International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I2P209
Volume 69 Issue 2, 59-67, February 2021

Performing Uni-variate Analysis on Cancer Gene Mutation Data Using SGD Optimized Logistic Regression

Ashok Reddy Kandula[1], Dr. R. Sathya [2] and Dr. S. Narayana[3]

[1]*Research Scholar, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu-608002, India.*

[2]*Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu-608002, India.*

[3]*Professor, Department of Computer Science and Engineering, Gudlavalleru Engineering College, Seshadri Rao Knowledge Village, Gudlavalleru-521356, India.*

[1] ashokreddy.gec@gmail.com, [2]sathya.aucse@gmail.com, [3]satyala1976@gmail.com

Abstract: *There exists a problem in selecting the appropriate machine learning model for any given domain-specific data. Still, researchers are having issues over the model selection in solving the business problem. Along with model selection issues, researchers also face problems in the dataset. Provided all features separating important features and unimportant features in predicting the target class is a challenging task. This paper resolves these issues by using univariate data analysis through machine learning classification techniques as a basic analysis in the process of learning about the data. The objective of the paper is to perform a multi-class classification technique on different classes of mutation effects for the discussed genes. An advanced machine learning-based univariate analysis is performed on each dependent feature to get information about the data. In this paper, we proposed an optimized logistic regression technique using a stochastic gradient optimizer to perform the prediction of target classes. The model prediction is evaluated with a multiclass log loss metric.*

Keywords: *Univariate analysis, Prediction, Mutation changes, Logistic regression, Stochastic Gradient Descent.*

I. INTRODUCTION

A developing methodology in personalized medical analysis named accuracy-based medication considers information and propensities for every patient to decide the disease treatment. This methodology is conceivable even in malignant growth treatment because of huge advances in genome sequencing. Improved genome sequencing has empowered ordering the changes of all the significant tumor types as drivers and passengers.

With the on-going advances in genetic sequencing, genetic testing is considered a significant part of personalized medications. However, genetic testing and identifying the appropriate solutions was always a time-consuming process because of performing through manual work despite everything required to understand the nature of genomics.

These ordering of the tumor type are generally done by domain specialists and clinical pathologists who comment on revealed changes dependent on the proof from text-based clinical writing. It is a time-consuming and tedious procedure. On a fundamental level, a machine learning model could be utilized to make the objective task simpler and more effective by subsequently classifying the mutations. For this reason, Memorial Sloan Kettering Cancer Centre (MSKCC) collected much data from cancer patients, where the information consists of several numbers of mutations, commented by world-class domain specialists and oncologists. This paper focuses on using machine learning techniques to classify genetic changes/mutations that explain cancer/malignancy tumor development, usually called drivers, within the context of neutral mutations that do not influence/affect the tumors, usually called passengers. The objective of the paper is to perform a multi-class classification of different classes of mutation effects for the discussed genes. From the view of many collected records regaurding clinical articles, the work tries to precisely classify among 9 classes of mutations impacts of the examined genes.

From the research analysis, we found that without the need to understand the classes in detail or domain knowledge about the data, we still could apply many machine learning models and Natural Language Processing (NLP) techniques in the proposed materials to achieve our classification. The paper performs some univariate analysis that explained the data on how it is distributed, class balancing, overfitting issues, stability, and usefulness of each feature in classifying the target class. Further, the data is vectorized using a one-hot encoding technique and utilized SGD optimizer with logistic regression to classify the multi-class classification problem. The results are then compared with the multi-class log loss metric.

II. LITERATURE REVIEW

A well-detailed study has been performed to understand the essentials of each technique that were contributed to the paper. In [1] examined the association among antimicrobial resistance with essential contributing features using correlation mechanisms for approaching

univariate analysis along with built a logistic regression model for approaching multivariate analysis. The work has shown that univariate models are considered as an effective descriptive method that highly utilizes one variable every time to analyze the model to test "what if" scenarios.

The comparisons of a univariate model with multivariate models are always performed to get the importance of features in predicting target classes in [2] utilized mean-squared forecast error in comparing the multivariate and univariate models. The paper utilized the Ordinary Least Squares (OLS) model to classify the target classes.

In [3] presented a random intercept-based multilevel logistic regression model technique. The multilevel logistic regression model allowed the clustering of data within the clusters of higher-level of units. A PubMed database was utilized in demonstrating the use of multilevel and hierarchical regression models.

In [4], the Entropy-based gradient-based learning algorithm Entropy-SGD was proposed that explained that it provided better generalization performance error rate under several empirical values. The entropy-based SGD is a PAC-Bayes bounded under the Gibbs a posterior classifier, which was a randomized classifier obtained by a risk-sensitive perturbation towards the weights which has been learned as a classifier. The entropy-based gradient algorithm works by optimizing the bound's priority. A machine learning classifier called deep in memory scheme was proposed [5] with stochastic gradient descent (SGD) with the help of chip trainer 16-kB 6T SRAM array. Deep In-Memory Architecture (DIMA) improves energy efficiency and throughput by reading multiple bits per bit line per reading cycle over conventional digital architectures. This scheme employs mixed-signal processing in the periphery of the bit-cell array.

In [6] describes a natural language text classification of BDNews24 documents data using the SGD classifier. It holds three stages feature Vectorization using Term Frequency and Inverse Document Frequency (TFIDF), classification of the model using SGD classifier, and the performance of the model is evaluated using F1 score measures.

In [7] presented a simple Variance Reduced Stochastic Gradient Descent (VR-SGD) approach that tackles non-smoothing and non-strongly convex problems along with founding that VR-SGD attains linear convergence the problem. In [8] [9] designed a trend smoothing algorithm in accelerating the training process to train with Asynchronous Stochastic Gradient Descent (A-SGD) machine learning model.

In [10] presented a frequentist statistical inference approach using SGD considering fixed size for addressing M-estimation issues. It used the average SGD sequences for addressing statistical inference after performing proper scaling.

In [11] addressed a multi-class classification problem by presenting parallelized logistic regression that classified large datasets holding high-dimensional images of signatures as classes. It utilized balanced batch stochastic gradient descent for logistic regression optimization with an on-versus-all strategy for approaching the multi-class classification model.

In [12] presented a data classification using multinomial logistic regression with a maximum entropy classifier approach that minimized the memory consumption of highly scalable data specifically for sparse-based document matrices. In [13] presented a quasi-newton method, which was a vertical framework type of learning through logistic regression for communication-based classification.

An efficient distributed stochastic optimization method [14] was proposed by combining adaptively with variance reduction techniques in order to yield a linear speedup in the number of machines with a logarithmic number of communiqué rounds. It is a black-box reduction mechanism that parallelizes serial online learning protocol, and optimal convergence rates can be achieved using adaptive algorithms.

In [15] designed a stochastic gradient descent technique Flex Compress SGD that trained neural networks for the distributed datasets through multiple workers and server-based data using ImageNet dataset. However, the SGD approach gave a lesser scalable cost. To handle sensor-based Cleveland Heart Disease Data (CHDD) in cloud computing [16], incorporated MapReduce based online stochastic gradient descent that optimized logistic regression.

In [17] utilized Stochastic Gradient Descent based Logistic Regression (SGD-LR) that classified image samples into two separate categories, i.e., raveling and non-traveling types based on extracted features. This implementation is performed in visual C#.Net applications. In [18] evaluated the Geographically Weighted Regression (GWR) model for addressing landslide susceptible mapping to analyze special geo data along with the incorporated chi-square-based feature selection method. The comparison is done among GWR, SGD-LR, SGD-SVM, and Support Vector Machine (SVM) models.

In [19] scaled-up the terabyte data by Logistic Regression (LR) or L1 regularized loss minimization approach through optimizing by SGD and Stochastic Coordinate Descent (SCD) algorithms. For addressing genomic cancer classifications [20] utilized SVM, it discovers new biomarkers, new drug targets, and well-classified cancer driver genes.

In [21] [22] SVM algorithm was utilized to predict chemotherapeutic drugs among the gene-expression profiles (RNA-Seq or microarray) in the patient tumors. To address personalized or precision medicine for patient diagnosis care of phenotype categories and population size and statistical analysis [23-25] performed pattern recognition of patient profiles. To address breast cancer diagnosis in patients, specifically, radiologists that conduct the Fine Needle Aspirate (FNA) process on breast tumor [24-] utilized sophisticated classifiers like LR, K-Nearest Neighbours (KNN), and SVM. To address this, it utilized a concrete relationship of precision, recall, and several feature engineering processes on the cancer malignant dataset.

The order of the research is as follows in section I, the paper explains the data and techniques used while in section II, a complete review of literature on various techniques utilized has been addressed in section III materials and methodologies are highlighted, and section IV experiments and appropriate results are mentioned. In section V each model evaluation is performed, while in section VI conclusion about the work is given.

III. MATERIALS AND METHODS

The paper performs some univariate analysis before applying machine learning models. The research analyses the cancer medical data gathered from the MSKCC dataset collected from thousands of cancer patients. The dataset was built with the feature gene, variants, text (clinical evidence), and target class. Among the feature gene and variants are categorical features, while the text is the text feature constructed in the form of a sequence of words and a multi-class target class holding nine classes ranging [1 to 9] that indicate the type of cancer.

A. PROPOSED APPROACHES

The goal of the paper is to analyze which feature is useful in predicting $y_i's$ a target class and identifying the feature stability. Many methods exist to evaluate the effectiveness of the feature in the prediction of $y_i's$ a target class. The research process follows by building the proper machine learning model using those features. The proposed approach has two phases firstly, performing univariate analysis on each feature secondly applying classification techniques.

B. UNIVARIATE ANALYSIS

The univariate analysis is the simplest technique in statistical analysis that takes only one variable to perform any analysis. Just like other forms of statistics, it could be inferential or differential in type. The univariate analysis techniques utilized in this paper highly describes the information about the data. The probability density function and Cumulative Distributive Function (CDF) are used in the paper. The PDF and CDF are two important and closely related statistical functions in reliability/ when these two functions are known, then other reliability measures can be easily obtained. Other statistical techniques like distributions are performed by histograms that explain the data distributions along with many other simple statistical calculations that are performed using many scientific libraries available in python.

C. MACHINE LEARNING MODELS

Machine learning models like SGD optimization of Logistic regression and Randomized multivariate classification models were proposed in this paper through python Jupyter notebook.

a) RANDOMIZED MULTIVARIATE CLASSIFICATION MODEL

The underlying benchmark model: randomized multivariate classification model has been utilized to predict the target class where further the individual feature models or univariate models are compared with it. When the univariate model provides better results than the benchmark model, then it could be said that the feature is good at predicting the results, and we could make those features primary in upcoming advancements. The classification by a randomized multivariate classification model is performed on the whole dataset with all features.
The randomized multivariate classification model is a classification technique that uses the Dirichlet distribution approach, which is a continuous multivariate probability distribution that has been parameterized with a vector representation. The Dirichlet approach is a concept usually uses values of previous data and enriches it at the current prediction, which mathematically given in Dirichlet function defined in equation 1 and analytically defined as given equation 2.

$$D(x) = \begin{Bmatrix} c \text{ for } x \text{ irrational} \\ d \text{ for } x \text{ irrational} \end{Bmatrix} \quad (1)$$

$$D(x) = \lim_{m\to\infty} \lim_{n\to\infty} \cos^{2n}(m!\pi x) \quad (2)$$

Where c and d ≠ c are real numbers, we take c and d values as 1 and 0, respectively.

b) LOGISTIC REGRESSION

Logistic regression is a statistical methodology used for analyzing a large-scale dataset. The logistic regression was effective for obtaining the relationship between one dependent variable and one or more independent variables. In the data frame, every independent feature is multiplied with weights and summed up among them, which is further inputted to the sigmoidal function to obtain the results. In logistic regression, it is important to obtain the best weights and regression coefficients, which can be obtained by optimization techniques like SGD. Optimization techniques were used to get the best and weights.

c) STOCHASTIC GRADIENT DESCENT

SGD is a basic yet effective approach in dealing with the fitting of classifiers and regressors under convex loss functions, for example, (linear) Support Vector Machines and Logistic Regression. SGD helps analyze large-scale and high sparse data especially specialized in text classification and NLP-based models. SGD, on the upper hand it is an optimization technique that does not fall under any machine learning models as a base. It is considered an effective way to train a model.

d) SGD OPTIMIZED LOGISTIC REGRESSION

Let us consider a two-class classification task for dataset D that has x_i points, $x_i \in R^d$ where x_i belongs to d-dimensional rational numbers, d = 3, and target class $y_i, \text{ and } y_i \in R$. For binary classification $y_i \pm 1.$, Which can be further extended to multi-class by one-versus rest approach. While logistic regression learns classification models, i.e., parameter vector w ϵ R^d for maximizing the likelihood of the data. The probability of a

data point whose target class belongs to positive classes is given in equation 3.

$$p(y_i = {}^{+1}/_{x_i}) = \frac{1}{1+e^{-(w.x_i)}} \quad (3)$$

Where y_i is target class point, w is weight x_i is a data point in dataset d. And further, the probability of the data point when the target class is negative is defined in equation 4.

$$p(y_i = {}^{-1}/_{x_i}) = 1 - p(y_i = {}^{+1}/_{x_i}) = 1 - \frac{1}{1+e^{-(w.x_i)}} = \frac{1}{1+e^{(w.x_i)}} \quad (4)$$

The probabilities from Equations 3 and 4 aftern rewritten are given in equation 5.

$$p({}^{y_i}/_{x_i}) = \frac{1}{1+e^{-y_i(w.x_i)}} \quad (5)$$

Equation 5 is extended by log-likelihood and rewritten in equation 6.

$$log\left(p({}^{y_i}/_{x_i})\right) = log\left(\frac{1}{1+e^{-y_i(w.x_i)}}\right) = -log\left(1+e^{-y_i(w.x_i)}\right) \quad (6)$$

The Logistic Regression in this paper utilizes Tikhonov regularization [25-28] as a trading-off factor that tends to increase the likelihood and reduces the error rates. The reason for utilizing regularization is to inflict a penalty on the magnitude over the weights parameter w. The proposed Logistic regularization technique achieves a parallel process in maximizing the log-likelihood for the data points at the same time minimizes the L2 regularization for weight parameter vector w. After applying regularization, the defined method is given in equation 7.

$$\min \varphi(w,[x,y]) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \log\left(1+e^{-y_i(w.x_i)}\right) \quad (7)$$

The derived Logistic Regression at base uses the logistic loss function $L(w,[x_i,y_i]) = log\left(1+e^{-y_i(w.x_i)}\right)$. The solution can be optimized and solved by using a stochastic gradient descent optimization process. The proposed Logistic Regression on every epochs or iteration (t) updates the weight with a given learning rate η. For every iteration t, the proposed Logistic Regression picks up a data point randomly (x_t, y_t) and computes sub-gradient factor $\nabla\varphi(w,[x_t,y_t])$ to update the weights w_{t+1}, which is shown in equations 8 and 9.

$$w_{t+1} = w_t - \eta_t \nabla_t \varphi(w,[x_t,y_t]) = w_t - \eta_t\left(\lambda w_t + \nabla_t L(w,[x_t,y_t])\right) \quad (8)$$

$$\nabla_t L(w,[x_t,y_t]) = \nabla_t \log\left(1+e^{-y_t(w.x_t)}\right) = -\frac{y_t x_t}{1+e^{y_t(w.x_t)}} \quad (9)$$

There the approach is extended to a multi-class classification problem where classes more than 3. In this proposed approach, we planned to use multi-class classification in one optimization problem combining the one-versus-one approach, which was mentioned as one of the parameters along with weights.

IV. EXPERIMENTATION AND RESULTS ANALYSIS

The experiments are performed in the Windows Operating system with RAM sized 4 GB and using the high-level programming language Python; the implementations of the approaches are done using Anaconda environment with Jupyter Notebook as a user interface. The Anaconda is an open-source and free application for python and R languages that facilitate performing scientific computing, classifications, etc... Anaconda aims to simplify package management and deployment. It is suitable for windows, Linux, and macOS systems.

A. DATA MODELLING

The dataset used in this whole research process is the MSKCC dataset collected from thousands of cancer patients. The classification is done on genetic mutations (a target class or independent features) based on the given data holding features like gene, variations, and (clinical evidence) text (three dependent features). The given data has nine distinct target classes of genetic mutations where classification needs to be performed. The total no of data points (D) in our data set is 3321, with 4 main features and an ID feature. The main features build with two categorical features (gene and variations), one text feature (text), and a multi-class target variable holding nine classes indicated by integers (1, 2, .. 9).

a) DATA PRE-PROCESSING

Utilizing NLP-based pre-processing on text data so that the raw data feature "text" is pre-processed into a suitable form so that Vectorization can be performed effectively on these data. Some effective processes involved in the paper are handling of special characters, spaces, case constraints, stop word removals.

b) TRAIN TEST SPLIT

The data D is broken into 80:20 ratio where whole data are broken into 80% of train data and 20% of test data D_Te. While further, the training data is broken into 80:20 ratios where 80% is broken into train data D_Trand 20% into CV data D_CVi.e. [train: CV: test:] as [64: 16: 20]. Among the whole data points, 3321, the number of data points in train data is 2124, the test is 665, and CV is 532.

1) UNIVARIATE ANALYSIS OF INDEPENDENT VARIABLE

The univariate analysis is performed on each individual feature considering and applying classification technique to it. From statistical analysis results shown in Table 1 among the train data D_{Tr}, class 7 has more data points around 609 holding 28% of train data, and class 4 counts 439 data points holding 20% of train data, class 1 counts 363 data points holding 13% of train data, class 2 has 289 of data points holding 13% of train data others falling under 10% of train data.

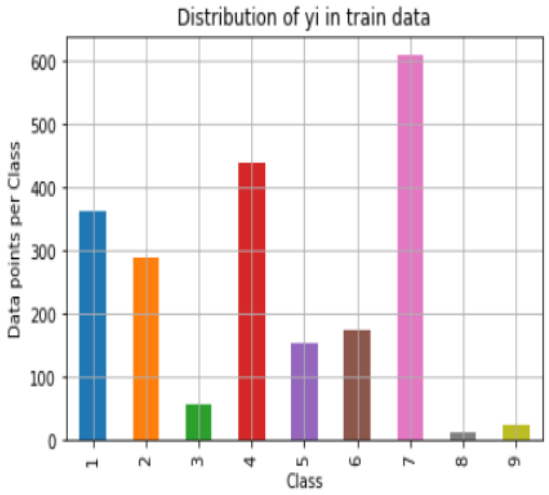

Figure 1(a): Distribution of target class in train data

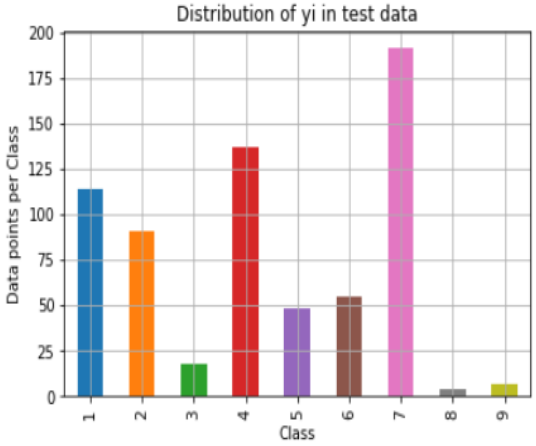

Figure 1(b): Distribution of target class in the test data

Similarly, in test and CV data, class 7, 4, 1, 2 have occupied more counts leaving others to fall under 10%; thus, train, test, and CV data have equal distribution among the whole data. This clearly explains class 7, 4, 1, and 2 are dominating the other classes 3, 5, 6, 8, 9 in the train, test, and CV as shown in figure 1 (a), 1(b), 1(c) and the dataset have the problem of class im-balancing.

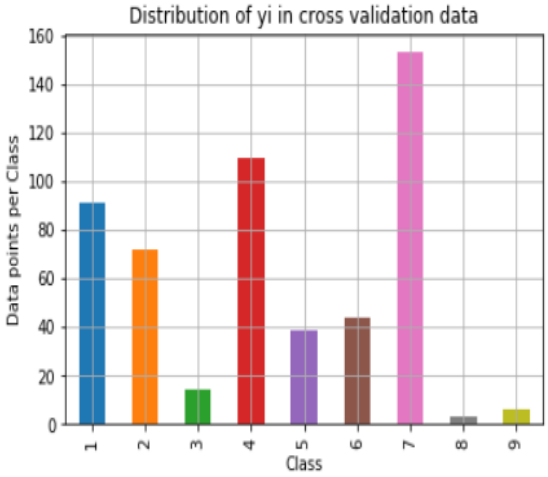

Figure 1(c): Distribution of target class in the CV test

The data that having class balancing improves the results ineffective way as the distribution of each class has equal weightage on the results otherwise the classes that have more data points dominates the fewer data point's classes in the results leading to failure of the machine learning models.

Table 1: Statistical Analysis of Distribution of Train, Test and CV data

Target Class	Train data		CV		Test	
	Data points (Counts)	Percentage of data points	Data points (Counts)	Percentage of data points	Data points (Counts)	Percentage of data points
7	609	28%	153	28%	191	28%
4	439	20%	110	20%	137	20%
1	636	17%	91	17%	114	17%
2	289	13%	72	13%	91	13%
6	176	8%	44	8%	55	8%
5	155	7%	39	7%	48	7%
3	57	2%	14	2%	18	2%
9	24	1%	6	1%	7	1%
8	12	0.5%	3	0.5%	4	0.6%

2) UNIVARIATE ANALYSIS OF DEPENDENT VARIABLES

The univariate analysis takes each feature separately to build a model (data frame) where on top of it, machine learning techniques are applied. The univariate analysis applied on gene a categorical variable, variants a categorical variable, and text a text-based variable containing the sequence of words. The objective of the univariate analysis is to address several questions like whether the feature is suitable for predicting the target class, what are the featuring techniques we could employ in it, and how stable the feature is in the data set.

a) DISTRIBUTION OF GENE CATEGORIES

From the statistical analysis, it was found that there are 240 distinct categories in the gene feature of whole 2125 train data points where the top 5 gene categories counts are BRCA1 – 169, TP53 106, EGFR – 91, PTEN – 81, BRCA2 – 81, etc.., the figure 2(a) explains that only topmost gene BRCA1 occurs 80% of the whole gene category while smaller genes occur only fewer times.

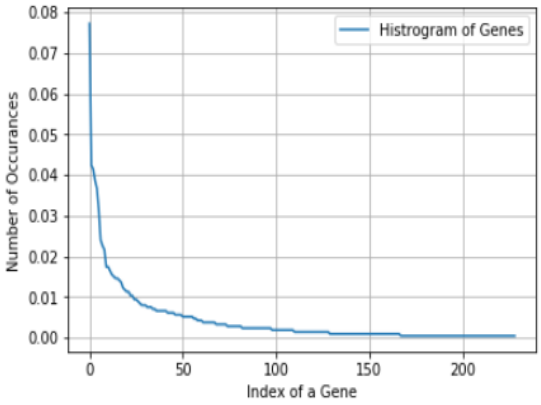

Figure 2(a): PDF of gene categories

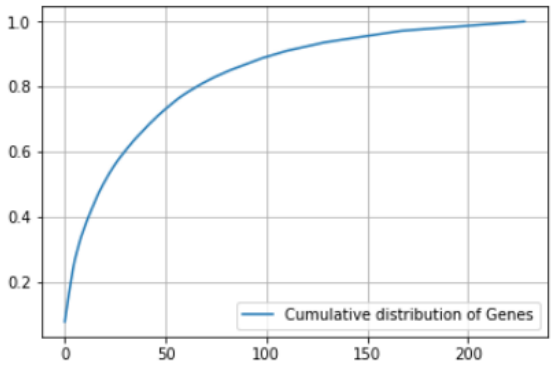

FIGURE 2(B): CDF OF GENE CATEGORIES

Figure 2(b) explains about almost all the top 50 gene categories contribute around 75% of whole gene categories in train data. The rest of the gene categories contribute only 25% of the whole train data. This explains in train data top 50 genes are frequently occurring gene categories while others occur rarely.

b) DISTRIBUTION OF VARIATION CATEGORIES

From the statistical analysis, it was found that there are 1917 unique variation categories in the training data among these, Truncating Mutations occur 67 times, Deletion occurs 46 times, Amplification occurs 45 times, Fusions occur 23 times rest of the various categories occur lesser than 5 times in the whole variation data points.

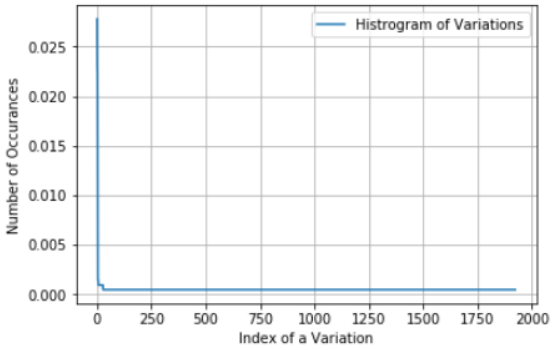

Figure 3(a): PDF of variation categories

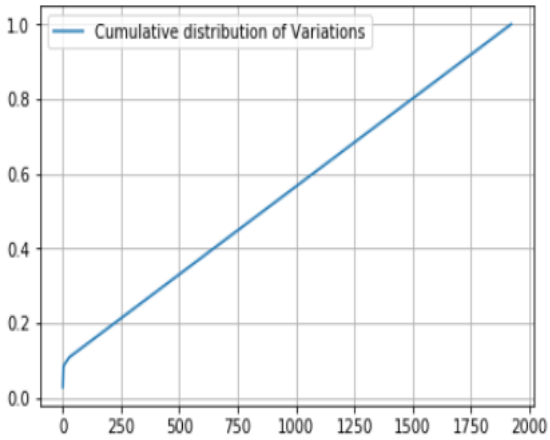

Figure 3(b): CDF of Variation category

Figure 3(a) explains the frequency of the variation falling sharply, i.e., only fewer categories of variations occur around 2.5% of data others occur lesser than 0.5% of times in the whole training data of variants. Figure 3(b) explains that the topmost 1500 variant's frequencies account for 80% of training data.

c) DISTRIBUTION OF TEXT FEATURE' WORDS

The statistical analysis is applied to the text feature after pre-processing the texts by removing stop words, special characters, and spaces to get efficient results. From the statistical analysis, it was found that the text feature has 54850 unique words in it. Among 54850 words, 6168 words occur 3 times in the whole train data, while 3775 words occur 4 times of whole data, while 3025 words occur 5 times in whole train data, 2977 words occur 6 times, 2140 occur 9 times. The number of words and their occurrences differs slightly when the analysis is restarted as the train test, and CV split is performed by randomly picking upon the data points.

B. FEATURE VECTORIZATION TECHNIQUES

The features gene, variants, and text are vectorized using one-hot encoding with Laplace smoothing in it. From table 2, the one-hot encoding for gene feature in train data results into a vector with 2124 data points and 239 dimensions while variants result in 2124 data points and 1945 dimensions and test results in 2124 data points with 52851 dimensions.

TABLE 2: DIMENSIONS OF THE DATA AFTER FEATURE VECTORIZATION THROUGH ONE-HOT ENCODING

Features	Train data		CV data		Test data	
	Data Points (Counts)	One-hot encoding (dimensions)	Data Points (Counts)	One-hot encoding (dimensions)	Data Points (Counts)	One-hot encoding (dimensions)
Gene	2124	239	532	239	665	239
Variants	2124	1945	532	1945	665	1945
Text	2124	52851	532	52851	665	52851

Similarly, in CV, 532 data points in gene and 239 dimensions, while variants have 532 data points and 1945 dimensions, and text has 532 data points and 52851 dimensions. Similarly, in test data, the gene feature results in 665 data points and 239 dimensions, the variants results in 665 data points and 1945 dimensions, and text results in 665 data points and 52851 dimensions.

V. UNIVARIATE AND MULTIVARIATE CLASSIFICATION

The research follows the classifying the target class and evaluating the model by comparing the benchmark results obtained by randomized multivariate model classification with each univariate model classification result.

A. MODEL I: RANDOMIZED MULTIVARIATE CLASSIFICATION MODEL

The benchmark model utilized here is the randomized multivariate classification model that uses Dirichlet Distribution based classification technique. The randomized model is trained with train data points and randomly predicts the target class for every training data and finally testing the model with CV and test data. The randomized multivariate classification is performed with all the features in the dataset, and the model is evaluated with a multi-class log loss metric.

The multi-class log-loss error for the randomized model for test data obtained is 2.5930, and CV data obtained is 2.5248. The precision matrix for the randomized model is given in figure 4.

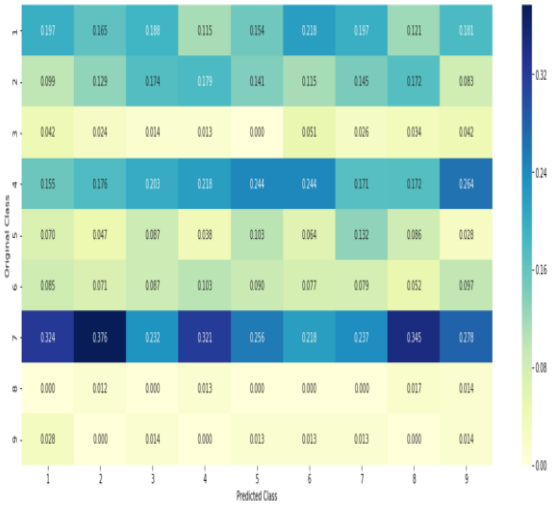

FIGURE 4: PRECISION MATRIX OF THE RANDOMIZED MULTIVARIATE MODEL

B. MODEL II: UNIVARIATE CLASSIFICATION MODEL ON GENE FEATURE

The data model built with a single feature, i.e., gene feature, is trained using SGD based logistic regression. The appropriate smoothing factor α is tuned with different values using L2 regularization. The best alpha obtained in model II is 0.0001 with train log loss as 1.0096, CV log loss as 1.1780, and test log loss as 1.1940. The precision matrix is shown in figure 5.

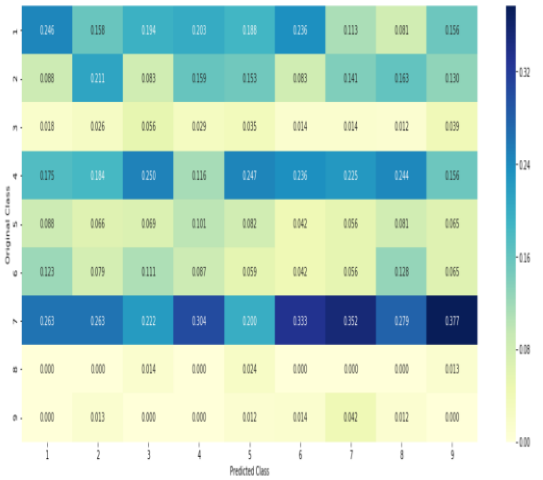

FIGURE 5: PRECISION MATRIX OF UNIVARIATE CLASSIFICATION MODEL ON GENE FEATURE

C. MODEL III: UNIVARIATE CLASSIFICATION MODEL ON VARIANT FEATURE

The data model built with a single feature, i.e., variant feature, is trained using SGD optimized logistic regression. The appropriate smoothing factor α is tuned with different values using L2 regularization.

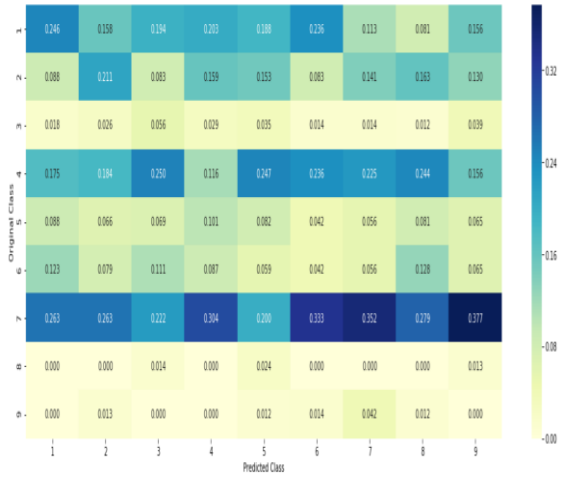

FIGURE 6: A PRECISION MATRIX FOR VARIANT FEATURE

The best alpha obtained in model III is 0.0001 with train log loss as 0.7558, CV log loss as 1.7344, test log loss as 1.7143. The precision matrix for the variant feature is shown in figure 6.

D. MODEL IV: UNIVARIATE CLASSIFICATION MODEL ON TEXT FEATURE

The data model built with a single feature, i.e., text feature, is trained using SGD classifier logistic regression. The appropriate smoothing factor α is tuned with different values using L2 regularization. The best alpha obtained in model III is 0.0001 with train log loss as 0.7528, CV log loss as 1.1137, test log loss as 1.1774. The test feature precision matrix is shown in figure 7.

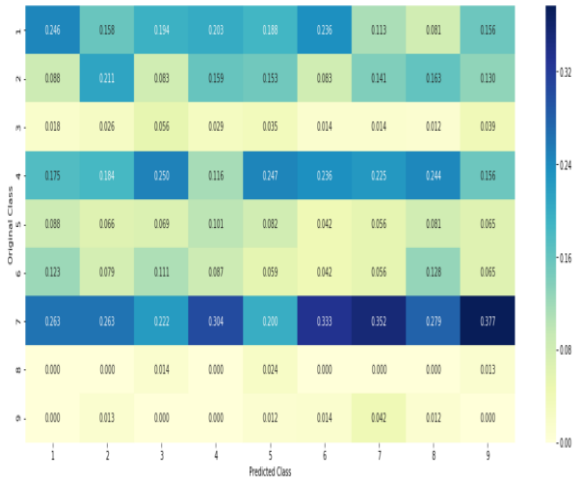

FIGURE 7: A PRECISION MATRIX FOR TEXT FEATURE

VI. MODEL EVALUATION

Table 3 shows the results obtained from all four models. The model classified using Gene feature, variant feature, and text feature gave lesser error than random model's test log loss error of 2.593, explains that these three features make better in prediction. While gene feature is far lesser compared to the variant and text feature, thus variant feature and text feature may be prone to increase the prediction error.

TABLE 3: MULTI-LOG-LOSS EVALUATION

Feature	Train (log-loss)	CV (log-loss)	Test (log-loss)	Stability
Randomized model	1.0030	2.5248	2.5930	-
Gene	1.0425	1.2325	1.2009	Stable
Variant	0.8255	1.6898	1.7362	Less Stable
Text	0.7614	1.3062	1.1902	Less Stable

From the statistical analysis, it was found that in gene features, 97% of data present in test data also present in train data while 98% of data present in CV data also present in train data. Similarly, for variant features, 90% of data present in test data also present in train data, while 97% of data present in CV data also present in train data. In the text feature, 96% of data present in test data also present in train data, while 97% of data present in CV data also present in train data. This indicates overlapping of the train, test, and CV data is more than 90%; thus, the data does not have overfitting issues.

The stability of the feature found from comparing the training CV and test of gene features were a train, CV, and test log loss 1.0425, 1.2325, and 1.2009 are almost given lesser difference thus gene feature is stable. While for a variant train, CV and Test 0.8255, 1.6898, and 1.7362, the error of test and CV falls far from the train data invariant feature; thus, the variant feature is less stable. For text feature train, CV, and test, 0.7614, 1.3062, and 1.1902, the error rates of CV and test falls somewhat near to train error; thus, text feature is also considered a stable feature. It could be said that the gene is stable while variant feature and text features are less stable in predicting the target classes but still gave lesser rates than the randomized model, so the features can be very useful in prediction by approaching different mechanisms.

VII. CONCLUSION

The paper successfully performed a univariate analysis and multivariate analysis of personalized cancer medical data in classifying the type of gene mutation. From the univariate analysis, it was found that the data has class im-balancing issues, features are not stable in prediction. From model classification performed by optimized logistic regression models, it was found that every feature gave a lesser error rate compared to the randomized model; thus, in future applications of the machine learning model, the gene, variant, and text features will be considered for analyzing. Further advanced Vectorization techniques like response coding will be incorporated in future works, along with probability measures of the target class will be considered to overcome these issues.

VIII. REFERENCES

[1] Collignon, P., Beggs, J. J., Walsh, T. R., Gandra, S., & Laxminarayan, R., Anthropological and socioeconomic factors are contributing to global antimicrobial resistance: a univariate and multivariable analysis. The Lancet Planetary Health, 2(9)(2018) e398-e405.

[2] Beanland, K., Roberts, J. W., & Stevenson, C., Modifications of Thomae's function and differentiability. The American Mathematical Monthly, 116(6)(2009) 531-535.

[3] Austin, P. C., & Merlo, J., Intermediate and advanced topics in multilevel logistic regression analysis. Statistics in medicine, 36(20)(2017) 3257-3277.

[4] Dziugaite, G. K., & Roy, D. M., Entropy-SGD optimizes the prior of a PAC-Bayes bound: Data-dependent PAC-Bayes priors via differential privacy., (2018).

[5] Gonugondla, S. K., Kang, M., & Shanbhag, N. R., A variation-tolerant in-memory machine learning classifier via on-chip training. IEEE Journal of Solid-State Circuits, 53(11)(2018) 3163-3173.

[6] Kabir, F., Siddique, ‘l;” stochastic gradient descent (sgd) classifier. In 2015 International Conference on Cognitive Computing and Information Processing (CCIP) (2015) 1-4. IEEE.

[7] Shang, F., Zhou, K., Liu, H., Cheng, J., Tsang, I. W., Zhang, L., ... & Jiao, L., VR-SGD: A simple stochastic variance reduction method for machine learning. IEEE Transactions on Knowledge and Data Engineering, 32(1)(2018) 188-202.

[8] Cui, G., Guo, J., Fan, Y., Lan, Y., & Cheng, X., Trend-Smooth: Accelerate Asynchronous SGD by Smoothing Parameters Using Parameter Trends. IEEE Access, 7, (2019) 156848-156859.

[9] Stewart, D. J., & Batist, G., Redefining cancer: a new paradigm for better and faster treatment innovation. Journal of Population Therapeutics and Clinical Pharmacology, 21(1) (2014).

[10] Li, T., Liu, L., Kyrillidis, A., & Caramanis, C., Statistical inference using SGD. arXiv preprint arXiv:1705.07477. (2017).

[11] Do, T. N., & Poulet, F., Parallel multiclass logistic regression for classifying large-scale image datasets. In Advanced Computational Methods for Knowledge Engineering ., (2015) 255-266. Springer, Cham.

[12] Jurka, T. P., MAXENT: an R package for low-memory multinomial logistic regression with support for semi-automated text classification. The R Journal, 4(1)(2012) 56-59.

[13] Yang, K., Fan, T., Chen, T., Shi, Y., & Yang, Q., A quasi-newton method-based vertical federated learning framework for logistic regression. arXiv preprint arXiv:1912.00513. (2019).

[14] Cutkosky, A., & Busa-Fekete, R.. Distributed stochastic optimization via adaptive SGD. In Advances in Neural Information Processing Systems (2018) 1910-1919.

[15] Phuong, T. T., Distributed SGD With Flexible Gradient Compression. IEEE Access, 8 (2020) 64707-64717.

[16] Manogaran, G., & Lopez, D., Health data analytics using scalable logistic regression with stochastic gradient descent. International Journal of Advanced Intelligence Paradigms, 10(2018) (1-2), 118-132.

[17] Hoang, N. D., Automatic detection of asphalt pavement raveling using image texture-based feature extraction and stochastic gradient descent logistic regression. Automation in Construction, 105(2019) 102843.

[18] Hong, H., Pradhan, B., Sameen, M. I., Chen, W., & Xu, C., Spatial prediction of rotational landslide using geographically weighted regression, logistic regression, and support vector machine models in Xing Guo area (China). Geomatics, Natural Hazards, and Risk, 8(2)(2017) 1997-2022.

[19] Kang, D., Lim, W., Shin, K., Sael, L., & Kang, U., Data/feature distributed stochastic coordinate descent for logistic regression. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (2014) 1269-1278.

[20] Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W., Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics-Proteomics, 15(1)(2018) 41-51.

[21] Xu, J., Yang, P., Xue, S., Sharma, B., Sanchez-Martin, M., Wang, F. & Parikh, B., Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges, and future perspectives. Human Genetics, 138(2) (2019) 109-124.

[22] Adlung, L., Elinav, E., Greten, T. F., & Korangy, F., Microbiome genomics for cancer prediction. Nature Cancer, 1(4)(2020) 379-381.

[23] Emmert-Streib, F., & Dehmer, M.. A machine learning perspective on Personalized Medicine: an automized, comprehensive knowledge base with ontology for pattern recognition. Machine Learning and Knowledge Extraction, 1(1)(2019) 149-156.

[24] Srinivasa Reddy, K., Suneela, B., Inthiyaz, S., Kumar, G.N.S., Mallikarjuna Reddy, A., Texture filtration module under stabilization via random forest optimization methodology.International Journal of Advanced Trends in Computer Science and Engineering, 8(3)(2019) 458-469.

[25] A.Mallikarjuna, B. Karuna Sree., Security towards Flooding Attacks in Inter-Domain Routing Object using Ad hoc Network. International Journal of Engineering and Advanced Technology (IJEAT), 8(3)(2019) 545-547.

[26] Mallikarjuna Reddy, A., Rupa Kinnera, G., Chandrasekhara Reddy, T., Vishnu Murthy, G. Generating cancelable fingerprint template using triangular structures, Journal of Computational and Theoretical Nanoscience, 16(5)(2019) 1951-1955(5).

[27] Sharma, A., Kulshrestha, S., & Daniel, S., Machine learning approaches for breast cancer diagnosis and prognosis. In 2017 International Conference on Soft Computing and its Engineering Applications (icSoftComp) (2017) 1-5. IEEE.

[28] Ashwini S. Savanth, Dr. P.A.Vijaya ,Artificial Neural Networks for fMRI Data Analysis: A Survey, International Journal of Engineering Trends and Technology (IJETT), 49(8) 487-494 2017.

[29] Edwards, T. H., & Stoll, S., Optimal Tikhonov regularization for DEER spectroscopy. Journal of Magnetic Resonance, 288(2018) 58-68.