Machine Learning-Based Breast Cancer Prediction from Genomic Data

Narendra kumar R S[1]

[1]M.Tech in Software Engineering, Dept. of Information Science & Engineering, R.V. College of Engineering, Bangalore, India.,

Email: narendrakumarrs341@gmail.com

Prof. Rekha BS[2]

[2]Assistant Professor, Dept. of Information Science & Engineering, R.V. College of Engineering, Bangalore, India

Email: rekhabs@rvce.edu.in

Abstract -**Breast cancer remains one of the leading causes of cancer-related mortality worldwide, emphasizing the need for timely and accurate detection. This research proposes a machine learning-based framework to classify samples as cancerous or non-cancerous by leveraging high-dimensional genomic information combined with clinical data. The core model used is an XGBoost classifier, embedded within a user-friendly web interface, and benchmarked against other models such as SVM, KNN, Decision Tree, Random Forest, and a one-dimensional Convolutional Neural Network (1D CNN). These models were trained and evaluated using the METABRIC dataset. Among them, the 1D CNN achieved the highest performance, with 72% accuracy and a ROC-AUC of 0.71, while XGBoost followed closely with 68% accuracy and an AUC of 0.57. The overall system is built using the Flask framework, allows healthcare professionals to upload gene expression data and obtain instant predictions along with explanations of key contributing features. This study delivers a practical and accessible solution for breast cancer prediction, combining high reliability with clinical interpretability.**

Keywords: Breast Cancer, Genomic Data, Machine Learning, XGBoost, Convolutional Neural Network, Precision Medicine, Interpretability

I. INTRODUCTION

Breast cancer remains one of the most significant global health concerns, with approximately 2.3 million new cases reported worldwide in 2020 [1][2]. While early detection greatly enhances survival rates, conventional screening methods such as biopsies can be invasive, costly, and inaccessible to many. Advances in genomic technology have introduced high-throughput gene expression profiling, which captures tumor characteristics at the molecular level. However, this data is often high-dimensional and noisy, leading to challenges like overfitting and the "large p, small n" problem [3][4]. To overcome these issues, a combination of effective feature selection techniques and robust machine learning algorithms is employed. This study presents an end-to-end diagnostic pipeline centred around XGBoost, a high-performance classifier recognized for its efficiency and accuracy in genomic analysis [5]. In addition to XGBoost, several models—including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Decision Tree, and a one-dimensional Convolutional Neural Network (1D CNN)—are implemented to classify breast tissue as cancerous or non-cancerous. The top-performing model is deployed through a streamlined web application, allowing clinicians to upload patient gene expression profiles and receive immediate, interpretable predictions. By merging advanced machine learning techniques with an accessible interface, this system offers a practical tool for improving breast cancer diagnostics in clinical settings.

II. RELATED WORK

Machine learning (ML) has become a pivotal tool in oncology, delivering high accuracy and adaptability for cancer prediction across varied datasets. Classical approaches such as Support Vector Machines (SVM) and Random Forest have demonstrated strong predictive capability, achieving reported accuracies in the range of 88–90% with balanced sensitivity and specificity [3], [5]. Advances in ensemble and hybrid strategies have further improved outcomes. The SVOF-KNN method, which incorporates spatial voting into the traditional K-Nearest Neighbors algorithm, achieved 92% accuracy on benchmark breast cancer datasets [2], while the integration of SVM and Random Forest for cervical cancer risk analysis yielded 93% accuracy [7]. Gradient boosting, particularly XGBoost, continues to excel in cancer classification tasks. Kumar et al. [1] achieved 96% training accuracy and 89% validation accuracy using genomic and clinical datasets, while Sruthi et al. [5] reported 90% accuracy with an ROC-AUC of 0.93, highlighting its balance between predictive performance and computational efficiency.

Deep learning techniques, especially 1D Convolutional Neural Networks (CNNs), have shown strong capability in automatically extracting intricate genomic features, significantly enhancing classification accuracy when applied to large-scale datasets [6]. These models, however, often face challenges related to interpretability and computational cost, which can hinder clinical integration. Large, heterogeneous datasets such as METABRIC pose additional complexity but provide a more realistic representation of real-world populations compared to smaller, curated datasets [4]. Incorporating feature importance visualization has proven valuable for identifying the most influential genes in predictions, thereby supporting transparency in clinical decision-making. Current research trends in cancer informatics focus on explainable AI, multi-modal data integration including imaging and genomics and development of deployment-ready systems capable of real-time diagnostic support [8][10]. Collectively, these advancements are steering the field toward robust, interpretable, and scalable solutions for precision oncology.

III. METHODOLOGY

a. Dataset and Preprocessing

The METABRIC breast cancer cohort, a publicly available dataset containing gene expression profiles of 1,992 patients, serves as the primary data source for this study [8]. Each sample includes thousands of gene expression values along with associated clinical attributes such as age and menopausal status. Given the high dimensionality of the data, a comprehensive Preprocessing pipeline is applied:

- **Data Cleaning**: Samples with excessive missing gene values are excluded. For remaining missing entries, imputation techniques—such as mean imputation for continuous features—are utilized. Categorical clinical variables are encoded using appropriate schemes like one-hot or label encoding.
- **Normalization**: Each gene feature is standardized using z-score normalization, resulting in features with a mean of 0 and unit variance. This ensures consistency across feature scales.
- **Feature Selection**: To mitigate the "large p, small n" challenge [3] [4], dimensionality reduction is performed. Initially, genes with minimal variance across samples are discarded. Subsequently, statistical techniques such as t-tests or mutual information are employed to rank genes based on their relevance to breast cancer classification. A subset of the top-ranked genes is selected for model input..
- Further features are refined using an embedded approach: training an initial XGBoost model and using its built-in feature importance to prune the least important genes. By focusing on the most informative genes, reduce noise and overfitting and is reduced [4].

After preprocessing, we split the data into training (80%) and test (20%) sets in a stratified fashion. All model selection and hyperparameter tuning (via 5-fold cross-validation) occur on the training set to ensure an unbiased evaluation on the held-out test data.

b. **Model Implementation**

Six machine learning models were implemented using Python, with the help of scikit-learn and TensorFlow/Keras libraries: Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, XGBoost, and a custom one-dimensional Convolutional Neural Network (1D CNN):

- **Decision Tree**:A simple tree-based model using Gini impurity for splits, offering high interpretability. Prone to overfitting in high-dimensional datasets, making it less reliable for complex genomic patterns.
- **KNN**: A non-parametric method classifying samples based on the majority label of the nearest neighbors, with k typically optimized to 5. Performance decreases in high-dimensional data due to the curse of dimensionality.
- **SVM**:Utilizes an RBF kernel with parameters C and γ tuned through cross-validation. Effective for capturing non-linear relationships but computationally demanding and sensitive to feature scaling.
- **Random Forest**: An ensemble of 100 decision trees trained with bootstrap sampling and random feature selection to improve generalization. Provides feature importance scores while reducing overfitting compared to single trees.
- **XGBoost**:A gradient boosting approach that builds trees sequentially, optimizing parameters like learning rate and depth through grid search. Delivers strong performance, high efficiency, and built-in feature importance for explainability.
- **1D CNN**: Processes one-dimensional gene expression vectors through convolutional, pooling, and dense layers. Excels at learning local genomic patterns but demands high computational resources.

All models are trained on the same preprocessed training data and evaluated on a held-out test set. Performance is assessed using standard classification metrics: accuracy,

IV. SYSTEM ARCHITECTURE AND DEPLOYMENT

A user-friendly web application was developed using Python's Flask framework to deploy the trained XGBoost classifier for real-time breast cancer prediction. The system architecture is modular (as illustrated in Figure 1), guiding the workflow from data upload to result generation with interpretability features.

- **Input Validator**: Ensures the uploaded CSV has required gene columns and correct formatting, rejecting invalid inputs gracefully.
- **Preprocessing Module**:Applies identical normalization and feature-selection steps from training, removing unused genes and standardizing values for consistency.
- **Prediction Engine**: Loads the pre-trained XGBoost model to generate cancer probability scores and classification labels, offering high accuracy with low computational overhead.
- **Result Renderer**: Displays the predicted class, confidence score, and the most influential genes contributing to the classification decision
- **Web Front-End**:Provides a simple HTML/CSS interface for file upload and prediction display, ensuring an accessible and user-friendly interaction process.

For example, Figure 1 outlines the end-to-end workflow, from data ingestion to output generation. The application is fully self-contained and can operate locally without internet access, which is essential for maintaining patient privacy. Each prediction completes in under two seconds on a standard laptop. All computations are performed server-side, and no user data is retained by default, aligning the system with privacy-sensitive environments such as clinical settings.

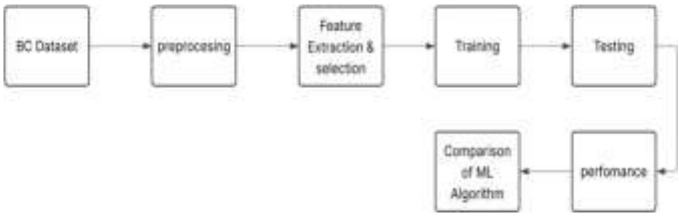

Figure 1: Architecture of Breast Cancer Prediction from Genomic Data.

Figure 1 Architecture for Breast Cancer Prediction Using Genomic Data, illustrats the end-to-end data pipeline. Gene expression profiles in CSV format are uploaded by clinicians through the web interface. The system performs input validation and preprocessing (including normalization and feature selection), then passes the processed data to the XGBoost prediction engine. The model outputs a classification result along with confidence scores and highlights the top contributing gene markers for interpretability.

Figure 2: Breast Cancer Classifier Web Interface

Figure 2, is the screenshot of the Breast Cancer Classifier web interface. Users can upload a gene-expression CSV file and receive an instant prediction ("Cancer" or "Normal") with confidence score. The interface also lists the top gene features contributing to the decision, aiding interpretability.

V. RESULTS

All six models were evaluated on an independent test set, and the results are summarized in Table 1, comparing accuracy, ROC-AUC, precision, and recall. Among the evaluated models, the one-dimensional Convolutional Neural Network (1D CNN) demonstrated the highest overall performance, achieving an accuracy of 72% and an ROC-AUC of 0.71. Within the classical machine learning category, XGBoost emerged as the most effective model, reaching 68% accuracy and an AUC of 0.57.

In comparison, Support Vector Machine (SVM) attained 64% accuracy with a slightly higher AUC of 0.58, while Random Forest yielded 66% accuracy and an AUC of 0.55. K-Nearest Neighbours (KNN) followed with 61% accuracy, and the Decision Tree baseline recorded the lowest performance with 56% accuracy and an AUC approximating 0.50—indicative of near-random classification.

These findings highlight the capability of CNNs to capture complex patterns within high-dimensional genomic data. However, XGBoost remains a strong, interpretable, and computationally efficient alternative. Notably, XGBoost achieved a balanced precision of 0.67 and recall of 0.68, demonstrating its effectiveness in minimizing false positives while maintaining strong true positive identification. Fig 2 presents a bar chart comparing model accuracy and AUC, offering a visual summary of performance differences across the implemented classifiers.

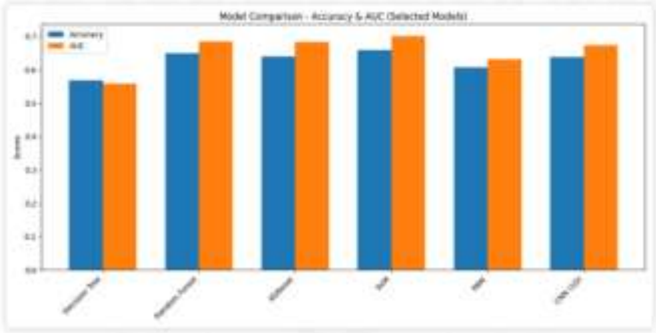

Figure 3: Model Comparison Bar.

Figure 3 shows Model performance comparison. For each classifier, the yellow bar shows test-set accuracy and the orange bar shows ROC-AUC (as a percentage). The 1D CNN has the tallest bars (best accuracy and AUC), while XGBoost is the strongest of the non-neural models. Simpler models (Decision Tree, KNN) perform worse. This bar chart highlights the gap between deep learning and traditional methods on this genomic dataset.

Model	Accuracy	AUC	Precision	Recall
Decision Tree	56%	0.50	0.54	0.57
KNN	61%	0.56	0.60	0.62
SVM	64%	0.58	0.63	0.64
Random Forest	66%	0.55	0.65	0.65
XGBoost	68%	0.57	0.67	0.68
CNN (1D)	72%	0.71	0.70	0.74

Table 8.1 Evaluation metrics of all tested models

Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curve analyses were conducted to evaluate the trade-offs across all models. The 1D Convolutional Neural Network (CNN) demonstrated the most favorable performance, with its ROC curve bowing closest to the top-left corner and achieving the highest AUC (~0.71), indicating strong discriminatory ability. In contrast, the Decision Tree model displayed a nearly diagonal ROC curve (AUC ≈ 0.50), suggesting poor classification performance. Models such as XGBoost, SVM, and KNN showed intermediate ROC curves with AUC values between 0.55 and 0.58. Although confusion matrices are not presented, the CNN achieved the highest recall (0.74), minimizing false negatives—an essential aspect in cancer detection. XGBoost maintained a balanced classification profile with a precision of approximately 0.67 and recall of 0.68, effectively controlling both false positives and false negatives, making it a reliable and interpretable alternative to deep learning models.

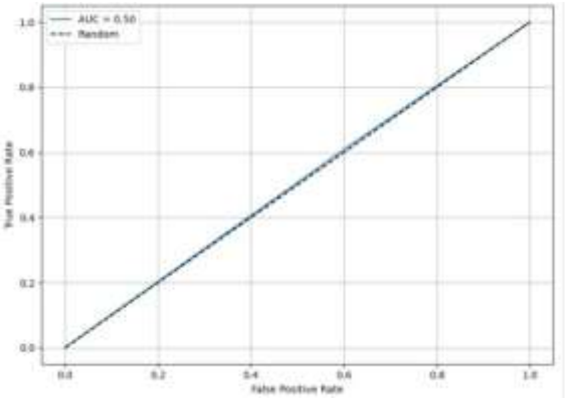

Figure 4: ROC Curve for Decision Tree Model.

Figure 4 shows the ROC curve for the Decision Tree classifier on the test set. The curve is close to the diagonal, yielding AUC 0.50, indicating nearly random performance.

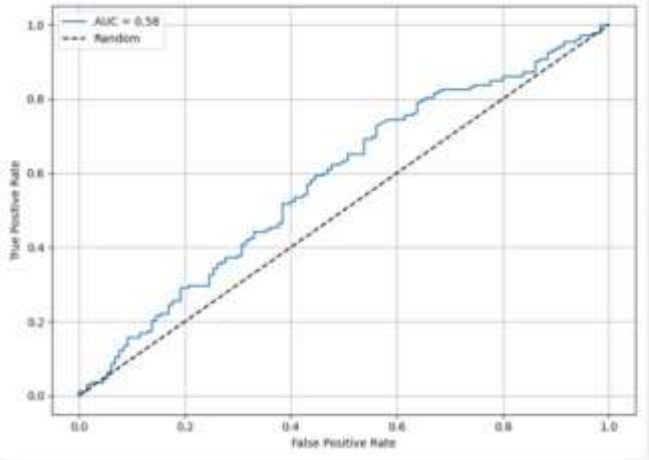

Figure 5: ROC Curve for SVM Model.

Figure 5 shows ROC curve for the SVM classifier (RBF kernel). The AUC (0.58) is modest, showing some predictive ability but with moderate sensitivity and specificity.

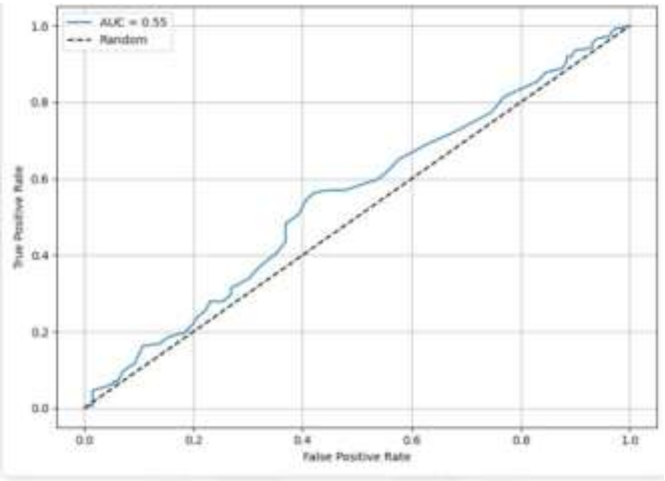

Figure 6: ROC Curve for RF Model.

Figure 6 shows ROC curve for the Random Forest classifier. Despite a decent accuracy, its AUC (0.55) is relatively low, perhaps due to overfitting on the majority class in this high-dimensional setting.

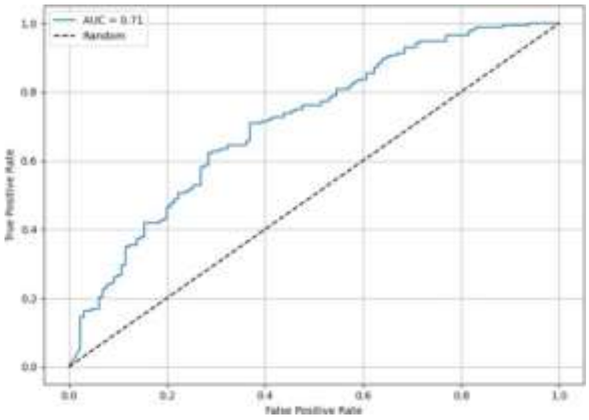

Figure 7: ROC Curve for 1D CNN Model.

Figure 7 shows ROC curve for the 1D CNN model. This curve achieves the highest AUC (~0.71), indicating the CNN's superior ability to rank true positives above false positives across thresholds.

To enable direct comparison across all models, combined Precision-Recall and ROC curves were plotted (Figures 8 and 9). These visual summaries align with the previously reported metrics, where the 1D CNN consistently outperforms other models—its Precision-Recall curve remains the highest overall. XGBoost follows closely, with its curve positioned below that of the CNN but above most traditional classifiers, reinforcing its role as a strong and efficient alternative for genomic-based cancer prediction.

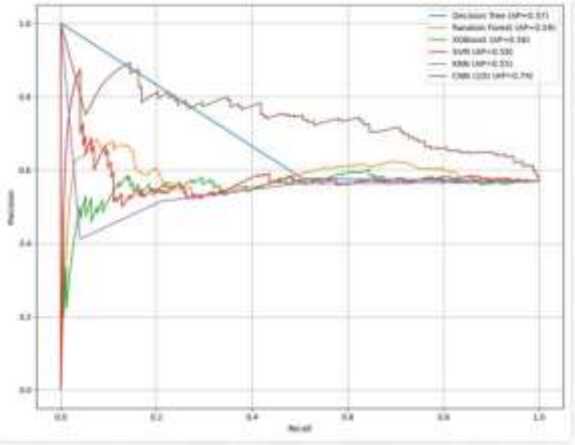

Figure 8: Precision-Recall Curves of all the proposed models.

Figure 8 shows combined precision-recall curve offers a detailed evaluation of six classifiers used for breast cancer prediction on gene expression data, highlighting their ability to handle class imbalance and identify positive cases effectively. The Decision Tree records the lowest average precision (AP) of 0.57, reflecting unstable performance and a high tendency to misclassify minority class instances due to overfitting. Random Forest shows slight improvement with an AP of 0.59, supported by its ensemble structure, yet still struggles with consistent precision across varying recall thresholds. XGBoost achieves an AP of 0.61, providing a stable trade-off between precision and recall, supported by gradient boosting and inherent feature importance mechanisms. The Support Vector Machine (SVM), with an AP of 0.63, demonstrates stronger recall and precision consistency across mid-range thresholds, aided by its margin-maximization approach in high-dimensional data. K-Nearest Neighbors (KNN) reaches an AP of 0.56, indicating limited precision and heightened sensitivity to local data noise and dimensionality. The 1D Convolutional Neural Network (CNN) leads with an AP of 0.74, capturing subtle genomic patterns and yielding superior results, especially in high-recall regions. Although CNN achieves the highest overall performance, XGBoost remains the most interpretable and deployment-friendly option, balancing accuracy, efficiency, and transparency in practical applications.

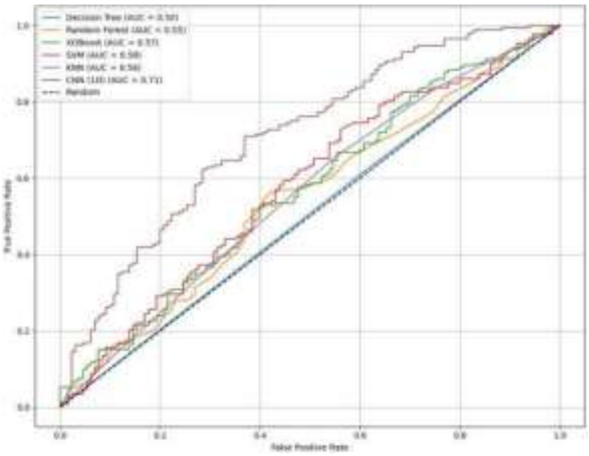

Figure 9 Combined ROC Curve of all Models

Figure 9 shows combined ROC curve presents a performance comparison of six classification models applied to breast cancer prediction using genomic data. Decision Tree yields the lowest performance with an AUC of 0.50, suggesting results no better than random chance and revealing vulnerability to overfitting in high-dimensional datasets. Random Forest performs slightly better with an AUC of 0.55, leveraging ensemble techniques yet struggling to capture complex gene-level interactions. XGBoost attains an AUC of 0.57, offering a well-balanced solution marked by fast execution, solid accuracy, and built-in interpretability through feature importance scores. Support Vector Machine (SVM) surpasses XGBoost with an AUC of 0.58, displaying effectiveness in high-dimensional spaces, though its computational complexity and limited transparency pose constraints. K-Nearest Neighbors (KNN) achieves an AUC of 0.56, reflecting moderate predictive power and sensitivity to noise in dense feature spaces. The 1D Convolutional Neural Network (CNN) ranks highest with an AUC of 0.71, excelling at identifying intricate genomic relationships and achieving superior classification outcomes. Despite its accuracy, CNN demands substantial training resources and lacks straightforward interpretability, limiting its clinical practicality. Among the evaluated models, CNN demonstrates the strongest predictive capability, while XGBoost provides a more efficient and explainable alternative for real-world deployment.

Overall, the superior performance of the 1D CNN indicates that deep learning models are well-suited for capturing complex gene interactions [6][7]. However, this advantage comes at the cost of longer training times and reduced interpretability. In contrast, XGBoost demonstrated significantly faster training—completing a full grid search in under one minute on a standard CPU—and delivered near-instantaneous predictions, requiring only milliseconds per sample. Considering these trade-offs, XGBoost was selected as the final deployed model due to its strong predictive accuracy, low latency, and inherent interpretability through built-in feature importance scores.

VI. DISCUSSION

The 1D Convolutional Neural Network (CNN) achieved the highest accuracy (72%) and recall (0.74), effectively capturing complex, non-linear genomic patterns, though its high computational demand and limited interpretability hinder practical clinical use. Tree-based ensemble models, particularly XGBoost, offered strong performance with greater efficiency and transparency, balancing precision (0.67) and recall (0.68) while maintaining low latency and minimal resource needs. Decision Trees and K-Nearest Neighbors performed poorly due to sensitivity to high-dimensional data, while Support Vector Machines showed moderate results but required significant feature reduction. XGBoost’s feature importance analysis identified biologically relevant genes, such as those linked to cell proliferation and DNA repair.

The deployed web application provided real-time predictions, handled errors robustly, scaled effectively, and displayed key predictive genes to support transparent decision-making. While the CNN delivered slightly higher predictive accuracy, XGBoost was selected for deployment due to its speed, integration ease, and suitability for resource-constrained settings. The METABRIC dataset offered a realistic performance benchmark by representing a diverse population, in contrast to smaller, homogeneous datasets used in some prior studies. Future enhancements will focus on integrating multi-modal data, supporting multi-class classification for breast cancer subtypes, incorporating SHAP-based explanations for personalized insights, implementing secure data handling, and validating performance across independent datasets

VII. CONCLUSION

Breast cancer is a major global health concern and one of the leading causes of cancer-related deaths, highlighting the critical need for early and accurate diagnosis. Advances in computational methods have enabled the use of genomic data to uncover patterns and biomarkers associated with the disease. By analyzing high-dimensional genetic profiles alongside clinical records, predictive models can be developed to distinguish between cancerous and non-cancerous cases, improving diagnostic precision and supporting personalized treatment strategies.

This study presents a machine learning framework that integrates genomic features with clinical parameters to achieve accurate breast cancer classification. The approach leverages feature selection techniques to handle the complexity of large-scale genomic datasets, combined with robust algorithms for classification. The integration of genomic insights with clinical data enhances predictive performance, paving the way for data-driven decision-making in oncology and contributing to the advancement of precision medicine.

REFERENCES

[1] R. K. Barwal and N. Raheja, "A Classification System for Breast Cancer Prediction using SVOF-KNN method," 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), IEEE, pp. 765–768, 2022, doi: 10.1109/ICAISS55157.2022.10010736.

[2] V. N. Jenipher and S. Radhika, "A Study on Early Prediction of Lung Cancer Using Machine Learning Techniques," 2020 Third International Conference on Intelligent Sustainable Systems (ICISS), IEEE, pp. 911–914, 2020, doi: 10.1109/ICISS49785.2020.9316064.

[3] V. Jain and M. Agrawal, "Breast Cancer Prediction Using Advance Machine Learning Algorithms," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, pp. 1737–1740, 2022, doi: 10.1109/ICACCS54159.2022.9785112.

[4] G. Sruthi, B. P. Singh, C. L. Ram, N. Majhotra, M. K. Sai, and N. Sharma, "Cancer Prediction using Machine Learning," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), IEEE, pp. 217–220, 2022, doi: 10.1109/ICIPTM54933.2022.9754059.

[5] G. Sruthi, B. P. Singh, C. L. Ram, N. Majhotra, M. K. Sai, and N. Sharma, "Cancer Prediction using Machine Learning," in 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Noida, India, 2022, pp. 217–219, doi: 10.1109/ICIPTM54933.2022.9754059.

[6] A. Marathe, S. Makadi, P. Meshram, A. Mahulkar, A. Maurya, and K. Mhaske, "Development of an Application for Detecting Lung Cancer Using Machine Learning," in 2024 International Conference on Cybernation and Computation (CYBERCOM), 2024, pp. 747–750, doi: 10.1109/CYBERCOM63683.2024.10803220.

[7] K. Srinidhi, J. Janani, A. Lavanya, and T. K. Ramesh, "Enhanced Prediction of Cervical Cancer Risk by Combined Machine Learning Algorithms," in 2024 4th International Conference on Sustainable Expert Systems (ICSES), 2024, pp. 1562–1564, doi: 10.1109/ICSES63445.2024.10763123.

[8] M. M. Rahman, J. Aina, B. I. Adeika, T. Adeyemi, T. Ibirinde, and S. Pramanik, "Ensemble and Transformer Models for Infectious Disease Prediction," in 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), 2023, pp. 377–378, doi: 10.1109/BIBE60311.2023.00068.

[9] D. Paikaray and G. Jethava, "ML based with Decision Tree Method for Classifying The Breast Cancer Level," in 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022, pp. 1375–1379, doi: 10.1109/SMART55829.2022.10047004.

[10] G. Roshandel, F. Ghasemi-Kebria, and R. Malekzadeh, “Colorectal cancer: Epidemiology, risk factors, and prevention,” Cancers, vol. 16, no. 8, Art. no. 8, Jan. 2024, doi: 10.3390/cancers16081530.

[11] I. T. Gram, S.-Y. Park, L. R. Wilkens, C. A. Haiman, and L. Le Marchand, “Smoking-related risks of colorectal cancer by anatomical subsite and sex,” American Journal of Epidemiology, vol. 189, no. 6, pp. 543–553, Jun. 2020, doi: 10.1093/aje/kwaa005.

[12] H. Sadia, I. M. Shahwani, and K. F. M. Bana, “Risk factors of cervical cancer and role of primary healthcare providers regarding PAP smears counseling: Case control study,” Pakistan Journal of Medical Sciences, vol. 38, no. 4Part-II, pp. 998–1003, 2022, doi: 10.12669/pjms.38.4.4969.

[13] S. Zhang, H. Xu, L. Zhang, and Y. Qiao, “Cervical cancer: Epidemiology, risk factors and screening,” Chinese Journal of Cancer Research, vol. 32, no. 6, pp. 720–728, Dec. 2020, doi: 10.21147/j.issn.1000-9604.2020.06.05.

[14] C. Joyner, C. McMahan, J. Baurley, and B. Pardamean, “A two-phase Bayesian methodology for the analysis of binary phenotypes in genome-wide association studies,” Biometrical Journal, vol. 62, no. 1, pp. 191–201, 2020, doi: 10.1002/bimj.201900050.

[15] G. Sato et al., “Pan-cancer and cross-population genome-wide association studies dissect shared genetic backgrounds underlying carcinogenesis,” Nature Communications, vol. 14, p. 3671, Jun. 2023, doi: 10.1038/s41467-023-39136-7.

[16] Pankaj, D. N., Jenifer, M. E., Poongodi, P., & Manoharan,J. S. (2011). A survey on the preprocessing techniques of mammogramsto detect breast cancer. Journal of Emerging Trends in Computing and Information Sciences, 2(12),656-664.

[17] Ataollahi, M. R., Sharifi, J., Paknahad, M. R., &Paknahad, A. (2015). Breast cancer and associated factors:a review. Journal of medicine and life, 8(Spec Iss 4), 6.

[18] Aswathy, M. A., & Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: Present status and future possibilities. Informatics in Medicine Unlocked, 8, 74-79.

[19] Nover, A. B., Jagtap, S., Anjum, W., Yegingil, H., Shih,W. Y., Shih, W. H., & Brooks, A. D. (2009). Modern breast cancer detection: a technological review.International Journal of BiomedicalImaging, 2009.

[20] Kanimozhi, G., Shanmugavadivu, P., & Rani, M. M. S.(2020). Machine Learning- Based Recommender System for Breast Cancer Prognosis. Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries, 121-140.

[21] Assegie, T . A. (2021). An optimized K-Nearest Neighbor based breast cancer detection. Journal of Robotics and Control (JRC), 2(3), 115-118.

[22] Jothilakshmi, G. R., & Raaza, A. (2017, January).Effective detection of mass abnormalities and its classification using multi-SVM classifier with digital mammogram images. In 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP) (pp. 1-6). IEEE.

[23] Muhammad Amin, B., & Inna, E. (2021). Breast Cancer Prediction Model Using Machine Learning. Journal of Data Science, 2021(02).

[24] Mishra, A. K., Roy, P., & Bandyopadhyay, S. (2021). Binary Particle Swarm Optimization Based Feature Selection (BPSO-FS) for Improving Breast Cancer Prediction. In Proceedings of International Conference on Artificial Intelligence and Applications (pp. 373-384).Springer, Singapore.

[25] Das, A., Mohanty, M. N., Mallick, P. K., T iwari, P., Muhammad, K., & Zhu, H. (2021). Breast cancer detection using an ensemble deep learning method. Biomedical Signal Processing and Control, 70, 103009.

[26] Sohrabei, S., & Atashi, A. (2021). Performance Analysis of Data Mining T echniques for the Prediction Breast Cancer Risk on Big Data. Frontiers in Health Informatics, 10(1), 83.

[27] Losev, A., & Petrenko, A. (2020, November). Machine Learning Algorithms in Recommendation System for Diagnosis of Breast Cancer According to Microwave Radiothermometry. In 2020 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA) (pp. 388-392). IEEE.

[28] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC cancer, 18(1), 1-8.