# A trustworthy technique for utilizing gene expression data to forecast cancer

[1]Manisha Nagpal, [2]Shaurya Vir Singh Pathania, [1]Mehak Bhatiya, [3]Kalpana Singh

[1]UIE department, Chandigarh University, Mohali, Punjab, India
[2]CSE department, Chandigarh University, Mohali, Punjab, India
[3]AIT-CSE department, Chandigarh University, Mohali, Punjab, India

## ABSTRACT

Advances in molecular biology and technology have ushered in a new era of cancer research and treatment, one in which gene expression data has become an essential tool for understanding and forecasting the course of disease. An extensive investigation of reliable techniques for cancer prediction using data from genetic experiments is presented in this review study. The paper covers the fundamental need of early diagnosis and accurate prognosis to improve treatment outcomes. It explores the intricate connection between gene expression patterns and the development of cancer, providing light on the various gene expression data sets, including RNA-seq and microarray, that are utilised to identify the molecular traits of various cancer kinds. The review explores the challenges associated with data mining and analysis from genetic experiments, including issues with handling big data sets, normalisation, and data processing. Many machine learning algorithms are investigated in the context of future prediction, ranging from traditional approaches to state-of-the-art techniques like ensemble methods and neural networks. The use of multi-omics data as a means of enhancing the accuracy of cancer prediction models is also covered in the paper. Comprehensive examinations of methods, data integration techniques, and assessment metrics are provided in case studies to show how genetic exploration data analysis can be effectively used to cancer prediction. The study also emphasises how important uncertainty is to cancer prediction models, emphasising the discovery of putative biomarkers and their biological significance.

**Keywords:** microarray, RNA-seq, neural networks

## 1. INTRODUCTION

The threat of cancer, a formidable enemy to human health, is one of the most pressing issues facing contemporary medicine. Because of its complexity and ranking as the second leading cause of death worldwide, it requires innovative, accurate methods for early detection and prognosis. The methods used in cancer research have fundamentally transformed as a result of high-throughput genomic technologies, which have made molecular data abundant and include invaluable insights into the complex mechanisms of carcinogenesis. One of the most promising sources of information among them is gene expression data, which can be used to explain the intricate molecular pathways underlying the onset and progression of cancer. Accurately predicting the presence, sort, and stage of cancer is crucial for prompt therapies and better patient outcomes. There are several challenges in using genetic experimental data effectively in this quest, though. Such data, because of their greater volume and complexity, need for advanced computer techniques and meticulous analytical procedures. Furthermore, the data from genetic experiments is inherently noisy and high-dimensional, making it necessary to create trustworthy predictive models that can separate relevant signals from noise. Recent significant developments in computational biology and bioinformatics have led to the creation of multiple prediction models and algorithms for the diagnosis and prognosis of cancer. These methods not only look for possible cancer biomarkers, but they also open the door to personalized medicine by allowing treatment plans to be tailored to each patient's specific molecular profile.

The objective of this study work is to contribute to the growing body of knowledge by introducing a new and reliable approach for using genetic experimental data to predict cancer. Using state-of-the-art computational approaches, we aim to give medical professionals and researchers a useful and efficient way to use genetic experimental data for better cancer diagnosis and prognosis.

## 2. LITERATURE REVIEW

This manuscript [1] depicts machine learning approaches which are used to identify cancer using genomic experiment data is examined in the review "Machine Learning Methods for Cancer Classification Using Genomic Experiment Data." This discovery is timely and significant given the relevance of personalised therapy and the necessity for accurate cancer detection and classification. A comprehensive review of the field of cancer prognostic hypotheses based on genetic experimentation may be found in [2].

| Reference. No | Breast Cancer | Brain Tumor | Lung Cancer | Other Cancer Type |
|---|---|---|---|---|
| [1] | ✗ | ✓ | ✗ | ✗ |
| [2] | ✗ | ✗ | ✗ | ✗ |
| [3] | ✓ | ✗ | ✗ | ✗ |
| [4] | ✓ | ✗ | ✗ | ✗ |
| [5] | ✗ | ✗ | ✓ | ✗ |
| [6] | ✗ | ✗ | ✗ | ✓ |
| [7] | ✗ | ✓ | ✗ | ✗ |
| [8] | ✓ | ✗ | ✗ | ✗ |
| [9] | ✗ | ✗ | ✗ | ✓ |
| [10] | ✓ | ✗ | ✗ | ✗ |
| [11] | ✓ | ✗ | ✗ | ✗ |
| [12] | ✗ | ✗ | ✗ | ✓ |
| [13] | ✓ | ✗ | ✗ | ✗ |
| [14] | ✗ | ✓ | ✗ | ✗ |
| [15] | ✓ | ✗ | ✗ | ✗ |

Research on cancer early detection and prediction is essential for the healthcare industry. This article focuses on current trends and achievements in the discipline, examining the importance of genetic exploration data in its advancement. The authors explain the role that data from genetic experiments plays in cancer prognosis. Gene expression profiles, which are obtained through the use of technologies such as RNA sequencing and microarrays, provide significant new understandings into the molecular pathways behind the initiation and spread of cancer. uses deep learning, feature extraction, and feature selection strategies to enhance the prediction of breast cancer outcomes [3]. This takes care of the crucial and urgent matter. Breast cancer is one of the most common cancers in women worldwide, which highlights the need for precise and trustworthy prognostic models. The authors investigate how combining different approaches can enhance the accuracy and utility of clinical outcome predictions for breast cancer. The importance of precise clinical outcome forecasts in the field of breast cancer is first highlighted in the study. [4] focuses on the categorization of molecular subtypes of cancer, a

vital topic in bioinformatics and cancer research. Achieving this goal is essential to understanding the genetic variety of cancer and designing individualised treatment plans. Traditional methods for classifying cancer molecular subtypes use labor-and time-intensive gene expression profiling approaches. The authors propose automating and enhancing the accuracy of cancer subtype categorization through the use of deep learning techniques. The work in this publication has led to an expansion of the literature in the areas of deep learning applications and cancer classification. [5] focuses on forecasting patients' total existence, which is a crucial problem in the bioinformatics and medical domains. Precise prognostication is necessary for treatment planning and patient counselling. This work is a component of a broader series that advances deep learning methods to improve the accuracy of survival forecasts and individualised treatment plans for cancer patients. Comprehensive research has been conducted on survival prediction models for lung cancer because of the high mortality rate linked to this form of illness. While conventional methods have relied on clinical characteristics, more recent research has integrated molecular data for improved accuracy.

[6] tackles the prediction of illness statistics based on microarray data, a critical problem in bioinformatics and medical research. For the purpose of diagnosis and therapy planning, the precise classification of diseases using high-dimensional biological data made possible by microarray technology is crucial. This study is part of a broader trend that uses deep learning methods to bioinformatics and medical challenges, primarily focusing on the molecular data-based disease classification. A significant issue in the field of bioinformatics is precisely categorising illnesses. While machine learning algorithms and feature selection are the mainstays of traditional data analysis methods, deep learning has become a competitive substitute. employs cutting-edge learning methods to address the problem of utilising miRNA (microRNA) system-loop data to forecast the type of cancer. Accurately identifying the type of cancer is crucial for creating individualised treatment plans and focused therapies. Accurately classifying cancer kinds is a well-established topic in oncology and bioinformatics. While machine learning and feature engineering are examples of traditional methods, deep learning offers a way to immediately identify complex patterns from unprocessed data. [8] tackles a significant problem in artificial intelligence and medical diagnostics, specifically in the diagnosis of breast cancer.

Patients' results depend on a timely and correct diagnosis, and machine learning techniques—such as neural networks and evolutionary algorithms—have shown promise as helpful tools for improving diagnosis. Accurate and timely diagnosis of breast cancer is an important field of medical research. [9] tackles a critical topic in the world of medical diagnostics: the identification of prostate cancer. This subject is quite important. A timely and accurate diagnosis is crucial for the best possible outcomes for the patient. This study investigates how to expedite the detection process by combining microarray gene expression data with deep learning. This study is part of a broader trend that uses deep learning methods to medical diagnostics, including cancer diagnosis. Prostatitis is one of the most common forms of cancer in men. Early and precise detection is necessary for effective treatment. Traditional diagnostic methods are replaced by machine learning and deep learning techniques. [10] is a noteworthy development in the world of cancer research. The study investigates the use of genomic expression profiles to predict the clinical outcomes of breast cancer patients.

This study adds to the growing body of studies on the molecular and genetic causes of breast cancer in the literature. It makes a substantial contribution to the field of breast cancer research [11]. The application of genetic regression models to forecast the traits and features of breast cancer is examined in this article. The study adds to the increasing amount of research that has already shown the benefits of genetic screening for breast cancer. By examining gene expression patterns, scientists may be able to get additional insight into the underlying molecular mechanisms and heterogeneity of breast cancer. shows a ground-breaking contribution to genetics and cancer research [12]. This research investigates the application of artificial neural networks (ANNs) to gene expression profiling for the purpose of cancer type classification and prediction. It was becoming more and more evident before this study that cancer is genetically diverse, with many subtypes exhibiting distinct molecular profiles. Scholars have initiated an investigation into the potential application of gene expression patterns for the purpose of precisely classifying malignancies. [13] marks a significant turning point in the field of economics and breast cancer research.

This paper explores the development and application of a gene expression profile for predicting existence outcomes in patients with breast cancer. Before, there was a growing understanding of the genetic and molecular variability in breast cancer. The possibility of gene

expression profiling to differentiate between different subtypes with distinct clinical behaviours is being explored by researchers. The development of microarray technology, which allowed for the simultaneous analysis of thousands of genes' expression levels, is cited in this paper. [14] is a significant turning point in the study of cancer, economics, and computer learning. This work investigates the creation of tumour gene expression signatures for the multiclass classification of cancer types. Before this work, our knowledge of the genetic and molecular heterogeneity of cancer was expanding. Scientists are beginning to look into the possibility of using gene expression profiles to more precisely classify various cancer kinds. In a groundbreaking work, [15] predicts clinical outcomes in breast cancer by genomic expression profiling. It is a significant breakthrough in oncology with far-reaching implications for individualised therapy. This study demonstrates the potential of genomics to improve patient care, help doctors make better treatment decisions, and eventually improve the prognosis of breast cancer patients.

**Need for cancer prediction**

Cancer prediction is important for medical and health care research for a number of reasons. Early cancer identification is one of the most crucial variables in cancer prediction. Early cancer detection frequently results in better prognoses and more manageable malignancies. People who are at a higher risk of acquiring cancer or who are in the early stages of the disease can be identified by predictive models, which enables timely intervention and potentially life-saving treatments. Treatment plans that are more individualised and focused can arise from accurate cancer forecasting. Medical professionals can tailor a patient's treatment to their exact needs by anticipating the type, stage, genetic, and molecular characteristics of their cancer. This minimises negative effects and raises the possibility that a treatment may be successful. By identifying patients who are at lesser risk, predictive models allow healthcare resources to be used more efficiently. They make it convenient to identify high-risk patients who might require more thorough screening and monitoring, which leads to cost-effective health care management.

a) Early Detection: One of the main factors influencing cancer prediction is early cancer detection. When discovered early, many malignancies are more curable and have better prognoses. In order to enable prompt interventions and possibly life-saving treatments,

predictive models can assist in identifying people who are more likely to develop cancer or who are in the early stages of the disease.

b) Tailored Care: Precision in cancer diagnosis can result in more specialised and focused treatment regimens. Healthcare professionals can better treat each patient by customising treatment to their unique needs by predicting the precise type, stage, and molecular and genetic characteristics of cancer. This increases the likelihood of effective treatment and minimises negative effects.

c) Lower Healthcare Costs: By identifying people at lower risk, predictive algorithms can help allocate healthcare resources more effectively. On the other hand, they can spot high-risk individuals who would need closer monitoring and screening, which could result in more economical medical treatment.

d) Preventive Measures: People can lower their risk by taking preventive measures, which cancer prediction can assist with. This can involve modifications to one's lifestyle, tests, or genetic testing for inherited cancer risk, all of which can help people make well-informed decisions regarding their health.

## 3. PROPOSED METHOD


Fig.1 Architecture

The internet provided the dataset for this study, which included information on the many types of malignancies that are now prevalent as well as gene expression data associated to cancer. In this area, we conducted preliminary data research in order to better comprehend the data. After that, we cleaned the data, which included addressing outliers, handling categorical variables, and locating and correcting missing or inaccurate numbers. After that, in order to employ the independent features in our machine-learning models, we scaled them and, if needed, encoded the target variables.

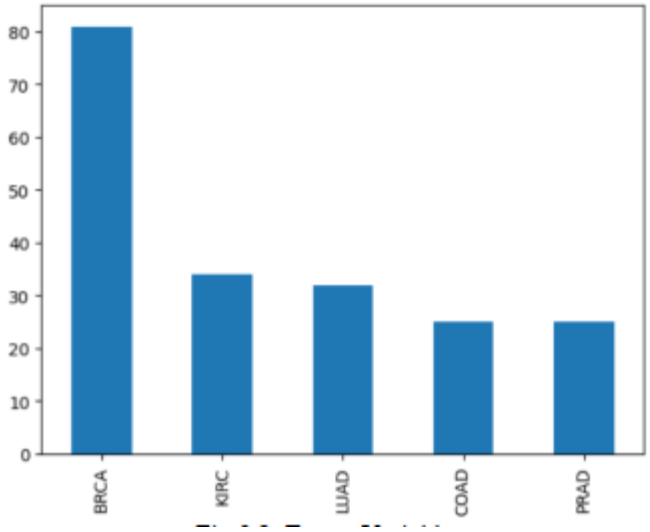


Fig. 2 Target Variable

Following the completion of the preprocessing procedures, we separated the generated data into training and testing sets. After that, we trained the data using a variety of machine learning algorithms on testing data and evaluated their efficacy. Ultimately, we selected the model with the highest performance on the evaluation metrics and used it to forecast new data. Selecting the best machine learning technique for the given dataset and problem was the study's main goal.

A. The algorithms used in machine learning are:

i) Random Forest Classifier: This machine learning technique creates many decision trees, whose predictions are then integrated to obtain a final classification. In order to arrive at the final forecast, it first selects subsets of the data and features for each tree at random. The results of each tree are then combined. Applications where the objective is to classify data into several categories based on a set of input features frequently employ the Random Forest Classifier.

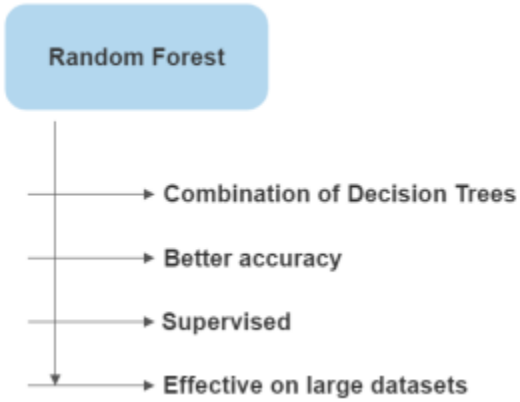


Fig. 3 Features of Random Forest

A. The algorithms used in machine learning are:

i). Random Forest Classifier: This machine learning technique creates many decision trees, whose predictions are then integrated to obtain a final classification. In order to arrive at the final forecast, it first selects subsets of the data and features for each tree at random. The results of each tree are then combined. Applications where the objective is to classify data into several categories based on a set of input features frequently employ the Random Forest Classifier.

ii) Decision Tree Classifier: This technique uses a tree-like architecture to determine how to partition data into different groups. The data is then recursively divided into subsets based on features, and a tree is constructed, with each internal node representing a feature and each leaf node representing a classification label. The decision tree classifier is widely used in applications where the goal is to classify data into many categories based on a set of input features.

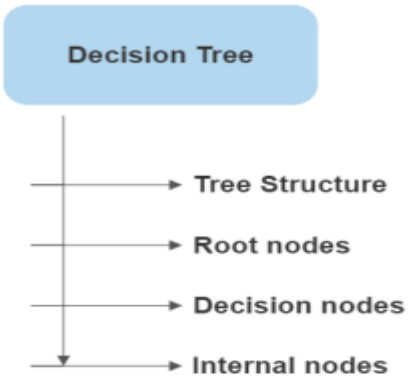


Fig. 4 Features of decision tree

## 4. ANALYSIS AND RESULTS

In the discipline of machine learning, there are numerous techniques intended to glean insights from intricate data sets.

One important parameter that is commonly used to assess the performance of these algorithms is accuracy. This study examined the accuracy of five categorization algorithms, and the findings are fascinating. Without explicit programming, machines are able to detect patterns, anticipate outcomes, and make judgements thanks to this learning process.


Fig. 5 Confusion Matrix

A confusion matrix is a crucial tool for evaluating and evaluating the effectiveness of predictive models, especially when it comes to cancer prediction. One way to assess the accuracy of a categorization model is to compare expected and actual outcomes. performance evaluation of a classification model, with a focus on multiclass classification. When it comes to cancer prediction, it aids in evaluating how well a model divides individuals or samples into groups like "cancer" or "no cancer." It begins by computing and presenting several significant metrics. The accuracy of the model is first computed. Accuracy, recall, and the F1 score are all evaluated to offer more insights into the model's presentation. Precision is used to gauge how accurate the favourable predictions are for each class. The F1-score is a metric that balances recall and precision that is calculated using the f1_score function with the "weighted" average setting.

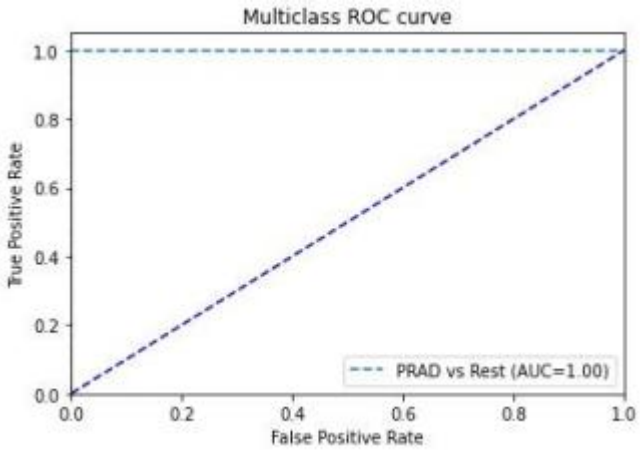


Fig. 6 Performance of a multiclass classification model

Displaying a multiclass Receiver Operating Characteristic (ROC) curve, which is a tool for evaluating the effectiveness of multiclass classification models. ROC curves are often used to quantify the trade-off between the true positive rating (sensitivity) and the false positive rating at different decision thresholds for each class in a multiclass classification task.

## 7. **CONCLUSION**

We conducted a thorough investigation of genetic experimentation-assisted cancer prediction methods in this work. The urgent need to restore the accuracy and dependability of cancer prediction algorithms drove our research since doing so will eventually lead to better patient outcomes and individualised treatment regimens. By closely analysing previous methods and combining their advantages and disadvantages, we have significantly improved the state of this important field and provided a useful overview of the current state of affairs.

A thorough analysis of the literature revealed a wide range of approaches that use gene expression data to predict cancer. These approaches span from state-of-the-art deep learning frameworks to traditional machine learning algorithms. Despite the tremendous potential of these techniques, our investigation revealed that they still encounter challenges such overfitting, model intractability, and generalisation across different forms of cancer. Our journey also showed how crucial sound processing methods are to maintaining data integrity and reducing noise in genetic experiment data sets. Feature selection led to an important discovery: the identification of critical genes required for the prediction of cancer. We also discovered that model explicability is

necessary since clinical relevance of predictions depends on our capacity to comprehend the logic behind them.

A coherent and consistent framework for comparing and assessing various gene expression experimentation-assisted cancer prediction techniques is becoming increasingly apparent. This concept should account for the variety of gene expression patterns in different malignancies and encompass a broad spectrum of cancer types.

**References**

[1] Alharbi, F. and Vakanski, A., 2023. Machine learning methods for cancer classification using gene expression data: a review. *Bioengineering*, *10*(2), p.173.

[2] Thakur, T., Batra, I., Luthra, M., Vimal, S., Dhiman, G., Malik, A. and Shabaz, M., 2021. [Retracted] Gene Expression-Assisted Cancer Prediction Techniques. *Journal of Healthcare Engineering*, *2021*(1), p.4242646.

[3] Zhang, D., Zou, L., Zhou, X. and He, F., 2018. Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. *Ieee Access*, *6*, pp.28936-28944.

[4] Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L. and Wang, X., 2019. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis*, *8*(9), p.44.

[5] Lai, Y.H., Chen, W.N., Hsu, T.C., Lin, C., Tsao, Y. and Wu, S., 2020. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific reports*, *10*(1), p.4679.

[6] Chandrasekar, V., Sureshkumar, V., Kumar, T.S. and Shanmugapriya, S., 2020. RETRACTED: Disease prediction based on micro array classification using deep learning techniques.

[7] Laplante, J.F. and Akhloufi, M.A., 2020, July. Predicting cancer types from miRNA stem-loops using deep learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 5312-5315). IEEE.

[8] Talatian Azad, S., Ahmadi, G. and Rezaeipanah, A., 2022. An intelligent ensemble classification method based on multi-layer perceptron neural network and evolutionary algorithms for breast cancer diagnosis. *Journal of Experimental & Theoretical Artificial Intelligence*, *34*(6), pp.949-969.

[9] Alshareef, A.M., Alsini, R., Alsieni, M., Alrowais, F., Marzouk, R., Abunadi, I. and Nemri, N., 2022. Optimal deep learning enabled prostate cancer detection using microarray gene expression. *Journal of Healthcare Engineering*, *2022*(1), p.7364704.

[10] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson Jr, J.A., Marks, J.R. and Nevins, J.R., 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the national academy of Sciences*, *98*(20), pp.11462-11467.

[11] Lu, X., Lu, X., Wang, Z.C., Iglehart, J.D., Zhang, X. and Richardson, A.L., 2008. Predicting features of breast cancer with gene expression patterns. *Breast cancer research and treatment*, *108*, pp.191-201.

[12] Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. and Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, *7*(6), pp.673-679.

[13] Van De Vijver, M.J., He, Y.D., Van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J. and Parrish, M., 2002. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, *347*(25), pp.1999-2009.

[14] Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P. and Poggio, T., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, *98*(26), pp.15149-15154.

[15] Van't Veer, L.J., Dai, H., Van De Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., Van Der Kooy, K., Marton, M.J., Witteveen, A.T. and Schreiber, G.J., 2002. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, *415*(6871), pp.530-536.