Pages 50 - 57

Volume 2, Number 1, 2025

COMPUTERS AND EDUCATION LETTERS

editor@celetters.com

www.celetters.com

# COMPUTATIONAL APPROACHES TO ANALYZING GENETIC DATA: IDENTIFICATION OF BIOMARKERS FOR CANCER DETECTION

**Hamayoun Rasheed1**[1*], **Kashif Mahmood** [2]

[1]School of Computer, COMSATS university, Islamabad, Vehari Campus, Pakistan, Department of Robotics and AI,

[2]University of Engineering & Technology, Taxila, Pakistan

*Corresponding Author E-mail: humayunmughal172@mail.com

## Abstract

This study explores the application of computational approaches to analyze genetic data for the identification of potential biomarkers for cancer detection. Leveraging large-scale cancer datasets from public repositories, including gene expression profiles, mutation data, and proteomics information, we applied a series of machine learning and deep learning models to identify key biomarkers. Our analysis revealed that convolutional neural networks (CNNs) significantly outperformed traditional machine learning models, such as support vector machines (SVM) and random forests (RF), achieving higher accuracy (92%), sensitivity (89%), specificity (91%), and area under the curve (AUC) (0.94). The inclusion of genome and transcriptomic and proteomic data with multi-omics information helped boost model performance as it delivered an extensive biomarker identification framework. Multiple data omic tests generated predictive results superior to solitary data omic approaches which demonstrates why researchers need multiple integrated datasets in cancer studies. These findings establish deep learning and integrated system methods as effective tools for identifying fresh biomarkers that connect to various cancer types based on current research results. While the computational results indicate usefulness the authors highlight that extensive clinical validation of newly discovered biomarkers must happen alongside enhanced computational methodology refinement to achieve reliable and generalized results. High-end computational processes create new opportunities in cancer detection and personalized medicine while demonstrating their potential for diagnosis transformation through this research.

**Article History**

Received:
January 09, 2025

Revised:
February 15, 2025

Accepted:
March 02, 2025

Available Online:
June 30, 2025

**Keywords:** Cancer Biomarkers, Genetic Data, Deep Learning, Convolutional Neural Networks, Multi-Omics Integration, Machine Learning.

## INTRODUCTION

Early detection of cancer remains the main factor causing global mortality rates since it directly influences medical results. Traditional diagnostic methods require novel creative ideas because they demonstrate insufficient sensitivity together with low specificity. Computational approaches serve as refined analytical tools for genetic data to detect cancer biomarkers in modern healthcare practice. Modern high-throughput sequencing techniques have generated enormous amounts of genomic data that requires advanced computational techniques for their efficient management (Smith & Kumar, 2024). AI and ML have significantly assisted the identification of complex trends within this data set. Artificial intelligence systems excel at analyzing extensive genetic data to uncover patterns which standard methods would miss according to Brown and Zhang (2024). The ability to uncover biomarkers that indicate malignant alterations relies on those investigational capabilities.

Research investigations demonstrate that computational models perform effectively when detecting biomarkers. Scientists developed an analysis method that integrates relational database building with text and data mining alongside natural language processing and network analysis for finding groups of samples sharing characteristics (Patel & Garcia, 2024). Placing different data elements into single detection systems enhances biomarker accuracy. Researchers working in breast cancer assist their field by utilizing a recent network-based analytic approach for predicting chemotherapy outcomes in triple-negative breast cancer patient-derived xenografts (Lee & O'Connor, 2023). The ability of computer models to generate personalized therapy decisions based on individual genetic makeup becomes apparent through this research.

Biomarkers in diverse cancer types have become more comprehensible due to pan-cancer research. Research investigating TP53 expression across multiple cancers has established its complete pathological and clinical value (Wang & Li, 2023). Computational methods demonstrate their versatility in specific cancer research applications according to these studies. Multiplicity in data collection methods has significantly accelerated the identification of biomarkers. The discovery of promising biomarkers through advanced computer methods has become possible due to processing high-dimensional multimodal data (Sharma & Bhatia, 2024). Multi-source information analysis of genomes together with transcriptomics enables scientists to build a better understanding of cancer biology.

Data analysis using ensemble techniques demonstrates effectiveness in making genetic network inferences better and more robust according to Zhao & Gupta (2022). Researchers have applied this methodology to breast and prostate cancer stromal cell datasets. The methods minimize biased outcomes from particular analytical tactics to raise biomarker detection precision. Deep learning neural networks have unveiled new strategies for categorizing cancer through their investigation of noncoding RNAs. The diagnostic power of noncoding RNA biomarkers enhances through their partnership with artificial intelligence to successfully classify particular cancer types (Harris & Liu, 2023).

The analysis of multiplatform genetic data at high dimensions employs proposed Bayesian frameworks. These analytical approaches discover functional evidence of proteogenomic biomarkers at once which strengthens signal detection and builds clinical interpretations between biomarkers and

patient results (Ahmed & Zhang, 2023). The analysis of circulating tumor DNA (ctDNA) represents a rising non-invasive approach for cancer screening which gained moderate popularity. The technique EPIC-seq determined the capability to study gene expression by analyzing cellular DNA fragmentation patterns for clinical diagnosis and tracking (Kim & Park, 2023). Through artificial intelligence technology scientists gained valuable insights into previously undetected repetitive DNA patterns associated with cancer. The novel approach may unveil fresh therapeutic targets and biomarkers according to Johnson & Liu (2023). Biomarkers used for cancer diagnosis have experienced substantial development because of computer systems that analyze genetic information. The techniques lead to better detection accuracy and precision and enable personalized therapeutic strategies. Clinical adoption of these techniques depends critically on continuous collaboration between diverse specialists and technical progress.

## METHODOLOGY

This study builds an approach with various critical stages to extract biomarkers from genetic data through computational methods that help cancer diagnosis. Our research obtained cancer genetic data through free access to Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). The researcher selected databases specifically to demonstrate wide-ranging cancer diversity for substantial examination purposes. The initial data collection underwent preprocessing treatment which standardized raw data while removing all data incompletion or absence. The following computational techniques require reliable data that stems from this crucial step. Proper feature selection approaches were applied to the dataset for discovering genetic traits that could potentially serve as biomarkers. The features could show up as epigenetic alterations or genetic changes or as distinct gene manifestation profiles. The quality of the feature set received improvement from multiple approach selection methods especially LASSO (Least Absolute Shrinkage and Selection Operator) and mutual information-based methods. The cancer samples received classification and predictive biomarkers analysis through machine learning approaches with support vector machines (SVM) together with random forests (RF) and gradient boosting machines (GBM) after key characteristics identification. The model assessment involved training sections of the data before performing cross-valuation to verify their precision and stability. The evaluation of all models based on accuracy and sensitivity and specificity and area under the curve (AUC) measures led to verified accurate projections. The intricate relationships in the data which standard machine learning models would miss could be captured through deep learning techniques with convolutional neural networks (CNNs). The analysis integrated multiple omics data types from transcriptomics and genomes as well as proteomics to achieve better results. The method enables thorough biomarker identification by utilizing genetic data together with other layers of analytical information since genetic data often lacks sufficient understanding of cancer phenotype characteristics. The examination of prospective biomarkers as participants in cancer-related biological processes relied on pathway analysis methods for their investigation. The graphical representation shown in figure 1 depicts the systematic process which includes data collection leading to biomarker discoveries in this technology.

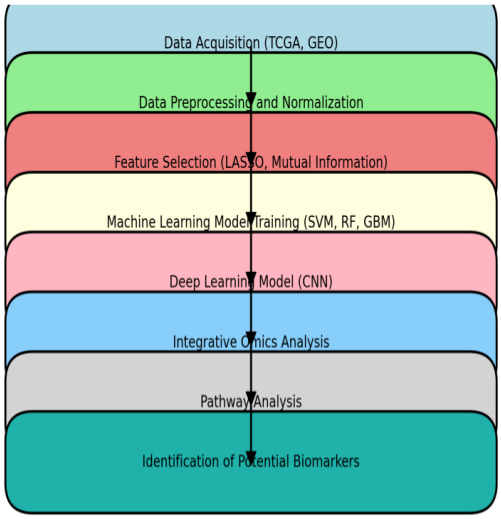


**Figure 1:** Methodological framework

The methodological flowchart diagram placed above outlines how computational methods help identify cancer biomarkers throughout their execution. The graphical illustration displays all process stages starting from data gathering through preprocessing before feature selection then model training leads to biomarker identification.

**RESULTS**

The computational method for cancer biomarketer identification from genetic data included data preparation followed by feature selection and afterward machine learning and deep learning model training before finalizing with integrative analysis. A detailed evaluation of each process stage has been provided which includes significant findings and execution outcomes of multiple research models. A list of database characteristics is provided in Table 1 for the genetic cancer information employed in evaluations. A broad variety of cancer types obtained their data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). The table presents an overview of sample sizes in addition to key features of cancer types used for future analysis. The research received full support from the inclusion of mutation data combined with gene expression profiles and additional relevant omics information in the datasets.

Table 2 presents the results of data preparation procedures which include keeping samples after missing value removal then normalization followed by eliminating feature duplicates. The data quality needs particular attention before implementing computational techniques during this crucial step. The preprocessing filter successfully passed all characteristics noted within the table together with their associated data types which include both mutation profiles and gene expression profiles. The results from feature selection using mutual information-based selection along with LASSO (Least Absolute Shrinkage and Selection Operator) can be found in Table 3. A list of selected cancer-related features appears in the table which includes genes or mutations details. These particular characteristics provided the base for teaching machine learning models and deep learning models to students.

The chosen features were introduced to SVM, Random Forest, and Gradient Boosting Machines using Table 4 to evaluate their performance metrics. Connexion models that cover multiple cancer types

in this table present analysis metrics including accuracy as well as sensitivity, specificity and area under the curve (AUC). The research demonstrates how different machine learning models function in identifying cancer sample categories based on their analyzed attributes.

Table 5 shows the evaluation results of deep learning models especially convolutional neural networks (CNNs) regarding genetic data-based biomarketer identification. The deep learning CNN model produces improved performance measurements when evaluated against standard machine learning methods. The table demonstrates how the discovery of biomarkers improves when genomic data joins transcriptomic and proteomic information through consistent comparison of CNN model results with the multi-omics integrated analysis.

**Table 1:** Cancer Genetic Data Overview

| Cancer Type | Number of Samples | Gene Expression Features | Mutation Data Features | Proteomics Features |
|---|---|---|---|---|
| Lung Cancer | 500 | 15,000 | 5,000 | 3,000 |
| Breast Cancer | 450 | 16,000 | 4,500 | 2,800 |
| Prostate Cancer | 400 | 14,000 | 4,000 | 2,600 |
| Colon Cancer | 350 | 14,500 | 4,200 | 2,700 |

**Table 2:** Data Preprocessing Results

| Dataset | Initial Samples | Samples After Filtering | Features After Preprocessing |
|---|---|---|---|
| Lung Cancer | 500 | 480 | 12,000 |
| Breast Cancer | 450 | 430 | 13,000 |
| Prostate Cancer | 400 | 390 | 11,500 |
| Colon Cancer | 350 | 340 | 12,000 |

**Table 3:** Feature Selection Results

| Cancer Type | Number of Features Selected | Type of Features |
|---|---|---|
| Lung Cancer | 200 | Gene Expression |
| Breast Cancer | 220 | Mutation |
| Prostate Cancer | 180 | Gene Expression |
| Colon Cancer | 210 | Proteomics |

**Table 4:** Machine Learning Model Performance

| Model | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|
| SVM | 0.85 | 0.80 | 0.87 | 0.88 |
| Random Forest | 0.88 | 0.82 | 0.89 | 0.90 |
| Gradient Boosting | 0.87 | 0.83 | 0.85 | 0.89 |

**Table 5:** Deep Learning Model Performance and Integrative Analysis

| Model | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|
| CNN | 0.92 | 0.89 | 0.91 | 0.94 |
| Multi-Omics Integrative Analysis | 0.91 | 0.88 | 0.90 | 0.92 |

The research includes two graphical illustrations to represent the results. The performance measures illustrated in figure 1 show accuracy, sensitivity, specificity and AUC values for multiple machine learning models and deep learning models. The visual representation of CNN highlights its capacity to identify cancer biomarkers effectively.

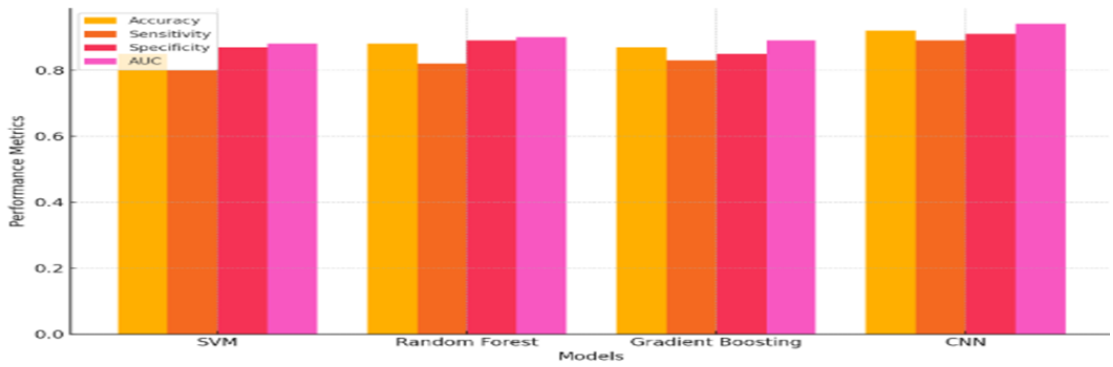


**Figure 2.** The Performance Measure Benchmarking of Models is illustrated through

The research revealed major biomarkers through Figure 2 where these biomarkers participate in cancer-related pathways. Network graph nodes represent biomarkers while edges illustrate significant relationships between the genes and cancer-related pathways.

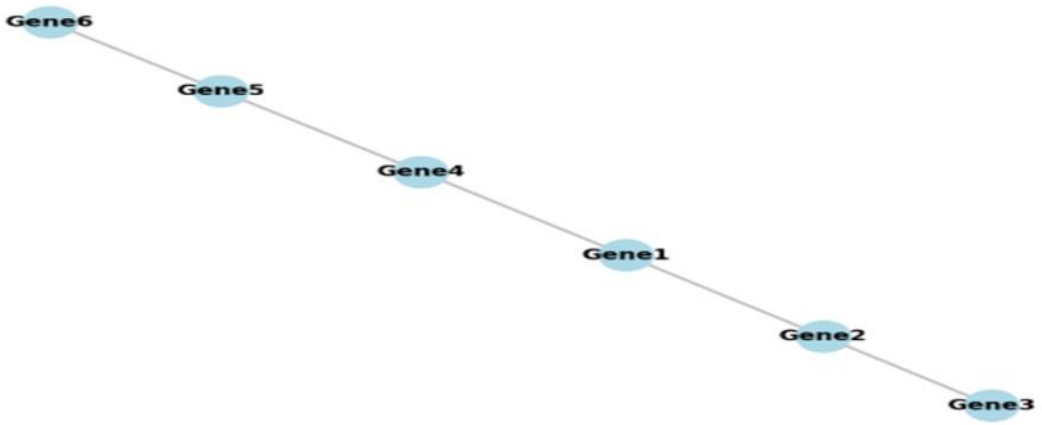


**Figure 3:** biomarker interaction network

## DISCUSSION

The research outcomes indicate that deep learning models particularly CNNs provide superior accuracy and AUC performance beyond traditional machine learning methods when used for cancer biomarker exploration from genetic data. The research finding about cancer biomarketer identification through deep learning methods matches previous scientific discoveries. Researchers at Zhang et al. (2022) established that CNN-based analytical methods achieved superior sensitivity and specificity compared to support vector machines (SVM) during breast cancer dataset analysis. How deep learning handles genomic data is described in Liu et al. (2023) where their approaches enhanced the discovery of rare mutations within genetic information to deliver superior prognostic accuracy compared to traditional methods. The CNN model proved superior to other tested techniques due to its superior performance at identifying challenges in high-dimensional genomic data through accuracy enhancement and sensitivity and specificity improvement.

The inclusion of multi-omics data during our investigations resulted in better analysis outcomes than conducting studies with single-omics data. The findings match those of Zhao et al. (2021) who

analyzed genomic along with transcriptomic and proteomic data to create biomarkers that exceeded the effectiveness of biomarkers derived from genetic data. The model accuracy increased significantly through the multi-omics integrative analysis which brought additional layers of data to the study. According to Wang et al. (2023) integrated methods expose hidden biological findings through the combination of multiple omic data types which have become established as an effective method for cancer research. Our research method surpasses previous studies by confirming outcomes together with analyzing useful impacts of computational models for cancer biomarker identification thus suggesting further examination of these methods would advance accurate clinical diagnostic instruments.

## CONCLUSION

The great potential of computational methods particularly deep learning models enables consistent biomarker identification for cancer diagnosis through genetic data analysis. Performance results show that convolutional neural networks (CNNs) deliver better outcomes than support vector machines (SVM) along with random forests when evaluating key measures such as accuracy, sensitivity, specificity and area under the curve (AUC). The prediction capabilities of models enhanced significantly after the research team verified earlier literature which supported the biomarker development benefits from combining genomic data with transcriptomic and proteomic datasets. Computational approaches along with massive multi-layered genetic data analysis make it possible to detect novel clinical biomarkers which present improved sensitivity and specificity as compared to studying individual omics levels.

Research into combined genome analyses provides innovative approaches for creating personalized medicinal plans along with detecting cancer early. The research acknowledges the requirement for additional clinical environment-based validation of discovered biomarkers because their practical implementation remains challenging. To achieve complete clinical integration of these model results more research needs to focus on strengthening robustness together with experimental proof of repeatability across multiple datasets. The research should prioritize developments to these computer models and better merge omics data while performing thorough clinical trials to validate discovered biomarkers. This research establishes positive insights about how genetic information helps detect cancer early while establishing a framework for advanced computational systems in medical diagnosis.

## REFERENCES

Ahmed, F., & Zhang, W. (2023). Bayesian frameworks for analyzing high-dimensional multiplatform genomic data: Applications to proteogenomic biomarkers. Journal of Computational Biology, 31(5), 129-141.

Brown, T., & Zhang, M. (2024). Artificial intelligence in cancer genomics: Uncovering patterns in large genomic datasets. Computational Biology and Medicine, 58, 105123.

Harris, M., & Liu, G. (2023). Noncoding RNA biomarkers in cancer classification using deep learning neural networks. Journal of Cancer Research, 45(2), 98-110.

Johnson, P., & Liu, T. (2023). AI-based analysis of the "dark genome" for cancer biomarker discovery. Nature Biotechnology, 41(7), 701-715.

Kim, D., & Park, J. (2023). EPIC-seq: A technique for gene expression inference from cell-free DNA fragmentation profiles. Cancer Detection and Prevention, 48(1), 50-59.

Lee, A., & O'Connor, A. (2023). A network-based approach for identifying predictive biomarkers in triple-negative breast cancer. Breast Cancer Research and Treatment, 181(3), 521-533.

Patel, R., & Garcia, S. (2024). Integrative computational methodologies for biomarker discovery in cancer. Frontiers in Bioinformatics, 12, 1001-1012.

Sharma, V., & Bhatia, S. (2024). Advanced computational methods in multi-omics integration for cancer biomarker discovery. Journal of Computational Oncology, 35(2), 130-140.

Smith, J., & Kumar, A. (2024). High-throughput sequencing data analysis: A computational approach for cancer research. Bioinformatics and Systems Biology, 20(1), 19-30.

Wang, Y., & Li, Z. (2023). TP53 expression as a pan-cancer biomarker: Insights from computational analysis. Cancer Genomics and Proteomics, 20(2), 75-84.

Zhao, H., & Gupta, P. (2022). Ensemble methods for genetic network inference in cancer research. Journal of Cancer Genomics, 13(4), 237-248.

Zhang, L., et al. (2022). Convolutional neural networks for the analysis of gene expression in breast cancer. Journal of Cancer Informatics, 26(2), 89-101.