IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CANCER DETECTION USING MACHINE LEARNING

Shivam Gupta
Mohd Saquib
Mohd Monis
Nitin Kr. Pandey
Department of Electronics and Instrumentation Galgotia's College of Engineering and Technology, GCET, Greater Noida, India

Abstract: Machine Learning algorithms are being widely used in the prediction of cancer type based on the past data. The machine learning algorithm is trained with past data that can be used to predict the type and severity of cancer, which can help the medical professional in treating cancer. The early and quick diagnosis of a cancer type have become a necessity in cancer research. In this study, the past data is pre-processed and was used to train several different machine learning models such as Logistic Regression, Support Vector Machine, K-Nearest Neighbour with the comparison of each model using Log-loss and Accuracy as performance metric. The result obtained after comparison of each model was the Logistic Regression with feature engineering attained the highest accuracy and least log-loss in both training and testing of our data.

***Keywords*—Machine learning (ML), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naive Bayes (NB).**

I. Introduction

Cancer is a disease in which abnormal cells divide and invade nearby tissues. Cancer cells can also spread to different body parts through blood and lymph systems. One of the main reasons for cancer is the gene mutation. When a patient seems to possess cancer, a tumour sample is taken from the patient which then undergoes through genetic sequencing of DNA. Once sequenced, a tumour can have thousands of genetic mutations. Now with the help of gene and its variation, we must classify which class it belongs to. The data set is taken from Memorial Sloan Kettering Cancer Centre which consists of the following features- genes, variation, clinical literature, and class to which the gene variation belongs to. In a manual diagnosis of cancer, a pathologist first selects a list of genetic variations to analyse, then he/she searches for evidence in the medical literature that is relevant to the genetic variation. Finally, the molecular pathologist spends a huge amount of time analysing the evidence related to each of the variations to classify them which is a very time-consuming task. So, the goal is to replace the final step with a machine learning model. Some of the constraints of this problem is the Interpretability of the algorithm which is a must because in the end pathologist should understand why the model is given a class. Next is no low-latency requirement which means result is not needed in seconds or minutes, the patient can wait for some hours [2]. As there's no low-latency requirement complex machine learning models can be applied. The last constraint is the errors which can be very costly. Also, the probability of belonging to class is needed.

II. Related Work:

In cancer detection field there are many studies with many concept and methods. Some of them are discussed below. This method was presented by Zhang et al. Using Ultra Wide Band antenna to get microwave images of part with cancerous information. They used gelatine-oil to get experimental results to manifest the efficiency of microwave images.

They are being widely used in detection of Cancer types.
A logistic Generalized Additive Model which was proposed by Roca-Pardinas et al. They used kernel smoothers with logistic GAM, and they speeded up their system using many techniques. In this simulation model they used odds-ratio curves. [1]

Machine Learning Problems:

Reading Data:
2 files of data are collected in csv form for Training variants & Training Text.
Train Variants consists of following feature- Id, Genes, Variation, and Class.
Text Variants data file consists of two features- Id and Text Literature

	ID	Gene	Variation	Class
0	0	FAM58A	Truncating Mutations	1
1	1	CBL	W802*	2
2	2	CBL	Q249E	2
3	3	CBL	N454D	3
4	4	CBL	L399V	4

Fig1: Training Variants

	ID	TEXT
0	0	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	Abstract Background Non-small cell lung canc...
2	2	Abstract Background Non-small cell lung canc...
3	3	Recent evidence has demonstrated that acquired...
4	4	Oncogenic mutations in the monomeric Casitas B...

Fig2: Training Texts

Imported these files using pandas' package.

A flowchart depicting the machine learning process involved here is depicted below as follows:

.

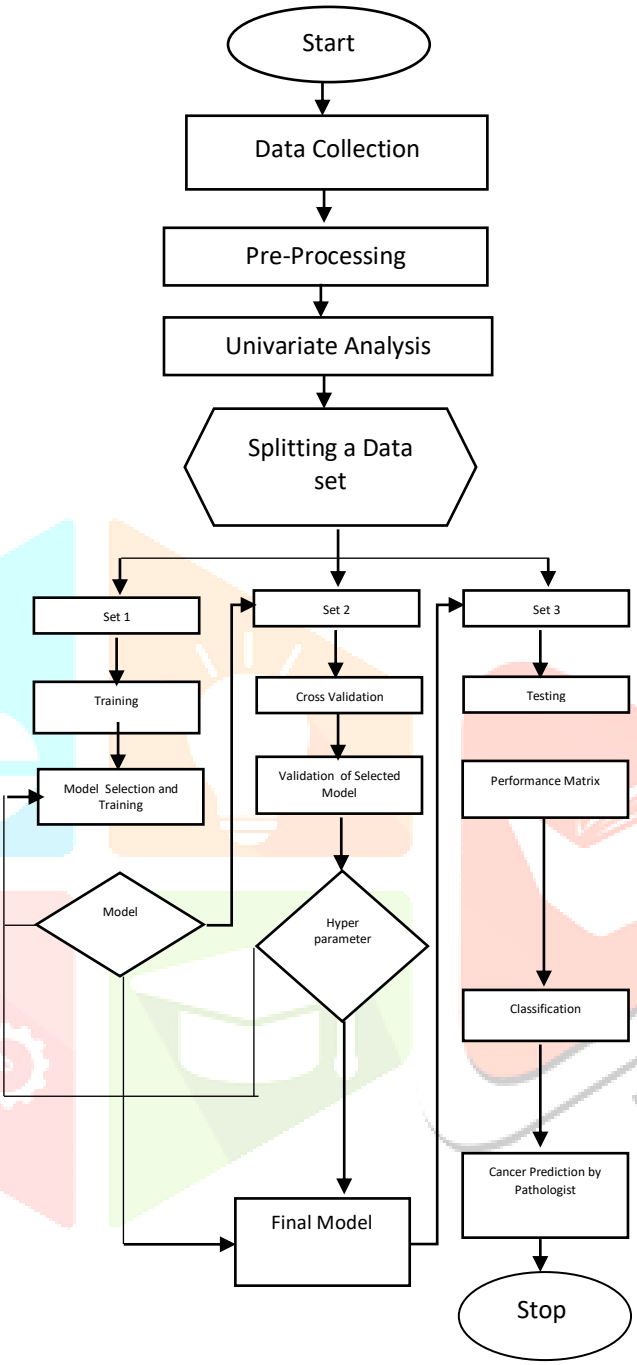

Fig3: Flowchart of the process involved

Pre-processing-

After reading data, the very first step is to perform pre-processing step, that is removing stop words, converting upper case text to lower case, and removing the punctuations.

Splitting the data-

As the data is not time based in nature which means the data is not changing with time so we split the data randomly into training, cross validating and testing to train the model using training data set, using cross validation data for finding appropriate hyperparameter after the model is trained.

Exploratory Data Analysis –

After performing the exploratory data analysis, it was found out that training and test data have similar distribution and from distribution and it is found that data is imbalanced. Some of the classes are present more as compared to other classes. It is also found

out that 3221 data points were present in the dataset. The class 7 had the highest distribution of y_i in train, test, and cross-validation dataset. Also, class 1, 2, 4, 7 had higher distribution than class 3, 5, 6, 8, 9. It was also observed that class 8, 9 were very rarely found in dataset which means that these classes were not uniformly distributed as some classes were present more in comparison to other classes. The distribution of class y_i is plotted for test data, CV data, and train data as shown:

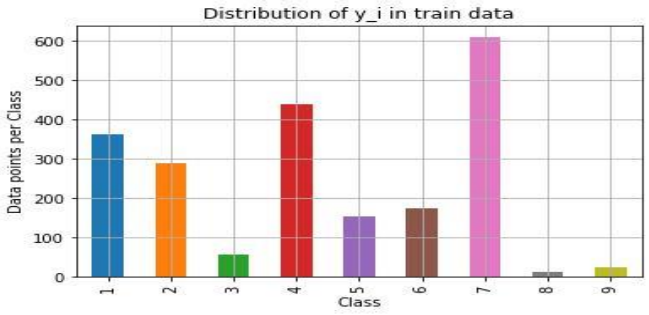

Fig4: Distribution in test data

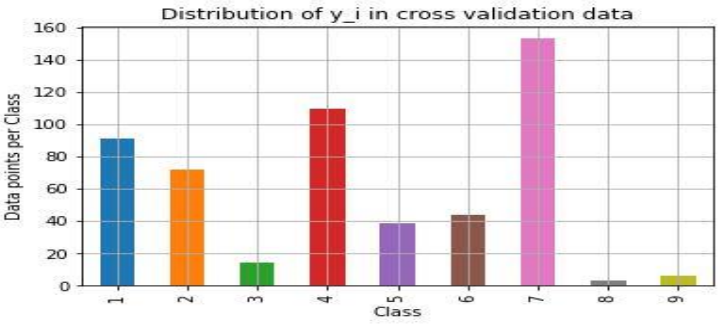

Fig5: Distribution in Cross Validation Data

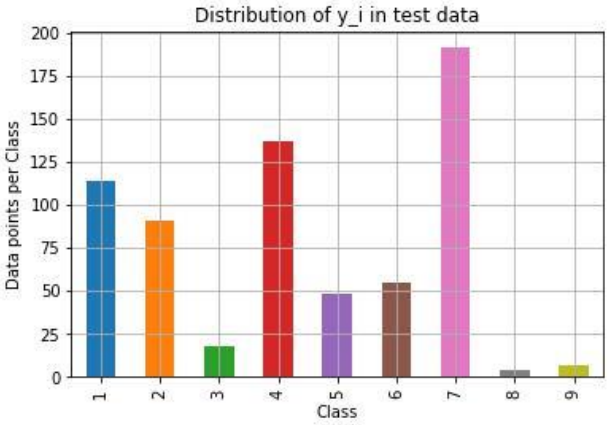

Fig6: Distribution in Test Data

Now it is known that log-loss ranges from 0 to infinite. So firstly, a random model is defined so that if ML model has log loss less than the random model, then the model is considered as good. After giving data to random model, it gives a log loss of roughly 2.5.
Also, it was checked that the precision and recall matrix in which diagonal elements (which are precision and recall of all the classes) which came out to be very low because of a random model. Univariate Analysis- Each feature is taken and checked whether it is useful for predicting in class label by various ways so that it is known whether which features are useful, and which are not.

1. Gene Feature:

The gene is a categorical feature from which it is observed that there are 235 types of unique genes out of which top 50 most frequent genes nearly contribute to 75 percent of data.

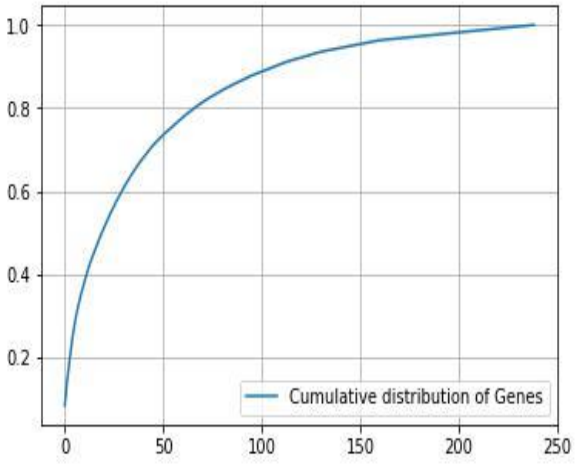

Fig7: Cumulative Distribution of Genes

After that, gene is featured into vector by both one hot encoding and response coding. Then a simple Logistic Regression model is built and gene feature and class labels to it are applied.
It is found that Train, CV, Test log-loss values are nearly same and also it is found out that log loss value is less than 2.5 that is random classifier values. Hence it is said that Gene is crucial feature for classification.

2). Variation Feature –

The variation feature is also a categorical feature and it is observed that 1927 unique variations out of 2124 were present in training data, which means most of variations occurred once or twice. Variation is featured into vector as done earlier for gene feature and a simple LR model is built and data is applied to it and it is found out that the loss values of Train, Test and CV is less than the Random Model, therefore Variation feature is also an important feature.

3) Text Feature –

The text data are total 53,000 unique words which are present in training data. It is also observed that most word occurs very few times which is common in text data. So, we convert the text data into vector by Bag of Words and Response Coding. As done in previous cases we then apply the model LR and log loss values of Train, CV, Test that are found to be less than Random Model. So, the test feature is also important feature.

Methods

1). Naive Bayes (NB)-

For the text data NB model is a baseline model. Training data is applied to the model and the CV data is used for finding the best hyperparameter that is (alpha).

After finding the best alpha, model is fit. The test data is then applied to the model and it is found that log-loss value is 1.23 which is quite less than random model. Here it is also found out that the total number of mis-classified cases is 42.4 percent. The probabilities of each class for each data interpreted each point are also checked. This is to check why it is predicting class randomly. It is then concluded that for misclassified points, the probability that point belongs to a predicted class is very low. From the precision and recall matrix it is found out that most of the points from class 2 predicted as 7. Similarly, most points from class 1 are predicted as 4.

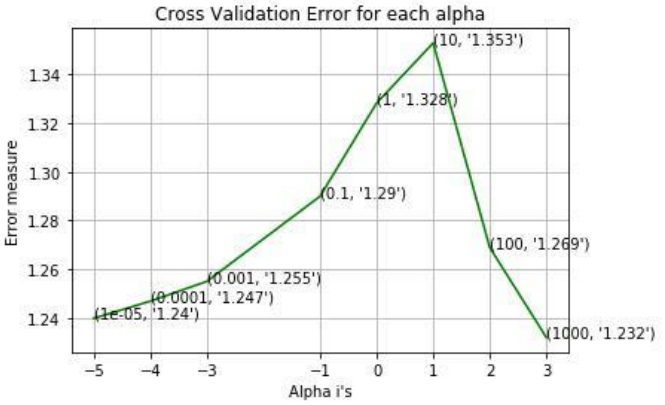

Fig8: Cross Validation for each alpha in NB

2). K Nearest Neighbours-

The KNN model is not interpretable but is still used just to find out the log-loss values.

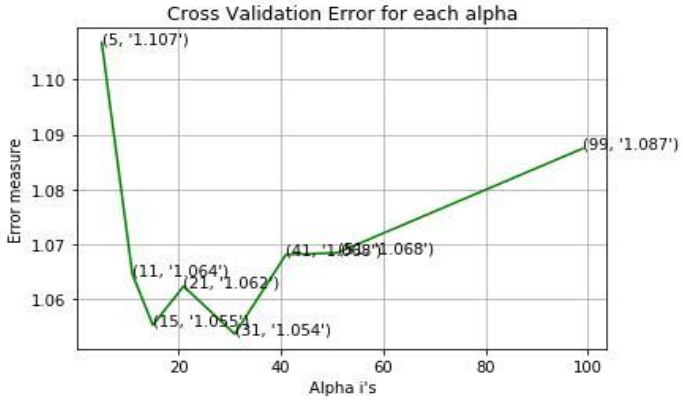

Fig9: Cross Validation error for each alpha in KNN

As KNN model suffers from curse of dimensionality, response encoding method is used instead of one-hot encoding. After applying the data to model best hyperparameter(k) is obtained.

With the simplest 'K' model is fit and test data is applied to the model. The log-loss value is 1.05 which is less that NB model and percentage of mis-classified point equals to 37%. In KNN model it is found out that most of points from class 2 are predicted as 7. Similarly, most of points from class 1 predicted as 4.

3. Logistic Regression (LR)-

The LR model worked very well with univariate analysis. So, some analysis of LR is done by taking both balanced and imbalanced data. With Class Balancing – It is known that LR works well with high dimension data and is also interpretable. So, oversampling of lower-class points are done and applied to the training data to the model and the CV data to find the hyper parameter lambda.

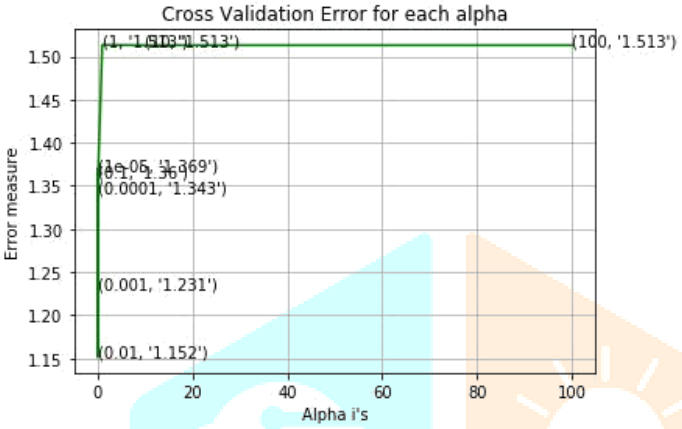

Fig10: Cross Validation error for each alpha in Logistic

With the simple lambda fitted to the model and test data is applied to the model. The log-loss value is 1.09(close to KNN). But number of misclassified points are 34.9 percent (which are less than NB and KNN). As LR is interpretable and misclassified points are less than other models that is KNN and NB it is better than KNN and NB. Without class balancing log loss and misclassified points are increased. Therefore, class balancing is used.

4.Support Vector Machine (SVM)-

Linear SVM with class balancing is used because it is interpretable and works very well with high dimension data. RBF Kernel SVM is not interpretable so it cannot be used. Training data is applied to the model and the CV data is used for finding the best hyper parameter C.

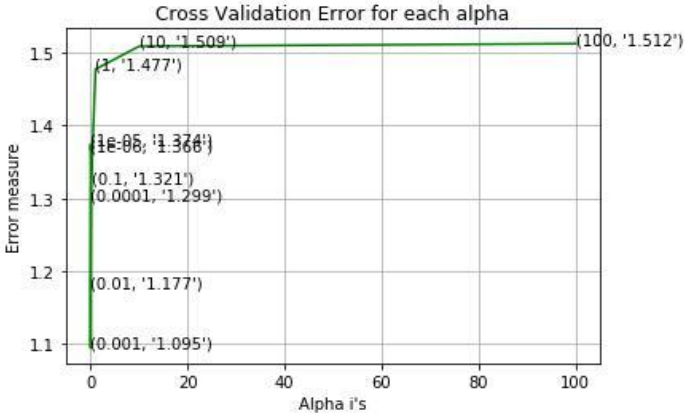

Fig11: Cross Validation error for each alpha in SVM

The simplest C model is fit, and test data is applied to the model. Now, the log-loss value is 1.15 that is near to LR, which is quite less than Random Model. Here, the total number of mis-classified cases is 34.96 percent (more than LR). Since class balancing is used, we got good performance for minor classes.

Results:

The data was used for training, test, and validating processes. The utilization percentage of data for testing, training, and validation is 70%, 20% and 10%.
First pre-processing step is performed, then exploratory data analysis, and a univariate analysis of each of features is done. After that it is applied to 5 different ML - LR, KNN, NB, and SVM. The log-loss for each of the model and the miss-classified percentage of all data points is calculated as shown in the Table: -

Sr. No	Model Name	Train log-loss	CV log-loss	Test log-loss	% miss-classified points
1	Naïve Bayes (NB)	0.91	1.20	1.20	36%
2	K-Nearest Neighbors (KNN)	0.65	1.09	1.05	36.6%
3	Logistic Regression with Class Balancing	1.117	0.95	0.95	29.4%
4	SVM	0.56	1.10	1.14	37.02%

Table 1: Comparison of all model used

The maximum accuracy obtained is by using Logistic Regression with Feature Engineering (One Hot Encoding) with the maximum accuracy of 71.6%.

Conclusion:

In this paper a disease is discussed that causes deaths every year and some models are proposed for early detection of this disease with high accuracy. The proposed method consists of mainly two parts one is training model and the other is testing model with machine learning techniques like logistic regression and support vector machine. It is observed that features by logistic regression had more accuracy of about 71% with only fewer features used. Techniques like feature engineering technique is used like stacking the features together, Stacking the model together and doing a Maximum Voting Classifier. Then it was tried to combine Gene and Variation into one list and apply Logistic Regression on both One Hot Encoding and Response Encoding by which it was able to achieve the accuracy of 71% which was better than most of the previous work done related to diagnosis of cancer using Machine Learning. A table is built for the comparison of few of the previous works done in this field as shown below in which it is mentioned that the different publication and the methodology they have used or the algorithm they have used like SVM, LR. The point to be noted here is that all the publications have used different datasets like SNPs, Clinical Colon Carino matosis, Cerviva 1 cancer, Breast Cancer and the study has been done on different type of cancers like, Muliple Myelona, Breast Cancer, Cervical 1 cancer so accuracy depends a lot upon data used and the type of cancer being diagnosed so the study performance cannot be compared accurately. However, it is mentioned so [1] the similar research on similar type of diagnosis of cancer using machine learning.

S. No	Publication	Method	Cancer Type	Type of Data	Accuracy
1	Wadell m. et. Al [1]	SVM	Multiple Myeloma	SNPs	70%
2	Listgarten J. et. Al	SVM	Breast Cancer	SNPs	69%
3	Stajadinovic et.al	BN	Colon Carinomatosis	Clinical Pathologic	71%
4	Tceng C-J et.al	SVM	Cervical cancer	Clinical Pathelogic	68%
5	Park k. et.al	Graph based SSL algorithm,	Breast cancer	SFER	71%

Table 2: Related Work done in the field of Machine Learning

References

[1] A. A. M. A. Moh'd Rasoul Al-Hadidi, *Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm",* 2016.

[2] S. J. Geetika Aggarwal, *Analysis of Genes Responsible for the Development of Cancer using Machine Learning,* ICISC, 2019.

[3] T. P. E. P. E. M. V. D. I. F. Konstantina Kourou, *Machine learning applications in cancer prognosis and prediction,* Computational and Structural Biotechnology Journal,, 2015.

[4] *Detection Analysis of Various Types of Cancer by Logistic Regression using Machine Learninng,* International Journal of Engineering and Advanced Technology, 2019.

[5] B. Y. Q. L. J. S. X. W. Qingguo Zhou, *Deep Autoencoder for Mass Spectrometry Feature Learning and Cancer Detection,* IEEE, 2020.

[6] "papers.ssrn.com," [Online]. Available: papers.ssrn.com.

[7] *Student paper, American Public University System.*

[8] pdfs.semanticscholar.org. [Online]. Available: pdfs.semanticscholar.org.

[9] *Submitted to University of Northumbria at Newcastle.*

[10] "worldwidescience.org," [Online].

[11] www.coursehero.com. [Online]. Available: www.coursehero.com.

[12] J. K. R. L. Y. Ravindra Kumar Yadav, "Dielectric Loading on Multi-Band Behaviors of Pentagonal Fractal Patch Antennas," no. Open Journal of Antennas and Propagation , 2013.

[13] M. S. Dung Tran, *Applying multilabel and multi-class classification to enhance K-anonymity in sequential releases,* Progress in Artificial Intelligence, 2016.

[14] *Student Paper, University of Edinburgh.*

[15] *Student Paper, National College of Ireland.*