Grenze International Journal of Engineering and Technology, June Issue

Utilization of Machine Learning Methods in Cancer Outcome Prediction

Prachi Damodhar Shahare[1] and Abha Mahalwar[2]
[1]Department of Computer Science and Engineering, YCCE, Nagpur, India
Email: prachi_shahare@yahoo.com
[2]Department of Computer Science and Engineering, ISBM University, Chhattisgarh, India
Email: tamrakar.abha@gmail.com

***Abstract*— Cancer is such a condition of the body in which the cells multiply uncontrollably. Fungal-like cells can damage and destroy normal tissues, including vital organs. There are more than two hundred varieties of ovarian carcinoma, and each of them is often misdiagnosed and insufficiently treated. Cancer is associated with several distinct forms of illnesses. The process of early identification and management of tumors is crucial, since early recognition aids in the treatment of the disorder. Numerous research organizations are exploring the application of Machine Learning techniques in biomedical science and bioinformatics to categorize patients of cancer into group of high-risk or low-risk. Hence, given methodology has been adopted as a model for therapy and prevention of malignancy. It is essential that ML tools can detect significant aspects within complex datasets. This strategy includes Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and other methods to forecast cancer therapy, and is widely utilized to compare the effectiveness of various machine learning algorithms for predicting cancer.**

***Index Terms*—Tumor, machine learning, decision tree, random forest, support vector machines.**

I. INTRODUCTION

Cancer is a condition in which some cells in the body multiply uncontrollably and spread to other parts of the body. Cancer is a global health concern and, according to the World Health Organization, remains a leading cause of mortality worldwide. With trillions of cells making up the human body, cancer can start practically anywhere. It is anticipated that the cancer patients in India will increase from 26.8 million to 29.8 million by 2027. Decreased mortality, improved survival rates and lower associated costs are all associated with early detection of cancer. Early diagnosis typically results in shorter hospital stays and less intensive treatment regimens, which lowers healthcare expenses. On the other hand, late-stage cancers put a significant financial strain on patients and healthcare systems because they require complicated multimodal therapies, long-term use of costly medications, and relapse management. Thus, the importance of early diagnosis cannot be overstated. Early management of any type of cancer is of considerable economic benefit. This is why predicting the disease at an initial stage is essential to saving more lives from this life-threatening condition. Machine Learning (ML) methods are increasingly being employed for disease forecasting. Previous studies have shown that prediction accuracy may vary across the same algorithm depending on the dataset used.

Grenze ID: 01.GIJET.12.2.170_22
© Grenze Scientific Society, 2026

II. LITERATURE REVIEW

- Alfayez, A. A., Kunz, H., & Lai, A.G. (2021). A systematic literature review on cancer associated risk prediction in adult population using supervised learning. BMJ Open, 11(9), e047755. The present study shows that ML frameworks can predict cancer prevalence, one of the most common diseases worldwide. Various methods including Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machines (SVM) were employed, producing accuracies of 0.6149, 0.5258, 0.5897, and 0.5166 respectively.
- Jasim, M., Machado, L., Al-Shamery, E. S., Ajit, S., Anthony, K., Mu, M., & Agyeman, M. O. (2022). In this study, a survey of machine learning approaches applied to gene expression analysis for cancer prediction. Proceedings in IEEE. This research addressed various aspects of cancer diagnosis with computational learning, including biomarker discovery, gene profiling, and microarray/RNA-Seq datasets. The achieved accuracies were 0.82, 0.79, 0.80 and 0.76 for MDNN, RF, SVM and LR respectively. Approaches of deep learning have been shown to surpass conventional ML methods in analyzing biological patterns for cancer prediction.
- MohitAgrawal (2020). International Journal of Science and Research (IJSR), Vol. 9, August. They used several algorithms such as K-Nearest Neighbours, Support Vector Machine, Logistic Regression and Naïve Bayes to name few. Performance comparison was done on Precision, Accuracy, Recall and F1-score. KNN with maximum validation accuracy of 98 % was the most effective algorithm in this study.
- Shah, F. J., & Rao, D. S. (2022). One was on cancer prediction using machine learning. Proceedings of Accoutrements, 50, 40–47. They reviewed Machine Learning and Deep Learning approaches for modeling cancer prediction in their work. The study's findings found that Decision Tree (DT) was a robust conditional classifier, with 93.6% accuracy on retained samples. Next in line, ANNs scored second with 91.2%, followed by Logistic Regression at 89.2%.
- Rafique, R., Islam, S. M. R., &Kazi, J. U. (2021). They used Machine Learning approach for Prediction of Tumor. The study aimed to distinguish malignant from benign tumors, using various cellular image features. Algorithms employed included KNN, SVM, and NB on a data set of 150 images. A CNN was also used for image recognition and classification, achieving an accuracy of 87.64%.

III. METHODOLOGY

This study employs a classification model, where the objective is to forecast a discrete value {0,1}, such as (yes, no) or (spam, not spam). In this case, the aim is to determine whether an individual is affected by cancer (yes or no). Classification problems can be categorized into three types:

- Binary: Problems that involve only two classes possible.
- Multi-Class: Problems that fall into categories that are more than two.
- Multi-Label: Each sample may have more than one label assigned to it at once.

Techniques applied for classification include:

- L R
- N B
- SVM
- D T
- R F
- KNN

A. Logistic Regression (LR)

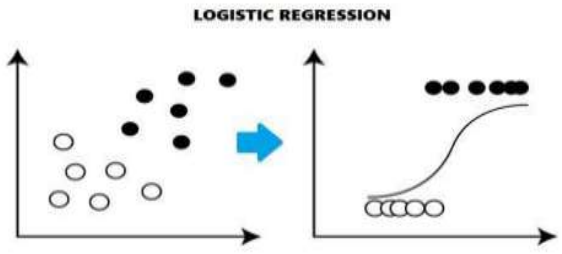

Figure 1. Logistic Regression

L R is a machine learning method you can use for problem classification. In which, the probability of outcomes which are possible of an experiment is represented using the logistic (sigmoid) function, as illustrated in Figure 1.

If you're classifying data and want to see how different things affect one key outcome, this method can be helpful. Just remember, it works best when what you're predicting has only two options, the things you're measuring don't affect each other, and your data has no gaps or missing information.

B. Naive Bayes (NB)

The Naive Bayes algorithm, which operates under the presumption of feature independence, is derived from Bayes' Theorem. Figure 2 illustratesThe effectiveness of the Naive Bayes classifier in many real-world scenarios, including text classification and spam detection. To calculate the required parameters, this method only needs a small amount of training data. Naive Bayes classifiers are also said to be poor estimators of probability, but they are extremely quick when compared to other sophisticated models.

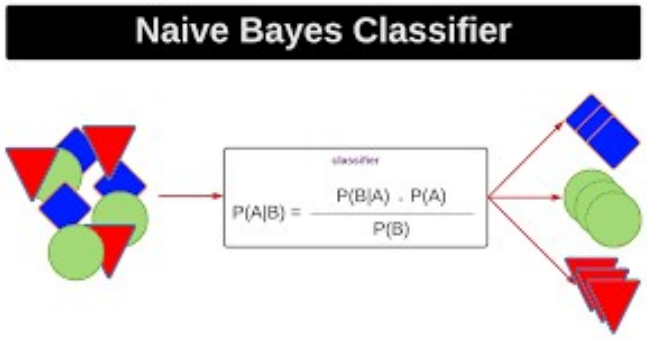

Figure 2. Naive Bayes

C. Support Vector Machine (SVM)

In a SVM, the training data is shown as points on a feature space; these points are split by the hyperplane of maximum margin. Recent events are plotted within the same region and categorized based on which side of the boundary they fall, as in Figure 3. This method saves memory by using only a portion of the training data, known as support vectors, when making decisions, making it especially useful for data with numerous dimensions. Despite their strength, SVMs do not automatically generate probability estimates; instead, these values are typically determined via additional cross-validation techniques such as five-fold testing.

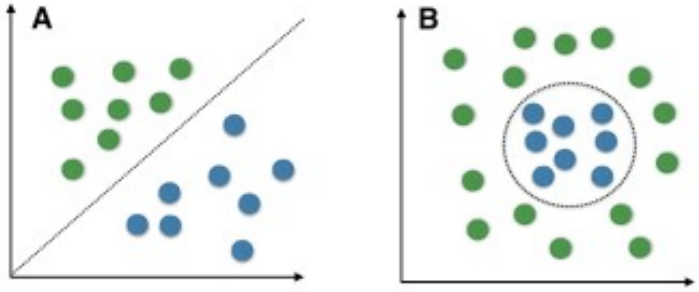

Figure 3. Support Vector Machine

D. Decision Tree (DT)

When data attributes and target classes are supplied, a DT generates a tree of rules that can further segment observations. Decision trees are capable of accommodating both numerical and symbolic values. It is visually appealing and intuitive, necessitating minimal data preparation on your part. Nevertheless, conventional decision trees are susceptible to overfitting and can develop excessively intricate trees that are less effective at generalizing to new data points. Additionally, they may be unstable due to the possibility of a completely different tree structure as a result of minor data variations.

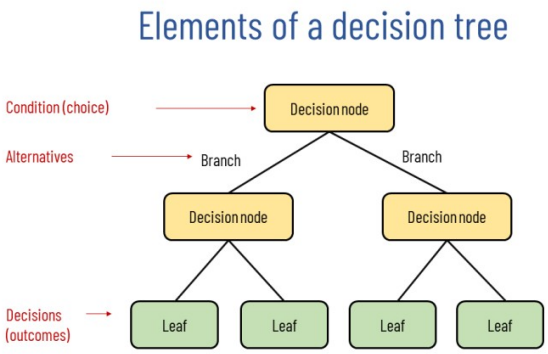

Figure 4. Decision Tree

E. Random Forest (RF)

Random Forest enhances prediction accuracy and mitigates overfitting by consolidating the capabilities of numerous Decision Trees. The model constructs distinct trees by selecting random subsamples from the dataset, each of which is approximately equivalent in size to the original data and is extracted with replacement. Afterward, it consolidates the predictions from all the trees to produce a final result. In contrast to relying on a single Decision Tree, Random Forest typically achieves accurate predictions and is perceived as dependable. However, it can be quite computationally intensive, which can make it difficult to use when you require immediate, real-time responses.

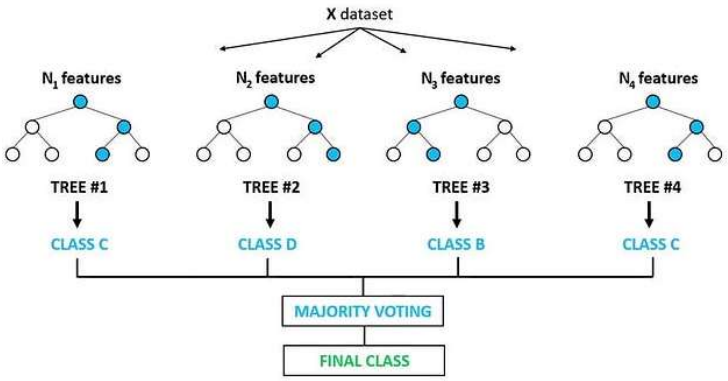

Figure 5. Random Forest Classifier

F. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors algorithm, keeps the training data for reference other than explicitly building a general internal model. This falls under the category of instance-based or lazy learning. Classification is achieved by analysing a sample's closest neighbors and among these neighbors, outcome is assigned based on majority of class, shown in Figure 6.

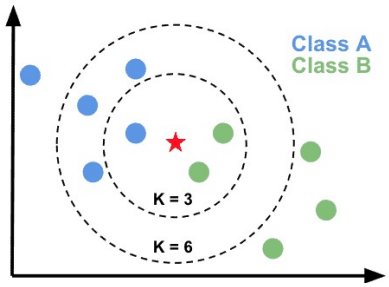

Figure 6. K Nearest Neighbour

IV. Result and Discussion

The different Machine Learning (ML) algorithms can be applied for cancer prediction,The study shows how computational methods can greatly improve medical diagnosis accuracy. Since Accuracy, Recall, Precision, Specificity, and F1-score are commonly regarded as trustworthy measures of classifier performance in biomedical research, they were employed as evaluation metrics.According to the comparison analysis, K-Nearest Neighbours and Random Forest have higher classification accuracy than the other models used. KNN had the best validation accuracy which can effectively categorised cancerous and non-cancerous samples with minimal error rates due to its ease of use and efficiency in dealing with non-linear data. Likewise, Random Forest shown resilience by reducing overfitting and providing reliable outcomes across many datasets. Naive Bayes and Logistic Regression both perform effectively in predictions; however, Naive Bayes is particularly distinguished when training data is limited. This is attributable to its management of probabilities. Nevertheless, no method is flawless, particularly when confronting complex, high-dimensional genetic data. What is a Support Vector Machine? It excels in extensive feature spaces. However, it is essential to select the appropriate kernel and adjust the settings, or it may fail to perform adequately. Decision Trees are relatively straightforward and easy to interpret; yet, even minor alterations in the information can significantly alter the structure of the tree, resulting in difficulties with generalization. In the presence of disordered or scattered data, ensemble methods such as Random Forest or instance-based techniques like KNN typically demonstrate superior performance in cancer prediction. Integrating various models or exploring deep learning can enhance accuracy and mitigate errors. Machine learning significantly enhances early detection and diagnosis in medicine. However, it relies on selecting the appropriate characteristics, necessitating high-quality data and an astute preprocessing strategy. The integration of machine learning with clinical experience yields enhanced and comprehensive healthcare outcomes.

Outcomes may be classified into three categories: low, medium, and high. This outcome is crucial in assessing illness severity. Table 1 displays the accuracy of each technique analyzed with each algorithm. Figure 7 additionally illustrates the accuracy of the observable structure.

Table I: ML Algorithms and Their Precision

S.No	Method	Accuracy
1.	Logistic Regression	95%
2.	Naïve Bayes	96%
3.	Support Vector Machine	95%
4.	Decision Tree	93%
5.	Random Forest	94%
6.	K-Nearest Neighbours	97%

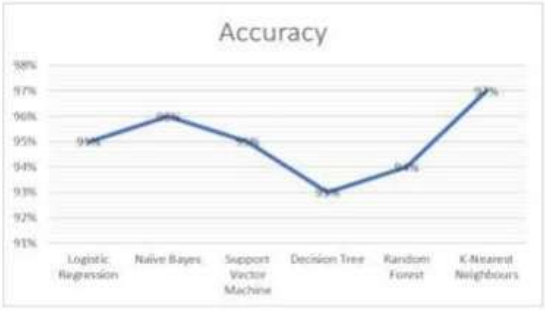

Figure 7. Algorithms with their accuracy

V. Conclusion

The effectiveness of various machine learning algorithms in predicting cancer was examined. It appears that intriguing trends emerged based on the final prediction type, the training data, and the learning methodologies employed. Among the algorithms, the K-Nearest Neighbors algorithm (KNN) was the most accurate at

predicting cancer. However, selecting the appropriate "K" value is crucial. That decision significantly impacts the outcome, necessitating caution.

Machine learning is a powerful instrument for the early detection of cancer. The study compared Support Vector Machine, Random Forest, KNN, Naive Bayes, Decision Tree, and Logistic Regression. There is a minor discrepancy in the dataset or methodology employed. KNN and Random Forest consistently outperformed Naive Bayes and Logistic Regression. In summary, machine learning facilitates early cancer diagnosis and aids in identifying cancer at a more treatable stage, hence lowering mortality and conserving financial resources. Nevertheless, these models yield optimal performance only when appropriate characteristics are selected, parameters are meticulously fine-tuned, and robust data is utilized from the outset. What are the subsequent actions? Future research should concentrate on integrating ensemble approaches, hybrid models, and deep learning to enhance predictive accuracy.

REFERENCES

[1] Alfayez, A. A., Kunz, H., & Lai, A. G. (2021). Cancer risk prediction in adults using supervised machine learning: A comprehensive review. *BMJ Open, 11*(9), e047755. https://doi.org/10.1136/bmjopen-2020-047755

[2] Jasim, M., Machado, L., Al-Shamery, E. S., Ajit, S., Anthony, K., Mu, M., &Agyeman, M. O. (2022). A survey of machine learning techniques applied to gene expression analysis for cancer prediction. *IEEE Access, 10*, 75234–75250. https://doi.org/10.1109/ACCESS.2022.3189346

[3] Agrawal, M. (2020). Cancer prediction using machine learning algorithms.*International Journal of Science and Research (IJSR), 9*(8), 1234–1238. https://www.ijsr.net/archive/v9i8/SR20727134059.pdf

[4] Shah, F. J., &Rao, D. S. (2022). Cancer prediction using machine learning.*Materials Today: Proceedings, 50*, 40–47. https://doi.org/10.1016/j.matpr.2021.04.123.

[5] Rafique, R., Islam, S. M. R., &Kazi, J. U. (2021). Machine learning in tumor prediction.*Cancers, 13*(3), 857.https://doi.org/10.3390/cancers13030857

[6] Raihan Rafique, S.M. Riazul Islam, Julhash U. Kazi “Machine Learning in the Cancer Prediction”, 2021.

[7] Mohit Agrawal, "Cancer Prediction Using Machine Learning Algorithms", International Journal of Science and Research (IJSR), Volume 9 Issue 8, August 2020.

[8] Waseem, M. H., Nadeem, M. S. A., Abbas, A., Shaheen, A., Aziz, W., Anjum, A., ...& Shim, S. O. (2019). On the feature selection methods and reject option classifiers for robust cancer prediction. IEEE Access, 7, 141072-141082.

[9] Saba, T. (2020). Recent advancement in cancer discovery using machine learning: Methodical check of decades, comparisons, and challenges. Journal of Infection and Public Health, 13(9), 1274-1289.

[10] Noor, M. M., &Narwal, V. (2017). Machine learning approaches in cancer detection and diagnosis: mini review. IJ Mutil Re App St, 1(1), 1- 8