iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** II **Month of publication:** February 2023

DOI: https://doi.org/10.22214/ijraset.2023.48926

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

Cancer Tumor Mutation Prediction

Bhumica Verma[1], Ankit Kumar Verma[2], Aditya Upadhyay[3], Shivam Parasher[4], Shobhit Saini[5]
[1, 2, 3, 4, 5]*IMSEC Ghaziabad*

Abstract: ***Every day new technologies are invented to overcome the problems faced by the people. Machine learning can be seen as a boon in the series of technologies inventions within the context of smart digital society. Before 2015, in medical departments, the diagnosis what is serious disease could take a two days or, maybe a week. After the advancements in machine learning the diagnosis time is minimized rapidly. Like the path AI machine learning algorithms helps diagnosis diagnostician in making quick and more correct diagnosis, as well as pinpointing people who may gain from state-of-the-art treatments or therapeutic, our model also could extend to such kind of self learning algorithms. The analysis of our proposed model shows that it is efficient and productive than existing techniques. We have achieved the efficiency of the model by developing certain methods that transforms a data from data frame do a highly correlated data frame. This helps random forest classifier to make a prediction faster while providing a much probable class as an output. Currently this model can assist the diagnosis process and reduces the search space by 90% of the total variance, this on average reduces the time by 1.5 days.***
Keywords: Machine Learning, cancer tumor variant, cancer mutation, random forest classifier, tumor

I. INTRODUCTION

Cancer is a harmful disease in which some cells in the body multiply uncontrollably and spread to other regions of the body. Cancer may start anywhere in the cells that comprises the human body. This well-ordered mechanism of the body occasionally fail, resulting in abnormal growth developing and producing when they shouldn't. The most common gene disorder in cancer patients is p53, also known as TP5 Most of the cancer patients are diagnosed missing or defective p53 gene. Although inherited p53 mutations are infrequent, those who have them are more likely to get cancer.

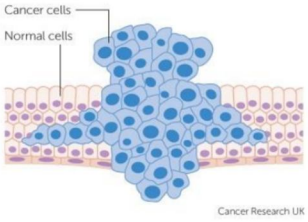


Fig 1.1 Cancer Cells Representation

A mass of abnormal tissue created when cells grow and develop more quickly than they should or do not die when they should. Tumors can be benign (not cancerous) or malignant (cancerous). Although benign tumors can get extremely large, they do not spread or infect surrounding tissues or organs.

II. LITREATURE REVIEW

Many machine learning models uses different techniques to make final prediction, here in this section we will discuss the most relevant below.

Related Work

Kourou [1] imposed various algorithms for predicting the survival rate. Most of the recent studies have focused on the development of predictive models using supervised ML methods and classification algorithms to predict valid disease outcomes. Based on their findings, it is clear that combining multidimensional heterogeneous data with different techniques for feature selection and classification can provide promising tools for inference in the cancer domain. We taken into account using the same methodology to predict tumor mutations because mutations can range from 100 to 1000 types. Because there are over 1000 categories to predict, the DNN concept will be extremely useful.

Knudson [2] explained in his study that classification is based on genetic disorders and inherited tumors, this will make feature engineering for our dataset much easier.

©IJRASET: All Rights are Reserved

Jacobs [3] discovered a black-box Al-based system that overrides well- established clinical guidelines that direct radiologists to base management decisions on nodule size and growth rate. However, this description lacks many technical details that could be critical for achieving a high level of performance.

D.J. McGrail1 [4] found in his study that due to the potential for tumor mutations to generate immunogenic neoantigens, high tumor mutation burden (TMB-H) has been proposed as a predictive biomarker for response to immune checkpoint blockade (ICB). TMB-H tumors had a 39.8 percent ORR to ICB [95 percent confidence interval (CI) 34.9-44.8) in cancer types where CD8 T-cell levels positively correlated with neoantigen load, such as melanoma, lung, and bladder cancers. Significantly higher than in low TMB (TMB-L) tumors [odds ratio (OR) 14 4.1, percent confidence interval (CI) 2.9-5.8, P 2 1016].

Joseph A. Cruz [5] discovered several trends in the types of machine learning methods used, the types of training data used, the types of endpoint predictions made, the types of cancers studied, and the overall performance of these methods in predicting cancer susceptibility or outcomes. For nearly two decades, artificial neural networks (ANNs) and decision trees (DTs) have been used in cancer detection and diagnosis. Machine learning methods are now used in a wide variety of applications, including detecting and classifying tumors using X-ray and CRT images. While ANNS continue to dominate, it is clear that a growing number of alternative machine learning strategies are being used and that they are being applied to many different types of cancers to predict at least three different types of outcomes. It is also clear that machine learning methods improve the performance or predictive accuracy of most forecasts, particularly when compared to traditional statistical or expert- based systems.

Yixuan Li [6] used five different classification models, including Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), and Logistics Regression (LR), to classify two different datasets related to breast cancer: Breast Cancer Coimbra Dataset (BCCD) and Wisconsin Breast Cancer Database (WBCD). The prediction results will aid in lowering the rate of misdiagnoses and developing appropriate treatment plans for therapy. This study makes use of two datasets. This study first collects raw data from the BCCD dataset, which includes 116 volunteers and 9 attributes, as well as raw data from the WBCD dataset, which includes 699 volunteers and 11 attributes. The raw data from the WBCD dataset was then preprocessed, yielding 683 volunteers with 9 attributes and an index indicating whether the volunteer has a malignant tumor. The accuracy, F-measure metric, and ROC curve of five classification models were compared, and the result showed that RF was chosen as the primary classification model in this study.

III. PRELIMINARY TECHNIQUES

In this section, we will go over all of the techniques used in implementation and determining results. This section assists in gaining a better understanding of all of the techniques used to carry out the proposed work.

A. Supervised Learning

You have input data called features and the expected result called label with Supervised Learning. It enables us to make predictions based on a model built from historical data and a predetermined algorithm.

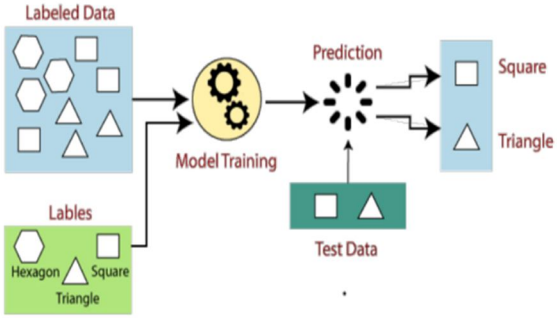


Fig 3.1 Supervised Learning

Supervised learning attempts to answer two questions:

- Classification: "which class?"; and
- Regression: "how many?".

B. Data Visualization

The visual representation of information and data in a graphical format is called data visualization (such as charts, graphs, and maps).

C. Data Manipulation

The Python Pandas library is an open source project that provides many simple data manipulation and analysis tools. Before building any models, any machine learning project will have to spend a significant amount of time, preparing the data and analyzing basic trends and patterns.

D. Train-Test Split

When using a machine learning algorithm to predict data that was not used to train the model, use a split procedure training test to evaluate its performance. This is a quick and easy procedure that allows you to compare the performance of machine leaning algorithms on predictive modeling problems.

E. Response Coding

It is a method for representing categorical data in a machine learning classification problem. We represent the probability of a data point belonging to a specific class given a category as part of this technique. So, for a K-class classification problem, we get K new features that embed the probability of a datapoint belonging to each class based on categorical data value.
Mathematically Speaking We Calculate -P(class X |category= A) = P(category A ∩ class=X)/ P(category=A)

F. One-Hot Encoding

integer encoding is insufficient for categorical variables where no such ordinal relationship exists. Using this encoding and allowing the model to assume a natural ordering of categories may result in poor performance or unexpected results. In this case, the integer representation can be encoded using a one-shot encoding. The integer encoded variable is removed at this point, and a new binary variable is added for each unique integer value.

G. Count Vectorizer and Normalization

In machine learning, data normalization is used to make model training less sensitive to feature scale. This enables our model to converge to better weights, resulting in a more accurate model. Normalization makes the features more consistent with one another, allowing the model to more accurately predict outputs

H. Logistic Regerssion

A fundamental classification technique is logistic regression. It belongs to the inear classifier family and is related to polynomial and linear regression. Logistic regression is quick and easy to understand, and the results are easy to interpret. Although it is primarily a binary classification method, it can also be applied to multiclass problemsGraph

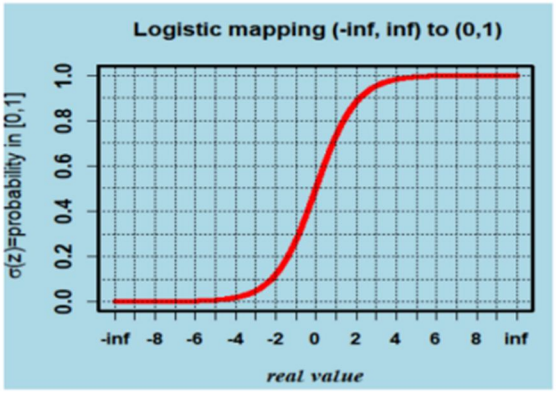


Fig 3.2 Logistic Regression Graph

I. Hyper-Parameter Tuning

A Machine Learning model is a mathematical model that must learn number of parameters from data. By training a model using existing data, we may fit the model parameters.

Some parameters, known as Hyper-parameters, cannot, however, be learned simply from the usual training procedure. They are generally resolved before the training process begins. These parameters reflect critical model aspects including complexity and learning speed.

J. Random Forest Classifier (RFC)

The random forest classifier is at the top of the classifier hierarchy. Here, we will explain how basic decision trees work and how individual DT are combined to form a random forest.

A tree-structured classifier (DT) is a tree-structured classifier in which internal nodes contain dataset attributes, branches represent decision rules, and each leaf node provides the outcome. The Decision Node and the Leaf Node are the two nodes of a Decision tree. Decision nodes make decisions and have several branches, whereas Leaf nodes, represent the results of those decisions and have no extra branches. As the name indicates, the random forest is composed of a huge number of individual decision trees that collaborate as an ensemble. Each tree in the random forest generates a class prediction, and the class with the most votes becomes our model's prediction.

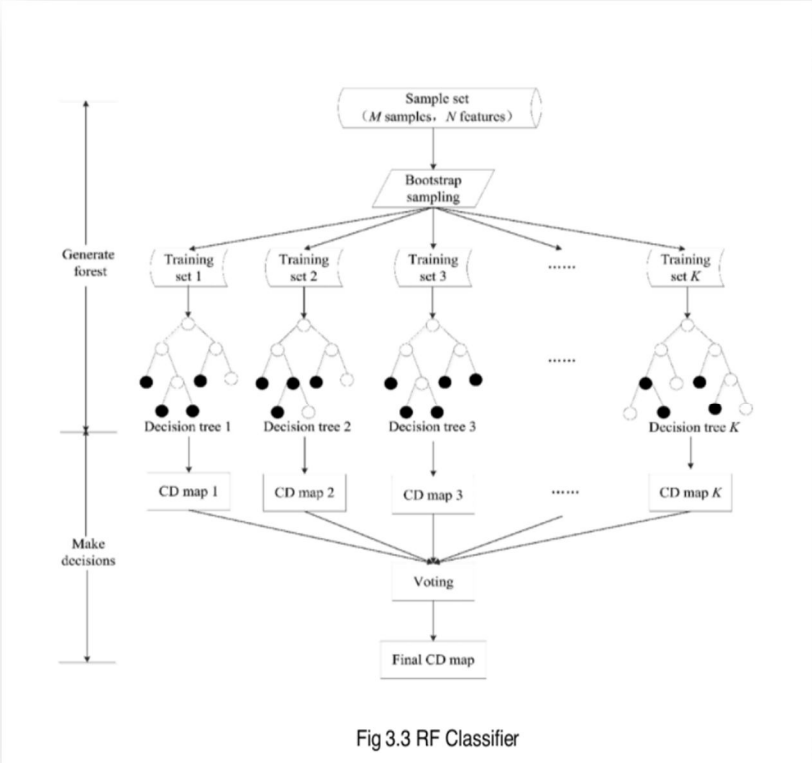


Fig 3.3 RF Classifier

IV. METHODOLOGY

A. Methodology Used

The entire implementation process is divided into three phases. The first phase is concerned with researching previous work done to achieve the current goal. The second phase is concerned with identifying some technique to improve on the existing traditional approach. The final phase focuses on implementation and improving the deployment model.. While researching previous models, we discovered a technological gap that is preventing us from achieving good accuracy for our model. As a result, we used a probability distribution approach to simplify the problem. Here, we will select the probability of different classes to get a better understanding of what we are dealing with and validate the model's performance using F1-score rather than accuracy.

B. Implementation Tools

1) *Python:* The implementation procedure is divided into three stages. The first stage involves researching previous work done to achieve the current goal. The second phase focuses on identifying a technique to improve on the existing traditional approach. The final phase focuses on implementation and improving the deployment model. We discovered a technological gap while researching previous models that is preventing us from achieving good accuracy for our model. As a result, to simplify the problem, we used a probability distribution approach. To get a better understanding of what we're dealing with, we'll select the probability of different classes and validate the model's performance using F1-score rather than accuracy.
2) *Google Colaboratory:* Colaboratory is a product of Google Research: Colab is perfect for machine learning, data analysis, and teaching since it allows anybody to create and run arbitrary Python code in the browser. Colab is a hosted Jupyter notebook service that requires no setup and otters free access to computer resources such as GPUs

C. Software Requirement

1) Operating System: Windows, Linux
2) Language: Python
3) IDE: Jyupter Notebook

D. Hardware Requirement

1) System: Intel 17 processor
2) Hard Disk: 10 GS
3) RAM: 16 GB

E. Scope of Research Work

New techniques and methods are emerging to help create a better machine learning model that can help in minimizing the computational overhead that we face due to a lack of technology, to reduce this overhead, people all over the world are working on improving these ML algorithms to help us step into Al.

In this case, the proposed technique will help reduce physical effort in locating the tumor mutation stage; additionally, because it provides a probabilistic prediction, it aids in a better understanding of all the other scenarios that could occur in the future. This model condenses the search space of 1900 variants into three classes. This technique can be extended to many other areas where the search space for categorical data is large and random.

V. PROPOSED WORK

A. Probabilistic Model

We propose a probabilistic model that describes mutation variation in different classes and predicts all possible outcomes. Because we are working with gene data, the prediction outcome is purely global, as gene data is not relevant in this context.

The Random Forest Classifier aided in the creation of n decision trees, which were then one-hot encoded to obtain the probability distribution of different classes. Because there are different types of patients with different gene information, the dataset with over 300 different types of genes is used here; there may be more genes to work with, but we only included the most common types for designing the model.

VI. CONCLUSION

We have mainly focus on improving the existing techniques and propose a efficient model that could be useful for real world problems.

By adopting the random forest classifier, we have successfully completed the first task of improving the existing techniques, and, with the techniques like class balancing, hyper parameter tuning and response coding we can deploy this model for medical checkups for predicting the cancer tumor variant.

VII. IMPLEMENTATION AND RESULT ANALYSIS

In this chapter, we have shown the results of implementation of the proposed techniques. We have also presented the performance analysis of the proposed techniques.

©IJRASET: All Rights are Reserved

A. Implementation Result

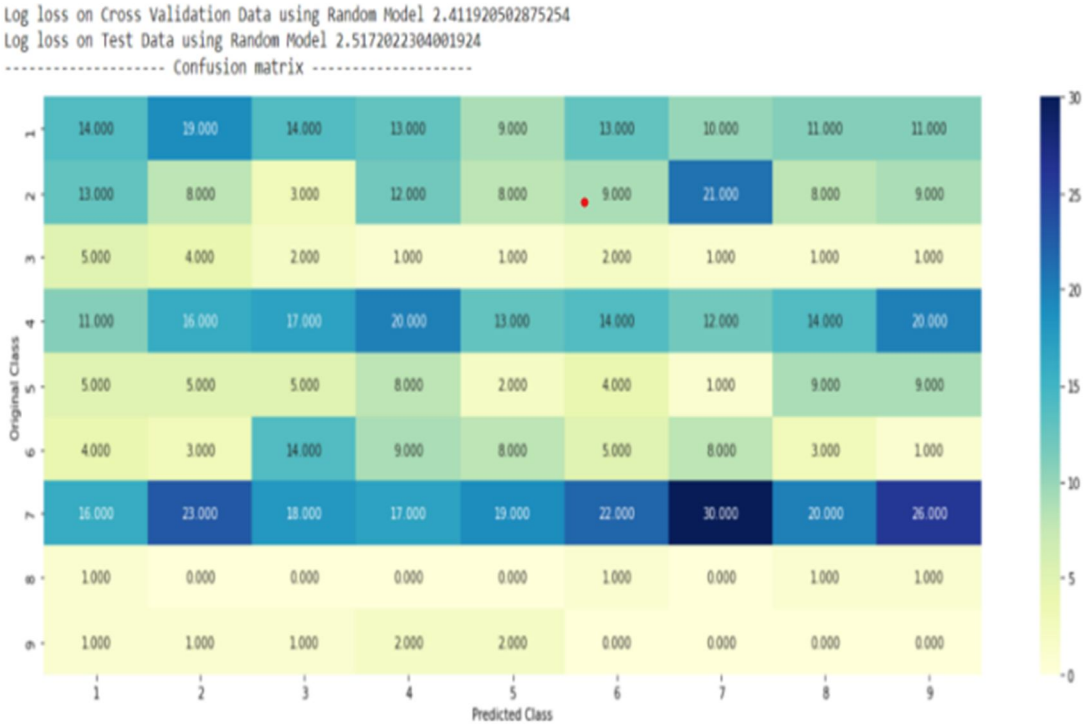


Fig 7.1 Random Model Confusion Matrix

1) In, this matrix higher the value more better the prediction, only diagonal values are true positive.
2) Low scores are obtained after intermediate model training**.**

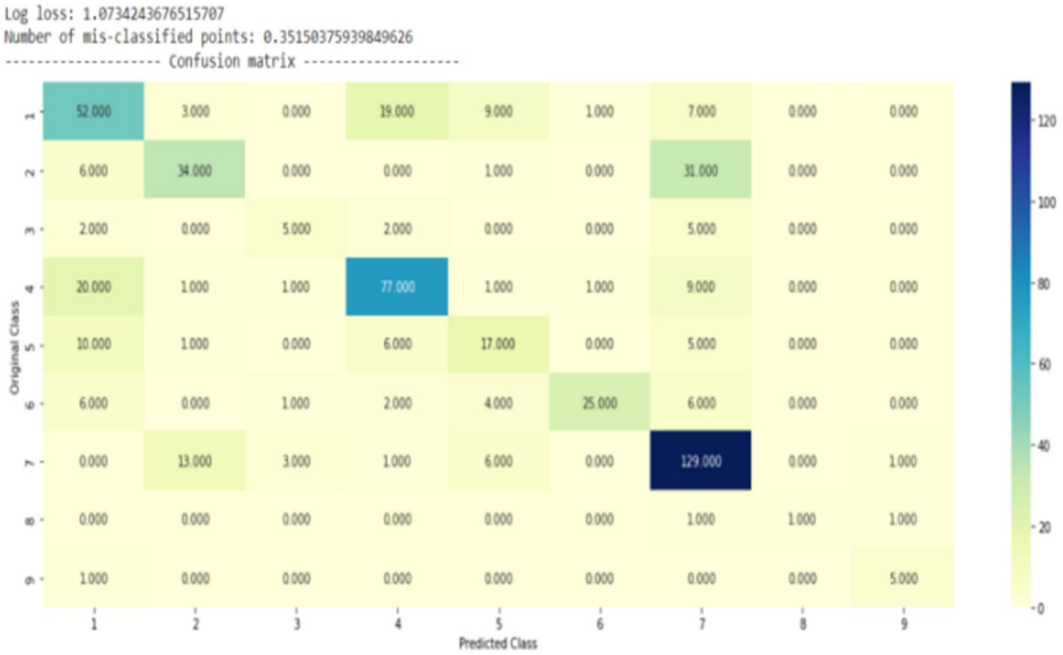


Fig 7.2 Logistic Regression Confusion Matrix after class balancing

After class balancing, true positive scores improved, but we still need to work on other classes with low scores.

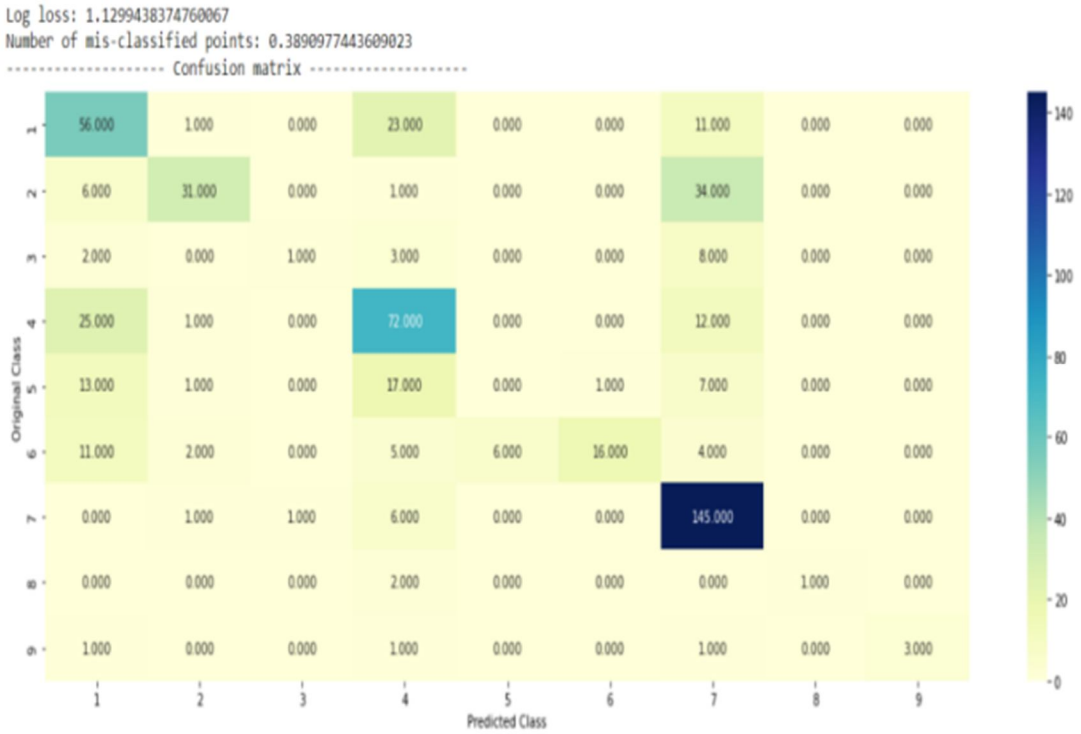


Fig 7.3 Random Forest Confusion Matrix

After using our own data processing methods the confusion matrix doesn't give much of a improvement. But in the figure 7.4 the recall matrix gives us more insight on our model performance.

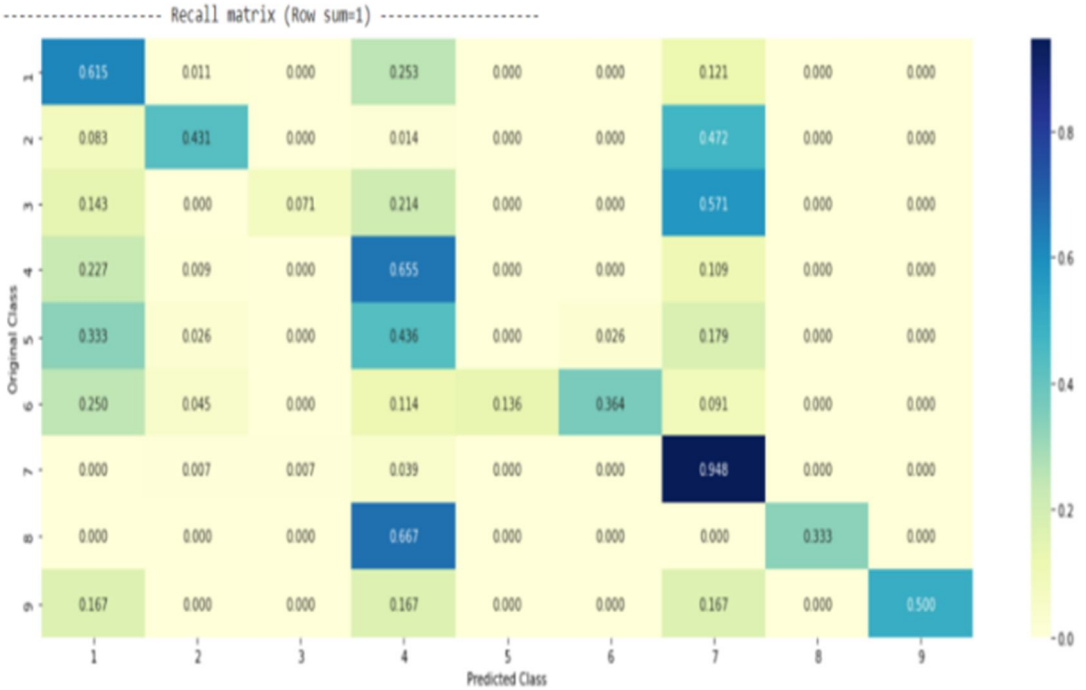


Fig 7.4 Recall Matrix for Random Forest Classifier

Model Recall matrix tells out of the total true values, what percentage are predicted true. This figure also tells us that the dataset is highly imbalance as recall score is very low for some classes.

©IJRASET: All Rights are Reserved



Fig 7.5 Random Forest Model with Hyper Parameter Tuning and Response Coding

Here in this figure we have to neglect class 8 because of its lower number of data points. After using our own methods we are able to achieve higher recall Score.

```
Predicted Class : 7
Predicted Class Probabilities: [[0.0218 0.0501 0.0113 0.0226 0.0302 0.025  0.8318 0.0031 0.0041]]
Actual Class : 7
--------------------------------------------------
```

Fig 7.6 Final Output for all 9 classes

This is the final output which gives a probabilistic class prediction as its output, here predicted class and actual class is same and the probability of class 7 is highest.

B. Performance Analysis

The confusion matrix, recall matrix, and log loss were used to assess the suggested system. The suggested approach employs the TF-IDF technology, a text vectorizer that transforms text to a vector. It brings together two concepts: term frequency (TF) and document frequency (DF) (DF). The term frequency is the number of occurrences of a certain phrase in a document. The frequency with which a phrase appears in a document reflects its value. Each text in the data is represented as a matrix by term frequency, with rows reflecting the number of documents and columns representing the number of different words across all documents.

C. Advantages of Proposed Work

We proposed a multi-class categorical predicting model that has the following advantages:

1) It reduces the work hour and errors during tumor mutation testing.
2) It offers potential classes to work with rather than variations, reducing the search space.
3) It is simpler to use because it only requires gene and patient history information.
4) It provides a usable prediction faster than any other existing model.

D. Limitations

Because this model can only provide probabilistic classes, it will fail to provide a solid prediction when working with similar genotypes. Furthermore, because accuracy cannot define performance for this particular model, real-world acceptance for this model is quite low. Because the current generation of neural networks cannot work with only text data, this model cannot be expanded to a neural network.

E. Applications

Our model provides a solution for the multi-class categorical problem through probabilistic approach. Some of the application areas where we can easily implement and work are listed below:

1) Quick Cancer Tumour Checkup
2) Online Assistance
3) Poor Countries

VIII. RESEARCH OBJECTIVE

Following the identification of the problem in the traditional approach, there is a need to research the cancer tumor mutation and improve prediction. The following are the goals of this research project:

1) To study about various works that have been done in cancer prediction field.
2) To find the problems in the previously proposed studies.
3) To find a practical solution to all the problems identified.
4) If a practical solution is not obtained for any problem, propose a theoretical approach.

A. Research Methodology

The research methodology describes the steps that must be taken in order to achieve the research goals. In this paper, we first examine the literature on multiple machine learning implementations in cancer prognosis, including their intricacies and possible improvements.

IX. MOTIVATION

The purpose of this is to differentiate reality from hype by proposing how artificial intelligence will change the field of medicine. In various forms and degrees, artificial intelligence (AI) has been used to advance and advance many fields including banking and financial markets, education, supply chain, manufacturing, marketing and e-commerce, and healthcare. Al has become a major source of business innovation in the tech industry. Web searches (eg Google), content recommendations (eg Netflix), product recommendations (eg Amazon), targeted advertising (eg Facebook), and private cars (eg Tesla). Significant progress has been made in the use of intelligent diagnostic procedures. For example, in the field of vision-oriented specialists such as dermatology, Esteva et al. Heckler et al developed a classification model using clinical thinking data to help clinicians diagnose skin cancer, skin lesions, and psoriasis. Esteva et al. trained a deep convolutional neural network (DCNN) model on 129,450 images to classify images such as keratinocyte carcinoma or seborrheic keratosis, or malignant melanoma or benign nevus (also called binary segregation problem). They also found that DCNN was working with 21 committee-approved dermatologists.

X. FUTURE WORK

In future, research work can be continued by applying the neural network, if some new technology emerges that can convert the text dataframe into a numerical vectors that can be used to work with neural network.

XI. ACKNOWLEDGEMENTS

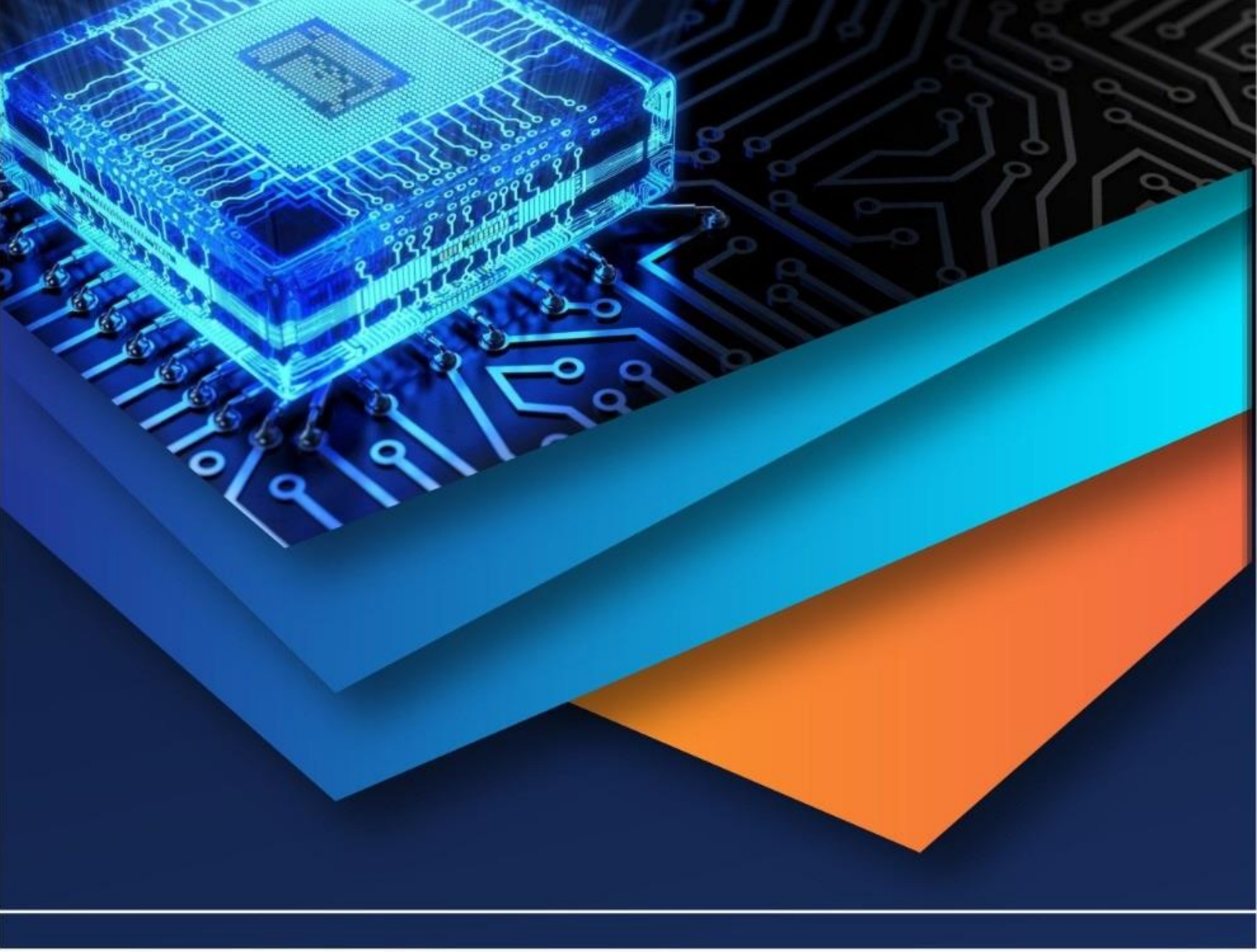
We are really thankful to Assistant Professor Mrs. Bhumica Verma from the IMS Engineering College in Ghaziabad's Computer Science and Engineering department for his assistance in assisting us with the application of our research to the real world. Its our privilege to express our sincere regards to our project guide, Assistant Prof. Mrs. Bhumica Verma for his valuable inputs, able guidance, encouragement, cooperation and constructive criticism throughout the duration of our project.
We sincerely thank the Project Assessment Committee members for their support and for enabling us to present the project on the topic. “Cancer Tumor Mutation Prediction”

REFERENCES

[1] Karamouzis, Michalis V., Fotiadis, Dimitrios I. “Machine learning applications in cancer prognosis and prediction.” Computational and Structural Biotechnology Journal, (2015): 8–17
[2] Knudson, Alfred G. “Mutation and Human Cancer.” Advances in Cancer Research Volume 17 (1973): 317–352.
[3] Jacobs, Colin van, Ginneken, Bram. “Google’s lung cancer AI: a promising tool that needs further validation.” Nature Reviews Clinical Oncology (2019)

[4] D. J. McGrail1, P. G. Pilié, N. U. Rashid, L. Voorwerk, M. Slagter, M.Kok, E Jonasch, M. Khasraw, A. B. Heimberger, B. Lim, N. T. Ueno, J.K.Litton, R. Ferrarotto, J. T. Chang, S. L. Moulder & S.-Y. Lin "High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. " Annals of Oncology (2021)
[5] Joseph A. Cruz, David S. Wishart "Applications of Machine Learning in Cancer Prediction and Prognosis" Cancer Informatics (2006)
[6] Yixuan Li1 , Zixuan Chen "Performance Evaluation of Machine Learning Methods for Cancer Prediction"Applied and Computational Mathematics (2018)

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)