IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007204|

Personalized Medicine: Redefining Cancer Treatment

Shivam Sharan [*1], Rutuja Sukede[*2], Mayank Prasad[*3], Manish Ranjan[*4], Kanchan Jadhav[*5]

Final Year Student, Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra, India[*1]

Final Year Student, Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra, India[2]

Final Year Student, Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra, India[3]

Final Year Student, Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra, India[4]

Assistant Professor, Computer Engineering, Sinhgad Academy of Engineering, Pune, Maharashtra, India[5]

A few years ago, a many research efforts are centred on genetics, understanding the disease and choosing treatment that is most suitable for patients. Genetic testing is one of the process of making medicines with new precision and method to treat Cancer. A major obstacle to use Genetic segregation is still a major craft and it is necessary. Memorial Sloan Kettering Cancer Centre (MSKCC) presented the competition, hosted by NIPS 2017 Competitive Track, as help was needed to make a medicine for you fully. MSKCC is a cancer treatment and research centre in New York. By the world's largest and oldest private cancer centre. A cancerous plant may have thousands of mutations in one set of genes. The purpose is to separate the conversion that you offer plant growth from neutral mutations.

ABSTRACT: Establish a diagnosis according to a specific sequence that can provide potential benefits to patients, insurance companies, and the health care system. Genetic data is available, it is produced in quantities that require careful planning. Therefore, the sequence of diagnoses is widely used. We have discussed the various methods of machine learning that accurately assess the clinical suitability of genetics relationships of new translation findings of clinical research context and also increases the level of diagnosis. The solution to dealing with big data successfully is to use a Machine to read. Medical science produces a large amount of information everyday research and development bases (R&D), physicians and clinics, patients, caregivers, etc. Mechanical diagnosis learning applies when the situation can be reduced to the function of distinguishing by physical details, in places where we are currently relying on the clinic to see visual patterns that indicate the presence or type of condition. A much has been said over the past several years about how drug accuracy and, more importantly, how much genetic testing will interfere with the treatment of diseases such as cancer. However this is only partial because a large number of manual labour is still required. This project aims to build a machine learning model to take personalized medicines with absolute power.

KEYWORDS: Genetic Mutation, Cancer Treatment, Machine Learning, Text Classification, Gene, Variation.

I. INTRODUCTION

Genetics is the sequence of nucleotides in DNA or RNA that involves the synthesis of a genetic product, either RNA or protein. Nucleotide is made up of 3 elements: Phosphate, Sugar and Nitrogen Base. Five bases are adenine, guanine, cytosine, thymine and uracil, with symbols A, G, C, T, and U, respectively. Differences in DNA among humans is known as Genetic Diversity, many sources of genetic variation including genetic mutations and genetic regeneration. Conversion is a modification of the nucleotide sequence of the genome of the creature. Once respectively, a cancerous tumour may have thousands of genes transformation. The translation of genetic mutations is very complex time-consuming work because it is done by hand. Accurate prognosis for survival in cancer patients remains challenging due to increased heterogeneity and the complexity of cancer, treatment options and patients. In the current practice, physicians have used data collected by bedside consultation, medical records or cancer registrations designed for the purpose to help predict and make decisions. If achieved, reliable predictions can be helpful personal care and treatment, and institutional improvement effectiveness in cancer management. This project aims to make Learning Machines an algorithm that uses the knowledge base described there world-class researchers and oncologists have done it by hand describe thousands of changes as a basis, so that automatically adjust genetic diversity. Gene sequences were immediately removed from the study domain in clinical setting.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007204|

II. RELATED WORK

1. Machine learning applications in cancer prognosis and prediction [1] authored by Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis.

This paper discusses the division of cancer into categories such as a circular disease that contains many different subtypes. Early diagnosis and prognosis of cancer types becomes a necessity in cancer research, as it can help subsequent clinical management of patients Importance of classifying cancer patients into higher or lower groups led many research groups, from biomedical and bioinformatics, studying machine use learning methods (ML). Therefore, these four methods have used as a model for continuous modeling and treatment of cancerous conditions. In addition, the ability to ML tools to detect key features from complex datasets reveal their importance. A variety of these methods, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Vector Machines (SVMs) and Decision Trees (DTs) widely used in cancer research development of predictive models, leading to re-operation making the right decisions.

2. Classification of Genetic Mutations for Cancer Treatment with Machine Learning Approaches , Gangmin Li and Bei Yao

This review focuses on our current understanding and advantages and disadvantages of various tumor cell-based mechanisms custom. Various culture matrices like biomimetic scaffolds and chemically defined media supplemented with essential nutrients, have been prepared for different tissues. These primary tumor cells redefine cancer therapies with high translational relevance.

3. Applications of Machine Learning in Cancer Prediction and Prognosis[2] by Joseph A. Cruz, David S. Wishart.

In this paper there are many trends noted for cancer prediction, including growing reliance on proteins biomarkers and microarray data, strong direct bias applications to prostate and breast cancer, as well as weight
on such "old" technologies of artificial neural artificial networks (ANNs) instead of newly built objects or machine translation methods. The number of published studies also seems to be lacking the appropriate level of validation or testing. Among the best studies designed and validated are clearly machine Learning methods can be widely used (15-25%) improve the accuracy of predicting cancer diagnosis, repetition and death. At the basic level, of course and the evidence that machine learning also helps to improve our basic understanding of cancer growth as well to continue.

III. METHODOLOGY

For this project, we have done Machine Learning an algorithm that uses informational information that is not defined they were the basis, which automatically distinguished genetic variation. Our approach to making the machine learning feature of this project to try a few ways to differentiate and feature combination as much as possible - in such a way that each method is possible to bring a new perspective that takes on different features from each category. The first step is to visualize the data and extract details as much as possible. After that we will have to check the file for the relationship between genetics and categories and as such knowing that a single gene can fall into many different categories, suggests that mutations within the same species are possible produces very different results. It is known that Most of the transformation of each situation is the transformation of points, in which amino acid is converted to another. It must be noted that genes are included in the training and testing database they were almost completely different, and from genetics / mutation data sets contain limited information, we found a lot of our details from text data.

This project aims to make Learning Machines an algorithm that uses the knowledge base described by world-class researchers and oncologists have done it by hand describe thousands of changes as a basis, so that automatically adjust genetic diversity.

The Machine Learning Model used is as follows:

i. Naïve Bayes

ii. K-Nearest Neighbors

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007204|

iii. Logistic Regression

iv. Linear Support Vector Machine

v. Random Forest Classifier

vi. Stacking Model

In our study, the task was to differentiate genetic mutations into enable a drug designed for you to treat cancer. Data comes from a publicly available source provided by Kaggle a competition called "Personalized Medicine: Redefining Cancer Treatment". Because of the text-based nature of the text data, modification was required for any separation skills to be adopted. We used TF-IDF, one hot code and feature-encoding features conversion; we have used the above separation models to make compare performance. Midnight test Dividing methods are performed according to the multi class log loss, described by Kaggle's website.

Dataset Analysis

The database has three parameters: genetics, differentiation and treatment text. In the training set, 9 classes of genetic modification are offered. Our aim was to use the three variables provided for prediction transformation classes. The training set has 3321 data entry nearly four times the test rate. Same Distribution can also be seen in a variety of variations. However, there were many genes on the test set compare the number of genres in the training set. Both training and testing data sets are provided in pairs separate files. A separate file provides details about genetic modification, and the test data file provides the clinical evidence in the format of the text the pathologists use edit genetic modification. Both of these data files are linked by using the ID field.

There are two training files i.e., training_text, training_variants.

1. Training_text has an ID, TEXT columns

2. Training_variants has an ID, GENE, VARIATION, CLASS columns

The training_text has 3321 rows and training_variants has 3321 rows. The dataset_variants dataset has 9 different classes that can be guessed based on the Gene, Variation, and Text columns. The output is fragmented and is therefore a Multi-Phase problem.

Data Preprocessing and Cleaning

Part of the text it actually contains relevant research papers for all classes that was specified in that case by hand. Text, therefore, it has many numbers and stops. As it is research paper where there may be gaps that should not be it needs to be addressed. These unnecessary parts were present removed using the NLTK library which is a platform used for Python programs that work with personal language data for to apply for Statistical Natural Language Processing (NLP) consists of token processing libraries, filtering, sorting, marking, marking and semantic to consult. We have processed the text by removing the numbers, which are not valid spaces and convert it to the bottom item to avoid mistakes using Regex. A decision was made to manage the lost data using Imputation by entering null values with GENE integration and the VARIATION column. Details are checked for testing prices do not apply after installation. Before data is segmented, all spaces in Geneva and The divergence column instead was taken _. Details are divided into training, testing, and certification for when Hyperparameter adjustment is performed, we try to improve accuracy in relation to a Test set that may lead to mischief sometimes models. The set of outcome details, therefore, was first broken down into 80% and 20% division of Trains and Tests and 80% of the train was also divided into 80% and 20% in the finals Train and confirmation set.

Implementation Strategy

As log loss is considered a primary test A random model parameter was created to compare loss of log against. Any successful model will return low log loss than this random Model. To check log loss, cross data verification

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007204|

Random Model, egg list made into the same length as the verification data. Random the opportunity is made for all nine Class values in all row also stored in this list in relation to length.

Problem with category data:

Many machine learning algorithms do not works on label directly. All inputs and outputs variables need to be numbers. This is a barrier to efficient implementation of machine learning algorithms instead of difficult limitations in the algorithms themselves. This means that category data must be converted to a numerical form, so a machine learning algorithm can work it. Conversion methods:

i. One hot coding

ii. Response Encryption (Mean Imputation)

Column checking:

The columns have been tested to see if they are correct contributes to the class column prediction which is variation depends.

According to the successful generation of log confusion matrix, matrix accurately, remember the matrix and determine the path of non-separate point functions were fixed ahead of time. All three variables are made using One-hot Encoding and Response Encoding were integrated respectively using the numpy function hstack. According to one hot code coding the features are:

1. Train (2124, 55129)
2. Test (665, 55129)
3. Verification of crossings (532, 55129)

According to Response encoding features are:

1. The train (2124, 27)
2. Test (665, 27)
3. Crossing confirmation (532, 27)

Therefore, we can conclude that Response Encoding will be limited number of columns in the best way.

IV. BUILDING MACHINE LEARNING MODELS

Naive Bayes

The Naive Bayes dividers are a family of simple "possible" things classifiers "based on the application of the theory firmly Independent (ignorant) assumptions of autonomy between factors. They are very scary, requiring a number of specific parameters to number of learning variables the problem. Training can be done by checking the closed form expression, which takes the time of a line. Separate with different features (e.g., word count for text separation), the international division of the Naive Bayes is suitable. Usually requires a full feature calculation. The Multinomial Naïve Bayes divider has been used in this project as we have detailed information in the form of 9 classes as well and because the model of the Naïve Bayes is very flexible.

K-Nearest Neighbors Algorithm (KNN)

The K-Nearest Neighbors algorithm (KNN) is not parametric method used for separation and reversal. In both cases, input contains the closest examples of training in the feature space. Depending on whether KNN is used for segregation or retreat, the effect changes.

i. According to the KNN section, the output of the class membership. Depending on the majority vote of its neighbors, the thing is separated by something given to the class too much common among its closest neighbors (k is a good thing total value, usually small). If k = 1, then the object is simply given a section of this nearest neighbor.

ii. In KNN deferred, the output is a property value of thing. The resulting value is an approximate value of k nearest neighbors.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007204|

Logistic Regression

It is an algorithm for the machine learning phase which is used to predict the potential for class dependence variable. In procrastination, flexibility depends on a binary conversion containing data listed as 1 (yes, success, etc.) or 0 (no, failure, etc.). Model of systematic order predicts P (Y = 1) as X function. The Logistic Regression model has been used because it is so advanced it is translated and can be used for many categories easily. In the LR model we use two methods:

i. Excessive sampling means classes have fewer data measure yourself in relation to other classes.
ii. Without measurement which means classes will not be measured get out.

Linear Support Vector Machine

Vector Machines support is based on the concept of finding a hyperplane that neatly divides the database into two categories. Data points closest to the hyperplane, is Support to carry. Because of this, they are considered critical data set elements. Linear Support Vector Machine was used because it is highly translated model.

Random Forest Classifier

Random Forests is a supervised learning algorithm, it can be used for partitioning and retrieval. It's over again a very simple and easy to use algorithm. The forest is like that containing trees. When it has many trees, in the forest. Creates decision trees on randomly selected details samples, find the forecast for each tree and choose the best one Solution by voting. Random Forest Planning is combine algorithm. Combine algorithms are those It includes more than one algorithm that is the same or different sort of sort of things. For example, using prediction over Naive Bayes, SVM, and Decision Tree and take the vote for the final consideration of the test item phase. The key word Bagging directs the group, i.e. means a group of objects that are considered as a whole. To ensure compliance, we need to do the following:

i. We have to build more models
ii. We have to cover their results.

V. WORKING OF THE SYSTEM

Frontend:

i. The user will input the genetic variation and gene in the system.
ii. As the user clicks the Prediction Button on the system these variation and gene are carried to the backend where they are processed to make an appropriate class prediction.

Backend:

i. At the backend side, all the necessary libraries including pandas, numpy, etc. are imported.
ii. After importing all the libraries, the first step is Data Preprocessing where the user input is processed by removing numbers, inappropriate spaces and converting it into the lower case to avoid errors.
iii. After Data pre-processing, the user data is cleaned and the missing values present in the data are handled by replacing the null values with the concatenation of GENE and VARIATION column. The data were crosschecked for null values after imputation.
iv. In the next step the Label Encoder is used which encodes categorical features as a one-hot numeric array. Label Encoder is used to normalize labels. Label Encoder transforms non-numerical user data to numerical labels. This is done in order to ensure that the user data is accepted by the model as the model can only process numerical data.
v. This user data is then passed to the previously trained and saved model. The model then makes the class prediction for the user input. vi. This predicted class is then displayed on the frontend.

VI. CONCLUSION AND FUTURE WORK

A well-organized gene database promotes accurate phenotype-driven gene analysis. It enables classification as new studies are published, providing patients with a diagnosis and opportunities for therapeutic options, and an end to their "diagnostic journey." This classification system is simple enough to be quickly implemented, and it can specifically guide reporting decisions at the important boundary of limited and moderate evidence, determining whether a gene is characterized.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007204|

Model	Train log-loss	Cross-Validation log-loss	Test log-loss	Miss-classified %	Log-loss
Naïve-Bayes	0.9	1.18	1.27	34.58	1.18
K Nearest Neighbors	0.60	0.95	1.03	31.95	0.95
Logistic Regression (With balancing)	0.54	1.02	1.05	29.69	1.02
Logistic Regression (Without balancing)	0.53	1.04	1.06	29.67	1.04
Linear Support Vector Machine	0.75	1.07	1.11	31.1	1.074
Random Forest Classifier (One-hot encoding)	0.67	1.13	1.15	36.84	1.12
Random Forest Classifier (Response encoding)	0.05	1.23	1.31	41.72	1.23

Figure : Log-Loss For Various Machine learning Models

The minimal value for the Log-Loss is obtained on implementing K Nearest Neighbors algorithm on the gene, variation,
and text data value points. Although the Log-Loss for KNearest Neighbors algorithm is less. We choose Logistic Regression (With balancing) because KNN gives an overfitted value (1.008).
As a future scope, the similar gene mutation and classification techniques can be extended to find a cure to diseases other than Cancer and can be a breakthrough technology in personalized medical space.

REFERENCES

[1] Abhishek Mitra, Lopa Mishra, Shulin Li.The technology for locating the basic cells of a tumor use ,Vol 6 , Release 3, July 2009

[2] Han-Ping-Zhu , Predicting Breast Cancer Mining Classification Algorithms Compare Study. Vol 2 , Release 4, June 2012

[3] By Sheng Sun, Zhao Zhao, , ”Danger features and prevention of breast cancer ” International Journal of Natural Sciences. Vol 3, Release 3, March 2011

[4] Alireza Osarech,”Computer Assistant Breast Cancer Screening Program ”, Worldwide Computer Science Problem Journal, Vol. 8, Release 2, March 2011

[5] Bita Shadgar, https: //www.kaggle.com/c/msk-redefining-cancertreatment Vol. 4, Release 3, March 2013

[6] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis.Machine learning applications in cancer prognosis and prediction , Vol. 2, Release 2, Oct 2018

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details