International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Detection Analysis of Various Types of Cancer by Logistic Regression using Machine Learning

Heena Nankani, Shruti Gupta, Shubham Singh, S. S. Subashka Ramesh

Abstract: ***Cancer is now a day's one of the main diseases which has widely affected among the peoples. A molecular pathologist selects a list of genetic variations of interest that he/she wants to analyze. The molecular pathologist searches for evidence in the medical literature that somehow is relevant to the genetic variations of interest Finally this molecular pathologist spends a huge amount of time detecting the evidence which is related to each of the variations to classify them. The ultimate goal is to replace step 3 by a machine learning model. The molecular pathologist will still have to decide which variations area of interest, and also collect the relevant evidence. In this paper, we apply machine learning methods especially logistic regression (which is more accurate) on the datasets to determine and examine whether there are any signs or possibilities of cancer and if the person is examined as cancerous then the stage of cancer is also determined. Cancer disease is classified into four types named type 1, type 2, type 3 and type 4. Id, Gene, variation, and class are the fields used.***

Keywords: Cancer, Gene, variations, class, pathologist, machine learning model.

I. INTRODUCTION

Detection and analysis of the disease Cancer have always been a major issue for the pathologists and the medical practitioners for treatment planning. Tumor grade is a tumor based on how abnormal the tumor cells and the tumor tissue look under a microscope. It provides us an indication of the growth of the tumor. If the cells of the tumor and the organization of the tumor's tissue resemble those of normal cells and tissue, then the tumor is called as "well-differentiated "[1-3]. As compared with the undifferentiated" or "poorly differentiated," these tumors tend to have less capacity to grow and spread. Based on differences in microscopic appearance, doctors assign a numerical "grade" to most of the cancers. The factors used to determine tumor grade grow rapidly and spread faster than tumors with a lower grade. If a grading system for a tumor type is not specified, then the following system is generally used GX: Grade cannot be assessed (undetermined grade)
G1: Well differentiated (low grade)
G2: Moderately differentiated (intermediate grade
G3: Poorly differentiated (high grade)
G4: Undifferentiated (high grade)

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Heena Nankani, Student of SRMIST, Ramapuram, Chennai, India.
Shruti Gupta, Student of SRMIST, Ramapuram, Chennai, India.
Shubham Singh, Student of SRMIST, Ramapuram, Chennai, India.
Dr. S. S. Subashka Ramesh, Assistant Professor of Computer Science And Engineering, SRMIST, Ramapuram, Chennai, India.

Breast and prostate cancers are the most common types of cancer as they have their grading systems[1,3] .tumors present. Different cancers are classified concerning the grading system.. Tumors are graded as 1, 2, 3, or 4, depending on the amount of abnormality In Grade 1 tumor, the tumor cells and the structure of the tumor tissue appear close to normal. These tumors tend to grow and spread slowly. Normal cells and tissues don't match with grade 3 and grade 4 cells and tissues. These tumors tend to grow at a higher pace and spread faster than tumors as compared with a lower grade.

Breast cancer. Doctors mostly use the Nottingham grading system (also known as Elston-Ellis modification of the Scarff-Bloom-Richardson grading system) for breast cancer. This system grades breast tumors based on the following factors[1,3,4].
Tubule formation, that is, how much of the tumor tissue has normal breast (milk) duct structures
Nuclear grade: It is an evaluation of the size and shape of the nucleus in the tumor cells
Mitotic rate: The number of dividing cells present, which results in the speed of growing and dividing the cells.
Each of the categories gets a score between 1 and 3; the most abnormal. The scores for the three categories are added, having a total score of 3 to 9. Three grades are possible:

Total score = 3–5: G1 denotes low grade or well-differentiated.
Total score = 6–7: G2 denotes intermediate-grade or moderately differentiated
Total score = 8–9: G3 denotes high grade or poorly differentiated

Prostate cancer. Prostate cancer is graded by the Gleason scoring system. The Gleason score is based on biopsy samples that are taken from the prostate. The pathologist checks the samples to see how much similar the tumor tissue resembles normal prostate tissue. Both the primary and the secondary pattern of tissue organization are identified.
The primary pattern represents and signifies the most common tissue pattern that is seen in the tumor, and the secondary pattern represents the next pattern. Each pattern is given a grade from 1 to 5, with 1 resembling the most like prostate tissue and 5 resembling the most abnormal. Gleason score is obtained by adding the two grades. The American Joint Committee on Cancer adviced to group Gleason scores into the following categories:
Gleason X: Gleason score cannot be determined
Gleason 2–6: The tumor tissue is well-differentiated
Gleason 7: The tumor tissue is moderately differentiated
Gleason 8–10: The tumor tissue is poorly differentiated.

Doctors often use tumor grade and other factors, such as cancer stage and a patient's age and general health, to develop a treatment plan and to determine a patient's prognosis.

Generally, a lower grade indicates a better prognosis. Higher-Grade cancer may grow and spread more quickly and require immediate aggressive treatment.

The importance of tumor grade in planning treatment and determining a patient's prognosis is more for certain types of cancer, such as soft tissue sarcoma, primary brain tumors, and breast and prostate cancer. Patients should consult with their doctor for more information about tumor grades and moreover their relation with their treatment and prognosis.

II. LITERATURE SURVEY

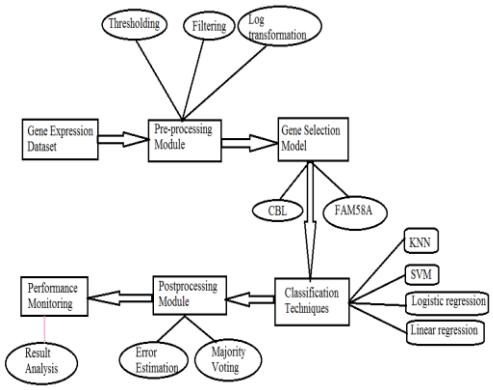
In recent years, few works have been reported in the literature for the design and development for the analysis and detection of the various types of cancer cells present. Various reports have been created and much hard work has already been performed for implementing the cancer detection model. In 2002[5], ISABELLE GUYON JASON WESTON STEPHEN BARNHILL says DNA micro-arrays now permit scientists to screen thousands of genes simultaneously and decide active hyperactive cancerous tissue. Because new micro-array devices generate bewildering amounts of raw data, new analytical methods must be developed to sort out whether cancer tissues have distinctive signatures of gene expression over normal tissues or other types of cancer tissues. In contrast with the baseline method, their method eliminates gene redundancy automatically and yields better and more compact gene subsets. In patients with leukemia(lack of white blood cells, (WBC)) our method discovered 2 genes that yield zero leave-one-out error, while 64 genes are necessary for the baseline method to get the best result (one leave-one-out error). Using 4 genes from the colon cancer database, this method is 98% accurate, while the baseline method is only 86% accurate. In 2003[2], Ain Choon tan and David gilbert In recent years, computational diagnostic tools andartificial intelligence techniques provide automate procedures for objective judgments by making use of quantitative measures and machine learning techniques. In this paper, they proposed a Support Vector Machines (SVMs) based classifier in comparison with Bayesian classifiers and Artificial Neural Networks for the prognosis and diagnosis of breast cancer disease. This paper was quite helpful for the classifiers provided with the implementation details and the corresponding results. Many comparative studies have been implemented regarding both the prognosis and diagnosis problem narrating the superiority of the proposed SVM algorithm in terms of accuracy. In 2007[7], Ilias Maglogiannis·Elias Zafiropoulos ·Ioannis Anagnostopoulos says In few years, computational diagnostic tools and artificial intelligence techniques provide automated procedures for objective judgments by making use of quantitative measures and machine learning techniques. In this paper, they proposed a Support Vector Machines (SVMs) based classifier in comparison with Bayesian classifiers and Artificial Neural Networks for the problem demonstrating the superiority of the proposed SVM algorithm in terms of accuracy. In 2009[1], Jagpreet Chhatwal1,2,3 Oguzhan Alagoz2 Mary J. Lindstrom says descriptors of the National Mammography Database using the algorithm logistic regression that can aid in decision making for the early detection of breast cancer. We created two logistic regression models based on the mammography features and demographic data for 62,219 consecutive mammography records from 48,744 studies in 18,270 patients had been reported using the Breast Imaging Reporting and Data System (BI-RADS) lexicon. State cancer registry outcomes were matched with the data served as the reference standard. Then it was found that the probability of cancer is found in both the models. Thus the variables of model 1 and the radiologist BI-RADs assessment were togetherly used for building Model2. We have used the algorithm 10-fold cross-validation to train and test the model and to calculate the area under the receiver operating characteristic curves (Az) to measure the performance. Both models were compared with the radiologists' BI-RADS assessments and accuracy was achieved. In 2015[4], Rajesh Kumar, Rajeev Srivastava, and Subodh Srivastava say a framework for automated detection and classification of cancer by microscopic biopsy images using clinically significant and biologically interpretable features is proposed and examined. The various stages involved in the methodology include enhancement of microscopic images, segmentation of background cells, features extraction, and finally the classification. An appropriate and efficient method is employed in each of the design cells, features extraction, and finally the classification. An appropriate and efficient method is employed in each of the steps of the proposed framework after making a comparative analysis of the commonly used method in each category. In 2017[3], Alaá Rateb Mahmoud Al-shamash, Ph.D. Unaizah Hanum Binti Obaidellah says that Cancer is the general name for a group of more than 100 diseases. Cancer is a mixture of different diseases but they all start because abnormal cells grow out of control. Without treatment, cancer can cause serious health issues and even loss of life. This paper represents a review of the detection methods for lung, breast, and brain cancers. Artificial intelligence techniques, such as support vector machine, linear and logistic regression, neural network, artificial neural network, fuzzy logic, and adaptive neuro-fuzzy inference system, with medical imaging like X-ray, ultrasound, magnetic resonance imaging, and computed tomography scan images. Imaging techniques are the most important and vital achievements for human cancer diagnosis. This technique was widely investigated to identify a method that provides accuracy and determine the best medical images for use in each type of cancer. In lung cancer, PET images presented a higher accuracy of 97%, which was achieved using an SVM classifier, and CT scan obtained the most accurate result of 96.04%, which was achieved using feedforward backpropagation. For brain cancer, an MRI scan yielded the most accurate result of 100%, which was achieved using a PPN. In 2018[6], Naresh Khuriwal Nidhi Mishra says According to Breast Cancer Institute (BCI), Breast Cancer is one of the most dangerous types of diseases which is very effective for women in this world. As per clinical experts detecting cancer in its first stage helps in saving lives.

As per cancer.net offers individualized guides for more than 120 types of cancers and related hereditary syndromes. Machine learning techniques are used mostly for analysis of cancer. In this paper, they have proposed an adaptive ensemble voting method for diagnosed breast cancer using the Wisconsin Breast Cancer Database. This work aims to compare and explain how ANN and logistic algorithm provide a better solution when its work with ensemble machine learning algorithms for diagnosing breast cancer even the variables are reduced. When compared to related work from the literature. It is shown that the ANN approach with a logistic algorithm is achieved 98.50% accuracy from another machine learning algorithm. In 2019[8], Neha Kumari and Khusbhu Verma say about the increased amount of breast cancer disease in women. Mammography is the technique that is used to identified the cancer tumor of the breast using x-rays. The paper discusses the comparative analysis of some of the existing cancer detection approaches. The paper discusses the early detection of cancer boosts the increase of survival chance to 98%. The algorithms used for the cancer diagnosis in this are Naïve Bayes classification, K nearest Neighbor Classification, Multilayer Perceptron and support vector machine. A Naïve Bayes classifier which is based on Naïve hypothesis. The advantage of this classifier is that it can be trained on a small data set and gives a result very fast.

III. METHODOLOGY

Figure 1 shows the flowchart illustrating the whole framework for the detection and analysis of different types of cancers using machine learning. In this process, the collection of gene data set is the foremost process that leads along with the pre-processing module which has been carried out in three processes i.e., filtering, thresholding, log transformation. Further, in the gene selection model, the genes that cause cancer are CBL and FAM58A[5]. Thus the machine is trained using 4 types of classifications techniques that are mainly KNN, Logistic Regression, SVM and linear regression[3]. In the postprocessing module, the techniques which gave the most accurate result were determined based on log values and thus linear and logistic regressions were proved to be most useful. The error estimation and majority voting were properly counted and determined. The performance of both the techniques was monitored and the logistic regression technique provided more accuracy as compared to linear regression with less error percentage.

ARCHITECTURE WORKFLOW DIAGRAM



Using the following dataset the machine was trained with these input samples. The dataset includes ID, the type of GENE, VARIATION, the type of CLASS or the stage of cancer[3,7].

The various machine learning techniques that we have Used are as follows :

LINEAR REGRESSION-

Linear Regression is based on supervised learning which stands as a great machine learning algorithm. Its basic work is to perform the regression task. It is mostly used to find out the relationship between variables and forecasting. There are different regression models that either shows the relationship between dependent and independent variables or the number of independent variables being used. Considering two values X and y, linear regression thus predicts the independent variable Y with respect to dependent variable X. So, the regression technique finds out a linear relationship between x (input) and y(output).

Hence, the term is Linear Regression. The equation of the linear regression is- $Y=\theta 1 + \theta 2.x$

x: input training data (univariate- one input variable (parameter)) , **y:** labels to data (supervised learning).

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.**θ_1:** intercept, **θ_2:** coefficient of x After getting the best θ_1 and θ_2 values, we get the best fit line. After using the model for prediction, it will give the output for y for each input of x,

ALGORITHM:

Step1: Load all the necessary libraries -numpy, pandas, matplotlib

Step 2:Load all the datasets in the machine.

Step 3:Input and output Data

Step 4: Perform normalization

Step 5:Split the dataset into the train set and test set.

Step 6: calculate z value.

Step 7: Perform prediction based on the given dataset and check the accuracy of the technique.

K-NN –

The KNN algorithm assumes that similar things do exist in close positions. In other words, similar things are close to each other.KNN Algorithm works on the concept of **feature similarity**: How closely out-of-sample features resemble our training set determines how we classify a given data point.

A selected object is classified through the vote given by its neighbors, which are the k-nearest neighbors. These neighbors thus can also be used for regression which gives the values of the object as the output. The final value is the average of its k-nearest neighbors.

ALGORITHM FOR KNN:

1. Load the data.
2. Initialize K to your chosen number of neighbor
3. For each example in the data

3.1 Distance between the query and the current example is calculated.

3.2 Distance and index of the example are calculated.

4. The ordered collection of the distance and index are arranged in ascending order.
5. From the collection first k entries are selected
6. Labels of the entries are selected.

7. The mean of k values is passed if regression is found.

LOGISTIC REGRESSION-

Logistic regression is the machine learning technique for the function used at the core of the method, the logistic function.

The logistic function, also known as the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. The logistic function equation is given as:

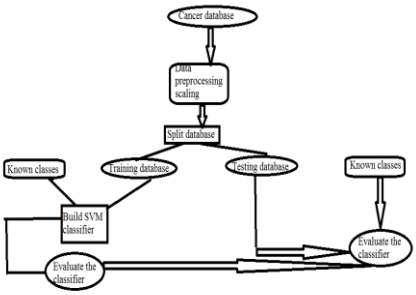
1 / (1 + e^-value)

Where e is the base of the natural logarithms and value is the actual numerical value that we want to transform.

SUPPORT VECTOR MACHINE (SVM)-

"Support Vector Machine" (SVM) is a machine learning algorithm which mainly focuses on classification challenges of the data. It is mostly used in classification problems. In n-dimensional space, each data item is plotted as a point. The two classes are differentiated bt finding a hyper-plane. It works as follows:

Step 1: Identify the right hyper-plane (Scenario-1)
Step 2: Identify the right hyper-plane
Step 3: Identify the right hyper-plane
Step 4:Can we classify two classes
Step 5:Find the hyper-plane to segregate to classes



IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed methodologies were implemented using classification techniques that mainly support vector machine (SVM), Logistic regression,knn mapping, and Linear regression. Firstly, the dataset is loaded into the machine.The dataset is basically an open CSV file. Then a graph is plotted between class and value count as X-axis and Y-axis respectively.

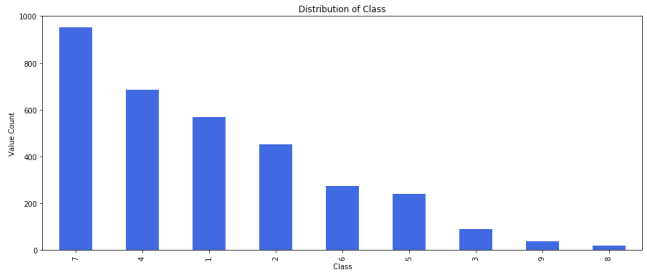


Figure 3. Value Count vs Class

Abstract background Non-small cell lung cancer (NSCLC) is a heterogeneous group of disorders with several genetic and proteomic alterations. c- CBL is an E3 ubiquitin ligase and adaptor molecule important in normal homeostasis and cancer. We determined the genetic variations of c-CBL, relationship to receptor tyrosine kinases and functionality in NSCLC. Methods and Findings Using archival formalin-fixed paraffin-embedded (FFPE) extracted genomic DNA, we show that c-CBL mutations occur in somatic fashion for lung cancers. c-CBL mutations were not mutually exclusive of MET or EGFR mutations; however, they were independent of p 53 and KRAS mutations. In normal/tumor pairwise analysis, there was a significant loss of heterozygosity (LOH) for the c-CBL locus (22%, nâ€Š=â€Š8/37) and none of these samples revealed any mutation in the remaining copy of c-CBL. The c-CBL LOH also positively correlated with EGFR and MET mutations observed in the same samples. Using select c-CBL'

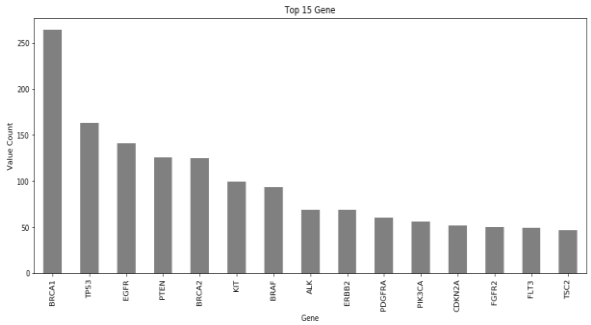


Figure 4. Valuecount vs gene

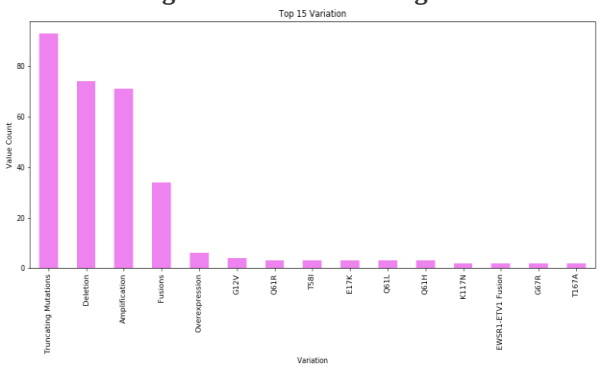


Figure 5.Value count vs variation

Fields are

- **ID:** the id of the row used to link to do mutation to the clinical evidence
- **Gene:** the genesis where this genetic mutation is located
- **Variation:** the amino acid change for these mutations
- **Class:** 1-9 the class this genetic mutation has been classified on

abstract background non small cell lung cancer nsclc heterogeneous group disorders number genetic proteomic alterations c cbl e3 ubiquitin ligase adaptor molecule important normal homeostasis cancer determined genetic variations c cbl relationship receptor tyrosine kinases met functionality nsclc methods findings using archival formalin-fixed paraffin-embedded ffpe extracted genomic

DNA show c cbl mutations occur somatic fashion lung cancers c cbl mutations mutually exclusive met egfr mutations however independent p53 Kras mutations normal tumor pairwise analysis significant loss heterozygosity loh c cbl locus 22 n 8 37 none samples revealed mutation remaining copy c cbl c cbl Loh also positively correlated egfr met mutations observed samples using select c cbl somatic mutations s80n h94y q249e w802 obtained caucasian Taiwanese African American samples respectively transfected SCLC cell lines increased cell viability cell motility conclusions taking overall mutation rate c cbl co' Further, we split the data set into a training set and testing set using the train_test_split library from SK learn. The data set is split into training and testing set on the test size of 0.2 or it can be said as that 80% of the data set belongs to the training set and 20% of the data set belongs to the testing data set. The following keywords were introduced as x train, ,y train, x test, and y test.

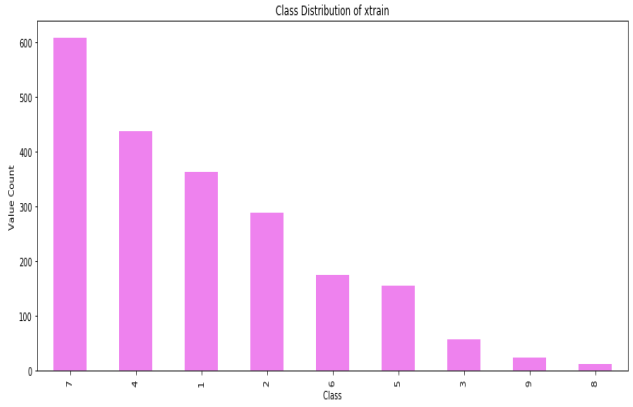


Figure 6. x train graph

% of data points belonging to Class 7 is : 28.6723 %
% of data points belonging to Class 4 is : 20.6685 %
% of data points belonging to Class 1 is : 17.0904 %
% of data points belonging to Class 2 is : 13.6064 %
% of data points belonging to Class 6 is : 8.2863 %
% of data points belonging to Class 5 is : 7.2976 %
% of data points belonging to Class 3 is : 2.6836 %
% of data points belonging to Class 9 is : 1.1299 %
% of data points belonging to Class 8 is : 0.565 %

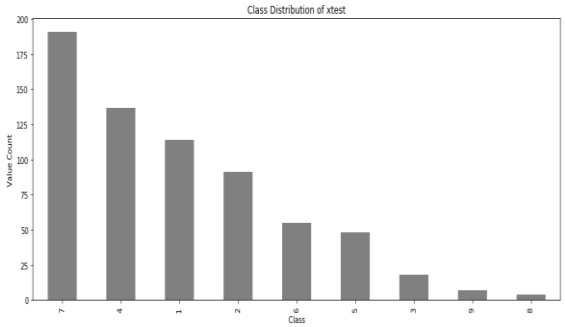


Figure 6.x test graph

% of data points belonging to Class 7 is : 28.7218 %
% of data points belonging to Class 4 is : 20.6015%
% of data points belonging to Class 1 is : 17.1429%
% of data points belonging to Class 2 is : 13.6842% % of data points belonging to Class 6 is : 8.2707 %
% of data points belonging to Class 5 is : 7.218 %
% of data points belonging to Class 3 is : 2.7068 %
% of data points belonging to Class 9 is : 1.0526 %
% of data points belonging to Class 8 is : 0.6015 %

All the data (test, train, cross validation) follow the same order of distribution of classes
(Highest data points) Class : 7 > 4 > 1 > 2 > 6 > 5 > 3 > 9 > 8 (Lowest data points)
Uni variate analysis of Genes
Number of Unique Genes: 234
BRCA1 172
TP53 112
EGFR 93
PTEN 86
BRCA2 85
KIT 70
BRAF 50
ERBB2 44
PDGFRA 43
PIK3CA 41
Name: Gene, d type: int64
% of data points in Test and CV data sets that are covered by the 234 genes in Train data set
Test data 645: 96.99248120300751 %
CV data 513: 96.42857142857143 %

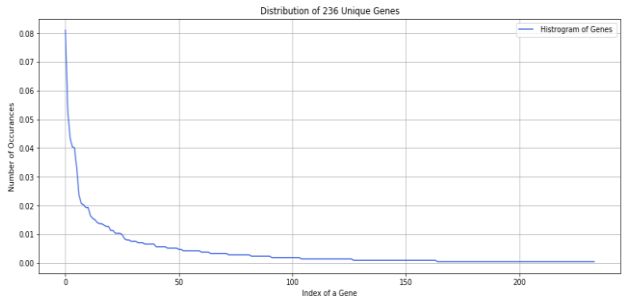
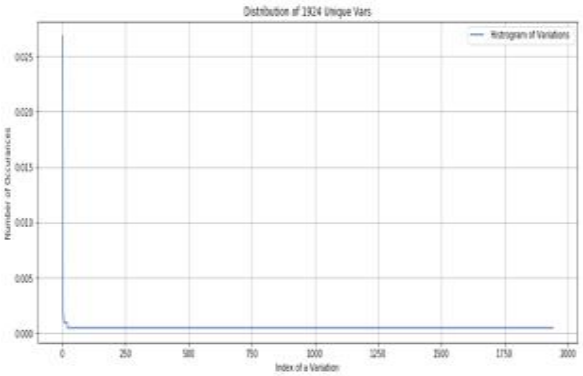


Figure 8. Distribution of genes

Uni variate Analysis Variation
Number of Unique Variation: 1942
Truncating_Mutations 57
Amplification 47
Deletion 41
Fusions 19
Over expression 5
Q61L 3
Q61R 3
Promoter_Hypermethylation 2
M1R 2
EWSR1-ETV1_Fusion 2
Name: Variation, d type: int64



Finally, after calculating all the train log loss and test log loss we found logistic regression is the best model with a Test log loss of 0.9257.

MODEL	TRAIN LOG LOSS	TEST LOG LOSS
Logistic regression	0.5120	0.9257
Logistic regression(balanced weight)	0.5179	0.9582
Linear SVM	0.5460	0.97

V. CONCLUSION

The contribution of our study is to prove the feasibility and effectiveness of machine learning for the detection and analysis of different types of cancer. In this paper, we employed simple classification techniques that are mainly KNN, SVM, Logistic Regression, Linear regression. While performing this machine learning model for cancer detection, we got the most accurate results in logistic regression. Thus in the future, we will perform and implement fewer ideas for more easy conductance for this Machine learning model.

REFERENCES

1. Neha kumara and Khushi Verma, Bansal institute of science and technology, Volume 10, May-June 2019." A survey on various machine learning approaches used for breast cancer detection."
2. Rajesh Kumar, Rajeev Srivastava, and Subodh Srivastava Department of Computer Science and Engineering, Indian Institute of Technology (BanarasHinduUniversity), Varanasi221005," Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features"
3. Alaá Rateb Mahmoud Al-shamash, Ph.D. Unaizah Hanum Binti Obaidellah, Ph.D. University of Malaya, Malaysia," Artificial Intelligence Techniques for Cancer Detection and Classification: Review Study"
4. Jagpreet Chhatwal1,2,3 Oguzhan Alagoz2 Mary J. Lindstrom4 Charles E. Kahn, Jr.5 Katherine A. Shaffer5 Elizabeth S. Burnside1,2,4Chhatwal J, Alagoz O, Lindstrom," A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis."
5. ISABELLE GUYON isabelle@barnhilltechnologies.comJASON WESTON STEPHEN BARNHILL Barnhill Bioinformatics, Savannah, Georgia, USAVLADIMIR VAPNIK vlad@research.att.com AT&T Labs, Red Bank, New Jersey, USA," Gene Selection for Cancer Classification using Support Vector Machines".
6. Naresh Khuriwal Nidhi Mishra Department of Computer Engineering Department of Computer Engineering Poornima University Jaipur, Poornima University Jaipur Jaipur, India Jaipur, India naresh.khuriwal89@gmail.com, nidhi.mishra@poornima.edu.in, "Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm"
7. Ilias Maglogiannis·Elias Zafiropoulos· Ioannis AnagnostopoulosPublished online: 12 July 2007 © Springer Science+Business Media, LLC 2007," An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifier".
8. Aik Choon tan and David Gilbert, Bioinformatics research center, Department of computing science, University of Glasgow, Glasgow, UK" Ensemble machine learning on gene expression data for cancer classification".

AUTHORS PROFILE

Heena Nankani : Student of SRMIST, Ramapuram, Chennai, persuing btech, CSE department III year. High school in Sarada Vidya Mandir, West Bengal. Currently in the progress of learning Machine learning and implementing more projects. Got Interned in Chennai developed projects and gained experience of working in II year of engineering. Mail ID- heenanankani32@gmail.com

Shruti Gupta Student of SRMIST, Ramapuram, Chennai, persuing btech, CSE department III year. High school in Brij Bhushan Lal Public School, Bareilly,Uttar Pradesh. Currently in the progress of learning Machine learning and implementing more projects. Got Interned in Motilal Nehru National Institute Of Technology,Prayagraj and developed projects based on big data and networking and gained experience of working in II year of engineering. Member of "MY CAPTAIN"(Sustainable Development Solutions Network by IIM,Bangalore) Mail ID- shruti0970@gmail.com LinkedIn-https://www.linkedin.com/in/shruti-gupta-631146168

Shubham Singh Student of SRMIST, Ramapuram, Chennai, persuing btech, CSE department III year. High school in Army Public School , Lucknow , Uttar Pradesh. Got Interned in IIT,BHU and developed project on machine learning named "DIABETIC PREDICTION" and gained experience of working in II year of engineering. Mail ID- shubham4181@gmail.com

Dr. S. S. Subashka Ramesh: Assistant professor of Computer Science And Engineering at SRMIST, Ramapuram ,Chennai .Persued M.E , Ph.D.(Anna University) MailId –subashka@gmail.com