Tumor Type Detection Using Machine Learning Algorithm on Gene Expression Cancer RNA-Seq Data Set

Rajkumar Patra[1], Debajyoti Dutta[2], Mitadru Datta[3], and Anupam Ghosh[4]

[1,2,3,4] Netaji Subhash Engineering College, Kolkata, India
rajkr.patra@gmail.com

Abstract. In the last couple of years, continuous and steady growth has been noticed in the domain of research related to tumors. A lot of new methods and approaches have been implemented in recent years like screening in the early stage, to detect various types of tumors because they show notable signs and symptoms. With the rapid growth in technology, especially in the health sector, an enormous number of tumor datasets have been reported, documented, and readily available. But the setback and challenge that physicians still face nowadays is the accurate prediction of the outcome of a disease. To overcome this hurdle, Machine Learning algorithms and tools are quite frequently used by medical researchers working in this domain. In this study, five different types of tumors: BRCA(Breast Cancer gene), KIRC(Kidney Renal Clear Cell Carcinoma), COAD(Colon Adenocarcinoma), LUAD(Lung Adenocarcinoma), and PRAD(Prostate Adenocarcinoma) in Gene Expression Cancer RNA-Seq Data Set are taken into account. We have used different kinds of feature selection methods and classification on the available data set to get the most accurate and desirable output. We have considered Gaussian Naïve Bayes, Multinomial Naïve Bayes, Recursive Feature Elimination, and Cross-Validation Selection methods to trim the data set according to their features. On the selected dataset, we have implemented classification algorithms to train the data and get the best data as an output. In this project, our model can take medical data and train it and keep account of the selected data and evaluate the accuracy of each classification algorithm. With this, we can understand which algorithm is suitable for a certain data set.

Keywords: Tumor classification, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree.

1 Introduction

With the growth in technologies, improvements, and research happening in the field of Machine Learning, in the form of new algorithms, tools, and software, research scientists and medical teams have been able to achieve significantly high accuracy for predicting tumors in the early stages for proper diagnosis and treatment. Nowadays getting hold of medical data like reports and finding is quite easy as a large number of

datasets containing thousands of data entries are easily available publicly for free on government and medical facilities websites. The real challenge lies in how early tumors can be detected accurately for proper and immediate treatment. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type.

Nowadays, we have witnessed significant growth in personalized medicine and rising demand and implementation of ML concepts and techniques. We are hereby using those methods and techniques that have been discovered in a review of studies that are currently using them in the domains of tumor prediction and prognosis. In these studies, predictive and prognostic features have been taken into account to be either independent of a certain treatment or integrated in order to guide therapy for tumor patients, respectively. In addition, we discuss the types of ML methods being used, the types of data they integrate, and the overall performance of each proposed scheme while we also discuss their pros and cons.

In recent years, we have witnessed a mixture of both clinical and genomic data in the proposed works in this field. But, we have observed a deficiency in terms of validating the accuracy of the various predictive models by an external source. There is no doubt at all regarding the fact that the ML models can achieve accuracy values for the detection of cancer, recurrence, and predicting chances of survival. We have seen that after applying ML models, the accuracy for the detection of cancer in its early stages has significantly improved by 15-20% in the last couple of years.

Recently, different methods and strategies are getting implemented and reports are generated for early cancer diagnosis and prognosis. The studies that are conducted are based on approaches related to the profiling of circulating RNA gene expressions that have been witnessed to be a promising class for tumor type detection and identification. However, these new methods and strategies have some drawbacks such as suffering from low sensitivity when they are being used for screening purposes at the premature stages and lacking the proper ability to differentiate benign from malignant tumors. Various aspects regarding the use of gene expression signatures for the prediction of the tumor are constantly discussed. These studies list the potential as well as the limitations of microarrays for the prediction of cancer outcomes. Though gene signatures have the full potential for improvement in the ability for prognosis in cancer patients, there has been a lack of progress or growth in the implementation on clinical grounds. However, first, the utmost necessity lies in studying and building models by training with large data samples for more accurate prediction and validation before gene expression profiling can be tested in clinical practice.

In our current work, only research studies that have deployed ML techniques for modeling tumor diagnosis and prognosis are shown. As we are witnessing a growing trend in the application of ML methods in the domain of tumor research, we hereby present the most recent publications that have incorporated these techniques and methods with the aim to build a better model to get superior accuracy over a large data set to predict better results in the future.

2 Related Works

Machine learning is gaining popularity in cancer research. To identify potential cancer biomarkers machine learning algorithms are used to correctly predict the cancer types/subtypes. Machine learning methods have the capability to identify key features [1,2] from complex datasets which makes them ideal in cancer studies. Some previous works [3] were done on cancer detection using machine learning. One of the previous works was done on the [4] Gene Expression Cancer RNA-Seq data set which is obtained from UCI Repository. The dataset is composed of five different types of tumors (BRCA, KIRC, COAD, LUAD, and PRAD) which are expressed by the gene expressions of the patients.

The data set [4] is an 801 patient samples collection where 136 samples are patients with PRAD tumor, 141 samples are patients with LUAD tumor, 300 samples are patients with BRCA tumor, 146 samples are patients with KIRC tumor and the remaining 78 samples are patients with COAD tumor. So, all samples are divided into 5 classes as 5 tumor types.

Table 1. RNA-seq dataset details.

Tumor classes	Samples per class
PRAD	136
LUAD	141
BRCA	300
KIRC	146
COAD	78

The data set [4] contains 20531 features of each sample. For each sample, the features are RNA-seq gene expression levels which are measured with the Illumina Hiseq platform.

In this study, the Naive Bayes method [5] on the WEKA tool was used to classify the tumor types with 98.7516% accuracy in 2.69 seconds by 10-fold cross-validation, 98.5% accuracy in 2.47 seconds by 50–50% train-test data split, 98.7526% accuracy in 2.63 seconds by 40-60% train-test data split, and 98.5294% accuracy on 2.65 seconds by 66-34% train-test data partition.

For diagnosing brain tumors interest in designing tools has been increasing in recent years. To classify images into a group that has a brain tumor and another group that does not, image processing clustering algorithms are used in the work of [6] Gopal and Karnan. 42 MRI images obtained from the KG hospital database are used in this work. The authors remove the film artifacts (labels and X-ray marks) in the preprocessing phase. To remove high-frequency components in the MRI image, a Median filter is used by them. The algorithm used here is Fuzzy C Means (FCM) as an image clustering algorithm, in addition to using Genetic Algorithm (GA) as an intelligent optimization tool. From this experiment, the result said that the classification accuracy of the classification algorithm FCM is 74.6% with less than 0.4% error rate. The authors used

an optimization technique called Particle Swarm Optimization (PSO) in order to enhance the accuracy. They achieved an accuracy level of 92%.

In [7], a new system for the automatic diagnosis of brain tumor is proposed by Othman and Ariffanan. For pattern classification problems the Probabilistic Neural Network (PNN) provides a good solution. A dataset from University Teknologi Malaysia (UTK) is used here in this paper. The dataset goes through a preprocessing phase as follows. The MRI images are converted into matrices by using MATLAB. To classify the MRI images the classification algorithm PNN is used then. The result said that more than 73% accuracy is achieved by the proposed system. Depending upon a smoothing factor [7] the accuracy level can even be higher.

Finally, a classifier to detect ab-normalities in CT brain images caused by the following diseases/cases: Atrophic, Hemorrhage, Hematoma, Infarct, and Craniotomy; is designed by Najadatat al. [8].

3 Proposed work

Based on the previous work [4] conducted using Naive Bayesian we have tried to improvise first to get a better result by using MultinomialNB [5]. First, we have encoded the categorical data into classes. Then we have applied GaussianNB [5] and MultinomialNB with all the features. We have observed a significant increase in accuracy level while using all the features available. But to make the model more simpler and to reduce execution time we have tried to reduce features and consider only the crucial ones contributing the most by using chi2 and RFECV(Recursive Feature Elimination with Cross Validation) [9]. But as chi2 has some limitations we have proceeded with RFECV [9]. We have used StratifiedKFold with cv values 5, 10, and 15. Within those cv values, 10 is found to give the best result. So, we have proceeded with cv value 10. We have used StratifiedKFold so that in each fold we get almost an equal percentage of samples belonging to every class level. After the feature selection process, we got 2661 features which are the most important for our prediction. Here we have noticed that the number of features has significantly reduced. The accuracy level is expected to be >98% so that we can conclude that the selection of features is successful according to the dataset.

Naïve Bayes Classifier (2) that is also called Simple Bayes, Idiot's Bayes, and Independence Bayes. This is an efficient supervised learning algorithm that is used for both binary and multiclass classification. As well as it is the statistical method based on Bayes' Theorem (1). Bayes theorem calculates the posterior probability using the prior probability.

$$P(c/x) = P(x/c)P(c) / P(x) \quad (1)$$

The assumption of the Naive Bayes classifier is that a particular feature present in a class is independent of any other features present in that class.

$$P(y|x1,\ldots,xn) = P(y)P(x1,\ldots,xn|y) \,/\, P(x1,\ldots,xn) \quad (2)$$

K-Nearest Neighbor is a simple Machine Learning algorithm based on a supervised learning technique. It finds out the similarity between the new data point and the available class's data points. The value of K is to be determined by the trial and error method. The principle behind this K-N-N method is to find the K nearest data points from the new unknown data point. In general, the distance can be any metric measure but standard Euclidean distance is the most common choice. Then it labels the new data point as a class in which the maximum nearest points are found. It is a non-generalizing machine learning method as it simply remembers all of its training data.

Support vector machine algorithm is a supervised learning algorithm. It finds a hyperplane in an N-dimensional space (where N is the number of features) that can distinctly classify the classes using the data points. There are many possible hyperplanes that can be found which can classify two classes. The objective of this algorithm is to find the hyperplane which has the maximum margin means the maximum distance between data points of both classes. Maximizing distance gives us more confidence to classify future data points.

Decision tree algorithm is a supervised learning algorithm that is used for both classification and regression problems. It is a tree-like structure. The attributes are represented by the internal nodes, a decision rule is represented by the branch and the outcome is represented by the leaf nodes. It first splits the data according to the attribute selection measure. Then it finds the best possible root node. There are two methods for doing this, the Gini index and Information gain. The Gini index measures the impurity of a node according to the incorrectly classified results. And it tries to reduce the Gini impurity in each split of the data. The attribute which has the lowest impurity becomes the root node. On the other hand, Information gain finds out the purity of an attribute and according to that, it decides the root node. In each and every step it finds the root node for each split until the leaf node. We can reach the leaf node or the outcome starting from the root node by asking a series of questions. Each time when we receive the answer a follow-up question is asked until we reach the leaf node which can classify the data at a class level. The series of questions and the possible answers form a tree-like structure which is a hierarchical structure made of nodes and direct edges.

After using RFECV [9] to get the proposed working features, the selected dataset is splitted in a 70:30 ratio using a train test split. But as this splitting function splits the dataset randomly, the accuracy value may differ according to the splitted dataset. So, it is better to use cross-validation techniques to get the average accuracy. We have moved on and used MultinomialNB and KNeighboursClassifier to fit the data into the model and used Cross Validation with fold value 10 to get the average accuracy value. We also experimented by using the Support Vector Machines and Decision Tree method to check which one is more suitable and fits the most. We have also tried the Random Forest model. All of these results were cross-validated by keeping the same fold value as a benchmark for a standard and unbiased comparison and noted the accuracy value. We have also tested various types of error values for each of the models used. Following is the flow chart of our work.

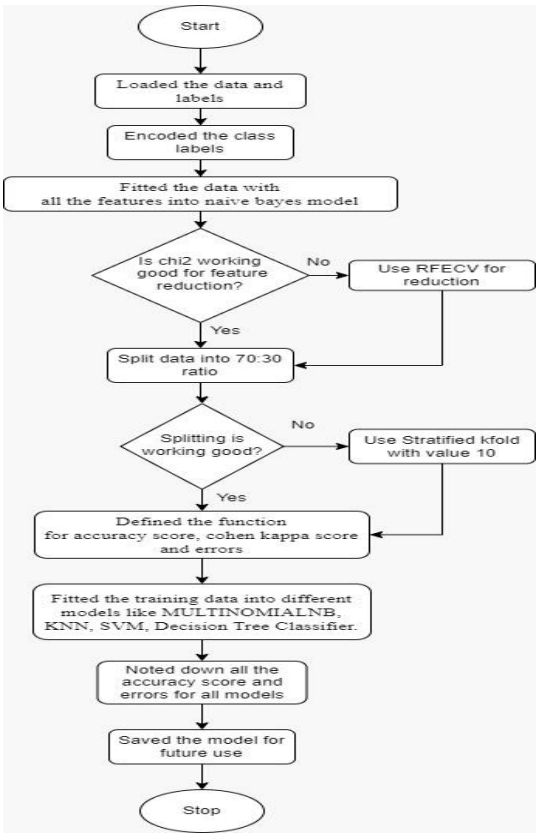

Fig. 1. Flowchart of the proposed method.

We have extended the previous research work on tumor type detection and to broader the research area we have used various types of classification algorithms so that we can compare the result among those and choose the best one. We are expecting a higher accuracy score than the previous one. Apart from that to reduce the computation time for early detection we have reduced the complexity of the algorithm by using different feature selection methods.

4 Result

4.1 Naïve Bayes Algorithm

From a report presented by [4] Gemci, Fahriye & Ibrikci, Turgay. (2017) on Tumor Type Detection using Naïve Bayes Algorithm on Gene Expression Cancer RNA-Seq Data Set, we can compare our result with this previous work. We have used the following metrics for the comparison:

Kappa Value: Cohen kappa gives us a scale to measure the accuracy of algorithms where 1.0 is the highest value.
Mean Absolute Error (MAE): This is the mean of all absolute errors. The less the value of MAE, the more the accuracy is.
Root Mean Square Error (RMSE): This gives us a more accurate error without deviation.

Table 2. Result of previous work vs. result of our work using Naïve Bayes Algorithm.

The Performance measurement of Naive Bayes	Result of Previous Work	Result of our Work
The Classification Accuracy	98.7526	99.6265
Kappa Value	0.9834	0.9945
MAE	0.005	0.0082
RMSE	0.0706	0.1288

4.2 Other Classifiers

When using K-Nearest Classifier, SVM, and Decision Tree, we expect a higher success rate of the model such as 95% - 99%. So, if the model gives us an accuracy level of 99% then we can conclude that the selection of features is successful according to the dataset.

Table 3. Accuracy percentage and different errors value using different algorithms.

The Performance measurement	Using KNN	Using SVM	Using Decision Tree	Using Random Forest
The Classification Accuracy	99.3796	99.3780	95.3796	99.6265
Kappa Value	0.9837	0.9837	0.9296	0.9945
MAE	0.0373	0.0248	0.1535	0.0082
RMSE	0.3347	0.2410	0.6967	0.1288

Apart from that, we have also tried with Random Forest classifier which has given an accuracy score of 99.62% with 10-fold cross-validation.

5 Discussions

From the experiment and result, we can understand that there are many classification algorithms to work with but the *Naïve Bayes Algorithm* gives us the best result with less amount of error (Naïve Bayes accuracy 99.6265% and error 0.008).

Other algorithms also give us satisfactory results but the *Decision Tree* is not so suitable for this kind of data set as it gives us significantly less accuracy (Decision tree accuracy 95.3796). *Support Vector Machine and KNN* algorithm give us 99.37% of accuracy which is less than Naïve Bayes. Though the accuracy score of the *Random*

Forest algorithm is very good and comparable to the *Naive Bayes Algorithm* with 10-fold cross validation still we have not considered it to be a good model because the Random Forest algorithm is based on the Decision Tree algorithm and we have seen that the accuracy score of Decision Tree algorithm is very low as compared to the other methods. So there is a high chance that the Random Forest Algorithm will predict the tumor type wrongly which can deviate the decision of the physician. As we have got our expected accuracy score from some of the algorithms that means our feature selection is successful.

Here, with the kappa value, we can also understand What is the best classification algorithm for this kind of dataset. Cohen kappa gives us a scale to measure the accuracy of algorithms where 1.0 is the highest and Naïve Bayes has got the best kappa value among the other classifiers.

6 Conclusion

It is very important to detect cancer early in a cancer control plan. So that the case can be detected in an earlier stage for more effective treatment which has a greater chance of cure.

In this work, we learned about the various techniques and algorithms used in tumor type detection (like Naïve Bayes), feature selection techniques (like Recursive Feature Elimination and Cross-Validation Selection algorithm), and classification techniques (like K nearest neighbors, Support Vector Machine, Decision Tree). We implemented our project using classification algorithms and using various feature selection methods for data trimming. Along with them, we used a confusion matrix (a table that describes the performance of a classification model.) as a tool to ensure that our classification algorithms are working and to provide features, like the cohen kappa score, to justify the accuracy of our algorithms over any provided data set.

It is certain that these accurate predictions will help us to predict future tumor-based problems and how to get desired results with better accuracy using classification algorithms of machine learning.

References

1. Renato Cordeiro de Amorim: Computational Methods of Feature Selection, Huan Liu, Hiroshi Motoda, CRC Press, pp. 440, Boca Raton, FL (2007).
2. Mahesh Pal, Giles M. Foody: Feature Selection for Classification of Hyperspectral Data by SVM. In: IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 5, pp. 2297-2307, (May 2010).
3. Tanzila Saba: Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. Journal of Infection and Public Health. Volume 13, Issue 9, 2020, pp. 1274-1289, (2020).
4. Fahriye Gemci, Turgay Ibrikci: Tumor Type Detection Using Naïve Bayes Algorithm on Gene Expression Cancer RNA-Seq Data Set. International Conference on Engineering Technologies (ICENTE'17), (Dec 07-09, 2017).

5. David D. Lewis: Naive (Bayes) at forty: The independence assumption in information retrieval. In: Nédellec, C., Rouveirol, C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science, vol 1398. Springer, Berlin, Heidelberg.
6. N. N. Gopal and M. Karnan: Diagnose brain tumor through MRI using image processing clustering algorithms such as Fuzzy C Means along with intelligent optimization techniques. In: IEEE International Conference on Computational Intelligence and Computing Research, pp. 1-4, (2010).
7. Othman, Mohd Fauzi Bin and Mohd Ariffanan Mohd Basri: Probabilistic Neural Network for Brain Tumor Classification. In: Second International Conference on Intelligent Systems, Modelling and Simulation, pp. 136-138, (2011).
8. Najadat H, Jaffal Y, Darwish O, Yasser N: A classifier to detect abnormality in CT brain images. In: International MultiConference of Engineers and computer scientists: IMECS 2011, Vol. 1, pp. 374–377, Hong Kong, China (March 2011).
9. Wang C, Xiao Z, Wu J: Functional connectivity-based classification of autism and control using SVM-RFECV on rs-fMRI data. pp. 99-105, Phys Med., (Sep 2019).