iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50276

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 IV Apr 2023- Available at www.ijraset.com

Cancer Gene Detection & Diagnosis

Harshal Uikey[1], Sumit Sundergiri[2], Manoj Shahare[3], Aman Patil[4], Prashant Katre[5], Prof. Vaishali Gedam[6]
[1, 2, 3, 4, 5]*B.E. Graduate (iv year), Department of Computer Science and Engineering, NIT, Nagpur*

Abstract: ***The sequence of excrescence of cancercontrols thousands of inheritable mutations. Now the most grueling task is to separate between the mutations which will further contribute to cancer growth and mutations. Interpretation of inheritable mutations ismanually done presently, which consumes a lot of time and may also lead to a squishy opinion that isn't tolerable in the healthcare sector. Clinical molecular biologists have to manually review textbook substantiation of clinical exploration literature for every singleinheritable mutation. Machine Learning (ML)helps in the precise and fast opinion of acomplaint and leads to effective decision- making. Once the excrescence is detected wego for testing whether it's cancerous or noncancerous. However, it goes for a gene panel test which takes a much longer time which is 3- 4 weeks, So using the patient's former medical records and using a vivisection report we're detecting for which gene he/she is positive, If set up to be cancerous. Using our ML model we will helpcases in early diagnosing which will also helpCroaker to go with the following treatment. Ittakes so long time for generating the gene panel cancer report roughly 3- 4 weeks. In these 14 days period cancer excrescence can surpass stage 3 or stage 4 which is veritablyparlous and occasionally may indeed lead to death. Using colorful machine literacy models and after resolving all business constraints this process will be done as soon as possible.***

I. INTRODUCTION

Cancer Gene Detection and Diagnosis is a major project that focuses on detecting and diagnosing cancer using gene analysis techniques. Cancer is a leading cause of deathworldwide, and early detection and diagnosisare crucial for the effective treatment and management of the disease. In recent years, advances in genomics and bioinformatics have revolutionized our understanding of cancer biology and enabled the identification of cancer-causing genes and mutations. The Cancer Gene Detection and Diagnosis projectaims to leverage these advances to develop new tools and methods for detecting and diagnosing cancer at an early stage. The project will involve the use of advanced geneanalysis techniques such as next-generation sequencing, microarray analysis, and gene expression profiling toidentify the genetic alterations that areassociated with cancer. The project will also involve the development of new algorithmsand computational tools to analyze the largeamounts of genomic data generated by thesetechniques. These tools will enable researchers to identify specific genes and genetic mutations that are associated with cancer and to develop personalized treatment plans for individual patients based on their unique genetic profiles. Dataset has nine different output classes, some corresponding to cancer, and some not. Theinput gene mutations (small changes) data must be classified into one class. A standardis defined to split the dataset in the train, CV,and test ratio. The size of the training datasetshould be large as ML models learn on the train data and data points should be stable across the train, CV, and test split also called"Overlapping". The higher the overlap, the higher the stability. The different ML algorithms are applied, and performance is evaluated on metrics such as AUC, log loss, sensitivity-specificity, precision-recall, and many more. The lower the numerical log loss value the higher the efficiency. In the ideal case, the log loss value is 1. Results vary on different disease dataset features. Consequently, standard steps are followed as data collection, data reading, pre- processing of text data, feature extraction, and then training models with the most important features to get higher prediction efficiency. Applying the emerging approach of Big Data Analytics to the healthcare sector will improve healthcare services. Overall, the Cancer Gene Detection and Diagnosis project has the potential to make a significant contribution to the field of cancer research and to improve the lives of millions of people around the world who areaffected by this devastating disease.

II. PROBLEM STATEMENT

Cancer is one of the deadliest diseases worldwide, causing millions of deaths every year. One promising approach to improve cancer treatment is to identify the genes that areresponsible for cancer development andprogression. Advances in high-throughput sequencing technologies have enabled the analysis of a large amount of genetic data from cancer patients. However, identifying cancergenes from such large-scale genomic data remains a major challenge.

The aim of this project is to develop a machine learning-based approach to identify the genes that are associated with cancer development and progression. Specifically, the project will focus on the following objectives:

1) Develop a pipeline to preprocess and integrate different types of genomic data, including gene expression, copy number variation, and somatic mutation data.
2) Apply various machine learning techniques, including feature selection,clustering, and classification algorithms, to identify the genes that are most likely to be associated with cancer.
3) Validate the identified cancer genes usingindependent datasets and functional analysis methods.

Expected outcomes: The project is expected to produce a set of candidate cancer genes that can be further investigated in future experimental studies. The developed machinelearning pipeline can also be applied to other types of cancer datasets and can help to improve cancer diagnosis and treatment in thefuture.

Significance: Identifying cancer genes is crucial for understanding the underlying mechanisms of cancer development and progression. The proposed machine learning- based approach can enable the identification of novel cancer genes from large-scale genomic data, which can ultimately lead to thedevelopment of better cancer treatments.

III. LITERATURE SURVEY

Multitudinous inquiries have been carried outin this exploration area.

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

We bandied the categorization of cancer as a miscellaneous complaint conforming of many different subtypes. The early opinion and prognostic of a cancer type have come a necessity in cancer exploration, as it can grease the posterior clinical operation of patients. The significance of classifying cancercases into high or low risk groups has led numerous exploration brigades, from the biomedical and the bioinformatics field, to study the operation of machine Learning (ML) styles. thus, these ways have been employed as an end to model the progression and treatment of cancerous conditions. In addition,the capability of ML tools to descry crucial features from complex datasets reveals their significance. A variety of these ways, including Artificial Neural Networks(ANNs), Bayesian Networks(BNs), Support Vector Machines(SVMs) and Decision Trees(DTs) have been extensively applied in cancer exploration for the development of prophetic models, performing in effectiveand accurate decision timber.

2) *Applications of Machine Learning inCancer Prediction and Prognosis by Joseph A. Cruz, David S. Wishart.*

We studied a number of trends are noted in cancer vaticination, including a growing dependence on protein biomarkers and microarray data, a strong bias towards operations in prostate and bone cancer, and aheavy reliance on " aged " technologies similar artificial neural networks (ANNs) ratherof more lately developed or more fluently interpretable machine literacy styles. A numberof published studies also appear to warrant an applicable position of confirmation or testing. Among the better designed and validated studies it's clear that machine literacy styles can be used to mainly(15- 25) ameliorate the delicacy of prognosticating cancer vulnerability, rush and mortality. At a more abecedarian position, it's also evident that machine literacy is also helping to ameliorate our introductory understanding of cancer development and progression.

3) *Technologies for deriving primarytumor cells for use in personalizedcancer therapy Authored by MitraA1, Mishra L, Li S.*

This review focuses on our current understanding and the pros and cons of different styles for primary excrescence cell culture. likewise, colorful culture matrices similar as biomimetic pulpits and chemically defined media supplemented with essential nutrients, have been prepared for different tissues. These well-characterized primary excrescence cells review cancer curatives withhigh translational applicability.

4) *Breast Cancer Prediction and Detectionusing Data Mining Classification Algorithms a Comparative Study.*

This paper, aims to prognosticate and descry bone cancer beforehand with non-invasive andeffortless styles that use data mining algorithms.

IV. METHODOLOGY

Comma separated train (CSV train) containingthe description of the inheritable mutations used for training.

Fields are ID(the id of the row used to link the mutation to the clinical substantiation), Gene(the gene where this inheritable mutation is located), Variation(the amino acid change forthese mutations), Class(1- 9 the class this inheritable mutation has been classified).

First step is to fantasize the data and excerpt asimportant information as possible.
Also we have to examine the connections between genes and classes and since it is known that one gene could fall into numerous different classes, suggesting that mutations within the same gene could produce extensively different goods.
It's known that the maturity of mutations ineach case are point mutations, in which oneamino acid is shifted to another.
It's to be noted that the genes included in the training and testing datasets were nearly entirely different, and since the gene/mutation datasets contain limited information, we acquired utmost of our information from the textbook data.

V. TOOLS & ALGORITHMS USED

1) NumPy
2) Pandas
3) Jupyter Notebook
4) Google Collaboratory
5) Machine Learning Algorithms
6) Seaborn

A. Algorithms Used

1) Naïve Bayes

Naive Bayes classifiers are a family of simple"probabilistic classifiers" grounded on applyingBayes' theorem with strong (naive) independence hypotheticals between the features. They're largely scalable, which requires a number of parameters direct in the number of variables in a literacy problem.
Training can be done by assessing an unrestricted- form expression, which takes direct time. For bracket with separate features (e.g., word counts for textbook bracket), the multinomial Naive Bayes classifier is suitable. It typically requires integer point counts. The Multinomial Naïve Bayes classifier was used inthis design as we've separate data in the form of9 classes and also because the Naïve Bayes model is largely interpretable.

2) K-Nearest Neighbors

K- Nearest Neighbors algorithm (KNN) is non- parametric system used for bracket and retrogression. In both cases, the input contains k closest training exemplifications in the point space. Grounded on whether the KNN is used for bracket or retrogression, the affair changes.
In KNN bracket, the affair is class. Grounded on a plurality vote of its neighbors, an object is classified with the object being assigned to the class most common among its k nearest neighbors(k is a positive integer, generally small). still, also the object is simply assigned to the class of that single nearest neighbor, If k= 1. In KNN retrogression, the affair is the propertyvalue for the object. The performing value is the normal of the values of k nearest neighbors.

3) Linear Support Vector Machine

Support Vector Machines are grounded on theidea of chancing a hyperplane that stylish divides a dataset into two classes. Data pointswhich are nearest to the hyperplane, are Support vectors. Because of which, they're considered as the critical rudiments of a data set. The Linear Support Vector Machine was usedbecause it's largely interpretable model.

4) Logistic Regression

It's Machine Learning bracket algorithm whichis used to prognosticate the probability of a categorical dependent variable. In logistic retrogression, the dependent variable is a double variable that contains dataenciphered as 1(yes, success, etc.) or 0(no, failure, etc.). The logistic retrogression modelpredicts P (Y = 1) as a function of X.
The Logistic Retrogression model was usedbecause it's largely interpretable and can beused for multi-class bracket fluently.

For the LR model we'd use two styles**:**
a) Over-sampling which means classeswith smaller data would be balancedout with respect to other classes.
b) Without balancing which meansclasses won't be balanced out.

5) Random Forest Classifier

Random timbers are a supervised literacy algorithm, which can be used for bothbracket and retrogression. It's also the most flexible and easy to use the algorithm. A timber is comprised of trees. The further treesit has, the more robust a timber is.

It creates decision trees on aimlessly named data samples, get a vaticination from each treeand selects the stylish result by means of voting.

Random Forest Classifier is an ensemble algorithm. Ensemble algorithms are those which combines further than one algorithm ofa same or different kind for classifying objects. For illustration, running vaticination over Naive Bayes, SVM, and Decision Tree and also taking a vote for final consideration of class for a test object.

The important term Bagging directs to the ensemble, which means that a group of thingsviewed as a whole. In order to ensure the ensemble, we need to do the following**:**

a) We should produce multiple models.

b) We should combine their results.

6) Stacking Model

Piled conception or mounding is an ensemble algorithm where new model is trained to combine the prognostications from two or further models formerly trained or your dataset.The prognostications from the being models or sub models are combined using a new model, and as similar mounding is frequently appertained to as blending, as the prognostications from sub-models are blended together. It's typical to use a simple direct system to combine the prognostications for submodels similar as simple averaging or voting, to a weighted sum using direct retrogression or logistic retrogression. Models that have their prognostications combined must have a skill onthe problem, but don't need to be the stylish possible models. This means as long as the model shows some advantage over a birth vaticination, you don't need to tune the sub models hardly.

The mounding classifier produces the train, cross-validation, and tests log- loss values as0.67,1.18,1.16 independently. The misclassified points were38.64 which means that the model predicts61.36 points rightly. The confusion, perfection and recallmatrix support these results.

VI. WHY IS THE PARTICULAR TOPICCHOSEN?

The topic of cancer gene detection and diagnosis is an important and relevant area of research and healthcare. Cancer is leading cause of death worldwide and early detection and accurate diagnosis are critical for effective treatment and improving patient outcomes.

Genetic testing can play critical role in cancerdetection and diagnosis as it can identify genemutations that increase the risk of cancer or indicate the presence of cancer cells.

This information can be used to develop personalized treatment plans and targeted andtargeted therapies, which can improve patient outcomes and reduce the likelihood of recurrence.

Additionally, advance in genomic sequencingtechnologies have enabled researches to identify new career genes and biomarkers, leading to the development of more effective diagnostic tools and treatment.

Overall, the topic of cancer gene detection isimportant because it has the potential to significantly impact cancer care and improve patient outcomes.

VII. WHAT CONTRIBUTION WOULD THEPROJECT MAKE?

In this project we worked with several machine learning algorithms to detect the types of cancer genes. This project could make several contributions to the field of cancer research and healthcare. Some potential contributions include:

A. Improving Early Detection

By identifying genetic mutations associated with cancer, a project focused on cancer gene detection and diagnosis could help improve early detection of the disease. Early detection is critical for effective treatment and improving patient outcomes.

B. Personalizing Treatment

Genetic testing can help identify specificgene mutations or biomarkers that may indicate a patient’s response to certain treatments. By personalizing treatment plans based on genetic information, patients may experience improved outcomes and fewer side effects.

C. Developing Targeted Therapies

Genetic testing can also identify potential targets for new cancer therapies. By targeting specific gene mutations or biomarkers, researchers may be able todevelop more effective and less toxic treatments.

D. Identifying New Cancer Genes and Biomarkers

As genomic sequencing technologies continue to advance, a project focused on cancer gene detection and diagnosis could help identify new cancer genes and biomarkers. This information could lead to the development of new diagnostic tools and treatments.

Overall, a project focused on cancer gene detection and diagnosis has the potential to make significant contributions to cancer research and healthcare, ultimately leading toimproved patient outcomes and better overallcancer care.

REFERENCES

[1] Ng PK, Li J, Jeong KJ, et al. Systematic functional annotation of somatic mutations in cancer. Cancer Cell. 2018;33(3):450- 462.e10. doi: 10.1016/j.ccell.2018.01.021

[2] Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-421. doi: 10.1038/nature12477

[3] Bailey MH, Tokheim C, Porta- Pardo E, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;174(4):1034-1035. doi: 10.1016/j.cell.2018.07.034

[4] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

[5] Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, Wong-Erasmus M, YaoL, Kasprzyk A. International Cancer Genome Consortium DataPortal--a one-stop shop for cancergenomics data. Database (Oxford). 2011;2011:bar026. doi: 10.1093/database/bar026

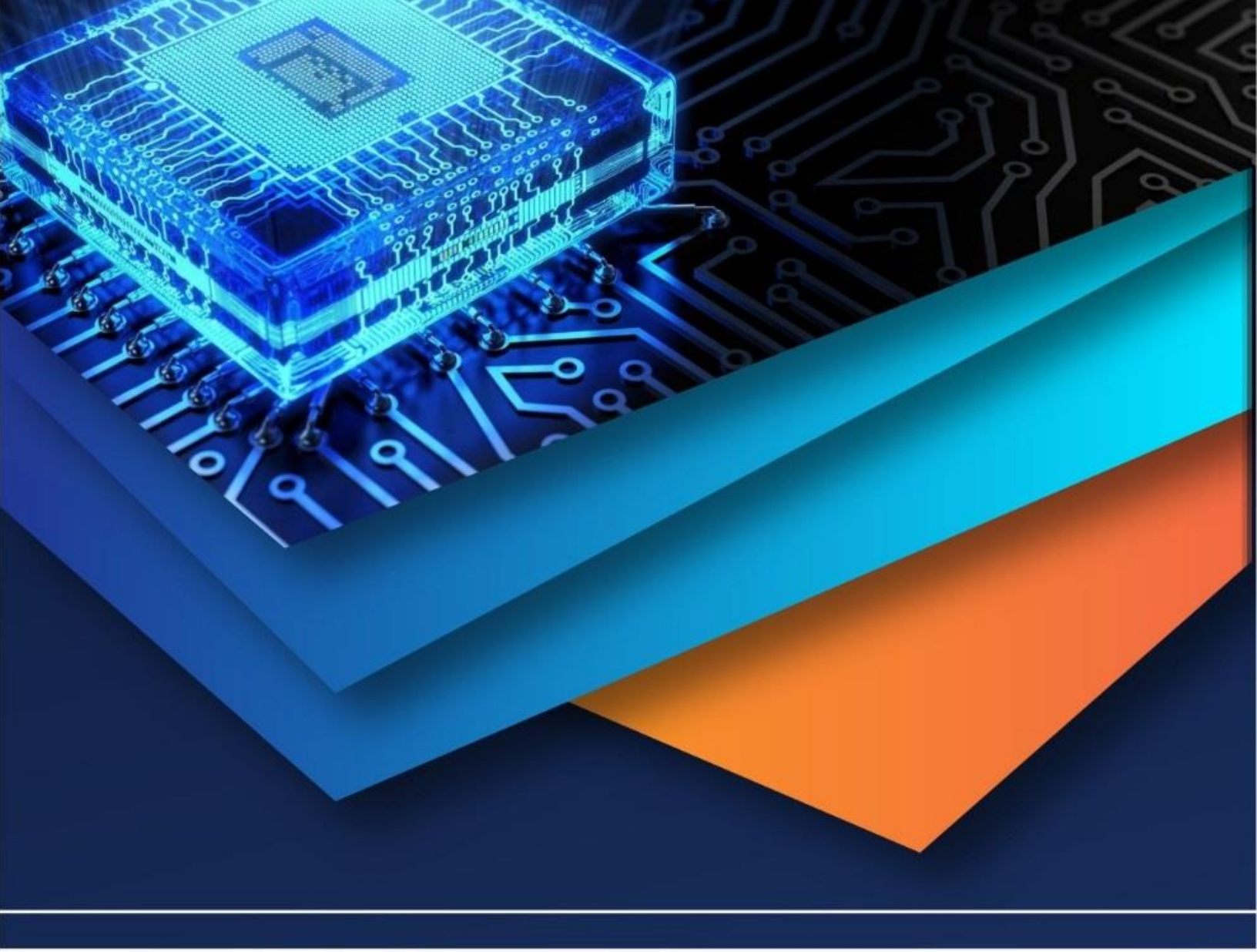
[6] Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations.PLoS Comput Biol. 2013;9(7): e1003153. doi: 10.1371/journal.pcbi.1003153

[7] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencingdata. Nucleic Acids Res. 2010;38(16): e164. doi: 10.1093/nar/gkq603

[8] Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-219. doi: 10.1038/nbt.2514

[9] Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol.2011;29(1):24-26. doi: 10.1038/nbt.1754

[10] Griffith M, Spies NC, Krysiak K, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer.Nat Genet. 2017;49(2):170-174. doi: 10.1038/ng.3774



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY