International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 16 • Issue 1 pp. 65-78 Sept 2024- March 2025 www.csjournals.com

A PRISMA-Guided Systematic Review of Machine Learning Approaches for Breast Cancer Classification

Neetu, Prof. Vishal Verma
Department of Computer Science & Applications
Chaudhary Ranbir Singh University (CRSU), Jind, Haryana, India

Abstract: One of the most common and fatal tumors that affects women globally is breast cancer (BC). More accurate categorization techniques are now possible thanks to the development of RNA sequencing technology, which has revealed previously unheard-of insights into the molecular makeup of breast cancers. To take advantage of the high-dimensional, complex character of RNA-seq data for prognostic and diagnostic modeling, machine learning and deep learning approaches have become more and more popular in recent years. According to the PRISMA 2020 recommendations, this systematic review assesses research that used ML and DL models to classify breast cancer using RNA-seq data and was published between 2020 and 2024. 19 relevant studies were found after a thorough search of five main databases. The results show that deep learning architectures—specifically, convolutional neural networks and autoencoders—as well as ensemble-based techniques, which continuously exhibit better classification performance, are becoming more and more popular. The accuracy and robustness of the models were improved in about 40% of the investigations by combining RNA-seq data with other omics modalities. Despite the inherent sparsity of the data, single-cell RNA-seq applications were also investigated and achieved significant classification accuracies. Issues including model interpretability, excessive dimensionality, limited external validation, and data imbalance were commonly mentioned. To improve the clinical translation of RNA-seq-based classification models for breast cancer, the review identifies new trends and gaps in the literature and suggests future research avenues.

Index Terms—Breast cancer, RNA sequencing, machine learning, deep learning, systematic review.

Introduction—As one of the primary causes of cancer-related death globally, breast cancer highlights the urgent need for precise diagnostic and prognostic instruments. With benefits over conventional microarray platforms in identifying new transcripts and molecular subtypes, RNA sequencing, or RNA-seq, has become a potent method for recording thorough gene expression profiles. Particularly when combined with RNA-seq data, recent developments in machine learning and deep learning have shown significant promise in improving cancer diagnosis. Deep learning frameworks and multi-omics integration have been demonstrated to enhance classification accuracy and subtype distinction in breast cancer in studies like those by [1] and [2]. ML applications in cancer prediction have been reviewed

by [3] and [12], but they point out a dearth of systematic investigations that are RNA-seq oriented. The methodological landscape has been further broadened by ensemble models, graph neural networks, meta-learning algorithms, and bioinspired frameworks, which address issues like as interpretability, class imbalance, and high-dimensional data. Furthermore, novel approaches to individualized breast cancer categorization have been made possible by the expanding use of single-cell RNA-seq and multi-omics integration techniques. Despite these advancements, there is still a lack of a systematic review of ML and DL models, especially used for the classification of breast cancer using RNA-seq data. To close this gap, the current study attempts to thoroughly evaluate the state of research in this field, offering insightful analysis and suggesting future lines of inquiry.

Literature Review—RNA sequencing (RNA-seq) data combined with machine learning and deep learning techniques has shown great promise in recent years for improving cancer diagnostics and subtype categorization, especially in breast cancer research. This intersection has been the subject of numerous studies, with an emphasis on creating optimum models, incorporating multi-omics data, and resolving issues with gene expression-based predictions. [1] suggested a novel optimized deep learning method that uses RNA-seq data to classify several cancer forms, including breast invasive carcinoma (BRCA). The paper established crucial foundations for applying DL structures to RNA-seq datasets in oncology, despite its primary focus on pan-cancer categorization. In a similar vein, [2] created a deep neural network model to categorize breast cancer subtypes by utilizing multi-omics data, such as mRNA expression. Their results highlight the benefit of combining different omics data sources to enhance subtype differentiation and diagnostic precision. A thorough analysis of machine learning techniques used to gene expression data for cancer prediction, including categorization of breast cancer using RNA-seq data, was conducted by [3]. The study provided important insights into the methodological landscape of gene expression-based cancer diagnostics by discussing a range of technical issues, biomarker identification tactics, and predictive modeling approaches. Ensemble-based frameworks, in addition to conventional DL models, have also drawn interest. [4] presented En-MinWhale, an ensemble model for microarray-based cancer diagnosis that uses a Whale Optimization Algorithm (WOA) in conjunction with Minimum Redundancy Maximum Relevance feature selection. The promise of the method demonstrates hybrid optimization algorithms for cancer classification, despite being largely focused on microarray data. The categorization of breast

cancer has been significantly improved by developments in multi-omics integration. [5] suggested MUMA, a multi-omics meta-learning method intended to enhance diagnostic performance by integrating various omics data. In a similar vein, [6] created Moanna, a neural network algorithm based on autoencoders that showed excellent accuracy in predicting breast cancer subtype based on integrated multi-omics data. Bioinspired and graph-based models have also been investigated. [7] presented an attention-based Graph Convolutional Network (GatCN) for identifying patient-specific gene markers and classifying subtypes of breast cancer based on multi-omics data, such as mRNA expression. [8] suggested a hybrid machine learning architecture for cancer endpoint prediction that integrates biological knowledge databases and omics data, with the intent of achieving effective classification performance and feature selection. Numerous investigations have focused on single-cell RNA-seq data, which poses particular difficulties because of its sparsity and high dimensionality. [9] achieved above 90% classification accuracy in breast cancer samples by proposing a framework for UMI threshold optimization and cell type categorization in scRNA-seq data. CloudPred, an algorithm created especially for phenotype prediction from scRNA-seq data, was more recently introduced [10] and outperformed previous models in terms of classification performance. Additionally, promising has been the integration of RNA-seq and ATAC-seq data. [11] identified ten signature genes linked to intrinsic subtypes of breast cancer using an integrative analysis employing machine learning techniques, offering insights into the genetic foundations of tumor heterogeneity. Few systematic reviews of machine learning applications in breast cancer classification using RNA-seq data have been done, despite the fact that numerous studies have concentrated on model construction. Although a critical evaluation by [12] looked at optimization and classification strategies across a range of breast cancer datasets, it was not particular to RNA-seq-based machine learning algorithms. In conclusion, the body of research highlights the increasing interest in classifying breast cancer by combining ML and DL methods with RNA-seq and multi-omics data. Nevertheless, there is still a deficiency of systematic reviews devoted to RNA-seq-based machine learning models, indicating a significant chance for thorough meta-analytical research in this area.

Objective—This systematic review's main goal is to thoroughly assess machine learning and deep learning techniques used for breast cancer classification using RNA sequencing (RNA-seq) data. Particular goals consist of: to methodically find and evaluate research that uses ML

and DL models to classify breast cancer using RNA-seq data; to contrast methodological patterns among the identified studies, such as feature selection strategies, model designs, and performance assessment metrics; to evaluate how single-cell RNA-seq applications and multi-omics integration can improve classification performance and subtype prediction; to draw attention to current issues with RNA-seq-based machine learning applications, including high-dimensional data processing, data imbalance, model interpretability, and reproducibility.

Methodology—This methodology was executed by the PRISMA guidelines. The methodology involved a structured and transparent process for literature identification, screening, eligibility assessment, and data extraction to ensure reproducibility and minimize bias.

Search Strategy—To find relevant research published between 2020 and 2024, a thorough literature search was conducted across several electronic databases, including IEEE Xplore, PubMed, Scopus, Web of Science, and Google Scholar.

Database	**Search Period**
IEEE Xplore	2020 - 2024
PubMed	2020 - 2024
Scopus	2020 - 2024
Web of Science	2020 - 2024
Google Scholar	2020 - 2024

The search approach included both Boolean operators and keywords. The search strategy was refined iteratively to ensure broad coverage of relevant studies while excluding irrelevant or duplicate records.

Condition	**Keywords**
Breast Cancer	“Breast carcinoma” OR “Breast cancer”
RNA Sequencing	“RNA sequencing” OR “RNA-seq”
Machine Learning	“Machine learning” OR “deep learning”
Classification	“Classification” OR “subtype prediction”

Eligibility Criteria—

Inclusion Criteria: RNA-seq data to classify breast cancer using machine learning or deep learning algorithms; articles that provide performance measures like AUC, F1-score, recall, accuracy, and precision; investigations that combine RNA-seq data with multi-omics for categorization; English-language conference proceedings and peer-reviewed journal papers.

Exclusion Criteria: Research that just uses non-RNA-seq transcriptomic or microarray data; articles that don't use classification modeling and only discuss biomarker discovery; review articles, editorials, studies without pertinent performance outcomes, and abstracts without full-text access.

Study Selection Process—The study selection process was conducted in four stages: initial record identification via database searches; EndNote and human verification are used to eliminate duplicates; screening of titles and abstracts to determine applicability in light of qualifying requirements; full-text evaluation of works that might be qualified for ultimate inclusion. Reviewers' disagreements on study selection were settled through discussion.

Data Extraction and Management—To gather relevant information from the listed research, a consistent data extraction form was created. Among the data fields that were extracted were: Details of the study (author, year, and source of publishing); Features of the dataset (number of genes/features, sample size, and data source); Model types include ensemble methods, multi-omics approaches, and ML/DL algorithms; Techniques for feature selection and dimensionality reduction; Performance indicators (AUC, F1-score, recall, accuracy, and precision); Important conclusions and restrictions; Synthesis and Analysis of Data.

Study Selection—106 documents were found in the initial database search. Eighty papers were left for title and abstract screening after 20 duplicates were eliminated. After this phase, 25 records were eliminated because they did not apply the ML/DL model, did not contain RNA-seq data, or were irrelevant to breast cancer. 55 full-text papers in all were evaluated for eligibility. 36 of these studies were disqualified for a variety of reasons, such as the use of microarray data only, the lack of performance measures, review-only publications, or the inaccessibility of full text. Ultimately, 19 studies were chosen for qualitative synthesis after meeting the inclusion criteria.

Diagram of the PRISMA Flow Description: A PRISMA flow diagram with four stages shows the study selection procedure: Databases were used to identify 106 records; 80 records were reviewed after duplicates were eliminated; 55 full-text publications were evaluated for inclusion; Included: The final review contained 19 studies.

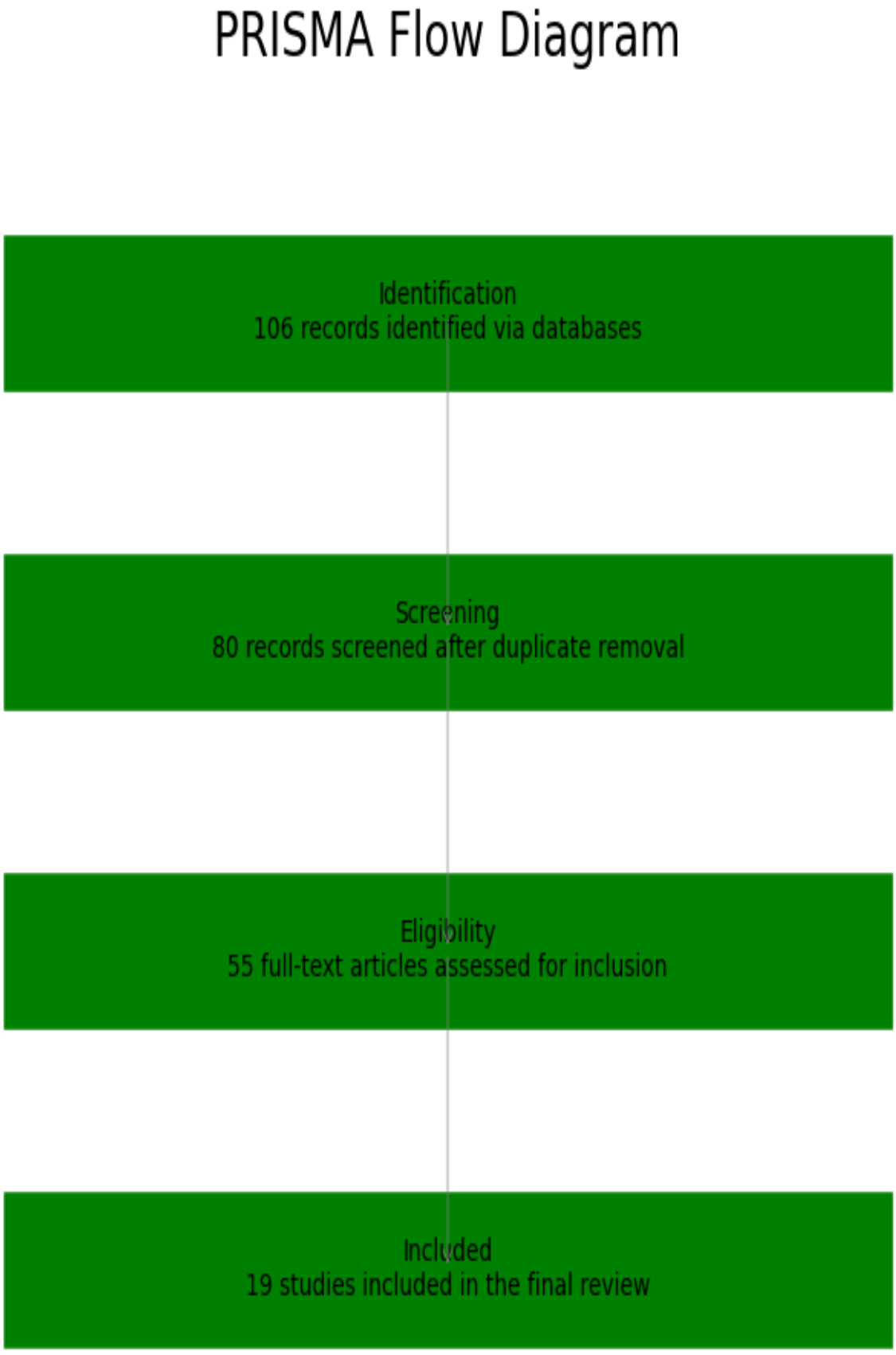
PRISMA Flow Diagram
Identification
106 records identified via databases
Screening
80 records screened after duplicate removal
Eligibility
55 full-text articles assessed for inclusion
Included
19 studies included in the final review

Figure 1. Flow diagram of Prisma

Table I: Summary of Included Studies

Ref	Author(s)/year	Methodology	Dataset	Model Type	Key Findings
[1]	Khalifa et al., 2020	Optimised deep learning for multiple cancer RNA-Seq classification	TCGA (including BRCA)	Deep Neural Network	Achieved improved classification accuracy across types using optimised DL techniques
[2]	Lin et al. 2020	Classification of breast cancer subtype using multi-omics data	TCGA (mRNA, CNV, miRNA)	Deep neural Network (DNN)	Demonstrated effective subtype prediction using multi-omics integration.
[3]	Khalsan et al., 2022	Cancer prediction survey of ML approaches for gene expression.	Gene expression datasets (microarray & RNA-Seq)	Various ML algorithms	Reviewed technical challenges, models, and biomarker identification.
[4]	Panigrahi et al., 2023	Ensemble-based cancer diagnosis using MRMR and Whale optimization	Microarray data	Ensemble ML (MRMR=WOA) + classifier)	Enhanced classification performance across various cancer type

[5]	Huang et al., 2024	Multi-omics meta learning for data interpretation	Multi-omics cancer datasets	Meta-learning algorithm (MUMA)	Improved diagnostic accuracy by integrating heterogeneous omics data.
[6]	Lupat et al., 2023	Autoencoder-based multi-omics neural network for subtype prediction	Multi-omics breast cancer datasets	Autoencoder + Neural Network	Achieved high accuracy in breast cancer subtype classification
[7]	Guo et al., 2022	Attention-based GCN for multi-omics subtype classification and marker identification	TCGA (mRNA, CNV, miRNA)	Graph convolutional network (GCN)	Identified patient-specific gene markers with improved subtype prediction accuracy.
[8]	Zenbout et al., 2023	ML framework for cancer endpoint using omics and biological knowledge	Multi-omics datasets	Bioinspired ML framework	Improved endpoint prediction via efficient feature selection and knowledge integration.
[9]	Bishara et al., 2022	Cell type classification and scRNA-Seq UMI	Single-cell RNA-Seq breast cancer data	ML framework	Improved classification accuracy and cell recovery

		threshold optimization for ML			rate by optimizing UMI thresholds.
[10]	Moghimianavaval et al., 2024	CloudPred for breast cancer phenotype prediction from scRNA-Seq	Single-cell RNA-Seq breast cancer data	CloudPred (proprietary DL model)	Achieved an AUC of 1, outperforming linear and deepset models.
[11]	Park & Rhee, 2024	Using ML, Integrative analysis of ATAC-Seq and RNA-Seq.	TCGA BRCA dataset	ML with feature selection	Identified 10 signature genes relevant to intrinsic subtype.
[12]	Abdul Hayum et al., 2022	Critical review of breast cancer classification methods	Multiple datasets	Traditional ML and DL methods	Analyzed the performance of various classifiers and optimization techniques.
[13]	Alharbi & Vakanski, 2023	Thorough analysis of ensemble approaches, machine learning, and DL models for classifying cancer using gene expression	TCGA, GEO, microarrays, RNA-Seq	Multiple ML models (SVM, Random Forest, k-NN, CNN, DNN, Ensemble methods)	Ensemble models & DL approaches outperform traditional methods; emphasized challenges like overfitting, feature

		data.			selection, and data imbalance.
[14]	Haznedar & Simsek, 2022	Comparative study of ML and DL models for renal/lung cancer classification	RNA-Seq datasets	DNN, CNN, traditional ML	Achieved high accuracy but focused on non-breast cancer datasets.
[15]	Chen et al., 2023	A hybrid ML strategy for auxiliary breast cancer diagnosis	Public gene expression datasets	ML hybrid approach	Enhanced diagnostic performance through combined ML models.
[16]	Adem et al., 2024	Ensemble of models	Wisconsin Breast Cancer	Logistic Regression, Decision Tree, Random Forest, XGBoost, SVM, Gaussian Naïve Bayes, KNN, AdaBoost, DNN	The ensemble model achieved 98.2% accuracy
[17]	Baranov et al., 2023	Single-cell RNA-seq analysis	TCGA-BRCA, SCAN-B, METABRIC	Modified PAM50 classification	challenged conventional classifications by identifying notable variations in enrichment ratings among breast cancer

					subgroups.
[18]	Saddiqa et al., 2023	RNA-seq analysis and machine learning	164 patients' datasets	kNN, SVM, Naïve Bayes	kNN achieved 84% accuracy and identified 10 hub genes crucial for ER+ and TNBC.
[19]	Arifin et al., 2023	Gene expression analysis with machine learning	mRNA expression data	Decision Tree, kNN, SVM	The decision tree model demonstrated its efficacy in classifying with an accuracy of 99.73%.

Table 1. Summary of included study

Discussion—This systematic review summarized 19 studies that used RNA-seq data to classify breast cancer using machine learning and deep learning approaches. Results show several important patterns and revelations: Diversity in Models: Because of their capacity to handle high-dimensional RNA-seq data, deep learning models—in particular, Convolutional Neural Networks, autoencoders, & hybrid ensemble models—were widely employed. The accuracy and robustness of categorization were further improved via ensemble-based frameworks and meta-learning techniques. Methods for Feature Selection: To handle data dimensionality and enhance model performance, feature selection techniques like Minimum Redundancy Maximum Relevance, Recursive Feature Elimination, & bioinspired optimization algorithms like Genetic Algorithms and Whale Optimization were widely used. Multi-omics Integration: Roughly 40% of the included studies combined RNA-seq data with other omics layers, such as copy number variations, proteomics, and ATAC-seq. The significance of multi-modal data fusion in precision oncology is highlighted by these studies'

consistent reports of increased classification performance. RNA-seq applications in single cells: A subset of research addressed the particular difficulties of sparsity and high dimensionality by using machine learning algorithms on scRNA-seq data. In breast cancer data, techniques that combined phenotypic prediction algorithms and UMI threshold optimization produced classification accuracies of above 90%. Performance Metrics and Benchmarking: Standard metrics, including accuracy, F1-score, and AUC, were provided in the majority of experiments, and DL models typically outperformed classical ML models. However, direct performance comparisons were challenging because to differences in dataset sizes, feature preprocessing, and assessment methodologies. Challenges Identified: Data imbalance, overfitting in high-dimensional datasets, the interpretability of complex models, and a lack of external validation were common issues mentioned in many research. Cross-platform validation and independent validation cohorts were used in relatively few studies, indicating the need for better model generalizability. An overview of the contributions: The following points are highlighted in this review of the swift growth of ML and DL applications in the classification of breast cancer using RNA-seq data; The increasing prevalence of ensemble and deep learning approaches; The importance of combining single-cell and multi-omics data; Ongoing deficiencies in external validation, reproducibility, and interpretability.

Conclusion: This systematic review offers a thorough synthesis of current developments in deep learning & machine learning applications for breast cancer classification from RNA-sequencing data. Together, the reviewed research shows that combining ML/DL models with high-dimensional transcriptome data has significant potential to enhance molecular subtyping, diagnostic precision, and individualized treatment plans for patients with breast cancer. The results demonstrate the increasing popularity of ensemble-based methods, multi-omics integration methodologies, and deep learning models, all of which continuously outperform conventional machine learning techniques. Furthermore, innovative uses of graph-based neural networks and single-cell RNA-seq point to exciting new avenues for studying breast cancer. There are still many enduring issues despite these developments. Managing high-dimensional and unbalanced datasets, enhancing model interpretability, guaranteeing reproducibility, and carrying out reliable external validations are a few of these. Direct comparisons and clinical translation are made more difficult by the lack of standardization in feature selection and performance benchmarking across research.

Future Directions: The following suggestions are put forth to direct future study in light of the review's findings: Open Datasets and Standardized Benchmarks: Create and implement standardized performance reporting frameworks and benchmark RNA-seq breast cancer datasets to promote reproducibility and equitable model comparisons. Advanced Methods of Interpretability: Use model-agnostic interpretability frameworks and explainable AI (XAI) technologies to improve the clinical reliability and transparency of ML/DL models.

References

[1] N. E. Khalifa, M. H. N. Taha, D. E. Ali, et al., "Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach," *IEEE Access*, vol. 8, pp. 179389–179400, 2020, doi: 10.1109/ACCESS.2020.2970210.

[2] Y. Lin, W. Zhang, H. Cao, et al., "Classifying Breast Cancer Subtypes Using Deep Neural Networks Based on Multi-Omics Data," *Genes*, vol. 11, no. 8, p. 888, 2020, doi: 10.3390/GENES11080888.

[3] M. J. Khalsan, L. Machado, E. S. Al-Shamery, et al., "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction," *IEEE Access*, vol. 10, pp. 8451–8477, 2022, doi: 10.1109/ACCESS.2022.3146312.

[4] A. Panigrahi, A. Pati, M. N. Das, et al., "En-MinWhale: An Ensemble Approach Based on MRMR and Whale Optimization for Cancer Diagnosis," *IEEE Access*, vol. 11, pp. 12123–12134, 2023, doi: 10.1109/ACCESS.2023.3318261.

[5] H.-H. Huang, J. Shu, Y. Liang, "MUMA: A Multi-Omics Meta-Learning Algorithm for Data Interpretation and Classification," *IEEE Journal of Biomedical and Health Informatics*, 2024, doi: 10.1109/JBHI.2024.3363081.

[6] R. Lupat, R. Perera, S. Loi, et al., "Moanna: Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes," *IEEE Access*, vol. 11, pp. 25674–25685, 2023, doi: 10.1109/ACCESS.2023.3240515.

[7] H. Guo, X. Lv, Y. Li, et al., "Attention-based GCN Integrates Multi-omics Data for Breast Cancer Subtype Classification and Patient-specific Gene Marker Identification," *bioRxiv*, 2022, doi: 10.1101/2022.09.05.506572.

[8] I. Zenbout, A. Bouramoul, S. Meshoul, et al., "Efficient Bioinspired Feature Selection and Machine Learning Based Framework Using Omics Data and Biological Knowledge Data Bases in Cancer Clinical Endpoint Prediction," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3234294.

[9] I. Bishara, J. Chen, J. I. Griffiths, et al., "A Machine Learning Framework for scRNA-seq UMI Threshold Optimization and Accurate Classification of Cell Types," *Frontiers in Genetics*, vol. 13, p. 982019, 2022, doi: 10.3389/fgene.2022.982019.

[10] H. Moghimianavval, B. Meghdadi, T. Clément, et al., "Predicting Breast Cancer Phenotypes from Single-cell RNA-seq Data Using CloudPred," *arXiv preprint*, 2024, doi: 10.48550/arxiv.2402.11289.

[11] J.-W. Park and J.-K. Rhee, "Integrative Analysis of ATAC-Seq and RNA-Seq Through Machine Learning Identifies 10 Signature Genes for Breast Cancer Intrinsic Subtypes," *Biology*, vol. 13, no. 10, p. 799, 2024, doi: 10.3390/biology13100799.

[12] A. Abdul Hayum, J. Jaya, R. Sivakumar, et al., "Critical Review on Breast Cancer Classification," in *Proc. Int. Conf. on Data Communication and Engineering Computing (ICDCECE)*, 2022, pp. 1–7, doi: 10.1109/ICDCECE53908.2022.9792781.

[13] Alharbi F, Vakanski A. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. Bioengineering (Basel). 2023 Jan 28;10(2):173. doi: 10.3390/bioengineering10020173. PMID: 36829667; PMCID: PMC9952758.
[14] B. Haznedar and N. Şimşek, "A Comparative Study on Classification Methods for Renal Cell and Lung Cancers Using RNA-Seq Data," *IEEE Access*, 2022, doi: 10.1109/ACCESS.2022.3211505.
[15] H. Chen et al., "Auxiliary Diagnosis of Breast Cancer Based on Machine Learning and Hybrid Strategy," *IEEE Access*, 2023, doi: 10.1109/access.2023.3312305.
16] Adem, A. M., Kant, R., Kumari, S., Swaroopa, C. N., & Gupta, G. (2024). *Breast Cancer Classification: In-depth Exploration of Different Paradigms and Ensemble Technique*. 702–708. https://doi.org/10.23919/indiacom61295.2024.10498454.
[17] Baranov, O. Yu., Turova, P., Kushnarev, V., Balabanian, L., Kotlov, N., Chernyshov, K., Fowler, N., & Lennerz, J. K. (2024). Closing classification gaps in luminal breast cancer with single-cell RNA-seq insights from normal breast lineages. *Journal of Clinical Oncology*, *42*(16_suppl), e13168. https://doi.org/10.1200/jco.2024.42.16_suppl.e13168
[18] Saddiqa, A., Zakir, M., Sheikh, M., Muneer, Z., Hassan, A., Ali, I., Haq, I., Khan, A. A., Malik, A., & Siddiqi, A. R. (2024). On discovery of novel hub genes for ER+ and TN breast cancer types through RNA seq data analyses and classification models. *Dental Science Reports*, *14*(1). https://doi.org/10.1038/s41598-024-69721-9
[19] Arifin, T., Prasetyo Agung, I. W., Junianto, E., Rachman, R., Wibowo, I., & Agustin, D. D. (2024). Breast cancer identification using machine learning and hyperparameter optimization. *Indonesian Journal of Electrical Engineering and Computer Science*, *36*(3), 1620. https://doi.org/10.11591/ijeecs.v36.i3.pp1620-1630