ISSN No:-2456-2165

Detection of Breast-Cancer by Logistic Regression Using Machine Learning

Juhi Seth
Department-computer science engineering
SRM Institute of Science and Technology,
Chennai, Tamil Nadu

Vishal Srivastava
Department-computer science engineering
SRM Institute of Science and Technology,
Chennai, Tamil Nadu

Abstract:- Cancer is one of the major leading problems in the world. Around millions of people are being infected with cancer. There is a large number of cancer patients in this world. Cancer is a group of various diseases involving abnormal cell clustering around a region which can further increase to other body parts rapidly. These can lead to tumors, which further doesn't spread in body. Pathologists selects a list of genetic variations from the area they have to analyze. After a huge span of time of detecting the piece or structure and analyzing the classes, the model is developed for predicting the cancer patients. In this paper, logistic regression model is used on the datasets to examine the possibility of the stages of the cancer.

Keywords:- *Breast-cancer, genetic classes, variations, Logistic Regression.*

I. INTRODUCTION

Cancer disease have been a major issue in the medical science for the pathologists in determining and analyzing of the disease and for treatment. Tumor is based on the abnormal cells and tissues that are observed under the microscope for the further detailing of the cell types. It provides an indication of the tumor cells growth. [1-3] Most of the cancers are caused due to intake of tobacco which causes around 22% of cancer deaths. While another 10% are due to the cause of obesity, poor diet, physical fitness lack, and excessive of drinking of alcohol. Cancer can also be due to genetic defects from parents to person. Many cancers can be detected by a certain signs and symptoms or screening test, even with the help of image processing by an expert of medical team. Early detection of cancer, can useful for cervical and colorectal cancer. Breast cancer are controversial. Breast cancer develops from breast tissues. A lump in the breast, or change in the breast shape, dimpling of skin are some of the signs of breast cancer.

Breast and prostate cancers are the one of the most common types of cancer as they have grading from 1 to 3 that is low grade, intermediate grade, high grade. Doctors mostly use Elston-Ellis modification grading system of Scarff-Bloom-Richardson for analyzing of breast cancer. The importance of tumor grade system is in determining the type of class or stage of the cancer in curing for it. Patients must be aware and must consult the doctor for further information about tumor cells and their treatment and curability.

II. LITERATURE SURVEY

[1] This paper proposes a logistic regression method which can aid in decision making for the early detection of breast cancer. This paper uses image processing techniques for the data and figures for comparison. Around 62,220 records from 48,745 studies in 18,200 patients have reported using breast imaging reporting and data system lexicon.

[2] This paper describes about the computational diagnostic tools and ai techniques which provide automate procedures for the objective results by making use of machine learning techniques. This paper uses support vector machine technique for classify with Bayesian classifier and artificial neural networks for the diagnostic of breast cancer. This paper provides a proper implementation details and results.

[3] This paper describes cancer as a general name for a group of more than 100 diseases. Cancer is described as a mixture of different diseases which start from an abnormal growth of cells and grows out of control. They say that without any proper treatment of cancer, it may cause to serious health problems and issues and can even be the cause of death. A review of how the lung, breast and brain cancers are detected is shown by this paper. Artificial intelligence techniques along with support vector machine, linear and logistic regression, neural network is used. Image techniques are the most common and important tool for human cancer diagnosis. This technique was widely investigated in the identification of cancer type.

[4] They proposed a framework for automatic detection and classification of cancerous tissues by microscopic biopsy images using clinically significant and biologically interpretable features is proposed and examined. There are number of stages in the detection of tissues are as follows enhancement of microscopic images, segmentation of the background cells, extraction of features visible, and finally the classification from the features extracted. An efficient approach is employed in each of the stages in the classification of the cancer. The model depicts the features that are necessary in the cancer tissues.

[5] In 2002, Jason Weston says that DNA micro-arrays permits scientists to the thousands of genes screens simultaneously and it describes and decides the active hyperactive tissues. A new analytical method must be designed for analyzing the cancerous tissues from the raw

data. If the patient is suffering from leukemia, 2 genes are discovered that yield zero leave-one-out error, while 64 genes are necessary for baseline to get effective results. An accurate of 98% efficiency is obtained using 4 genes.

[6] In this paper, according to breast cancer institute, breast cancer is defined as one of the most dangerous types of diseases which is most common in women in this world. Detecting cancer in its first stage helps in saving lives is proven by clinical experts. This paper proposes an adaptive voting method for the diagnostic breast cancer using breast-cancer database. This aims to provide an explanation of how ann and logistic algorithm provide a better solution when its works with ensemble machine learning algorithms for diagnosis breast-cancer.

III. METHODOLOGY

The flowchart in the figure 1, shows the overall framework for the detection and analysis of different types of cancers using machine learning algorithms. The process states that, the gene data set is collected as the first process and further leads with the data cleaning. Data cleaning is the process of removing useless data from the data set and removing null values from the data set. Next process involves of data preprocessing module, which is carried in three different steps, the first and formost step is filtering, next is thresholding and last is log transformation. Further, genes that causes cancer i.e CBL and FAM58A[5] is used for gene selection. To classify the types of cancer four algorithms are used, that are KNN, linear regression, logistic regression, and svm[3]. The model is trained and tested with various models and the best accurate model is used for traing the model based on the log values. The error estimation and majority voting were properly counted and determined.

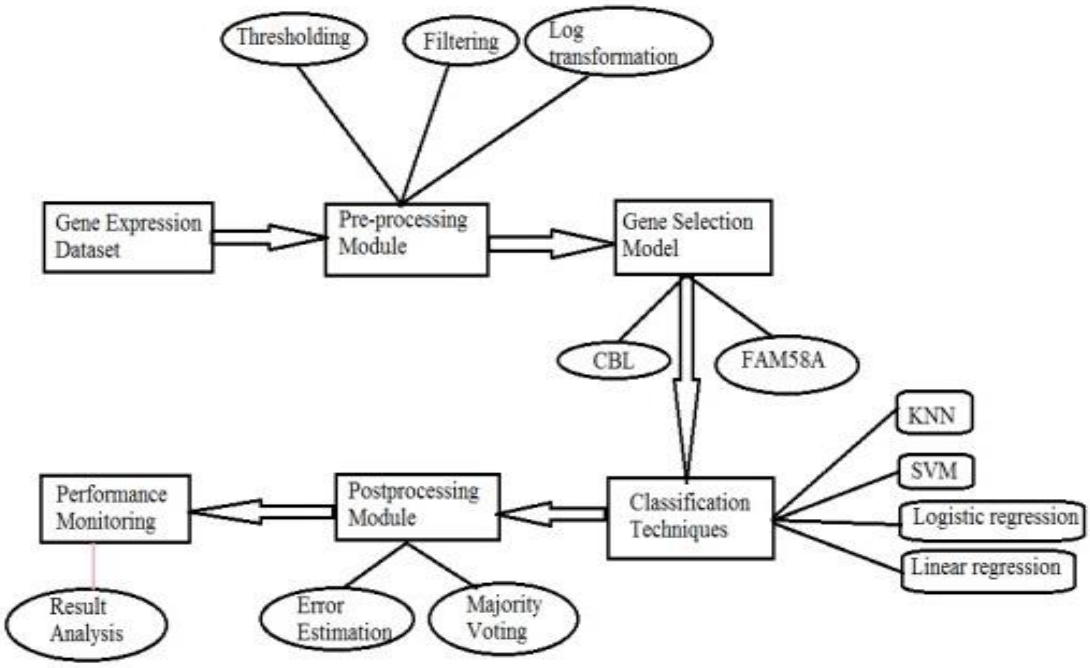

Fig 1:- Architecture Diagram

With the following dataset, the machine is trained with the input samples provided. The dataset consists of ID column, which describes the id of each patients and this ID is unique for all the patients, GENDER column, which describes the gender of the patient i.e. Male or Female, GENE describes the type of genes of cancer, CLASS describes the type of cancer stage [3,7]. Linear Regression algorithm is a type of supervised learning. Linear regression works on the linear relationship between two values x and y. Linear regression technique predicts that's the independent variable y with respects to dependent variable x. Here input is described by x variable while output is described as y variable. The linear regression divides the plane in two parts and it finds out the relation between the input and the slope. The equation of linear regression is, $y = m + cx$ where y is the output and m is the slope of the line, c is the intercept. KNN algorithm is used for the prediction of the close points in the graph. KNN algorithm works on the concept of feature similarity. It closely examines the features that resemble the training set and examine how feature is classified on a given data point. Logistic Regression is the technique for the function used at the core of logistic function. The sigmoid function was developed by statisticians to describe the property of the population growth in ecology. Its 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 actual numerical value that we want to transform.

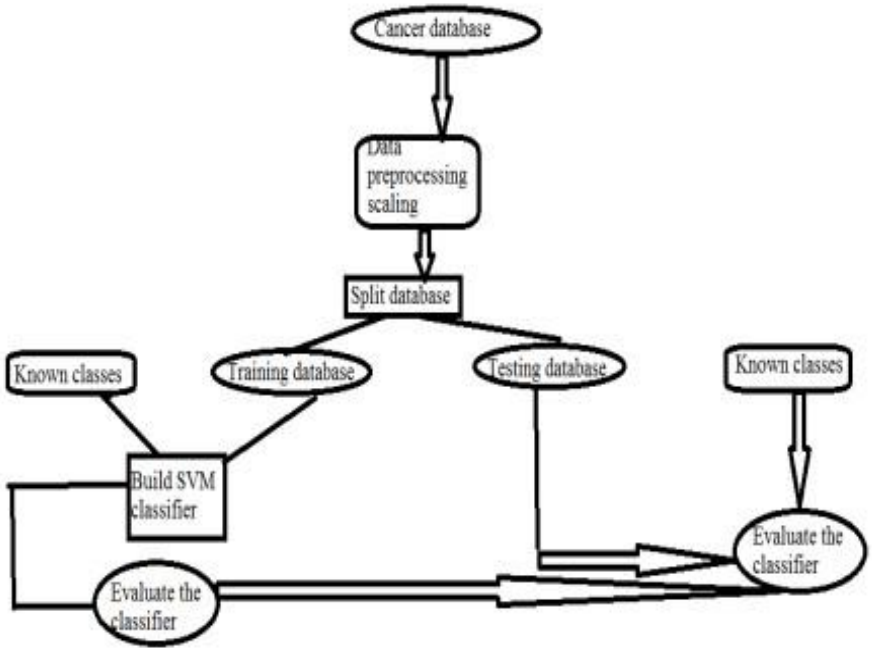

Fig 2:- Flow-Control of algorithm

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed paper was implemented using different types of algorithm like linear regression, logistic regression, support vector machine, knn mapping. The first step is data loading where data is loaded into the machine. The data is mostly in two forms, either csv file or image format. Here the data is in the form of csv file. Once the data is loaded into machine, the data is processed as data cleaning and then data is examined using plots and graphs for the various class and values.

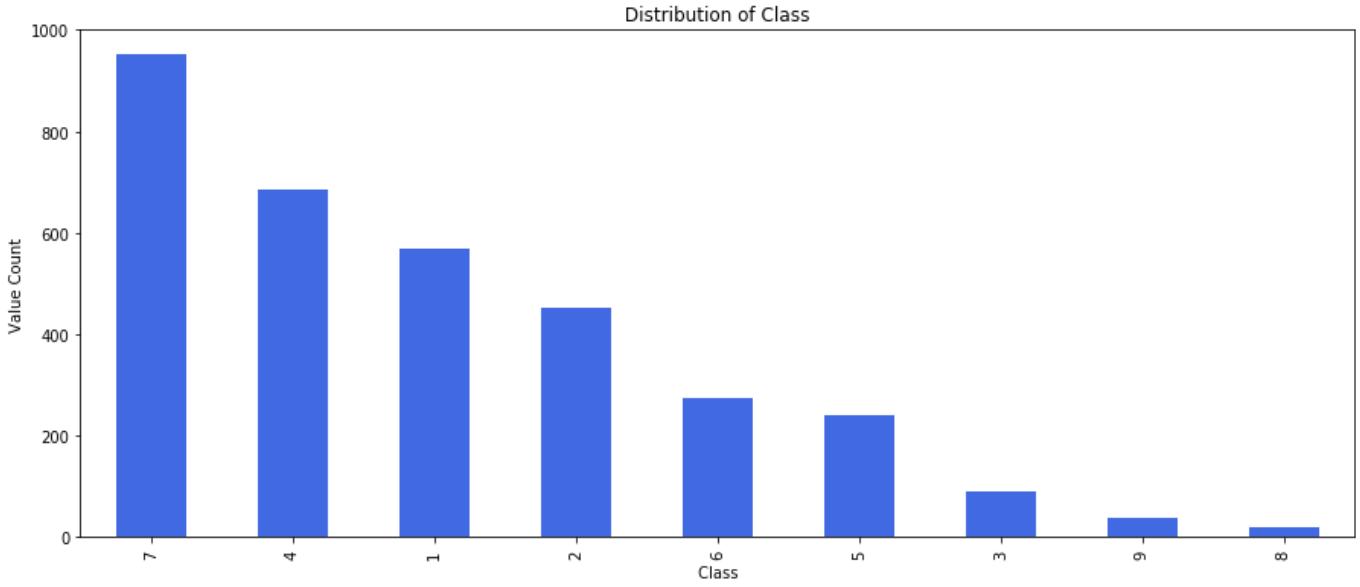

Fig 3:- Graph on class vs value count

ISSN No:-2456-2165

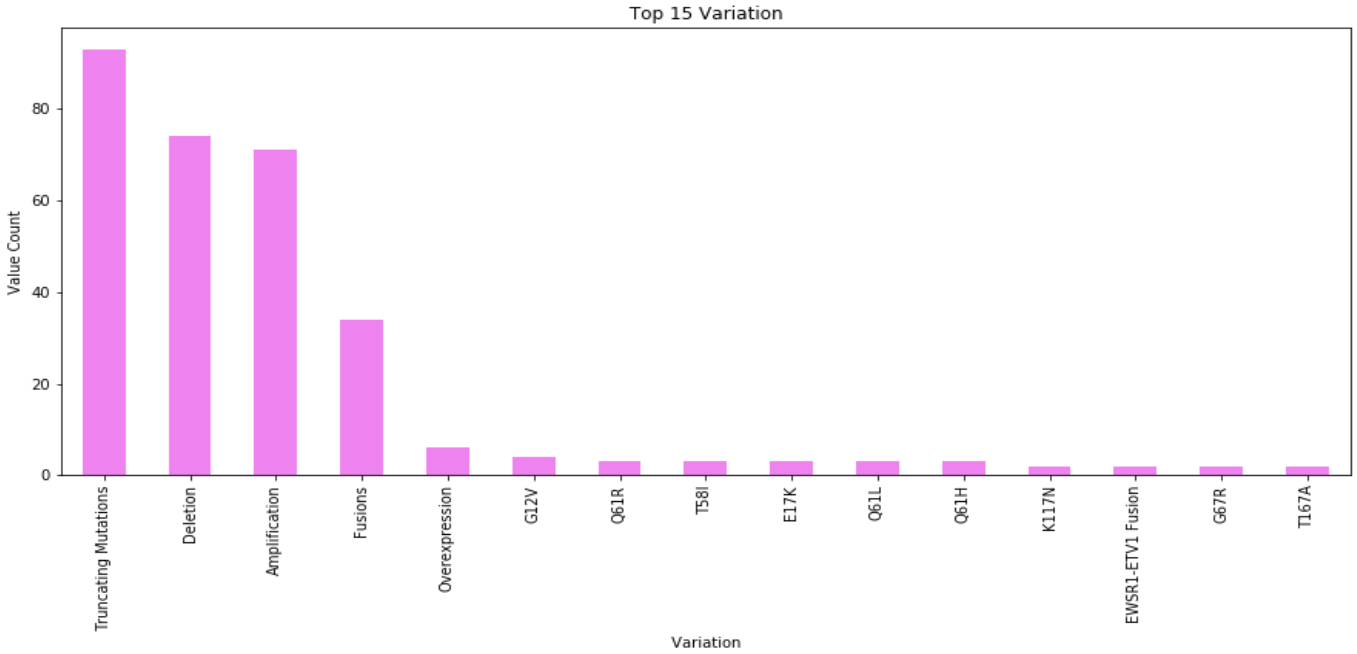

Fig 4:- Gene vs Value Count

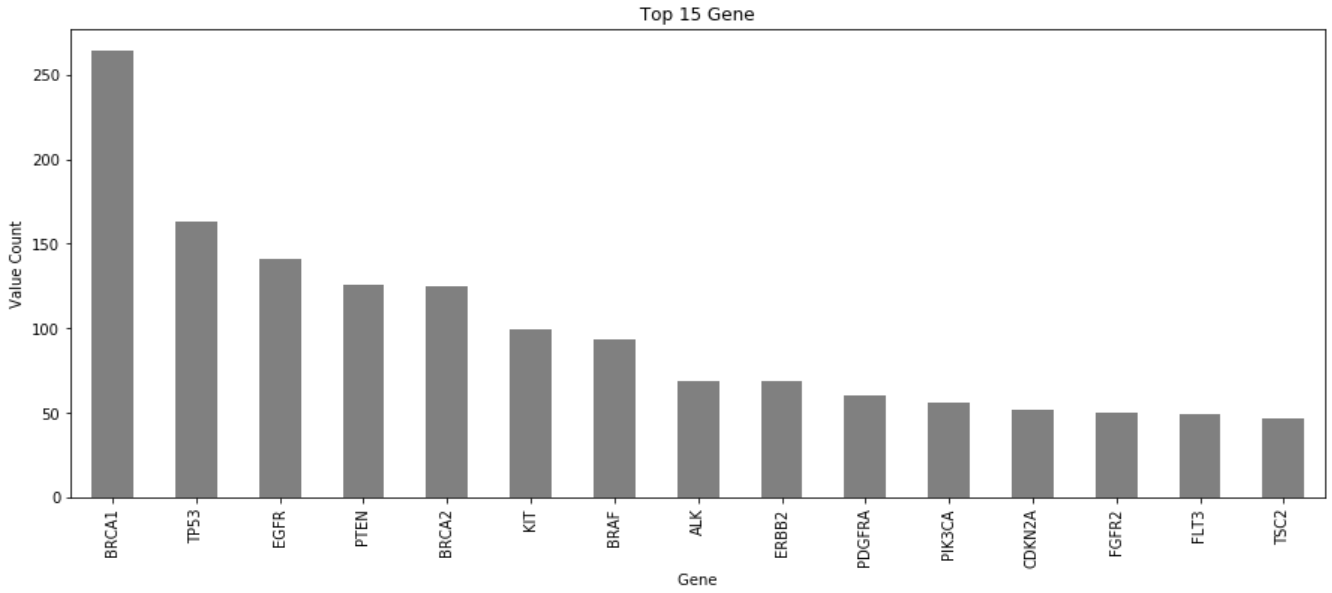

Fig 5:- Value Count vs Variation

The genetic variations of c-CBL relationship is mapped with the receptor tyrosine kinases. Different gene are evaluated on the base of value count. After examining with the different field of the data, the data is trained with the algorithm used and then the outcome is determined for that. Mainly breast cancer involves genetic history, the size of the tumor which describes the stage of the cancer, the radius and dimensions, the age factor is also one of the field of concerns. The model is trained in such a way that it depicts some accuracy and output for the data given after as a test data. During a model, various factors take role in determining the output. A model must be able to predict correctly with the train and test values. Accuracy factor describes the model efficiency and scope of using it further.

ISSN No:-2456-2165

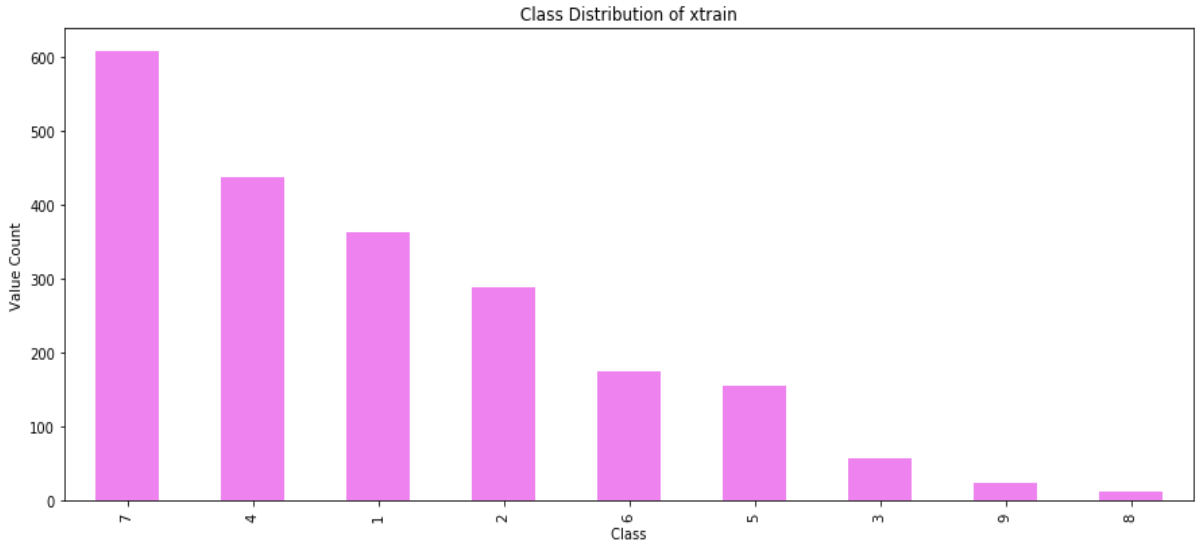

Fig 6:- X train graph

V. CONCLUSION

This paper focuses on the feasibility and effectiveness of the model in the analyzing different types of stages of breast cancer. This paper, different algorithms are used for analyzing the best algorithm suited for the data. Simple classification techniques are employed for the model. Different accuracy for different models- Logistic Regression 95.28%, Linear Regression 89.79% and SVM 92.43%. Thus in future, a better model for better accuracy with different models will be performed.

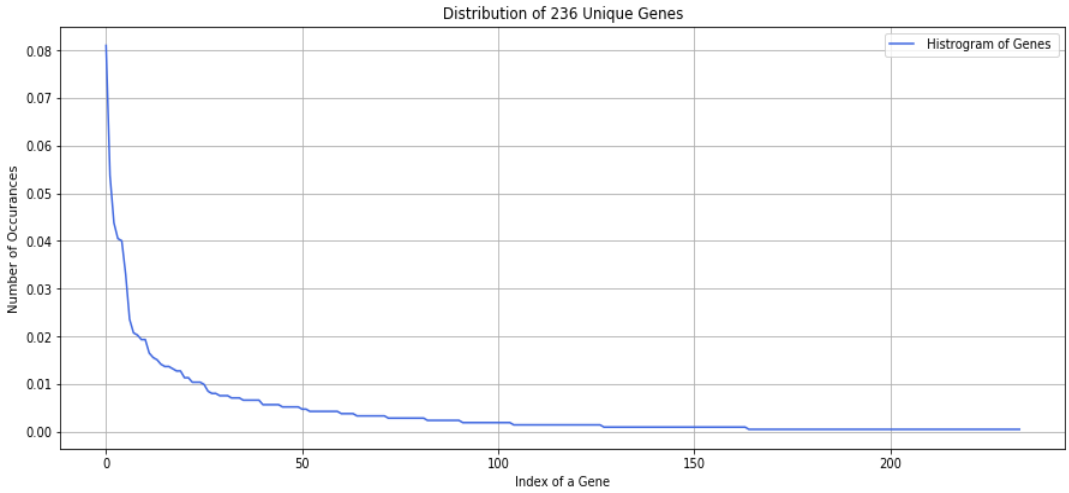

Fig 7:- Distribution of genes

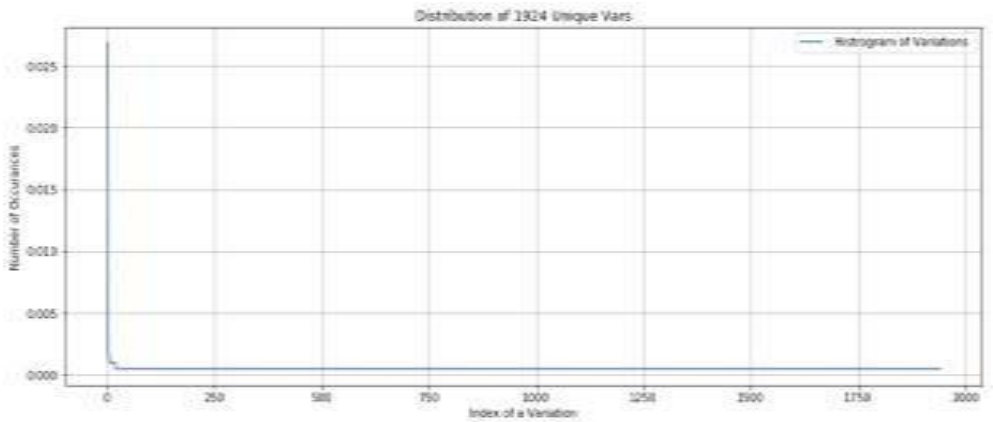

Fig 8:- Distribution of variations

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, Varanasi, **" 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 Chhatwal, Oguzhan Alagoz, Mary J. Lindstrom**, " A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis."**

[5]. Isabelle Guyon, Jason Weston Stephen Barnhill Bioinformatics, Savannah, Georgia, USA, **"Gene Selection for Cancer Classification using Support Vector Machines".**

[6]. Naresh Khuriwal, Nidhi Mishra , Department of Computer Engineering, Poornima University Jaipur, **"Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm".**

[7]. Ilia Kalogiannis,·Elia Markopoulos· Iohannis Anagnostopoulos **"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".**