecular classification of cancer: class dis-covery and class prediction by gene expression monitoring. Science. 1999; 286(5439) :531–537. doi: 10.1126/science.286.5439.531.

[2] J Pharm Bioallied Sci. 2012 Aug; 4(Suppl 2):S310-2. doi: 10.4103/0975-7406.100283

[3] J. Pati, "Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach," in IEEE Access,vol. 7, pp. 4232-4238, 2019, doi: 10.1109/ACCESS.2018.2886604.

[4] Liu, Wei, et al. "Identification of genes associated with cancer progression and prognosis in lung adenocarcinoma: Analyses based on microarray from Oncomine and The Cancer Genome Atlas databases." Molecular genetics & genomic medicine 7.2 (2019): e00528.

[5] G. Russo, C. Zegar, and A. Giordano, ``Advantages and limitations of microarray technol-ogy in human cancer," Oncogene, vol. 22, no. 42, pp. 64976507, 2003.

[6] A. A. Shanab, T. M. Khoshgoftaar, R. Wald and A. Napolitano, "Impact of noise and data sampling on stability of feature ranking techniques for biological datasets," 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), Las Vegas, NV, 2012, pp. 415-422, doi: 10.1109/IRI.2012.6303039.

[7] Y. Sun, S. Todorovic, and S. Goodison, “Local-learning-based feature selection for high-dimensional data analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1610–1626, Sep. 2010.

[8] Li. (2002). Kent ridge bio-medical data set repository. Institute Info- comm Research. [Online]. Available: http://sdmc.lit.org.sg/GEDatasets/ Datasets.html

[9] D. Pavithra and B. Lakshmanan, "Feature selection and classification in gene expression

cancer data," 2017 International Conference on Computational Intelligence in Data Sci-ence(ICCIDS), Chennai, 2017, pp. 1-6, doi: 10.1109/ICCIDS.2017.8272668.

[10] Battiti, R 1994, "Using mutual information for selecting features in supervised neural net learning", IEEE Transactions on Neural Networks,vol. 5, no. 4, pp.537-550.

[11] Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS ONE 14(11): e0224365. https://doi.org/10.1371/journal.pone.0224365.

[12] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. 18, 1 (January 2017), 559–563.

[13] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and mod-el selection." Ijcai. Vol. 14. No. 2. 1995.

[14] Ahmed MS, Shahjaman M, Rana MM, Mollah MNH. Robustification of Naïve Bayes Classifier and Its Application for Microarray Gene Expression Data Analysis. Biomed Res Int. 2017;2017:3020627. doi:10.1155/2017/3020627.

[15] Du, K.-L & Swamy, M.N.s. (2014). Multilayer Perceptrons: Architecture and Error Back-propagation. 10.1007/978-1-4471-5571-3_4.

[16] Paliwal, M., and Kumar, U. A., Neural networks and statistical techniques: A review of applications. Expert Syst. Appl. 36(1):2– 17, 2009.

[17] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. Mach.Learn.20,3(Sept.1995),273–297. DOI:https://doi.org/10.1023/A:1022627411411.

[18] Kecman, Vojislav. (2005). Support Vector Machines – An Introduction. 10.1007/10984697_1.

[19] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge

Discov-ery and Data Mining (KDD ’16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI:https://doi.org/10.1145/2939672.2939785.

[20] Li, Wei et al. “Gene Expression Value Prediction Based on XGBoost Algorithm.” Fron-tiers in genetics vol. 10 1077. 12 Nov. 2019, doi:10.3389/fgene.2019.01077.

[21] Seaborn.countplot:https://seaborn.pydata.org/generated/seaborn.countplot.html.

[22] https://www.bioline.com/workflows/gene-expression-analysis.