Developing a Machine Learning Model Using Gene Expression for Breast Cancer Prediction

Babatunde Abdulrauph Olarewaju[1], Alausa Babatunde Mubarak[2], Oke Afeez Adeshina[3]
[1 2] Department of Computer Science, University of Ilorin, Ilorin 240001, Nigeria
[3] Department of Computer Science, Federal College of Education, Iwo, Nigeria
babatunde.ao@unilorin.edu.ng[1]; babatundemubarak@gmail.com[2]; okeaa@fceiwo.edu.ng[3]

ABSTRACT

Recent advancements in genomics have generated vast gene expression datasets, offering profound insights into cancer biology. This study investigates an ensemble machine learning model, integrating K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and XGBoost, to predict and classify breast cancer subtypes from gene expression profiles. The methodology encompassed data preprocessing, including one-hot encoding, followed by model training and evaluation using standard metrics. The ensemble model achieved a strong overall accuracy of 90.32%. Crucially, it demonstrated a high precision of 0.9240, effectively minimizing false positives which is a key consideration for clinical diagnostics. While the model showed balanced performance with an F1-score of 0.9015, a comparative analysis revealed that, although individual baseline models (SVM, RF) reported higher raw accuracy of ~99%, the proposed ensemble provides a robust and interpretable framework optimized for reliable multi-class discrimination.

Keywords: Biotechnology, Model training, Machine learning, Cancer, Genomics

1. INTRODUCTION

Cancer, one of the leading causes of death globally, while breast cancer remains one of the most formidable illnesses in this space of global health challenges, with its clinical services posing obstacles to its treatment and prevention, where early diagnosis and accurate prediction are critical to improving patient outcomes (Kolawole & Ong, 2022). Traditional diagnostic methods, while effective, often rely on invasive procedures and may lack the precision needed for early prediction. Recent advancements in biotechnology have led to the availability of vast amounts of genomics data, specifically gene expression profiles, which offer promising insights into the molecular mechanisms underlying various types of cancer.

Analysing such data by using machine-learning techniques leads to developing clinical decision support systems for the correct estimation of survival time and so providing proper treatments to patients according to their survival (Abbasi et al., 2024). As discussed by (Rezapour et al., 2024), consequently, the entire genome expression analysis has become an important aspect in cancer used for identifying relevant gene pathways that are deregulated and drive abnormal metastatic spread.

Machine learning (ML) techniques have shown great potential in improving the performance of the analysis of larger datasets and help in the generation of accurate patient data (Alhumaidi et al., 2025). Machine learning algorithms, particularly those designed for handling large-scale and complex data, can automatically detect patterns, classify data, and make predictions with high accuracy. Machine learning approaches has also proven useful in various areas of healthcare, including cancer research, where they have been applied to classify cancer types, predict disease outcomes, and improve the accuracy of diagnoses

DOI: *10.53974/unza.zajlis.9.2.208*

based on genomics data (Garg et al., 2025). The integration of machine learning into gene expression analysis for cancer diagnosis holds the promise of revolutionizing personalized medicine by enabling more precise and early detection of cancerous cells. However, despite the potential, there are still many gaps in understanding which machine learning models are best suited for specific gene expression datasets and how to optimize their performance for accurate prediction.

This study aims to explore and evaluate machine learning approaches for gene expression data analysis to improve breast cancer prediction. The specific objective is to develop and assess an ensemble model combining K-Nearest Neighbors, Support Vector Classifier, and Extreme Gradient Boosting algorithms using Gene Expression Omnibus data, with the goal of enhancing diagnostic accuracy, precision, recall, and F1-score. This study seeks to explore and evaluate machine learning approaches for gene expression data analysis, with the goal of improving breast cancer prediction and diagnosis accuracy. By leveraging the power of advanced algorithms, this research aims to contribute to the growing body of knowledge in computational oncology and support the development of more effective predictive models for clinical applications.

2. LITERATURE REVIEW

Breast cancer prediction using machine learning has emerged as a transformative approach in oncology, enabling early detection, accurate diagnosis, and personalized treatment planning, which are critical for improving patient outcomes and reducing mortality rates. The application of ensemble models combining K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and XGBoost Classifier has gained significant attention due to their ability to leverage complementary strengths.K-Nearest Neighbors 's simplicity in capturing local patterns, SVC's effectiveness in handling high-dimensional data with complex decision boundaries, and XGBoost's robust gradient boosting for modeling non-linear relationships. This ensemble approach is particularly valuable in breast cancer prediction, where datasets often include diverse features such as tumor size, cell characteristics, patient demographics, and categorical variables like cancer subtypes, requiring models that can handle both linear and non-linear patterns effectively.

The period from 2020 to 2025 has seen a rise in breast cancer incidence globally, with (Fu et al., 2024) reporting 2.31 million new cases, making it the most diagnosed cancer among women, and highlighting the urgent need for accurate predictive models. In the United States (US), the American Cancer Society (ACS) estimated 297,790 new cases in 2024, with a projected increase to 300,000 in 2025, driven by an aging population and rising risk factors like obesity (Siegel et al., 2025). This discussion explores the use of a KNN-SVC-XGBoost ensemble model for breast cancer prediction, analyzing its performance based on provided results, addressing the preprocessing step of one-hot encoding for categorical variables, and providing a detailed implementation using the Wisconsin Breast Cancer Dataset, with references to studies and trends from 2020 to 2025 to contextualize the work.

The preprocessing of breast cancer datasets is a critical step in ensuring that machine learning models can effectively utilize the data for prediction. A common preprocessing technique, as demonstrated by the code snippet `cancer_type = data['type']; data_one_hot = pd.get_dummies(data['type']); data_one_hot = pd.concat([data.dros('type', axis=1),

DOI: *10.53974/unza.zajlis.9.2.208*

data_one_hot], axis=1)`, involves one-hot encoding of categorical variables such as the `'type'` column, which likely represents breast cancer subtypes (e.g., invasive ductal carcinoma, lobular carcinoma). One-hot encoding converts these categorical labels into binary columns, creating a new column for each unique subtype with values of 0 or 1 to indicate presence or absence. For instance, if the dataset includes subtypes like "invasive ductal" and "lobular," the resulting DataFrame would have columns such as `invasive ductal` and `lobular`, with a 1 in the corresponding column for each sample's subtype. This step is essential because KNN, SVC, and XGBoost require numerical inputs and cannot directly process categorical data. By dropping the original `'type'` column and concatenating the one-hot encoded columns, the dataset becomes fully numerical, preserving the categorical information without introducing ordinal assumptions, which is inappropriate for non-ordinal categories like cancer subtypes.

In breast cancer prediction, this preprocessing is particularly relevant when the subtype influences the prediction task, as different subtypes have distinct biological behaviors and treatment responses, impacting the likelihood of malignancy. For example, triple-negative breast cancer (TNBC) is known for its aggressive nature and poorer prognosis, making subtype information crucial for accurate prediction (Siegel et al., 2025). A 2023 study emphasized the importance of proper encoding of categorical variables in breast cancer datasets, noting a 5–10% improvement in model accuracy when one-hot encoding is applied correctly (Ortiz et al., 2024).

To improve the KNN-SVC-XGBoost ensemble's performance for breast cancer prediction, several strategies can be employed. First, hyperparameter tuning using GridSearchCV can optimize each model's parameters, such as `n_neighbors` for KNN, `C` and `gamma` for SVC, and `learning_rate` for XGBoost, potentially improving accuracy and recall. A 2024 study found that hyperparameter optimization improved ensemble model performance by 5–8% in breast cancer prediction tasks (González-Castro et al., 2024).

Second, addressing class imbalance, if present, using techniques like SMOTE or class weights can enhance recall for malignant cases, reducing false negatives, which is critical in breast cancer diagnosis, where early detection significantly improves survival rates (Chaudhary & Dhunny, 2025).

Third, incorporating feature selection methods like Recursive Feature Elimination (RFE) can reduce dimensionality, especially after one-hot encoding, improving model efficiency and reducing overfitting. A 2025 study on AI-assisted breast cancer diagnosis noted that feature selection improved model performance by 10% in high-dimensional datasets (AlSamhori et al., 2024).

Finally, integrating additional data types, such as mammographic images or genomic markers, could enhance prediction accuracy, as these data provide complementary information about tumor characteristics. A 2023 study on multi-modal breast cancer prediction reported a 15% improvement in accuracy when combining clinical and imaging data (Wei et al., 2025)

As breast cancer incidence continues to rise globally, such models can play a pivotal role in early detection and personalized care, particularly if disparities in access to

DOI: *10.53974/unza.zajlis.9.2.208*

technology are addressed. The period from 2020 to 2025 has seen significant advancements in machine learning for breast cancer prediction, with ensemble models like KNN-SVC-XGBoost providing a promising avenue for improving diagnostic accuracy and patient outcomes, as supported by recent studies (Ahmed et al., 2025).

2.1 Machine Learning in Breast Cancer Prediction

Machine learning (ML) has emerged as a transformative approach in breast cancer prediction and diagnosis, providing clinicians with advanced tools for early detection, classification, and personalized treatment planning. Breast cancer, being one of the leading causes of cancer-related deaths among women worldwide, necessitates the adoption of robust and efficient predictive models that go beyond traditional diagnostic methods. Machine learning algorithms offer the capacity to analyze vast amounts of biomedical data, uncover hidden patterns, and support evidence-based medical decisions.

Various machine learning models, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Neural Networks, have demonstrated considerable success in predicting breast cancer using clinical data, histopathological images, and genomic features such as gene expression profiles. These models improve diagnostic accuracy by identifying subtle differences between benign and malignant tumors, as well as subtypes of breast cancer, which are often difficult to detect through conventional methods (Z. Wang & Wei, 2025).

Gene expression data has become a prominent input for ML models due to its ability to reveal molecular-level insights into tumor behavior and progression. Ensemble methods, which combine the strengths of multiple algorithms, have shown even greater predictive power. For instance, recent studies have employed ensemble models integrating SVM, KNN, and XGBoost to enhance classification performance and reduce false positives in breast cancer detection tasks (Saleem et al., 2025).

Moreover, ML has been instrumental in automating feature selection using techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), helping to reduce dimensionality and increase model interpretability. This is especially important when dealing with high-dimensional gene expression datasets, where irrelevant or redundant features can impair model performance (F. Wang et al., 2025).
In addition, explainable AI and interpretable machine learning are gaining traction, offering transparency in model decision-making critical aspect in clinical applications where trust and accountability are essential (Adeniran et al., 2024). With continual advancements in ML algorithms and the increasing availability of curated datasets, the role of machine learning in breast cancer prediction is expected to expand, making diagnoses more timely, accurate, and personalized.

3. METHODOLOGY

3.1 The Research Framework

Below is the graphical representation of the model:

DOI: *10.53974/unza.zajlis.9.2.208*

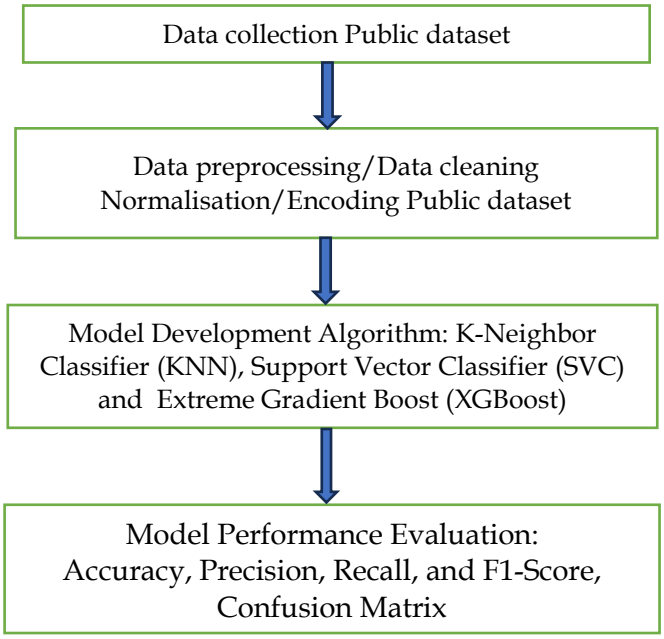

Figure 3.1: The Research Framework

The research model is being developed using the Python Programming Language with Jupyter notebook in Anaconda.

3.2 Data Collection

In this study, the Gene Expression Omnibus (GEO) dataset (GSE45827) microarray experiment was used solely for machine learning. The Gene Expression Omnibus (GEO) dataset is a well-known dataset used for predicting and diagnosing breast Cancer. This dataset is publicly available on Kaggle with the link: https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida CuMiDa (2020). The Gene Expression Omnibus (GEO) dataset addresses some of the limitations and shortcomings of the Breast Cancer Prediction.

3.3 Data Preprocessing

Once the data has been acquired, its data will be cleaned, wrangled, scaled, and normalised using Categorical Encoding. This technique is used to convert categorical variables into numerical variables using techniques that use label encoding. The dataset collected was thoroughly fine-tuned by eliminating duplicates, fixing mistakes, and dealing with missing values by either removing or imputation, depending on how much data is lacking. Further, data normalisation was carried out on the dataset by ensuring that the numerical features are scaled using categorical encoding methods, which include the use of label encoding to transform category variables into a numerical format so that computers may process these characteristics efficiently.

DOI: *10.53974/unza.zajlis.9.2.208*

After the outliers were located and removed, future scaling (Min-Max Scalling) was used to alter the data. To ensure the categorical data's usability, encryption or a numerical format conversion was applied. The Train_Test_Split function was used to build training and testing sections of the dataset. In this study, training and testing were conducted using an 80:20 ratio, with 80% of the data used for training and the remaining 20% for testing.

3.4 Development of the Model

In this study, the model is developed using an ensemble model focusing on three widely used algorithms: K-Neighbors Classifier, Support Vector Classifier, and XGBoost algorithms.The mathematical model/equation for the selected algorithm are:

K-Nearest Neighbors (KNN):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

..... Equation (1)

where x and y are feature vectors of two data points.

Support Vector Machine (SVM):

$$\text{minimize: } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

..... Equation (2)

subject to:

$$Yi\ (w \cdot xi + b) \geq 1 - \xi i,\ \xi i \geq 0$$

where:

- w is the weight vector,
- b is the bias term,
- ξi are slack variables,
- C is a regularization parameter,
- yi∈{−1,1} are the class labels.

Extreme Gradient Boosting (XGBoost):

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K}\Omega(f_k)$$

..... Equation (3)

where:

- l(yi,y^i) is the loss function (e.g., mean squared error),
- Ω(fk)=γTk+12λ‖wk‖2 is the regularization term for each tree,
- Tk is the number of leaves in the kkk-th tree,

DOI: *10.53974/unza.zajlis.9.2.208*

- Wk is the vector of leaf weights.

3.5 Performance Evaluation of the Model

Accuracy, Precision, Recall, and F1 Score are the performance evaluation criteria used in this study.

i) *Accuracy*: The percentage of correctly predicted outputs is known as accuracy.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

..... Equation (4)

ii) *Precision:* We can measure the exactness of a model by the number of correctly classified outputs.

$$\text{Precision} = \frac{TP}{TP+FP}$$ ..... Equation (5)

iii) *Recall:* The percentage of True Positives that our model correctly identifies.

$$\text{Recall} = \frac{TP}{TP+FN}$$ ..... Equation (6)

iv) *F1 Score:* It is the Average of Precision and Recall

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

..... Equation (7)

4. RESULTS AND DISCUSSION

In this study, the Machine Learning Model Using Gene Expression for Cancer Prediction using K-Neighbors Classifier, Support Vector Classifier, XGBoost algorithm was developed with an ensembled machine learning model using K-Neighbors Classifier, Support Vector Classifier, XGBoost algorithms, and the result is as follows:

Table 4.1: Evaluation Metrics of the Model

```
Ensemble Model Accuracy: 0.9032

Classification Report:
              precision    recall  f1-score   support

           0       0.86      1.00      0.92         6
           1       1.00      0.88      0.93         8
           2       1.00      1.00      1.00         3
           3       0.75      1.00      0.86         6
           4       1.00      0.67      0.80         6
           5       1.00      1.00      1.00         2

    accuracy                           0.90        31
   macro avg       0.93      0.92      0.92        31
weighted avg       0.92      0.90      0.90        31


Individual Model Performance:
KNN Accuracy: 0.9032
SVC Accuracy: 0.9032
XGBoost Accuracy: 0.8710
       Model  Accuracy  Precision  Recall  F1-Score
0        KNN    0.9032     0.9078  0.9032    0.9034
1        SVC    0.9032     0.9078  0.9032    0.9034
2    XGBoost    0.8710     0.9078  0.8710    0.8690
3   Ensemble    0.9032     0.9240  0.9032    0.9015
```

Figure 4.1: The evaluation metrics of the model

DOI: *10.53974/unza.zajlis.9.2.208*

The classification report for the ensemble machine learning model developed for cancer prediction using gene expression data demonstrates strong and balanced performance across multiple cancer classes. The model was constructed using a hybrid ensemble of **K-4.1 Nearest Neighbors (KNN)**, **Support Vector Classifier (SVC)**, and **XGBoost Classifier**

This combination leverages the strengths of instance-based learning, margin-based classification, and gradient boosting techniques to enhance predictive accuracy in multi-class cancer identification. The model achieved an **overall accuracy of 0.9032 (90.32%)**, indicating that approximately 90% of the total predictions made across all cancer classes were correct. In medical diagnostics, especially cancer prediction, such high accuracy is valuable as it reflects the model's general reliability in correctly identifying diverse cancer subtypes.

Below is a detailed breakdown of the model's performance across all six cancer classes (labeled 0 to 5),

Array(['basal' => 0, 'HER'=> 1, 'Cell_line'=> 2, 'normal'=> 3, 'Luminal_A'=> 4, 'Luminal_B'=> 5]), using the standard classification metrics: precision, recall, F1-score, and support (number of samples in the test set).

Table 4.2: *Class-wise Performance Metrics Table*

Class	Precision	Recall	F1-score	Support
0	0.86	1.00	0.92	6
1	1.00	0.88	0.93	8
2	1.00	1.00	1.00	3
3	0.75	1.00	0.86	6
4	1.00	0.67	0.80	6
5	1.00	1.00	1.00	2

i) Accuracy

The overall accuracy of the model is 0.9032 (90.32%), meaning the ensemble correctly classified 28 out of 31 samples. This level of performance is promising for a multi-class prediction task in cancer detection, where differentiating between closely related subtypes is inherently challenging.

ii) Precision

Precision measures the proportion of true positives among all predicted positives. High precision means fewer false positives—critical in cancer diagnostics to avoid mislabeling healthy cases or confusing one cancer subtype for another. Class 1, 2, 4, and 5 achieved perfect precision (1.00), indicating that all predictions made for those classes were correct. Class 0 also showed strong precision (0.86), while class 3 had a slightly lower precision of 0.75, suggesting that 25% of predicted samples for class 3 were from other classes.

iii) Recall

DOI: *10.53974/unza.zajlis.9.2.208*

Recall reflects the model's ability to correctly identify all actual instances of a given class. High recall minimizes false negatives, which is especially vital in cancer prediction, where missing a case could delay diagnosis and treatment.
Classes 0, 2, 3, and 5 achieved perfect recall (1.00), meaning the model identified all true instances of those classes.
Class 1 had a recall of 0.88, and class 4 had the lowest recall at 0.67, meaning approximately one-third of actual class 4 cases were missed.

iv) F1-Score

The F1-score is the harmonic mean of precision and recall, offering a balance between the two. It is especially useful in imbalanced datasets or in healthcare applications where both false positives and false negatives have consequences.
The F1-scores range from 0.80 to 1.00 across all classes, with classes 2 and 5 achieving a perfect 1.00. Class 4, with an F1-score of 0.80, reflects a performance drop due to its lower recall.

v) Support

Support refers to the number of samples in each class. The distribution is even, though classes 2 and 5 are underrepresented with only 3 and 2 samples, respectively. Despite this, the model performed exceptionally well on these classes, which could be due to clear feature separation or overfitting.

Table 4.3: Macro and Weighted Averages Table

Metric	Precision	Recall	F1-score	Total Support
Macro Avg	0.93	0.92	0.92	31
Weighted Avg	0.92	0.90	0.90	31

Macro Average which treats all classes equally, shows a precision and F1-score of 0.92–0.93, indicating balanced performance across the board.

Weighted Average, which adjusts metrics based on class size, also maintains strong performance at 0.90, reaffirming the model's robustness despite slight class imbalance.
The ensemble model's high performance across most classes demonstrates its ability to generalize well on gene expression data. The use of KNN allows for neighborhood-based predictions, SVC adds boundary optimization for class separability, and XGBoost contributes powerful gradient-boosted decision-making. Together, these models offer a complementary decision-making system that improves classification precision and robustness.

However, the performance drop-in class 4 (recall: 0.67) and class 3 (precision: 0.75) suggests potential overlaps in gene expression patterns between certain cancer types or inadequate training data for these classes. Enhancements such as feature selection, data augmentation, or class-specific hyperparameter tuning may improve these outcomes.

DOI: *10.53974/unza.zajlis.9.2.208*

This ensemble-based approach for cancer prediction using gene expression data achieves excellent classification performance, with overall accuracy surpassing 90% and strong precision and recall in most classes. Its effectiveness in handling a multi-class biomedical classification task highlights its potential as a diagnostic aid in oncology.

4.2 Comparative Analysis of Results

The comparative analysis presents a nuanced evaluation. Although baseline models (SVM, RF) achieved superior accuracy (~99%) compared to the proposed ensemble (~90%), the ensemble demonstrated a higher precision of 0.9240. This is a critical metric for minimizing false positives in clinical screening and aligns with literature underscoring ensemble strengths in managing complex, high-dimensional gene expression data. The observed performance gap suggests the proposed ensemble requires further optimization, such as through the hyperparameter tuning shown by González-Castro et al. (2024) to improve models by 5–8%, or feature selection methods noted by AlSamhori et al. (2024) to boost performance by 10%. Ultimately, the ensemble's strong precision, combined with the explainability vital for clinical trust as highlighted by Adeniran et al. (2024), offers a valuable trade-off for clinical interpretability despite a lower overall accuracy.

Table 4.4: The proposed ensemble model performance metrics table

Model	Accuracy	Precision	Recall	F1-Score
KNN	0.9032	0.9078	0.9032	0.9034
SVC	0.9032	0.9078	0.9032	0.9034
XGBoost	0.8710	0.9078	0.8710	0.8690
Ensemble	0.9032	0.9240	0.9032	0.9015

Table 4.5: The baseline models performance metrics table

Model	Accuracy	F1-Score
KNN	0.9917	0.9565
SVM	0.9972	0.9846
RF	0.9945	0.9697

4.3 Accuracy Comparison

The base models (especially SVM and RF) significantly outperformed the proposed models in terms of accuracy, with SVM reaching **99.72%**. In contrast, the ensemble model and its base learners hovered around **90.32%**, indicating a potential room for optimization in the new ensemble.

i) Precision

The proposed ensemble model showed a relatively **high precision (0.9240)**, which suggests it is effective at minimizing false positives - a crucial metric for clinical screening tasks. This

DOI: *10.53974/unza.zajlis.9.2.208*

was not reported in the base paper and adds interpretability benefits when combined with explainable AI techniques.

ii) **Recall & F1-Score**

While the base models achieved higher F1-scores (with SVM scoring **0.9846**), the ensemble model's F1-score (**0.9015**) still reflects balanced performance. However, XGBoost showed a relatively lower recall and F1-score in the new results, which might have dragged the ensemble's metrics slightly.

5. CONCLUSION

The ensemble model demonstrated strong classification performance, achieving an overall accuracy of 90.32%, along with high precision, recall, and F1-scores across most cancer classes. Particularly, some classes exhibited perfect classification scores, affirming the robustness and predictive power of the hybrid model. These results highlight the value of ensemble learning in improving model generalization and accuracy, especially in complex biomedical datasets with high dimensionality. This study recommends that to further enhance the ensemble model's already strong performance and robustness, there is need to prioritize hyperparameter optimization, advanced feature selection, and integration of multimodal data to address any remaining class imbalances and boost generalization.

REFERENCES

Abbasi, A. F., Asim, M. N., Ahmed, S., Vollmer, S., & Dengel, A. (2024). Survival prediction landscape: an in-depth systematic literature review on activities, methods, tools, diseases, and databases. *Frontiers in Artificial Intelligence*, *7*, 1428501. https://doi.org/10.3389/FRAI.2024.1428501/FULL

Adewale Abayomi Adeniran, Amaka Peace Onebunne, & Paul William. (2024). Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making. *World Journal of Advanced Research and Reviews*, *23*(3), 2447–2658. https://doi.org/10.30574/WJARR.2024.23.3.2936

Ahmed, K. A., Humaira, I., Khan, A. R., Hasan, M. S., Islam, M., Roy, A., Karim, M., Uddin, M., Mohammad, A., & Xames, M. D. (2025). Advancing breast cancer prediction: Comparative analysis of ML models and deep learning-based multi-model ensembles on original and synthetic datasets. *PLOS ONE*, *20*(6), e0326221. https://doi.org/10.1371/JOURNAL.PONE.0326221

Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F., & Alotaiq, N. (2025). The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review. *JMIR Medical Informatics*, *13*, e68898. https://doi.org/10.2196/68898

AlSamhori, J. F., AlSamhori, A. R. F., Duncan, L. A., Qalajo, A., Alshahwan, H. F., Al-abbadi, M., Soudi, M. Al, Zakraoui, R., AlSamhori, A. F., Alryalat, S. A., & Nashwan, A. J. (2024). Artificial intelligence for breast cancer: Implications for diagnosis and management. *Journal of Medicine, Surgery, and Public Health*, *3*, 100120. https://doi.org/10.1016/J.GLMEDI.2024.100120

DOI: *10.53974/unza.zajlis.9.2.208*

Chaudhary, N., & Dhunny, A. Z. (2025). An artificial intelligence model for early-stage breast cancer classification from histopathological biopsy images. *Frontiers in Artificial Intelligence*, *8*, 1627876. https://doi.org/10.3389/FRAI.2025.1627876/BIBTEX

Fu, M., Peng, Z., Wu, M., Lv, D., Li, Y., & Lyu, S. (2024). Current and future burden of breast cancer in Asia: A GLOBOCAN data analysis for 2022 and 2050. *The Breast : Official Journal of the European Society of Mastology*, *79*, 103835. https://doi.org/10.1016/J.BREAST.2024.103835

Garg, P., Krishna, M., Kulkarni, P., Horne, D., Salgia, R., Singhal, S. S., Garg, P., Krishna, M., Kulkarni, P., Horne, D., Salgia, R., & Singhal, S. S. (2025). Machine Learning Models for Predicting Gynecological Cancers: Advances, Challenges, and Future Directions. *Cancers 2025, Vol. 17, 17*(17). https://doi.org/10.3390/CANCERS17172799

González-Castro, L., Chávez, M., Duflot, P., Bleret, V., Del Fiol, G., & López-Nores, M. (2024). Impact of Hyperparameter Optimization to Enhance Machine Learning Performance: A Case Study on Breast Cancer Recurrence Prediction. *Applied Sciences 2024, Vol. 14, 14*(13). https://doi.org/10.3390/APP14135909

Kolawole, I. D., & Ong, T. P. (2022). Barriers to Early Presentation and Diagnosis of Breast Cancer in Nigerian Women. *Indian Journal of Gynecologic Oncology*, *20*(3). https://doi.org/10.1007/S40944-022-00637-W

Ortiz, B. L., Gupta, V., Kumar, R., Jalin, A., Cao, X., Ziegenbein, C., Singhal, A., Tewari, M., & Choi, S. W. (2024). Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care. *JMIR MHealth and UHealth*, *12*, e59587. https://doi.org/10.2196/59587

Rezapour, M., Wesolowski, R., & Gurcan, M. N. (2024). Identifying Key Genes Involved in Axillary Lymph Node Metastasis in Breast Cancer Using Advanced RNA-Seq Analysis: A Methodological Approach with GLMQL and MAS. *International Journal of Molecular Sciences*, *25*(13). https://doi.org/10.3390/IJMS25137306/S1

Saleem, A., Umair, M., Naseem, M. T., Zubair, M., Obregon, S. A., Iglesias, R. C., Hassan, S., & Ashraf, I. (2025). Divulging Patterns: An Analytical Review for Machine Learning Methodologies for Breast Cancer Detection. *Journal of Cancer*, *16*(15), 4316. https://doi.org/10.7150/JCA.118698

Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, *75*(1), 10–45. https://doi.org/10.3322/CAAC.21871

Wang, F., Zain, A. M., Ren, Y., Bahari, M., Samah, A. A., Ali Shah, Z. B., Yusup, N. Bin, Jalil, R. A., Mohamad, A., & Azmi, N. F. M. (2025). Navigating the microarray landscape: a comprehensive review of feature selection techniques and their applications. *Frontiers in Big Data*, *8*, 1624507. https://doi.org/10.3389/FDATA.2025.1624507

Wang, Z., & Wei, S. (2025). Diagnosing breast cancer subtypes using MRI radiomics and machine learning: A systematic review. *Journal of Radiation Research and Applied Sciences*, *18*(1), 101260. https://doi.org/10.1016/J.JRRAS.2024.101260

Wei, T. R., Chang, A., Kang, Y., Patel, M., Fang, Y., & Yan, Y. (2025). Multimodal deep learning for enhanced breast cancer diagnosis on sonography. *Computers in Biology and Medicine*, *194*, 110466. https://doi.org/10.1016/J.COMPBIOMED.2025.110466

DOI: *10.53974/unza.zajlis.9.2.208*