IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5900~5911
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5900-5911

Feature selection and class balancing for breast cancer classification using ribonucleic acid sequencing data

David Thuku Gana, Consolata Gakii Mwirichia, Faith Mueni Musyoka
Department of Computing and Information Technology, University of Embu, Embu, Kenya

Article Info

Article history:

Received Apr 19, 2025
Revised Aug 21, 2026
Accepted Sep 20, 2026

Keywords:

Breast cancer prediction
Classification
Imbalanced data
Machine learning
RNA-seq dataset
Sampling techniques
Synthetic minority over-sampling technique

ABSTRACT

Breast cancer remains a major health challenge and continues to create a need for dependable methods for disease prediction and diagnosis. Ribonucleic acid sequencing (RNA-seq) data provide gene expression profiles that can be used to predict cancerous and non-cancerous conditions, although the existing studies using RNA-seq datasets often face two challenges: high dimensionality and data imbalance issues. This study incorporated and systematically evaluated the impact of feature selection approaches (least absolute shrinkage and selection operator (LASSO) and principal component analysis (PCA)) and class balancing methods (synthetic minority over-sampling technique (SMOTE), random oversampling (ROS), and random undersampling (RUS)) on breast cancer prediction using an RNA-seq dataset of 103 samples (56 cancerous cases and 47 non-cancerous cases), resulting in a class ratio of 1.19:1 with 24,833 features. Four algorithms were applied and evaluated using metrics both with and without feature selection and class balancing techniques. According to the experimental findings, integrating LASSO with ROS produced the strongest performance for both logistic regression (LR) and support vector machine (SVM) models, resulting in the highest accuracy rate of 94%. Subsequently, the random forest (RF) and k-nearest neighbor (KNN) algorithms achieved an accuracy of 91%. In contrast, the combination of PCA with RUS demonstrated poor performance.

This is an open access article under the CC BY-SA license.

Corresponding Author:

David Thuku Gana
Department of Computing and Information Technology, University of Embu
P.O. Box 6-60100, Embu, Kenya
Email: ganadavidthuku@gmail.com

1. INTRODUCTION

Breast cancer affects both sexes, although women are more susceptible than men [1], [2]. Early intervention and treatment improve survival rates. However, owing to the complexity of breast cancer, personalized methods are required [3]. Conventional methods, such as mammography and biopsy, are important for identifying breast cancer. However, machine learning (ML) is an emerging trend that offers additional capabilities for analyzing large-scale data, like ribonucleic acid sequencing (RNA-seq) datasets. RNA-seq reveals the patterns of gene expression in cancer cells, allowing researchers to identify differentially expressed genes and potential biomarkers associated with breast cancer [4]. RNA-seq datasets face two major problems: a high number of gene features and differences in class distribution. However, the involvement of feature selection and class balancing techniques in breast cancer experiments improves both the interpretability and accuracy of the results [5].

High dimensionality refers to a situation in which the number of features is more than the number of samples. Feature selection techniques, including principal component analysis (PCA) and least absolute

shrinkage and selection operator (LASSO), are utilized to extract the most informative features and prevent overfitting [6]. PCA is an essential tool that helps reduce the large dimensions of RNA-seq datasets, retains useful information, and lowers computational complexity [7]. LASSO is a regularization tool that applies a penalty to irrelevant features by shrinking their coefficients to zero, resulting in a simpler and more interpretable model [8]. Another feature selection approach, the select k best method, selects the top "k" features using statistical tests, including chi-squared, analysis of variance, and mutual information, to enhance model efficiency and predictive accuracy [9].

Class imbalance occurs in a dataset when certain classes of data are significantly underrepresented or overrepresented, causing a model to favor the class with many instances, thereby reducing the prediction accuracy for the class with few instances [10]. Some techniques commonly used to resolve this problem include the synthetic minority over-sampling technique (SMOTE), which creates synthetic instances for the minority class; random oversampling (ROS), which adds cases of the minority class by duplicating existing cases; and the random undersampling (RUS) method, which reduces the instances of the class with more instances by randomly removing some samples [11].

ML algorithms such as support vector machine (SVM), logistic regression (LR), k-nearest neighbors (KNN), and random forest (RF) have been applied in health, agriculture, finance, fraud detection, and sentiment analysis to address challenges [12]–[16]. SVM performs classification by identifying a decision boundary that separates observations belonging to different classes within the feature space [17]. LR estimates the probability that an observation belongs to a particular category based on its predictor variables and is therefore suitable for binary disease classification [18]. RF uses an ensemble of decision trees (DTs) and combines their predictions to produce a final classification, making it useful for complex datasets containing many variables [19]. The KNN algorithm performs classification by comparing different data points to make a prediction based on the class with the most instances of its closest k neighbors, where k is a positive integer in the feature space, which makes it effective for problems with well-defined clusters [20]. This study aimed to incorporate and systematically evaluate different feature selection and class balancing techniques for breast cancer prediction using an RNA-seq dataset.

2. RELATED WORK

Akben [21] contributed to early breast cancer diagnosis using DTs to analyze blood test and body measurement data from the Coimbra University Hospital dataset. This approach involves creating value ranges that indicate the presence of cancer. Techniques such as the Gini index and entropy have identified important data attributes. The DT model exhibited an accuracy of 90.52%. The study also suggested that some algorithms, such as DT models, can improve early detection and decision-making. Botlagunta *et al.* [22] developed a non-invasive ML model for breast cancer classification using clinical data. Python was used within an Anaconda Jupyter Notebook environment to implement and evaluate several algorithms. Model performance was assessed using accuracy, area under the curve (AUC), and receiver operating characteristic (ROC) measures, while a Welch unpaired *t*-test was applied to investigate statistical significance. The evaluated algorithms included LR, KNN, DTs, RFs, SVMs, gradient boosting, and extreme gradient boosting (XGBoost). The decision-tree model recorded the highest accuracy and AUC, at 83% and 87%, respectively. This study did not use feature selection and class balancing together or evaluate their performance.

Binsaif [23] focused on creating a model to identify and detect breast cancer using a blood test and anthropometric dataset from the University of California, Irvine (UCI) dataset. The study compared various algorithms, such as DT, RF, KNN, artificial neural network (ANN), SVM, and LR models. Ten-fold cross-validation and an 80:20 training-testing division were used for model validation. Additional statistical analyses involved *p*-values, Pearson correlation, and the Kolmogorov-Smirnov test. RF produced the best results, reporting 83.3% accuracy, 100% sensitivity, 64% specificity, and an AUC of 0.881. Alfian *et al.* [24] investigated breast cancer prediction using SVM together with extra-trees-based feature selection. Data obtained from the gynecology department of the Coimbra University Hospital were used in the study. The extra-trees classifier was first used to remove less relevant variables before SVM classification was performed. Stratified 10-fold validation was used to assess the models. The combined approach achieved an accuracy of 80.23%, while feature selection improved prediction accuracy by as much as 7.29% compared with models that did not apply feature-selection methods.

Wang *et al.* [25] applied the approaches of ML to identify genes that show different expression patterns between original human breast tumors and patient-derived xenograft models. The study analyzed gene expression data from 831 breast cancer samples and used Monte Carlo feature selection and RF to select the most important genes. A total of 32 key genes were identified and used to build a prediction model with good performance. This study also identified biologically meaningful interactions between genes. However, the class imbalance between human and xenograft samples was not explicitly addressed, leaving room for further improvements. Yuan *et al.* [26] investigated the use of models for cancer detection based on gene

expression data. To identify the most informative genes, the authors employed a combination of Monte Carlo feature selection and incremental feature selection. Subsequently, SVM and RF classifiers were trained and evaluated. The experimental results demonstrated that the SVM model outperformed the RF model, achieving an accuracy of 100%. Overall, SVM showed superior performance compared to RF in this study. The study did not explain how they solved the imbalances or whether they evaluated feature selection and class balancing techniques. Wang *et al.* [27] investigated the prediction of breast cancer grades using RNA-seq data. The model was trained on data from 275 women and tested on 487 patients in The Cancer Genome Atlas (TCGA) database. The important genes were selected using elastic net regression and gene expression analysis. The model achieved an AUC rate of 0.97 for distinguishing between grades 1 and 3 tumors. It also helped classify difficult grade 2 tumors into low- and high-risk groups, respectively. Many studies applied class balancing and feature selection independently, without evaluating their effectiveness to determine the best combination of these methods. This study addresses this gap by incorporating and systematically evaluating the impact of feature selection and class balancing techniques using an RNA-seq dataset for breast cancer prediction.

3. METHOD

The experimental procedure consisted of several major stages: obtaining the dataset, preprocessing the observations, applying feature-selection techniques, dividing the data into training and testing sets, balancing the classes, training ML models, and evaluating their performance. The complete workflow is presented in Figure 1. The experiments of this study were performed using Python and the scikit-learn library. Scikit-learn is a widely used open-source library that offers several algorithms and tools for ML tasks. To support the implementation of the algorithms, other Python libraries, such as NumPy and SciPy, were used for the numerical computations and scientific processing. The main steps followed during the experiment and the overall workflow of this study are shown in Figure 1.

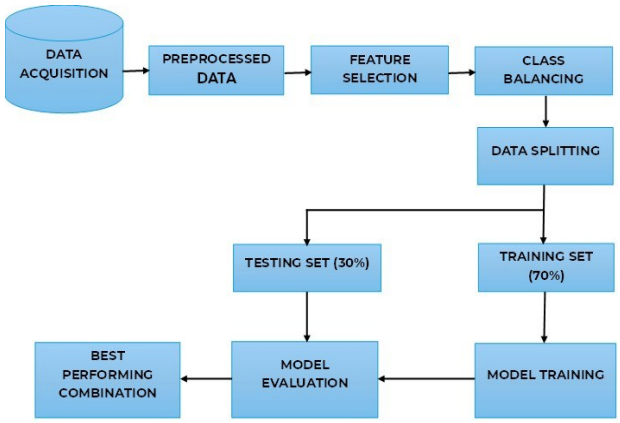

Figure 1. Breast cancer prediction model

3.1. Data acquisition

The dataset that was investigated was the GSE100925 breast cancer dataset obtained from the National Center for Biotechnology Information [28]. It contained 103 biological samples, including 56 cancerous and 47 non-cancerous samples. The dataset consisted of 28,088 gene expression features based on RNA-seq data.

3.2. Data preprocessing

Pandas was used to prepare the dataset before model development. The gene expression matrix was initially transposed to align the samples in the rows and features in the columns. The metadata were combined with the gene expression features dataset to include the target variable (cancerous and non-cancerous), then encoded into numerical values, assigning 1 to cancerous samples and 0 to non-cancerous samples. Columns containing only zeros were removed to reduce uninformative features. Subsequently, the dataset was checked for missing values using functions such as isna() and isnull(). No missing values were identified, indicating that the dataset was complete and appropriate for the analysis. Standardization was then performed using the StandardScaler function from scikit-learn. This process scaled each gene feature to have a mean of 0 and a variance of 1, ensuring that all features contributed equally

during model training. The final dataset contained 103 samples and 24,833 gene expression features. No other techniques were used in this study.

3.3. Feature selection

Feature selection methods were introduced to reduce the large number of gene expression variables and retain informative information for classification. PCA was utilized to reduce dimensionality by converting the original correlated gene characteristics into a set of a lower number of uncorrelated principal components while retaining the highest amount of variance in the data [7]. PCA was set to select five principal components (n_components = 5). LASSO was applied to shrink less important coefficients toward zero, retaining only the relevant genes. The combination of feature selection and regularization (L1 penalty) prevents overfitting while improving the model interpretability [8]. The regularization parameter (alpha) of LASSO was fixed to 0.01 and up to 1,000 iterations to ensure appropriate convergence when selecting features.

3.4. Class balancing

This study quantitatively assessed the distribution of classes, indicating the proportion between the majority and minority classes. This was achieved by computing the percentage of samples in each class, which indicated the proportion between the majority and the minority. Of the 103 samples, 56 were cancerous, and 47 were non-cancerous. This resulted in a class imbalance ratio of 1.19:1, meaning that the cancerous category contained slightly more observations than the non-cancerous category. Three balancing strategies were evaluated. SMOTE (sampling_strategy = 'auto' and random_state = 42) was applied to generate new artificial samples for the minority class to balance the dataset. ROS (sampling_strategy = 'auto', random_state = None, and shrinkage = None) involved duplicating the existing minority sample class to increase its representation, whereas RUS (sampling_strategy = 'auto', random_state = None, and shrinkage = None) worked by reducing the majority sample class to achieve a better balance using Pandas. These techniques were chosen because they are effective, widely used, and help improve the classifier's performance by preventing bias toward the majority class.

3.5. Data splitting

The dataset was split into two parts, namely training and testing subsets, using the train_test_split function provided by scikit-learn, with 70% assigned to the training set and the remaining 30% to the testing set. A random seed of 42 was used to ensure consistency and reproducibility of the experiments. The splitting was performed after applying the class-balancing techniques.

3.6. Model training

Ten-fold cross-validation was used to create robust models. Using this method, the training data were split into ten equal subsets. The model was trained on nine subsets in each round, and the remaining subset was used for validation. This was performed ten times, and each data point was included in the validation set. Cross-validation minimized the chances of overfitting and provided a more reliable model performance. Several ML classifiers were implemented using scikit-learn. Each algorithm was fitted using a training set and examined using a testing set to measure its performance. The goal was to select a model that provided the best results across all the employed performance metrics. The hyperparameters were set based on standard recommendations without exhaustive tuning. The ML algorithms used include the following:

LR was set to run for a maximum of 1,000 iterations, with a regularization parameter of 1.0, using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) solver, and a random state of 42. It relies on log loss to estimate the prediction error and L2 regularization to prevent overfitting. The SVM used a linear kernel, regularization of 1.0, and a random state of 42. It measures margin errors using hinge loss and prevents overfitting through regularization. The RF consisted of 100 trees with unrestricted tree depth, a minimum of two samples required to split a node, and a random state of 42. It was configured with Gini impurity to split nodes, and five neighbors were set in KNN with the Minkowski distance of $p = 2$.

3.7. Model evaluation and best model

The system was evaluated using several metrics, calculated using the Python library scikit-learn for both subsets to examine the generalization ability of the model. The different models were subsequently compared, and the strongest-performing configuration was identified. Additionally, to provide further insight into classifier performance, ROC and precision-recall curves, together with confusion matrices, were generated. Metrics used include as follows.

3.7.1. Accuracy

This measures the proportion of observations made correctly by the model, out of the total number of observations. Accuracy was calculated using (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where TP and TN represent correct positive and negative predictions, respectively, whereas FP and FN denote incorrect positive and negative classifications, respectively.

3.7.2. Recall

This describes the effectiveness of a classifier in locating all the positive cases. Recall was calculated using (2).

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

3.7.3. Precision

This shows how many positive observations are correct out actually positive predicted. Precision was calculated using (3).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

3.7.4. F1-score

These combined measures of precision and recall into a single measure by calculating the harmonic mean. F1-score was calculated using (4).

$$F1 - score = \frac{2*(Precision*Recall)}{Precision+Recall} \quad (4)$$

4. RESULTS AND DISCUSSION

This section reports the study outcomes under various experimental configurations and places the findings in the context of related research. It shows how the different ML models performed when combined with various feature selection and class balancing techniques. The performance of the models was measured using accuracy, precision, recall, and F1-score. The results were compared to identify which combination performed best. A statistical test was also conducted to determine whether the differences in performance were significant.

4.1. Without feature selection and class balancing techniques

Figure 2 illustrates the baseline performance of the evaluated classifiers before feature reduction and class-balancing techniques were introduced. The models were compared using accuracy, precision, recall, and F1-score. RF achieved the highest overall performance, recording 81% for each metric. KNN, however, obtained the lowest results, with an accuracy of 52% and an F1-score of 46%, indicating that it performed less effectively than the other classifiers under the baseline configuration.

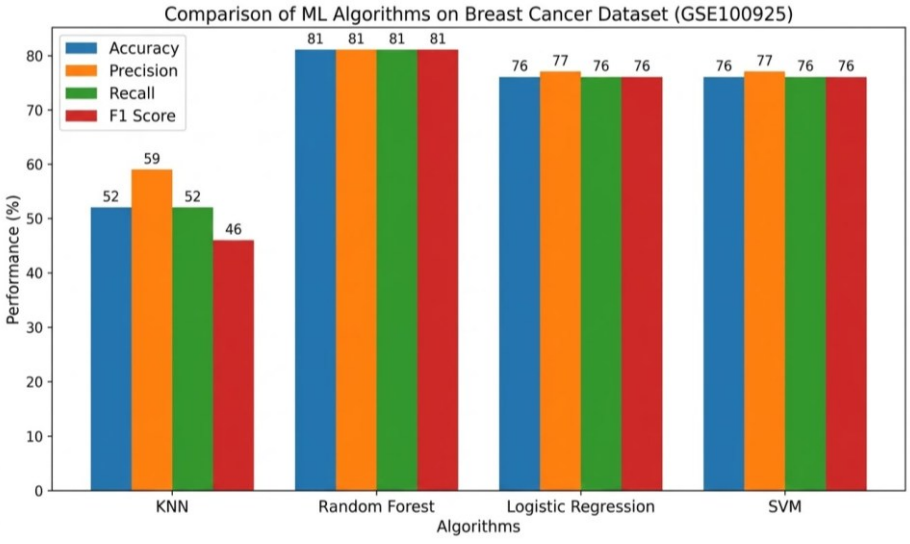

Figure 2. Comparison of algorithm performance under baseline conditions

4.2. With feature selection and class balancing techniques

Different feature selection and class balancing methods were tested, including PCA with SMOTE, LASSO with SMOTE, PCA with ROS, LASSO with ROS, PCA with RUS, and LASSO with RUS. The models were compared according to their classification performance across accuracy, precision, recall, and F1-score. The experiments were conducted to determine how each combination affected the performance of the ML models. The results helped identify which combination produced the best classification performance. The findings are presented in tables for easier comparison.

4.2.1. Logistic regression performance

Table 1 shows that a combination of LASSO and ROS with LR achieved a score of 94% across accuracy, precision, recall, and F1-score, indicating its effectiveness in enhancing the model's performance. PCA with SMOTE, ROS, and RUS ranges a rate of approximately 79%, showing no improvement with all class balancing techniques. Table 2 shows that across all models involving LR, LASSO with ROS performed significantly better than all other LR combinations ($p < 0.05$).

Table 1. Performance of LR

Feature selection	Class balancing	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PCA	SMOTE	79	80	79	79
LASSO	SMOTE	85	86	85	85
PCA	ROS	79	80	79	79
LASSO	ROS	94	94	94	94
PCA	RUS	79	79	79	79
LASSO	RUS	83	83	83	83

Table 2. Wilcoxon statistical analysis for LR with LASSO with ROS

Comparative analysis	p-value	Significant ($\alpha = 0.05$)
LASSO + ROS vs PCA + SMOTE	0.00195	Yes
LASSO + ROS vs LASSO + SMOTE	0.00195	Yes
LASSO + ROS vs PCA + ROS	0.00195	Yes
LASSO + ROS vs PCA + RUS	0.00195	Yes
LASSO + ROS vs LASSO + RUS	0.00195	Yes

4.2.2. Support vector machine performance

In Table 3, SVM with LASSO and ROS techniques achieved the highest performance, achieving a consistent 94% across accuracy, precision, recall, and F1-score, while other combinations yielded moderate percentage rates of accuracy, precision, recall, and F1-score ranging between 76%-82%. LASSO with RUS showed poor performance, with consistent results of 45% across accuracy, precision, recall, and F1-score, indicating that the combination was ineffective. The PCA models showed a moderate performance rate. Table 4 shows that across all models that involved SVM, LASSO with ROS was compared to all other feature selection and class balancing combinations, and it performed significantly better than the others.

Table 3. SVM performance results

Feature selection	Class balancing	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PCA	SMOTE	82	83	82	82
LASSO	SMOTE	82	82	82	82
PCA	ROS	76	81	76	76
LASSO	ROS	94	94	94	94
PCA	RUS	76	78	76	74
LASSO	RUS	45	45	45	45

Table 4. Wilcoxon statistical analysis for SVM with LASSO with ROS

Comparative analysis	p-value	Significant ($\alpha = 0.05$)
LASSO + ROS vs PCA + SMOTE	0.00195	Yes
LASSO + ROS vs LASSO + SMOTE	0.00195	Yes
LASSO + ROS vs PCA + ROS	0.00195	Yes
LASSO + ROS vs PCA + RUS	0.00195	Yes
LASSO + ROS vs LASSO + RUS	0.00195	Yes

4.2.3. Random forest performance

Table 5 summarizes the evaluation metrics for the RF. The highest accuracy rate of 91% was observed when using LASSO with ROS, suggesting that this combination enhances the RF's ability to handle class imbalances. LASSO with SMOTE also performed very well with scores between 79% and 80%. The lowest scores were obtained using PCA with SMOTE and PCA with RUS. Overall, LASSO was better than PCA, and ROS was more effective as a balancing technique than SMOTE. Table 6 shows that across all models that involved RF, LASSO with ROS consistently outperformed all other RF combinations.

Table 5. RF model performance measures

Feature selection	Class balancing	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PCA	SMOTE	71	71	71	70
LASSO	SMOTE	79	80	79	79
PCA	ROS	74	74	74	74
LASSO	ROS	91	91	91	91
PCA	RUS	66	67	66	66
LASSO	RUS	76	78	76	76

Table 6. Wilcoxon statistical analysis for RF with LASSO with ROS

Comparative analysis	p-value	Significant (α = 0.05)
LASSO + ROS vs PCA + SMOTE	0.00195	Yes
LASSO + ROS vs LASSO + SMOTE	0.00195	Yes
LASSO + ROS vs PCA + ROS	0.00195	Yes
LASSO + ROS vs PCA + RUS	0.00195	Yes
LASSO + ROS vs LASSO + RUS	0.00195	Yes

4.2.4. K-nearest neighbor performance

Table 7 indicates that the model was most effective with LASSO to select features and ROS to balance the data at 91% across all measures. PCA with SMOTE and ROS achieved average figures of approximately 70-76%, whereas LASSO with RUS achieved a better score of 79%. LASSO with SMOTE recorded the worst result (68%). In general, the best prediction was achieved using LASSO with ROS. Table 8 shows that LASSO with ROS outperformed all other KNN models.

Table 7. Performance of KNN

Feature selection	Class balancing	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
PCA	SMOTE	76	81	76	76
LASSO	SMOTE	68	69	68	68
PCA	ROS	74	77	74	73
LASSO	ROS	91	91	91	91
PCA	RUS	72	72	72	72
LASSO	RUS	79	79	79	79

Table 8. Wilcoxon statistical analysis for KNN with LASSO with ROS

Comparative analysis	p-value	Significant (α = 0.05)
LASSO + ROS vs PCA + SMOTE	0.001953	Yes
LASSO + ROS vs LASSO + SMOTE	0.001953	Yes
LASSO + ROS vs PCA + ROS	0.001953	Yes
LASSO + ROS vs PCA + RUS	0.001953	Yes
LASSO + ROS vs LASSO + RUS	0.001953	Yes

4.2.5. Best-performing combination

Optimal results were achieved by combining the LASSO feature selection with ROS, applied to LR and SVM models. Both models recorded the highest classification performance compared to the other combinations tested. This shows that LASSO with ROS effectively improved the models' ability to classify breast cancer samples. The results also showed that the choice of feature selection and class balancing techniques influenced the overall performance of the models.

i) Logistic regression

The model performed very well using the top genes selected by LASSO and data balanced using ROS. Figure 3 shows that the ROC curve has a high AUC (0.94), indicating that the model

effectively differentiates between cancer and non-cancer cases. The precision-recall plot in Figure 4 reveals that it maintains high precision at most recall values and false positives, achieving most of the true cancer cases.

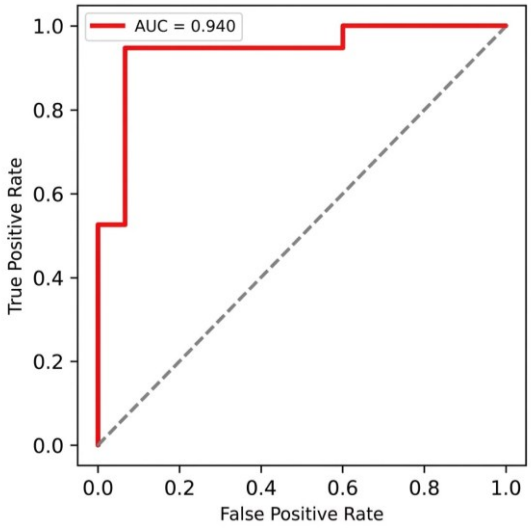

Figure 3. LR ROC curve

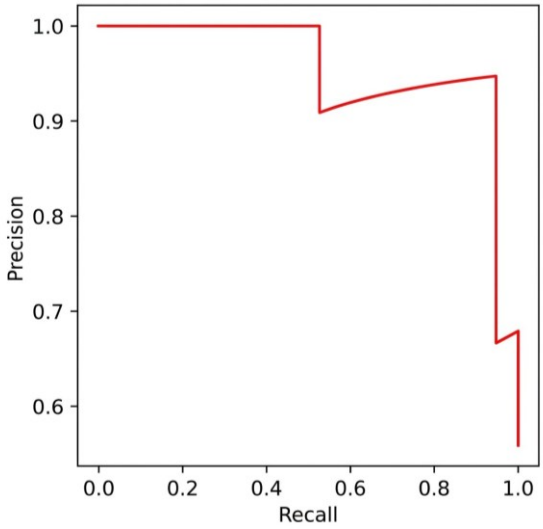

Figure 4. LR precision–recall plot

The confusion matrix in Figure 5 shows that 14 samples from class 0 and 18 samples from class 1 were correctly classified, with only two misclassifications. This indicates that LASSO and ROS significantly enhanced the model's precision and trustworthiness. The results also show that the model performed well in identifying samples from both classes.

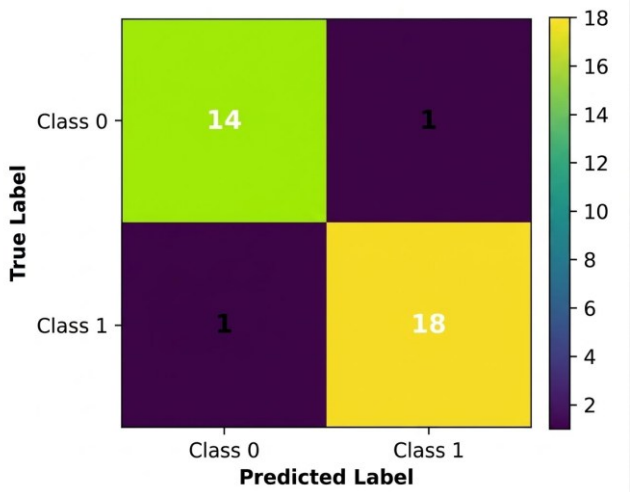

Figure 5. LR confusion matrix

ii) Support vector machine

The model performed well using LASSO and ROS. Figure 6 indicates that the ROC curve has a large AUC of 0.94, whereas Figure 7 indicates that it has high precision at most of the recall values. The confusion matrix, as shown in Figure 8 indicates that 14 samples from class 0 and 18 samples from class 1 were correctly predicted, with only two misclassifications in total. This proves that SVM and LR were accurate and reliable.

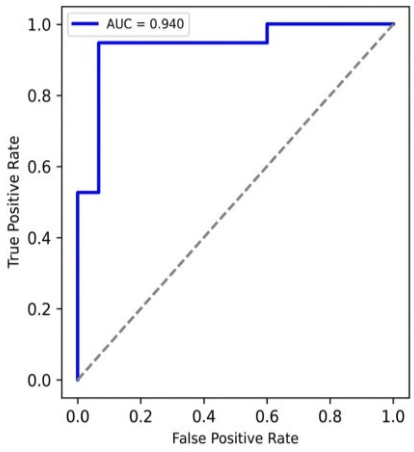

Figure 6. SVM ROC curve

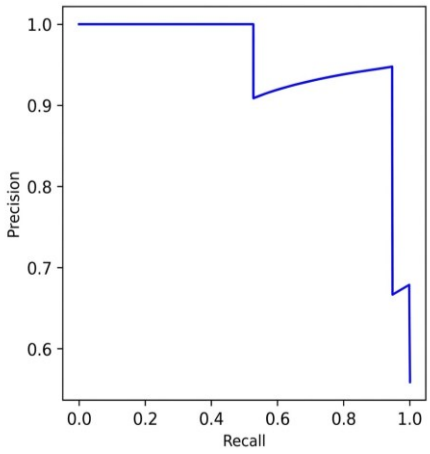

Figure 7. SVM precision–recall curve

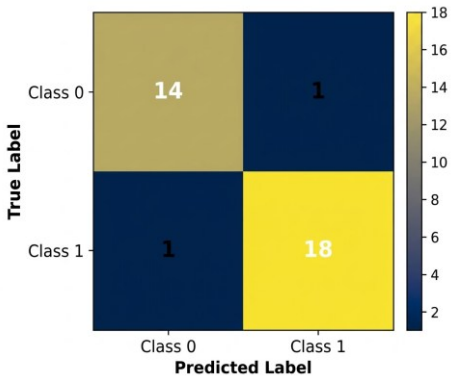

Figure 8. SVM confusion matrix

iii) Biological significance of selected genes

The top genes identified by LASSO play important roles in breast cancer. ATPase Na^+/K^+-transporting subunit beta-2 (ATP1B2) (gene 1) helps regulate the movement of substances in and out of cells, and its dysfunction can lead to abnormal cell growth. Small Cajal body-specific RNA 9 (SCARNA9) (gene 2) regulates RNA to ensure that genes function properly, and changes in this gene promote cancer

growth. Small nucleolar RNA, C/D Box 19 (SNORD19) (gene 3) is involved in protein synthesis and modification, and abnormal activity can disrupt normal cell functions. LOC10537544 (gene 4) regulates the expression of other genes and may be associated with cancer-related changes. Claudin 10 (CLDN10) (gene 5) maintains cell adhesion; in the case of a lack of its function, cancer cells may move freely. These genes demonstrate that the characteristics chosen by the model are not only helpful for making predictions but also for comprehending the way breast cancer develops.

4.2.6. Comparison with prior studies

Several studies have utilized RNA-seq data to predict breast cancer; however, they have done so in different ways and have encountered several limitations. An instance where this was done is in the case of Paulino and Sufyan [29], who compiled data from various breast cancer studies to form a dataset of more than 60,000 cells. They applied t-distributed stochastic neighbor embedding (t-SNE), UMAP, and RF methods to determine six major breast cancer subtypes, which included luminal, HER2-enriched, and basal-like. The most important genes formed the RF model and attained more than 90% accuracy in tumor sample classification. Similarly, Wu and Hicks [30] concentrated on the classification of triple-negative and non-triple-negative breast cancer. They chose some of the important genes and the models to be tested, including SVM, KNN, naive Bayes, and DT. The results indicate that the SVM achieved optimal accuracy of 90%. However, multiple studies did not use class balancing and feature selection techniques together or evaluate feature selection or class balancing techniques systematically. In contrast, this study incorporated feature selection and class balancing approaches and evaluated their effectiveness to determine the best combination. LASSO feature selection with ROS combined with LR and SVM models recorded a 94% accuracy. This demonstrates the benefit of applying feature selection and class balancing in breast cancer prediction

5. CONCLUSION

This experiment analyzed the impact of feature selection and class balancing techniques on breast cancer prediction using an RNA-seq dataset. LASSO with ROS achieved the best performance when applied to LR and SVMs, yielding 94% in all evaluation measures. These results indicate that combining feature selection and class balancing methods can considerably enhance diagnostic accuracy. However, this study was conducted on a single, relatively small dataset with no external validation. Hyperparameter optimization was not explored in depth, and the computational cost and scalability of the methods should also be considered. Future research should consider larger datasets, independent validation, training-only class balancing, hybrid methods, and deep learning techniques to improve breast cancer identification and support clinical decision-making.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the University of Embu (UoEm) for providing the resources and institutional support required to undertake this study.

FUNDING INFORMATION

The authors confirm that this research received no financial support and that no funding was provided for the preparation of this article.

AUTHOR CONTRIBUTIONS STATEMENT

The authors' contributions are reported using the Contributor Roles Taxonomy (CRediT), in accordance with the journal's authorship and contribution requirements.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
David Thuku Gana	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓
Consolata Gakii Mwirichia	✓	✓		✓			✓	✓		✓		✓	✓	
Faith Mueni Musyoka	✓	✓		✓			✓	✓		✓	✓	✓		

C: **C**onceptualization
M: **M**ethodology
So: **So**ftware
Va: **Va**lidation
Fo: **Fo**rmal analysis
I: **I**nvestigation
R: **R**esources
D: **D**ata Curation
O: Writing - **O**riginal Draft
E: Writing - Review & **E**diting
Vi: **Vi**sualization
Su: **Su**pervision
P: **P**roject administration
Fu: **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

INFORMED CONSENT

This study used existing data obtained from the NCBI Gene Expression Omnibus (GEO). The researchers did not recruit any participants themselves.

DATA AVAILABILITY

The dataset utilized in this study was accessed from the NCBI Gene Expression Omnibus (GEO) under the accession number GSE100925, reference number [28].

REFERENCES

[1] E. Heer, A. Harper, N. Escandor, H. Sung, V. McCormack, and M. M. F.-Benaoudia, "Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study," *The Lancet Global Health*, vol. 8, no. 8, pp. e1027–e1037, Aug. 2020, doi: 10.1016/S2214-109X(20)30215-1.

[2] L. Wilkinson and T. Gathani, "Understanding breast cancer as a global health concern," *The British Journal of Radiology*, vol. 95, no. 1130, Feb. 2022, doi: 10.1259/bjr.20211033.

[3] L. Wang *et al.*, "Breast cancer stem cells: signaling pathways, cellular interactions, and therapeutic implications," *Cancers*, vol. 14, no. 13, Jul. 2022, doi: 10.3390/cancers14133287.

[4] M. Hong *et al.*, "RNA sequencing: new technologies and applications in cancer research," *Journal of Hematology and Oncology*, vol. 13, no. 1, Dec. 2020, doi: 10.1186/s13045-020-01005-x.

[5] M. O. Arowolo, M. O. Adebiyi, C. Aremu, and A. A. Adebiyi, "A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00441-x.

[6] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: a review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/j.jksuci.2019.06.012.

[7] L. H. Nguyen and S. Holmes, "Ten quick tips for effective dimensionality reduction," *PLOS Computational Biology*, vol. 15, no. 6, Jun. 2019, doi: 10.1371/journal.pcbi.1006907.

[8] F. E.-Streib and M. Dehmer, "High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 359–383, Jan. 2019, doi: 10.3390/make1010021.

[9] N. R. A.-Althaqafi and H. A. Alsalamah, "The effect of feature selection on the accuracy of X-platform user credibility detection with supervised machine learning," *Electronics*, vol. 13, no. 1, Jan. 2024, doi: 10.3390/electronics13010205.

[10] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced big data," *Journal of Big Data*, vol. 10, no. 1, Apr. 2023, doi: 10.1186/s40537-023-00724-5.

[11] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *Journal of Biomedical Informatics*, vol. 90, Feb. 2019, doi: 10.1016/j.jbi.2018.12.003.

[12] P. Doupe, J. Faghmous, and S. Basu, "Machine learning for health services researchers," *Value in Health*, vol. 22, no. 7, pp. 808–815, Jul. 2019, doi: 10.1016/j.jval.2019.02.012.

[13] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: a review," *Sensors*, vol. 18, no. 8, Aug. 2018, doi: 10.3390/s18082674.

[14] N. Nazareth and Y. V. R. Reddy, "Financial applications of machine learning: a literature review," *Expert Systems with Applications*, vol. 219, Jun. 2023, doi: 10.1016/j.eswa.2023.119640.

[15] L. H. Aros, L. X. B. Molano, F. G.-Portela, J. J. M. Hernandez, and M. S. R. Barrero, "Financial fraud detection through the application of machine learning techniques: a literature review," *Humanities and Social Sciences Communications*, vol. 11, no. 1, Sep. 2024, doi: 10.1057/s41599-024-03606-0.

[16] M. Subramanian, V. E. Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan, "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models," *Alexandria Engineering Journal*, vol. 80, pp. 110–121, Oct. 2023, doi: 10.1016/j.aej.2023.08.038.

[17] D. Valkenborg, A.-J. Rousseau, M. Geubbelmans, and T. Burzykowski, "Support vector machines," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 164, no. 5, pp. 754–757, Nov. 2023, doi: 10.1016/j.ajodo.2023.08.003.

[18] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic regression model optimization and case analysis," in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Oct. 2019, pp. 135–139, doi: 10.1109/ICCSNT47585.2019.8962457.

[19] H. A. Salman, A. Kalakech, and A. Steiti, "Random forest algorithm overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[20] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: a review," *Journal of Data Analysis and Information Processing*, vol. 8, no. 4, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[21] S. B. Akben, "Determination of the blood, hormone and obesity value ranges that indicate the breast cancer, using data mining based expert system," *IRBM*, vol. 40, no. 6, pp. 355–360, Dec. 2019, doi: 10.1016/j.irbm.2019.05.007.

[22] M. Botlagunta *et al.*, "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Scientific Reports*, vol. 13, no. 1, Jan. 2023, doi: 10.1038/s41598-023-27548-w.

[23] N. Binsaif, "Application of machine learning models to the detection of breast cancer," *Mobile Information Systems*, vol. 2022, pp. 1–8, Mar. 2022, doi: 10.1155/2022/7340689.

[24] G. Alfian *et al.*, "Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method," *Computers*, vol. 11, no. 9, Sep. 2022, doi: 10.3390/computers11090136.

[25] D. Wang, J.-R. Li, Y.-H. Zhang, L. Chen, T. Huang, and Y.-D. Cai, "Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms," *Genes*, vol. 9, no. 3, Mar. 2018, doi: 10.3390/genes9030155.

[26] F. Yuan, L. Lu, and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1866, no. 8, Aug. 2020, doi: 10.1016/j.bbadis.2020.165822.
[27] M. Wang, D. Klevebring, J. Lindberg, K. Czene, H. Grönberg, and M. Rantalainen, "Determining breast cancer histological grade from RNA-sequencing data," *Breast Cancer Research*, vol. 18, no. 1, Dec. 2016, doi: 10.1186/s13058-016-0710-8.
[28] National Center for Biotechnology Information, "GEO DataSets," ncbi.nlm.nih.gov. Accessed: Jan. 15, 2026. [Online]. Available: https://www.ncbi.nlm.nih.gov/gds/
[29] P. J. I. V. Paulino and M. Sufyan, "Identifying molecular subtypes of breast cancer using single cell RNA-seq data integration and random forest classification," *International Journal of Bioinformatics Research and Applications*, vol. 20, no. 5, pp. 468–494, 2024, doi: 10.1504/IJBRA.2024.141751.
[30] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *Journal of Personalized Medicine*, vol. 11, no. 2, Jan. 2021, doi: 10.3390/jpm11020061.

BIOGRAPHIES OF AUTHORS



David Thuku Gana holds a Bachelor of Science degree in Information Technology from the University of Embu, Kenya, awarded in 2021. He is currently pursuing a Master of Science degree in Information Technology at the same institution. His research interests lie in the areas of machine learning and bioinformatics, with a strong focus on leveraging computational approaches to analyze biomedical data for disease prediction and improved healthcare outcomes. In the current study, he was responsible for data preprocessing, feature selection, model evaluation, and interpretation of results. His academic path is characterized by a deep interest in innovative problem-solving and collaboration across interdisciplinary domains. He is passionate about advancing research that integrates technology with healthcare for meaningful impact. He can be contacted at email: ganadavidthuku@gmail.com.



Consolata Gakii Mwirichia holds a Ph.D., a Master of Science, and a Bachelor of Science in Information Technology from Jomo Kenyatta University of Agriculture and Technology, Kenya. She is a lecturer in the Department of Computing and Information Technology at the University of Embu, Embu, Kenya. She is a specialist in machine learning and big data analytics, with a strong passion for using data-driven technologies to solve real-world challenges. Her current research focuses on the application of artificial intelligence in health-related data, particularly within the One Health framework. She has contributed to several interdisciplinary projects that use AI to address public health issues. Notably, she is involved in initiatives across Embu and Murang'a counties aimed at raising awareness of HIV/AIDS through AI-driven systems. Her scholarly contributions include research on data modelling, predictive analytics, and the development of intelligent systems for health interventions. In the current paper, she contributed to the conceptualization, data curation, analysis, and critical review of the manuscript. She remains committed to using advanced computing methods to promote public health and support evidence-based decision-making. She can be contacted at email: gakii.consolata@embuni.ac.ke or at gakiiconsolata@gmail.com.



Faith Mueni Musyoka holds a Ph.D. in Information Technology from Kabarak University, Kenya, a Master of Science in Information Technology, and a Bachelor of Science in Computer Science from Masinde Muliro University of Science and Technology, Kenya. She is currently a lecturer in the Department of Computing and Information Technology at the University of Embu. Her research interests include the application of internet of things (IoT) technologies in healthcare, software quality, software metrics, digital mental health technologies, and AI-driven health interventions. In addition, she has led multiple research and innovation projects focused on digital solutions in healthcare and education. Her work, recognized nationally, has contributed to advancements in early detection, monitoring, and support systems for youth mental health. She also serves in institutional initiatives promoting mental wellness among students and staff. In the current paper, she contributed to the conceptualization, formal analysis, data interpretation, and critical review of the manuscript. She is passionate about advancing smart healthcare solutions and continues to explore innovative ways to bridge technology and health for societal benefit. She can be contacted at email: mueni.faith@embuni.ac.ke.