

A machine learning based framework for breast cancer prediction using biomarkers

Apurva Vashist, Anil Kumar Sagar, Anjali Goyal

Department of Computer Science and Engineering, Sharda University, Greater Noida, India

Article Info

Article history:

Received Nov 17, 2023

Revised Apr 3, 2024

Accepted Apr 17, 2024

Keywords:

Biomarkers

Breast cancer prediction

Feature selection

Gene

Genetic algorithm

RNA

ABSTRACT

Breast cancer is the most frequent cancer in women and the second-leading cause of cancer-related deaths globally. The main problems in managing breast cancer are high heterogeneity and the formation of therapeutic resistance. White blood cells, omics and large Wisconsin diagnostic breast cancer datasets present the three-decade genomic revolution and advance the understanding of cellular function. The precision of cancer diagnosis has also increased over the past decades. High throughput sequencing, screening, and artificial intelligence technologies have significantly improved and increased the methodologies used for diagnosis, prognosis, and therapy. This paper follows several phases of breast cancer, studies datasets and evaluate many algorithms of machine learning (ML) used for analysis and feature selection i.e. k-means, similarity correlation, genetic algorithm, and principal component analysis, have been used to recognize the subset of proteins with the highest significance for breast cancer prediction by using different biomarkers. The best correlation, as determined by Pearson correlation, between copy number and protein is 0.014, and the accuracy achieved by the genetic algorithm is 93.5% using multi-omics datasets.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Apurva Vashist

Department of Computer Science and Engineering, Sharda University

Greater Noida, Uttar Pradesh, India

Email: apurva.vashist@gmail.com

1. INTRODUCTION

Breast cancer (BC) is the most well-known malignancy that drastically reduces the wellness of women throughout the world [1]. Uncertainty surrounds the contribution of exogenous female hormones to the emergence of BC. BC risk increases following menopause. According to DeSantis *et al.* [2], BC happens to approximately 17 million cases annually, which is a worrying increase. In the United States, there were approximately 268,600 new cases of BC reported in 2019; there were also 41,760 reported fatalities [3]. Cancer diagnosis has improved over the past few decades. Technology for high throughput sequencing and screening, as well as artificial intelligence (AI), have had a significant positive impact on the approaches used for diagnosis, prognosis, and therapy. As the top reason of death for women in industrialized countries, BC has exceeded all other malignancies [4]. However, BC's recurrence, metastasis, and quick spread have not yet been totally managed and have become a major challenge for clinical management. So, it is imperative to look for more reliable prognostic biomarkers as prospective targets to comprehend probable pathways, enhance therapeutic effectiveness, and decrease distant metastasis, all of which will increase the survival rate. Many biomarkers have been used to date for screening, diagnosing, and keeping track of BC recurrence.

Biomarkers are molecules that indicate whether a physiological activity occurs normally or abnormally. They can also be signs of an underlying condition or illness. Different types of molecules,

including RNA (miRNA and mRNA), hormones, proteins, and DNA (genes), can act as biomarkers. Biomarkers are discovered to distinguish cancer via knowledge-based gene selection, gene expression profiling, or protein profiling. There are various factors that contribute to BC, the majority imply alterations in the expression of genes, like microRNAs (miRNAs). miRNAs can regulate signaling pathways, which has an impact on carcinogenesis and other aspects of cancer progression [5]. Numerous biomarkers have been employed in the detection, prediction, diagnosis, and monitoring of BC recurrence. For instance, human epidermal growth factor receptor 2 (HER2), which was discovered as a predictor of poor prognosis 25 years ago [6], is over expressed in 15% of BC patients. A crucial cell proliferation-related biomarker called antigen KI-67, which is encoded by the MKI67 gene, has been employed in clinical settings as a predictive indication of tumor recurrence and clinical prognosis [7].

Furthermore, co-transcriptional regulator miRNAs in BC are expected to interact with their target mRNAs in both favorable and unfavorable ways. miR-1307-3 p, miR-940, and miR-340-3 p were found to have negative effects on overall survival in BC patients [8]. The hsa-miR-503, hsa-miR-1307, hsa-miR-212, and hsa-miR-592 expression changes are significantly correlated with BC prognosis [1]. Due to the lack of consistency, the approximation of KI-67 and these miRNAs in BC has not yet been widely used as a biomarker in the clinic. So, to increase the precision of BC diagnosis, more precise biomarkers must be found. We have emphasized the value of omics datasets and biological information for enhancing the predictive capacity of various machine learning (ML) methods. Also, the benefits of ML in metabolic applications like protein engineering, designing gene circuits and pathways, and optimizing bioprocesses are presented. The challenges of developing ML techniques for growing designer microbes with enhanced production have also been emphasized.

2. LITERATURE REVIEW

A thorough assessment of the literature on BC prediction is presented in this section. This section surveyed 28 papers published in similar domain. Taghizadeh *et al.* [5] used 762 BC patients and 138 solid tissue normal participants to investigate relevant BC characteristics. There were three categories of ML algorithms being used: feature selection techniques, feature extraction approach: principal component analysis (PCA), 13 classification algorithms along with automated ML hyper parameter adjustment. Ayoola and Ogunfunmi [9] found a model that is most suited for forecasting type of tumor cell. Genetic algorithm is used to choose the subcategories of input features that are most pertinent to the target variable on the Wisconsin BC dataset. 5 ML regression classifiers were compared including support vector regression, logistic regression (LR), random forest (RF), and decision trees (DT).

In accordance with recent research by Nouria *et al.* [10] the performance of classifiers can be improved by removing noise and unimportant data during data preparation using a feature selection strategy, such as the genetic algorithm. The comparatively high accuracy of some ML regression approaches was also highlighted as a result. In this work, mRNA gene expression data were subjected to several feature selection/collection models established for cancer classification and medical outcome prediction. Research by Alcudia and Rodrigues [11], hybrid bioinspired models-based algorithms have been developed to choose a selection of relevant genes with cancer prediction performance relevancy. It blends instructional models with an artificial bee colony (ABC), first using a ranking approach to condense the available space and then using ABC to select the most relevant gene subset.

Zenboud *et al.* [12] uses the protein-protein interaction (PPI) for correlating proteomics and the cluster-based grey wolf optimization algorithm (CB-GWO) method for feature selection. To forecast clinical outcomes, use a deep stacked canonical relationship autoencoder. For feature selection, a CB-GWO and a deep stacked canonical correlation autoencoder (DSCC-AE) for clinical endpoint prediction are the essential components of this design. Isik and Ercan [13] employed the PPI system to find the maximum associated proteins and the coding genes of those proteins to predict the medical outcomes of patients, served as the basis for our reverse phase protein array (RPPA)-based omics biomarker identification method. 2018 put special emphasis on the ML model's classification technique for cancer prediction. They found good accuracy using support vector machines, linear regression, and k-nearest neighbors (KNN) on the Wisconsin BC dataset. Data source, data comprehension and preparation, feature training and selection, and classification methodology applied. Through BC survival prediction based on pathway activity inference, Kim and Tagkopoulos [14] explored biological methods and implications of learnt features. Omics dataset served as an inspiration for this effort.

Recent decades have seen a thorough investigation of numerous artificially assisted systems for cancer analysis, whether through omics data analysis [15] or medical image analysis. Numerous studies have also been conducted to incorporate omics data to build models that can predict medical outcomes and enhance cancer-related medical decisions Biswas and Chakrabarti [16]. The integrative framework presented in this paper generates a finding model built on a bioinspired feature choice method, along with trained ML models

that could be used as cancer prediction tools. Table 1 presents an overview of past studies in the broad domain of BC prediction using ML techniques.

Table 1. Comparison of ML based approaches

Reference	Technique used	Experimental dataset	Parameter	Method summary and merits	Remarks
[4]	ML (KNN, C4.5, SVM)	WBC and fine needle aspirate (FNA)	Early-stage prediction	BC diagnosis and prognosis	Classification accuracy better than clustering 81%
[17]	Support vector classifiers, RF, KNNs, and LR	Wisconsin Hospital dataset from Kaggle.com	Breast temperature monitoring, MRI	Historical image analysis and use of haematoxylin and eosin	96.23% in KNN, 96.28% in RF, 98.11% in SVM, and 98.18% in LR
[18]	SVM, decision tree, RF, and LR	Wisconsin BC dataset	Summary and description	Vessels of plasma and lymph vessels	Highest accuracy given by RF 98.24%
[19]	ML algorithms with IoT	WDBC	Summary and description	Age, glucose and resistin effective biomarker for BC	As compared to LR and RF, MLP produces greater accuracy with lower fault rate
[20]	RF, LR, SVM, and Gaussian process (GP) classifiers	Novel biomarkers	Summary and description	The complexity of genes and the interconnections between them need for advanced AI models	GP performed =90%
[21]	SVM, LR, RF, and Bayesian classification	Cohort comprised 302 patients	Summary and description	Human epidermal growth factor receptor	To forecast BC metastases at least three months in advance, use a RF model
[22]	ML	Raw data	Summary and description	combining radiomic variable, clinical data, and pathological data on MRI	most accurate prediction was made by RF
[23]	ML algorithm (RF, LR, ANN, and SVM)	GSE20271 and GSE22093, used 134 genes as diagnostic indicators	Summary and description (correct discrimination of patients)	Gene wise eigenvector	Cancer staging prediction
[24]	ML	Gene expression, DNA, miRNA expression, multi-omics data	Summary and description	ML pipeline survival predictions in comparison to single-modality data-based forecasts	SVM performed accuracy 92.21% AUC accuracy 87%

3. ROLE OF miRNA IN BREAST CANCER PREDICTION

One of the most researched molecules with the potential to serve as cancer biomarkers is miRNA. Deregulated miRNA expression is a hallmark of all cancer types, and it has been linked to the onset, progression, and response to treatment. Moreover, miRNAs are stable, brief, and their expression in tissues and body fluids may be easily found. These features of miRNAs have made them potentially useful biomarkers for cancer research, including diagnostic, prognostic, and predictive purposes. Although differential expression is the method used in many cancer research biomarker studies to identify biomarkers, there is an increasing tendency towards the usage and use of ML techniques for predictive modelling and data mining. Yet, some miRNAs that have been linked to BC have been found to have dysregulated expression, which may have an impact on the onset and spread of the disease. These include, among others, miR-21, miR-155, miR-10b, miR-210, and miR-146a. miRNAs have received substantial research as potential indicators for the diagnosis and prognosis of BC. Little non-coding RNA molecules known as miRNAs are crucial for post-transcriptional gene control, and deregulation of these molecules has been linked to several illnesses, including cancer.

The development of ML algorithms that can predict outcomes of metabolic engineering depends heavily on experimental datasets and omics data. ML offers statistical techniques for data exploration that can be used to train computers to predict the outcomes of genetic therapy. The three paradigms of reinforcement learning, unsupervised learning, and supervised learning are typically used to categorize ML research. In the last ten years, various interactive databases have been created for organizing biological data in such enormous quantities [25]. Finally, miRNAs have the potential to be important components of BC predictive modelling. To translate miRNA research into clinically useful applications, it is imperative to solve the difficulties related to their usage as biomarkers. For better BC diagnosis, prognosis, and treatment planning, ongoing research

attempts to improve experimental strategies, validate miRNA-based predictive models, and improve data processing techniques.

3.1. Applications of machine learning for biological functions

Various ML techniques covered in the preceding section produce tools to solve diverse difficulties in creating biological systems. ML models can be used to create system-level cells by employing quantifiable input and output variables. Because they are trained using experimental data, these data-driven models possess predictive ability without necessitating a thorough grasp of mechanisms. A well-framed model might foresee the effects of future *in silico* trials, which could upgrade experimental approaches for optimum bio system design in a way that is both economical and labor-intensive [26]. Models are often trained and verified using existing information and data gathered from experiments. Yet choosing the best model to use is a critical undertaking that mostly depends on domain knowledge and an assessment of alternative promising models. The model with higher predictive power is utilized for experimental validation after selection procedure. The newly acquired data is incorporated into the model to improve it and give it better forecasting power.

In numerous bacterial and yeast genera, ML has been used to precisely predict the RNA activities. On target movement predictors have been developed using a variety of ML methods, from straightforward linear regression to intricate, convoluted neural networks. Moreover, non-linear parameter SVM models with single guide RNA (sgRNA) scorer and clustered regularly interspaced short palindromic repeat DNA targeting (CRISPR-DT) [27] have been constructed. Genetic circuits are specialized groups of DNA sequences that enable intercellular communication to perform a specific purpose and code for a particular RNA or protein. These networks can control a variety of cell behaviors based on inputs from the surrounding environment. A conventional gene circuit design strategy involves putting forth a system, using computational methods to identify a workable role, and adjusting the parameters as necessary. In terms of ML, an ideal purpose can be created by iteratively examining the best fit from a library of potential system arrangements.

3.2. Protein engineering

In human body, proteins perform a wide range of functions, such as preserving the physical integrity of cells, acting as energy storage units, assisting in the movement of molecules across membranes, and taking part in a variety of molecular processes like DNA replication and transcription. The amino acid components that make up the entire protein are contained within these complexes. Sequencer-based protein function predictions have been the subject of much study and have found widespread use in the scientific, technological, and medical fields [28], [29].

4. DATA SOURCE

Omics biomarkers dataset is a collection of data from different biological levels, such as genomics, transcriptomics, proteomics, and metabolomics. The molecular properties of BC can be thoroughly viewed using these databases, which can also be used to find possible biomarkers or treatment targets. The patterns that distinguish malignant tissue from healthy tissue can be found by examining omics data from BC patients. With this data, prediction models can be created to help identify individuals who are at a high risk for BC and to guide treatment choices. Omics datasets can be an effective tool for predicting the development of BC overall, and further study is expected to find further applications for these datasets that will enhance patient outcomes [13].

The dataset contains 1937 features and a target variable. The dataset contains instances of 705 patients (611 patients survived, 94 died). Each feature has a prefix according to the data type: mu: somatic mutation (yes, no) [somatic mutation: a post-conceptual change in DNA. Somatic mutations can happen in any type of cell in the body, except for germ cells (egg and sperm), which are not transferred to progeny].cn: copy number variation as calculated by gistic (-2, -1, 0, 1, 2), rs: RNA (Ribonucleic acid) sequencing i.e gene expression, pp: phosphor-protein levels. The overall size of dataset is 705 rows *1941 columns.

5. MACHINE LEARNING METHODS AND ALGORITHMS IN BC PREDICTION

Algorithms that aid in prediction and classification can be created and assessed using ML techniques. Four processes form the foundation of ML: gathering data, processing it, training the model, and testing it. Four categories of algorithms were used in this investigation: protein similarity measured using Pearson correlation technique, feature extraction, feature selection, and clustering shown in Figure 1.

ML is the study of programming computers to automatically learn from and adapt to new situations based on available data. By utilizing statistical correlations from any dataset, predictive models may be developed that can be trained to predict a variety of events [30], [31]. ML is a technique for optimizing several

metabolic engineering components, including gene circuit strategy, flux prediction, strain generation, target product yield enhancement, route proposal, and optimization. This is due to the easy availability of data on microbial metabolism in various cellular and physiological states. Thoughtful consideration of the particulars of the data, the nature of the problem, and the study objectives goes into choosing ML algorithms and evaluation criteria. When evaluating and contrasting the performance of various algorithms, cross-validation, relevant metrics, and domain expertise are essential components.

Methodology

- Recover all possible proteins (proteins list) from multi omics dataset.
- Build the correlation interaction matrix of the proteins to identify protein biomarkers with the direct association to cancerous patients.
- Selected best protein by using feature selection algorithm (genetic algorithm), PCA, and clustering algorithm.
- The main objective of the projected method is to identify the subset of proteins with the highest relevance for prediction of BC.
- Collect the genes corresponding to the K selected proteins from omics dataset.

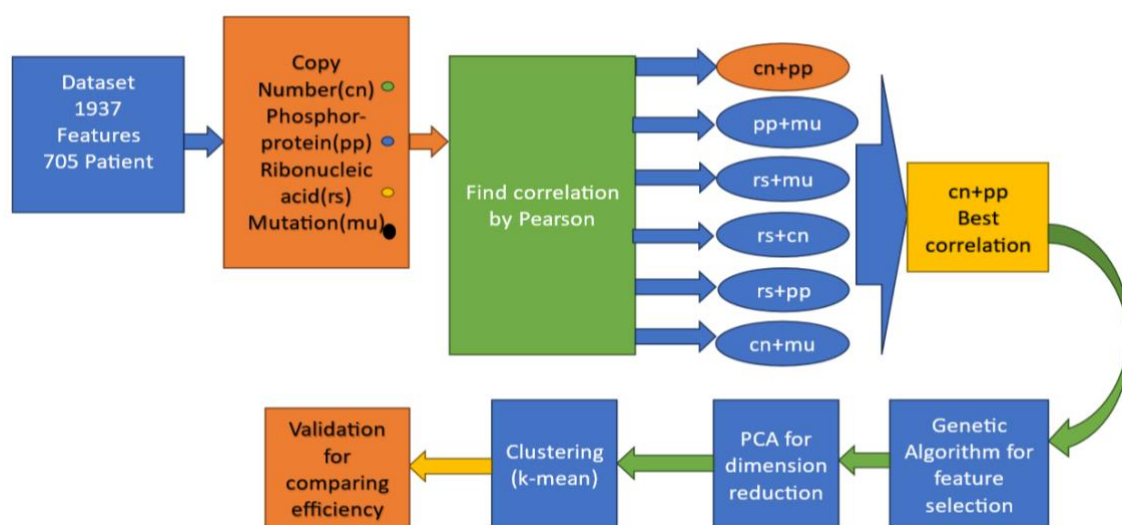


Figure 1. Workflow diagram for BC prediction

6. RESULT AND DISCUSSION

The two key parts of this suggested strategy are feature selection and clustering. The best features in the illness prediction dataset are found in the first step using clustering algorithm and using a ML classifier. One way to assess the model's performance is using the confusion matrix (CM). Some common measures used in ML studies to assess the effectiveness of models containing confusion matrices are: recall or sensitivity, specificity, F1-score, area under the curve (AUC) curve, accuracy, precision, and receiver operating characteristics (ROC) curve.

Clustering using multi omics dataset: It is important to consider statistical significance in addition to correlation magnitude. If the sample size is large enough, a modest connection could nevertheless be statistically significant. To ascertain whether a connection deviates significantly from zero, statistical tests like p-values are usually employed. Table 2 shows the highest correlation between copy number and phosphoprotein expression in BC. This can provide valuable insights into the molecular mechanisms underlying the disease. Such correlations can help to understand how genetic alterations (CNAs) may influence the activation or inactivation of specific signaling pathways represented by phosphoprotein expression. These findings can potentially lead to the identification of biomarkers and therapeutic targets for more personalized BC treatments. Table 3 compares the accuracy of the various models/methods, providing a visual representation of their performance. The accuracy performance of each method is shown in the graph or chart either side by side or superimposed. This comparison shows that various models used in the study differ in their predictive ability.

Table 2. Correlation matrix

	Phosphor-protein (PP)	Ribonucleic acid (RS)	Mutation (Mu)	Copy number (CN)
Phosphor-protein (PP)	1	-0.004	0.013	0.014
Ribonucleic acid (RS)	-0.004	1	-0.036	-0.136
Mutation (Mu)	0.013	-0.036	1	-0.076
Copy number (CN)	0.014	-0.136	-0.076	1

Table 3. Feature selection approaches

Algorithm Name	Dataset	Accuracy (%)
Cluster based grey wolf optimization algorithm	Multi omics (BRCA)	91
Genetic algorithm	Multi omics (BRCA)	93.5

A PCA analysis of integrated copy number, protein expression, and somatic mutation data is probably depicted in Figure 2(a), which provides a visual depiction of the molecular landscape of BC samples. Based on the combined molecular characteristics of copy number variations and protein expression levels, Figures 2(b)-2(d) probably offers a thorough visualization of BC samples, using the knowledge from PCA and K-Means clustering analysis. In the context of BC research, this method helps to uncover patterns, subgroups, and possible therapeutic implications.

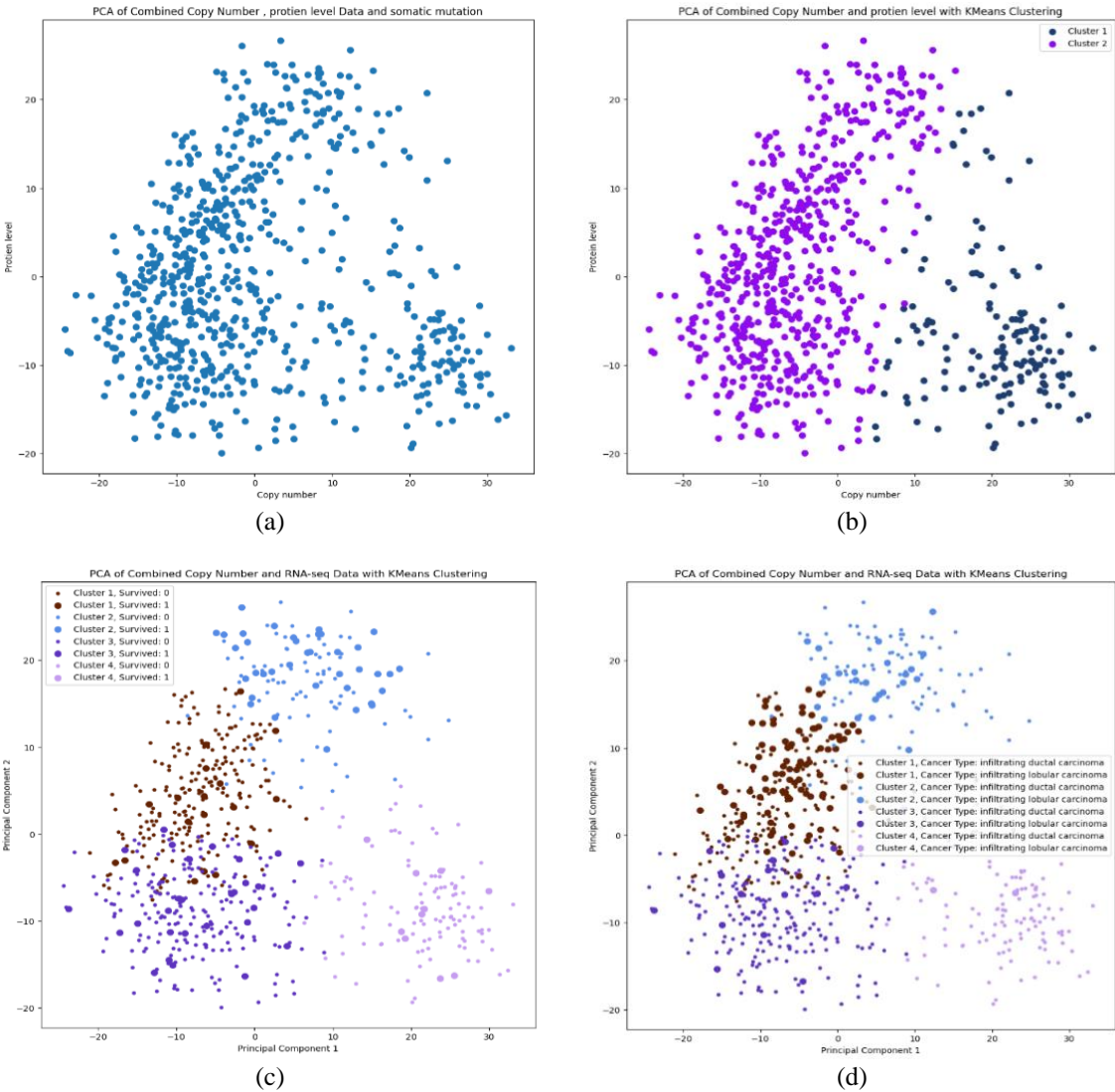


Figure 2. PCA of (a) combined copy number, protein, and somatic mutation, (b) combined copy number and protein level with k-means clustering, (c) combined copy number and rna-seq data with k-means clustering, and (d) combined copy number and rna-seq data with k-means clustering

7. CONCLUSION AND FUTURE WORK

The presented omics data must be standardized using an appropriate knowledge engineering technique to produce complete training datasets for ML. Hence, combining knowledge engineering with ML offers a setting that promotes ongoing learning for developing improved explanations for BC problem prediction. This review emphasizes the broad range of ML applications in several biological fields that have proven to be extremely helpful. Even the possibility of automated procedures involving robots being used in ML has been theorized. When compared to manual analysis, ML has shown to be effective at substantially reducing processing time and improving the accuracy of expected results. The best correlation, as determined by Pearson correlation, between copy number and protein is 0.014, and the accuracy achieved by the genetic algorithm is 93.5% using multi-omics datasets. In the future, various feature selection techniques, such as variance threshold and whale optimization algorithm can be employed to further study the effect on predictions.




REFERENCES

- [1] S. Yerukala and S. Y. Ho, "Identifying a miRNA signature for predicting the stage of breast cancer," *Scientific Reports*, vol. 8, no. 1, 2018, doi: 10.1038/s41598-018-34604-3.
- [2] C. E. DeSantis, J. Ma, A. G. Sauer, L. A. Newman, and A. Jemal, "Breast cancer statistics, 2017, racial disparity in mortality by state," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 6, pp. 439–448, 2017, doi: 10.3322/caac.21412.
- [3] J. Tang, W. Ma, Q. Zeng, J. Tan, K. Cao, and L. Luo, "Identification of miRNA-based signature as a novel potential prognostic biomarker in patients with breast cancer," *Disease Markers*, vol. 2019, 2019, doi: 10.1155/2019/3815952.
- [4] J. K. Sandhu, A. Kaur, and C. Kaushal, "Analysis of breast cancer in early stage by using machine learning algorithms: a review," *Proceedings of 2022 IEEE International Conference on Current Development in Engineering and Technology, CCET 2022*, 2022, doi: 10.1109/CCET56606.2022.10080757.
- [5] E. Taghizadeh, S. Heydarheydari, A. Saberi, S. J. Nesheli, and S. M. Rezaei, "Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods," *BMC Bioinformatics*, vol. 23, no. 1, 2022, doi: 10.1186/s12859-022-04965-8.
- [6] A. Yuryev, "Gene expression profiling for targeted cancer treatment," *Expert Opinion on Drug Discovery*, vol. 10, no. 1, pp. 91–99, 2015, doi: 10.1517/17460441.2015.971007.
- [7] J. Li, R. Liang, C. Song, Y. Xiang, and Y. Liu, "Prognostic value of Ki-67/MIB-1 expression in meningioma patients: A meta-analysis," *Critical Reviews in Eukaryotic Gene Expression*, vol. 29, no. 2, pp. 141–150, 2019, doi: 10.1615/CritRevEukaryotGeneExpr.2019025430.
- [8] A. D. M. Gutierrez *et al.*, "Identification of mirna master regulators in breast cancer," *Cells*, vol. 9, no. 7, pp. 1–20, 2020, doi: 10.3390/cells9071610.
- [9] J. Ayoola and T. Ogunfunmi, "A comparative analysis of regression algorithms with genetic algorithm in the prediction of breast cancer tumors," *2022 IEEE Global Humanitarian Technology Conference, GHTC 2022*, pp. 143–149, 2022, doi: 10.1109/GHTC55712.2022.9911033.
- [10] K. Noura, Z. Maalej, F. B. Rejab, L. Ouerfelly, and A. Ferchichi, "Analysis of breast cancer data: A comparative study on different feature selection techniques," *Proceedings of 2020 International Multi-Conference on: Organization of Knowledge and Advanced Technologies, OCTA 2020*, 2020, doi: 10.1109/OCTA49274.2020.9151824.
- [11] V. C. Alcudia and M. A. V. Rodríguez, "Artificial bee colony algorithm based on dominance (ABCD) for a hybrid gene selection method," *Knowledge-Based Systems*, vol. 205, 2020, doi: 10.1016/j.knsys.2020.106323.
- [12] I. Zenboud, A. Bouramoul, S. Meshoul, and M. Amrane, "Efficient bioinspired feature selection and machine learning based framework using omics data and biological knowledge data bases in cancer clinical endpoint prediction," *IEEE Access*, vol. 11, pp. 2674–2699, 2023, doi: 10.1109/ACCESS.2023.3234294.
- [13] Z. Isik and M. E. Ercan, "Integration of RNA-Seq and RPPA data for survival time prediction in cancer patients," *Computers in Biology and Medicine*, vol. 89, pp. 397–404, 2017, doi: 10.1016/j.compbiomed.2017.08.028.
- [14] M. Kim and I. Tagkopoulos, "Data integration and predictive modeling methods for multi-omics datasets," *Molecular Omics*, vol. 14, no. 1, pp. 8–25, 2018, doi: 10.1039/c7mo00051k.
- [15] Q. Xiao *et al.*, "High-throughput proteomics and AI for cancer biomarker discovery," *Advanced Drug Delivery Reviews*, vol. 176, 2021, doi: 10.1016/j.addr.2021.113844.
- [16] N. Biswas and S. Chakrabarti, "Artificial Intelligence (AI)-based systems biology approaches in multi-omics data analysis of cancer," *Frontiers in Oncology*, vol. 10, 2020, doi: 10.3389/fonc.2020.588221.
- [17] M. Agrawal and V. Jain, "A machine learning based approach for breast cancer prediction," *International Conference on Automation, Computing and Renewable Systems, ICACRS 2022*, pp. 623–626, 2022, doi: 10.1109/ICACRS55517.2022.10029256.
- [18] Jamal, J. H. Antor, R. Kumar, and P. Rani, "Breast cancer prediction using machine learning classifiers," *5th IEEE International Conference on Advances in Science and Technology, ICAST 2022*, pp. 456–459, 2022, doi: 10.1109/ICAST55766.2022.10039656.
- [19] V. N. Gopal, F. A. -Turjman, R. Kumar, L. Anand, and M. Rajesh, "Feature selection and classification in breast cancer prediction using IoT and machine learning," *Measurement: Journal of the International Measurement Confederation*, vol. 178, 2021, doi: 10.1016/j.measurement.2021.109442.
- [20] S. Tonmoy, H. K. Hiya, M. Z. Hasan, K. M. Z. Hasan, and N. Zahan, "Breast cancer prediction with gaussian process using anthropometric parameters," *2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021*, 2021, doi: 10.1109/ICCCNT51525.2021.9579704.
- [21] Y. J. Tseng *et al.*, "Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies," *International Journal of Medical Informatics*, vol. 128, pp. 79–86, 2019, doi: 10.1016/j.ijmedinf.2019.05.003.
- [22] W. Sheng *et al.*, "Invasive ductal breast cancer molecular subtype prediction by MRI radiomic and clinical features based on machine learning," *Frontiers in Oncology*, vol. 12, 2022, doi: 10.3389/fonc.2022.964605.
- [23] K. Athira and G. Gopakumar, "Breast cancer stage prediction: a computational approach guided by transcriptome analysis," *Molecular Genetics and Genomics*, vol. 297, no. 6, pp. 1467–1479, 2022, doi: 10.1007/s00438-022-01932-z.
- [24] J. Mitchell, K. Chatlin, L. Tong, and M. D. Wang, "A translational pipeline for overall survival prediction of breast cancer patients by decision-level integration of multi-omics data," *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM*




- 2019, pp. 1573–1580, 2019, doi: 10.1109/BIBM47256.2019.8983243.
- [25] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, 2016, doi: 10.1093/nar/gkv1070.
- [26] N. A. Abujabal and A. B. Nassif, “Meta-heuristic algorithms-based feature selection for breast cancer diagnosis: A systematic review,” *International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2022*, 2022, doi: 10.1109/ICECCME55909.2022.9988285.
- [27] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins, “Next-generation machine learning for biological networks,” *Cell*, vol. 173, no. 7, pp. 1581–1592, Jun. 2018, doi: 10.1016/j.cell.2018.05.015.
- [28] M. Hosseinzadeh, A. Gorji, A. F. Jouzdani, S. M. Rezaei, A. Rahmim, and M. R. Salmanpour, “Prediction of cognitive decline in parkinson’s disease using clinical and DAT SPECT imaging features, and hybrid machine learning systems,” *Diagnostics*, vol. 13, no. 10, 2023, doi: 10.3390/diagnostics13101691.
- [29] S. Heydarheydari, M. J. T. Birgani, and S. M. Rezaei, “Auto-segmentation of head and neck tumors in positron emission tomography images using non-local means and morphological frameworks,” *Polish Journal of Radiology*, vol. 88, no. 1, pp. e364–e369, 2023, doi: 10.5114/pjr.2023.130815.
- [30] S. M. Rezaei, N. Chegeni, F. B. Naeini, D. Makris, and S. Bakas, “Within-modality synthesis and novel radiomic evaluation of brain MRI scans,” *Cancers*, vol. 15, no. 14, 2023, doi: 10.3390/cancers15143565.
- [31] H. Khanfari *et al.*, “Exploring the efficacy of multi-flavored feature extraction with radiomics and deep features for prostate cancer grading on mpMRI,” *BMC Medical Imaging*, vol. 23, no. 1, 2023, doi: 10.1186/s12880-023-01140-0.

BIOGRAPHIES OF AUTHOR






Apurva Vashist    is a postgraduate student in Department of Computer Science and Engineering, Sharda University. She has research interests in the application of ML in medical domain. She can be contacted at email: apurva.vashist@gmail.com.



Anil Kumar Sagar    is currently working as Professor in Department of Computer Science and Engineering in School of Engineering and Technology, Sharda University, India. He obtained his doctorate from JNU, New Delhi in Ad-hoc Networks. He obtained his B.E. in Computer Science & Engineering from G. B. Pant Engineering College Pauri Garhwal, and M.Tech. from JSSATE Noida. He can be contacted at email: aksagar22@gmail.com.



Anjali Goyal    is an Assistant Professor in the Department of Computer Science and Engineering at School of Engineering and Technology, Sharda University, Greater Noida. She possesses 7 years of teaching and research experience. Her research interests are in the areas of mining software repositories, text analytics, and machine learning. She has published several research papers in various international conferences and journals of repute. She can be contacted at email: anjali1@sharda.ac.in.