Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/190517

Transcriptomic Pattern Analysis in Breast Cancer Patients: A Machine Learning Approach

Natália Padre [1], *Monique Borges Seixas* [2], *Paulo Victor dos Santos* [3], *Glaucia Maria Bressan* [1], *Heron Oliveira dos Santos Lima* [2] *and Marcella Scoczynski*[1,*]

[1] Programa de Pós-Graduação em Bioinformática,Universidade Tecnológica Federal do Paraná, Cornélio Procópio, PR, Brazil
[2] Universidade Tecnológica Federal do Paraná- UTFPR Brazil
[3] Hospital Albert Einstein, SP, Brazil

Received: 22 Mar. 2025, Revised: 22 Jul. 2025, Accepted: 12 Aug. 2025
Published online: 1 Sep. 2025

Abstract: Breast cancer remains one of the leading causes of mortality among women and is one of the most prevalent cancers worldwide. Advancing accurate diagnosis, treatment, and prevention requires a deeper understanding of the genetic alterations involved in tumorigenesis. Integrating genomic and transcriptomic data offers a powerful approach to uncover the molecular mechanisms underlying the disease. Transcriptomic profiling, which involves sequencing RNA to analyze gene expression, enables the identification of biomarkers for disease progression and supports the discovery of novel therapeutic targets. However, transcriptomic studies often include a mix of cancerous and non-cancerous cells, requiring robust analytical methods to distinguish between them and ensure meaningful interpretation. In this study, we propose a comprehensive pipeline for transcriptomic analysis using gene expression data from The Cancer Genome Atlas (TCGA). The dataset is pre-processed and normalized using the TCGAbiolinks package within the R software environment. Machine learning algorithms are employed to classify samples as tumor or normal tissue. Seven models are evaluated, with Random Forest and Radial Basis Function Support Vector Machine (RBF SVM) achieving the highest performance. RBF SVM reached an accuracy of 99.55%, precision of 99.55%, recall of 99.64%, and F1-score of 99.64%, while Random Forest obtained an accuracy of 99.38%, precision of 99.38%, recall of 99.50%, and F1-score of 99.50%. Stratified 5-fold cross-validation confirmed the models' robustness, showing low variance across folds. Feature selection is performed to enhance interpretability, and five key genes were identified: ENSG00000152256.14 (PDK1), ENSG00000155875.15 (SAXO1), ENSG00000165194.15 (PCDH19), ENSG00000176884.16 (GRIN1), and ENSG00000180910.8 (TTTY11). These genes are further investigated using Ensembl for biological interpretation highlighting PDK1, involved in cancer metabolism; SAXO1, linked to cytoskeletal stability; PCDH19, associated with cell adhesion; GRIN1, related to glutamate signaling; and TTTY11, a pseudogene with potential regulatory roles. This study highlights the potential of machine learning in transcriptomic data analysis and offers a framework for identifying key biomarkers, contributing to precision oncology in breast cancer research.

Keywords: Supervised classification, transcriptomic analysis, predictive models, gene expression, breast cancer.

1 Introduction

Cancer remains one of the most feared and impactful diseases worldwide, primarily due to its high mortality rate and substantial burden on public health. According to the Brazilian National Cancer Institute (INCA), approximately 704,000 new cancer cases are projected in Brazil between 2023 and 2025. Globally, the Global Cancer Observatory (GLOBOCAN) reported 19.3 million new cases in 2020 alone, with breast cancer being the most prevalent—accounting for over 2.3 million diagnoses in the past four years—underscoring its critical global health significance[6].

The process of *carcinogenesis*, wherein normal cells acquire malignant characteristics, results from a multifactorial interplay of internal and external factors. Genetic mutations, hormonal imbalances, and immune dysregulation, combined with environmental exposures to carcinogens, contribute to the onset of *neoplasia*—the abnormal, uncontrolled proliferation of cells. In breast tissue, this can give rise to tumors with diverse

* Corresponding author e-mail: marcella@utfpr.edu.br

morphological and molecular profiles, complicating both diagnosis and treatment [2].

Technological advances in genomics and transcriptomics have transformed our understanding of cancer biology [**?**]. In particular, transcriptomic profiling, which sequences RNA to analyze gene expression, enables the identification of biomarkers linked to disease progression and therapy response [**?**]. Integrative analyses of cancerous and non-cancerous transcriptomic data reveal insights into tumor heterogeneity and interactions with the microenvironment [7].

To centralize and standardize molecular cancer data, the U.S. National Institutes of Health launched The Cancer Genome Atlas (TCGA), a comprehensive resource comprising clinical and multi-omics data from over 33 cancer types [8]. However, its volume and complexity demand specialized tools for processing. To address this, the R/Bioconductor package TCGAbiolinks was developed, facilitating efficient querying, downloading, preprocessing, and analysis of TCGA data [5].

In parallel, machine learning has emerged as a transformative approach in cancer research, enabling automated discovery of complex patterns across large-scale datasets. Techniques such as supervised classification, clustering, and neural networks are now widely applied for patient stratification, tumor subtype identification, and image-based diagnosis [3].

Building on these developments, the present study proposes an automated pipeline using TCGAbiolinks for breast cancer transcriptomic data analysis. By applying advanced machine learning models, including deep learning, we aim to classify tumor and normal samples, identify key biomarkers, and support personalized cancer diagnosis and treatment strategies.

2 Methodology and Experiments

The methodology proposed in this study aims to perform a comprehensive transcriptomic analysis using gene expression data from The Cancer Genome Atlas (TCGA), evaluated through various machine learning classifiers. The pipeline begins with data acquisition using the TCGAbiolinks package in the R programming environment. The analysis includes breast cancer samples, comprising primary tumors, metastatic tumors, and non-cancerous solid tissue samples.

After data retrieval, the workflow applies preprocessing steps to ensure compatibility with the machine learning pipeline. These steps include data transposition, dataset merging, and normalization to reduce technical variability. In the model training phase, seven supervised learning algorithms classify the samples into tumor and non-tumor categories. A cross-validation strategy evaluates the robustness of the models. Finally, the *SelectFromModel* method performs feature selection to identify genes most relevant to classification. The overall process is illustrated in Figure 1.

2.1 Preprocessing

This study uses breast cancer gene expression data from TCGA, including three groups: primary tumor, metastatic tumor, and non-cancerous solid tissue. The datasets contain expression profiles of 60,661 genes across 1,110 primary tumor samples, 87 metastatic cases, and 7 normal samples.

The *TCGAbiolinks* package facilitates data querying, downloading, and preprocessing through the following functions:

- ***GDCquery()*** – searches for relevant transcriptomic datasets.
- ***GDCdownload()*** – downloads the selected data.
- ***GDCprepare()*** – preprocesses and normalizes the data using FPKM-UQ (Fragments Per Kilobase Million Upper Quartile) normalization [4].

After retrieval, the workflow transposes the data so that rows represent individuals and columns represent genes. We then merge the datasets into a single matrix. Due to the limited number of metastatic cases—and their clinical relevance—these samples are grouped with primary tumor data for normalization purposes.

Normalization scales gene expression values to the $[0, 1]$ range using min-max normalization:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where X' is the normalized value, X is the original gene expression, and $X_{\min}$, $X_{\max}$ are the minimum and maximum observed values [9].

Labels are assigned as follows: non-cancerous tissue is labeled as 0, and tumor samples (primary or metastatic) are labeled as 1.

2.2 Model Training

For classification, this study evaluates seven machine learning models implemented using the *scikit-learn (sklearn)* library in Python [1]:

- Decision Tree
- Gaussian Process Classifier
- Neural Network
- Logistic Regression
- Random Forest
- K-Nearest Neighbors (KNN)
- Radial Basis Function Support Vector Machine (RBF SVM)

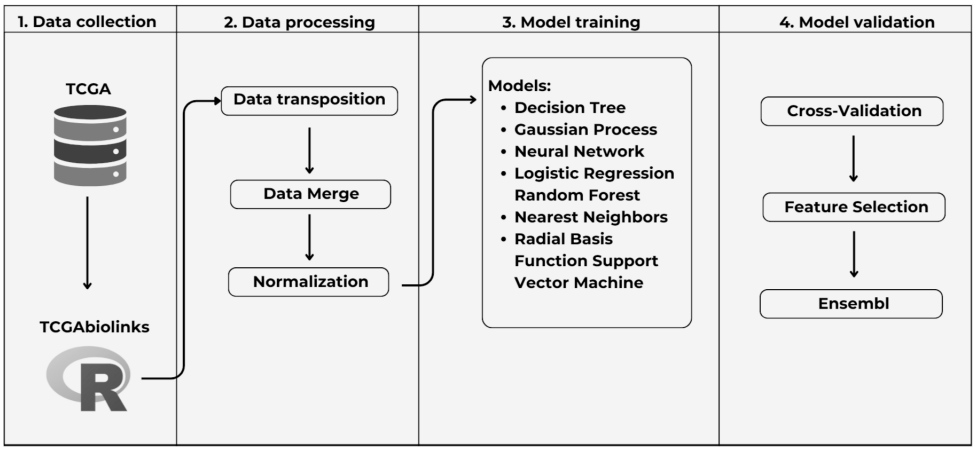

Fig. 1: Methodology flowchart

Each model undergoes cross-validation to assess generalization and avoid overfitting. The analysis computes standard performance metrics—accuracy, precision, recall, and F1-score—and averages them across all validation folds. Finally, the *SelectFromModel* method identifies the most relevant genes contributing to classification, refining the models' interpretability and potential biological significance.

2.2.1 Decision Tree

The decision tree model operates by recursively partitioning the feature space into smaller regions, making decisions at each node based on feature values. In *sklearn*, the Gini index is used as the criterion for splitting the attribute space. The Gini index is defined as:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{2}$$

where p_i represents the proportion of samples belonging to class i. While decision trees are intuitive and interpretable, deep trees are prone to overfitting. To mitigate this, key hyperparameters such as the maximum depth and the minimum number of samples per node were carefully tuned[10].

The hyperparameters for the decision tree model are configured as follows:

–Split criterion: Gini index.
–Split strategy: Best split at each node.
–Minimum samples to split an internal node: 2.
–Minimum samples in a leaf node: 1.

2.2.2 Gaussian Process (GP)

The Gaussian Process classifier considers the model output (or predicted value) as a function that follows a Gaussian distribution. For each input point, the model output is treated as a random variable that follows a normal distribution. The parameter considered in *sklearn* for this model is the RBF *kernel* (1.0), which is used to define the covariance between different input data points, specifying the outputs and how these points are correlated.

$$K(x_i, x_j) = k(x_i, x_j) \tag{3}$$

where x_i and x_j are samples, and σ is a scale hyperparameter[11].

2.2.3 Neural Network

For the neural network classifier, we consider a *Multilayer Perceptron* (MLP) architecture, defined by multiple hidden layers and ReLU (*Rectified Linear Unit*) activation[12]:

$$f(x) = \max(0, x) \tag{4}$$

The training is performed using backpropagation and the Adam optimizer, minimizing the Cross-Entropy loss function for binary classification:

$$\mathscr{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{5}$$

The hyperparameters defined for this model are as follows:

–Penalty control: *L2* ($\alpha = 1$);
–Maximum number of iterations: 1000.

2.2.4 Logistic Regression

The Logistic Regression classification algorithm is based on two fundamental components: the logistic (sigmoid)

function and the logit. This model estimates the probability of a binary outcome—in this context, the presence or absence of a tumor—by mapping predicted values to a probability range between 0 and 1. The algorithm applies the sigmoid function to a linear combination of input features, producing a probability estimate that facilitates binary classification. The sigmoid function is defined as:

$$P(Y=1|X)=\frac{1}{1+e^{-z}} \tag{6}$$

where z represents the linear predictor (i.e., the weighted sum of input features),

$$z=\beta_0+\sum_i \beta_i X_i$$

. This output can then be interpreted as the likelihood of the sample belonging to the positive class. The Logistic Regression has a maximum number of iterations of 1000.

2.2.5 Random Forest

Random Forest is a machine learning algorithm based on an ensemble of decision trees. It combines the predictions of multiple trees to produce a more robust and accurate final decision. The model constructs N decision trees, each trained on a random subset of the dataset. At each node, a random subset of features is selected to determine the best split, allowing each tree to make an independent prediction. The final prediction is obtained through a weighted average of the individual tree outputs:

$$\hat{y}=\sum_{i=1}^{n} w_i h_i(x) \tag{7}$$

Random Forest also facilitates feature selection by identifying the genes with the greatest influence on classification [17]. The model improves generalization by mitigating the effects of *overfitting* that individual decision trees typically exhibit. The hyperparameters used in this study are:

- `max_depth` = 5
- `n_estimators` = 10
- `max_features` = 1

2.2.6 Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm classifies samples based on their similarity to other samples in the dataset. It assigns the class that is most common among the K nearest neighbors. This model is particularly useful for identifying samples with similar gene expression profiles and grouping them accordingly, aiding in the detection of tumor subtype variations. The Euclidean distance metric is used to measure similarity:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{8}$$

Euclidean Distance, with x and y representing feature vectors. To measure the distance between the unknown sample and all known samples, the model was configured with $k=3$. where x and y represent the feature vectors of two samples. In this study, the model is configured with $k=3$.

2.2.7 Radial Basis Function Support Vector Machine (SVM RBF)

The Support Vector Machine (SVM) aims to find the optimal hyperplane that separates data points of different classes with the maximum margin. When the data is not linearly separable, the model applies a kernel function—such as the Radial Basis Function (RBF)—to project the input data into a higher-dimensional space, where a linear separator may exist. This allows SVM to learn complex, nonlinear relationships in gene expression that are indicative of cancer. The RBF kernel is defined as:

$$K(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{9}$$

$||x_i-x_j||^2$ being recognized as a squared Euclidean distance between the two feature vectors.σ is a free parameter. An equivalent definition involves a parameter $\gamma=\frac{1}{2\sigma^2}$.

The final expression being equivalent to:

$$K(x_i,x_j)=\exp(-\gamma||x_i-x_j||^2)$$

The parameter γ controls the influence of individual training samples [16].

This study evaluates two variations of SVM:

- **Linear SVM**: Configured with $C=0.025$ to control the penalty for misclassification and reduce overfitting.
- **RBF Kernel SVM**: Configured with $\gamma=2$ and $C=1$, allowing greater flexibility for capturing nonlinear gene expression patterns.

2.2.8 Model Evaluation and Validation

Model performance is assessed using the `train_test_split` function from the `scikit-learn (sklearn)` library, combined with cross-validation techniques. Cross-validation partitions the dataset into training and testing subsets multiple times to provide a reliable estimate of model generalization. In this study, we use stratified k-fold cross-validation with $k=5$ and allocate 25% of the data for testing [13].

The final accuracy is computed as the average across all k folds:

$$\text{Acc}_{\text{Average}} = \frac{1}{k}\sum_{i=1}^{k}\text{Acc}_i, \quad k=5 \tag{10}$$

To balance precision and recall, we also calculate the F1-score using the Equation:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

2.3 Feature Selection

To identify the most relevant features for classification, we apply two methods from the *scikit-learn* library: `SelectFromModel` and `SelectKBest`.

The `SelectFromModel` method selects features based on importance scores derived from a pre-trained estimator. By setting `prefit=True`, the selector uses an already fitted model to rank and retain only the most informative features. The `transform`$(X_{\texttt{train}})$ function then reduces the dimensionality of the training data by removing less relevant columns, streamlining the input for subsequent modeling.

Complementarily, the `SelectKBest` method identifies the top k features with the highest statistical association with the target variable, based on a univariate scoring function. In this study, we use the chi-square (χ^2) test, which evaluates the degree of dependence between each feature and the output class. This test is particularly suitable for categorical or non-negative continuous data and supports effective dimensionality reduction while preserving the most predictive variables [18].

2.4 Gene Identification Tool

To annotate and interpret the selected genes, we use the *Ensembl* genome browser—a comprehensive platform for automated genome annotation and integration of genomic data. *Ensembl* currently supports over 50,000 genomes, including vertebrates and a wide range of other organisms, and continues to expand in scope to meet evolving research needs.

Ensembl facilitates detailed gene exploration, with human gene identifiers beginning with "ENSG" followed by a unique numerical sequence. For example, the identifier `ENSG00000152256.14` corresponds to a specific human gene. Researchers can input this identifier into the Ensembl search tool to access a dedicated gene page containing information on its genomic location, known variants, transcript isoforms, and expression profiles. This resource enables thorough functional characterization of genes of interest and supports biological interpretation of machine learning results [15].

3 Results and discussion

Several classifiers were evaluated for the analysis of the breast cancer dataset, including decision trees, Gaussian processes, neural networks, logistic regression, random forests, k-nearest neighbors (KNN), and radial basis function (RBF) support vector machines (SVM). The results showed that the random forest and RBF SVM models achieved the highest performance, with accuracies exceeding 98%. In contrast, the neural network and Gaussian process models exhibited the lowest performance, each reaching an accuracy of 92.20%. The decision tree, logistic regression, and KNN classifiers demonstrated intermediate performance, with accuracies ranging between 96% and 98%. These results provided a general assessment of model accuracy when the dataset was split into training and testing sets.

To further evaluate model performance, the confusion matrix was analyzed to offer a detailed breakdown of the classification outcomes. The dataset contained 1,231 records and a high-dimensional feature space comprising 60,661 gene expression features. Among the records, 87 were labeled as class zero (normal tissue), and 1,117 were labeled as class one (tumor tissue). A total of 923 records were used for training, while 308 were set aside for validation.

The dataset showed a significant class imbalance, with the tumor class being heavily overrepresented in comparison to the normal class. This imbalance had the potential to skew predictions in favor of the majority class, thereby reducing sensitivity to minority class cases. Table 1 presents the confusion matrices for the main models evaluated, illustrating their performance in classifying tumor versus normal samples.

The confusion matrix is a crucial tool for evaluating classifier performance, as it provides a detailed breakdown of correct and incorrect predictions. The Random Forest and RBF SVM models demonstrated the best performance, with minimal classification errors (false positives and false negatives). The Random Forest model correctly identified 282 tumor cases and 23 normal cases, making only three misclassifications (1 false positive and 2 false negatives). This reflects its high precision and recall, making it an efficient predictor for both classes.

The Decision Tree model exhibited slightly lower performance, misclassifying eight tumor cases as normal (false negatives) and two normal cases as tumors (false positives). Although its accuracy remained high, the increased false negative rate is concerning in medical applications, where precise tumor detection is critical.

The KNN and Logistic Regression models delivered intermediate performance, with some tumor misclassifications (5 false negatives in KNN and 7 in Logistic Regression), indicating slightly lower sensitivity in detecting positive cases. The Gaussian Process and Neural network models showed distinct misclassification patterns. The Gaussian Process model produced 24 false positives and no false negatives, indicating that it

Table 1: Model Confusion Matrix

Classifier	Normal (Predicted)	Tumor (Predicted)	Total
Nearest Neighbors	23	1	24
	5	279	284
RBF SVM	22	2	24
	3	281	284
Gaussian Process	0	24	24
	0	284	284
Decision Tree	22	2	24
	8	276	284
Logistic Regression	19	5	24
	7	277	284
Neural Net	0	24	24
	0	284	284
Random Forest	23	1	24
	2	282	284

classified nearly all samples as tumors. This led to very low specificity, reflecting a poor ability to correctly identify normal cases. Similarly, the K-Nearest Neighbors (KNN) classifier entirely failed to detect normal samples, misclassifying all instances as tumors, which also resulted in zero specificity.

To ensure a robust evaluation of model performance, a *5-fold cross-validation* technique was employed. This method partitioned the dataset into five equal subsets while preserving class proportions. In each iteration, four subsets were used for training and one for testing, allowing every sample to contribute to both training and validation. After all five folds were completed, performance metrics—including accuracy, precision, recall, and F1-score—were averaged to obtain overall performance estimates.

In terms of average accuracy and model stability, the best-performing models were the RBF SVM (0.9955 ± 0.0048) and the Random Forest (0.9938 ± 0.0080), both of which achieved high accuracy with low variance across the folds. The Decision Tree, Logistic Regression, and KNN classifiers also demonstrated good performance, albeit with slightly higher variance. In contrast, the Gaussian Process and Neural Network models yielded lower average accuracy (0.9082 ± 0.0019), indicating weaker generalization.

Regarding average recall—the ability to correctly identify tumor cases—the RBF SVM (0.9964) and Random Forest (0.9950) models again outperformed others, achieving near-perfect recall. Although the Gaussian Process and Neural Network models also attained high recall values (0.9518), this was attributed to their bias toward predicting all samples as tumors, which came at the cost of specificity.

The average precision—the proportion of true tumor predictions among all tumor classifications—was highest for RBF SVM (0.9955) and Random Forest (0.9938), indicating minimal false positives. Conversely, the Gaussian Process and Neural Network models had lower precision, reflecting their tendency to misclassify normal cases as tumors.

When considering the average F1-score, which harmonizes recall and precision into a single metric, Random Forest (0.9950) and RBF SVM emerged as the top performers, striking an effective balance between sensitivity and specificity. The Gaussian Process and Neural Network models exhibited lower F1-scores due to their imbalanced performance.

The Decision Tree, Logistic Regression, and KNN models also performed reasonably well but fell slightly behind the leading models. Moreover, their relatively higher standard deviations suggested greater variability depending on the specific train-test split, which may indicate reduced stability in real-world applications.

Table 2 summarizes the key performance metrics across all models.

Based on the results, the Random Forest and RBF SVM models emerged as the most suitable classifiers for the breast cancer dataset, demonstrating superior generalization capabilities and robustness in handling class imbalance.

To further investigate and refine these models, a feature selection step was carried out to improve classification efficiency and reduce dimensionality. In this phase, we analyzed the Decision Tree, Random Forest, and RBF SVM models, each employing a distinct machine learning strategy to identify the most relevant variables contributing to classification performance.

First, we applied a Decision Tree-based feature selection method with a maximum depth of five. This approach significantly reduced the number of features from 60,661 to just five, highlighting the most discriminative genes in the dataset. The selected genes were `ENSG00000152256.14`, `ENSG00000155875.15`, `ENSG00000165194.15`, `ENSG00000176884.16`, and `ENSG00000180910.8`, as illustrated in Figure 2.

Table 2: Model Performance

Classifier	Average Accuracy (± Deviation)	Average Recall	Precision Average	F1-Score Average
Nearest Neighbors	0.9937 ± 0.0066	0.9928	0.9938	0.9928
RBF SVM	0.9955 ± 0.0048	0.9964	0.9955	0.9964
Gaussian Process	0.9082 ± 0.0019	0.9518	0.9082	0.9518
Decision Tree	0.9883 ± 0.0086	0.9847	0.9883	0.9847
Logistic Regression	0.9910 ± 0.0027	0.9892	0.9910	0.9892
Neural Net	0.9082 ± 0.0019	0.9518	0.9082	0.9518
Random Forest	0.9938 ± 0.0080	0.9950	0.9938	0.9950

Subsequently, we employed a Random Forest classifier with 100 estimators for feature selection. This method retained a larger subset of 927 features, suggesting that the Random Forest approach captured a broader range of potentially informative genes. The increased number of selected features may be attributed to the model's ensemble nature and its sensitivity to the underlying class imbalance, which can affect the weighting and ranking of features with high expression variance.

Additionally, we applied an SVM classifier with a linear kernel for feature selection, which retained 16,432 variables. This high number suggests that the SVM may be capturing complex relationships between the predictor and response variables. A potential improvement for the analysis would be to reevaluate the classification models using the feature subsets identified by the different selection methods to assess their impact on model performance.

We applied the SelectKBest statistical method using the chi-square test to identify the 20 most relevant features, as shown in Figure 3, which illustrates the correlation between variables. These results underscore the substantial variability in feature retention across different selection techniques, reinforcing the need for additional statistical analyses to determine the true impact of these features on classifier performance.

To validate whether the genes selected by the decision tree model, which significantly reduced the number of analyzed genes, are genuinely associated with breast cancer, we utilized the Ensembl tool to cross-reference the selected genes with existing literature. As a result, Table 3 was developed, linking each identified gene to its corresponding description and classification, enabling a more comprehensive biological analysis.

The five genes selected through the decision tree model offer promising insight into breast cancer biology. Among them:

- **PDK1 (ENSG00000152256.14)** – Pyruvate Dehydrogenase Kinase 1 plays a critical role in metabolic reprogramming of cancer cells, inhibiting pyruvate entry into the TCA cycle and promoting the Warburg effect, a hallmark of cancer metabolism [19].
- **SAXO1 (ENSG00000155875.15)** – While primarily associated with microtubule stabilization in neuronal cells, emerging evidence suggests that SAXO1 dysregulation may affect cytoskeletal dynamics, impacting cancer cell motility and invasion [20].
- **PCDH19 (ENSG00000165194.15)** – This protocadherin is involved in cell-cell adhesion and signaling. Loss or mutation of cadherin genes, including PCDH19, has been linked to breast and ovarian cancer invasiveness [21].
- **GRIN1 (ENSG00000176884.16)** – Encodes a subunit of the NMDA glutamate receptor. GRIN1 is implicated in calcium signaling and neural plasticity but also contributes to tumor proliferation and resistance in several cancers, including breast cancer [22].
- **TTTY11 (ENSG00000180910.8)** – Although listed as a pseudogene on the Y chromosome, recent studies show that pseudogenes like TTTY11 may act as competitive endogenous RNAs (ceRNAs), modulating gene networks and influencing tumor progression [23].

These genes offer potential as biomarkers for diagnostic or therapeutic targeting. These findings highlight the importance of feature selection in optimizing predictive models and emphasize the need for further investigations to confirm the association of the selected genes with breast cancer development and progression. Future research may validate their expression in larger cohorts and investigate their mechanistic roles in breast tumorigenesis.

4 Conclusion

Cancer remains one of the most formidable global health challenges, primarily due to its high mortality rates and complex biological behavior. The integration of genomic and transcriptomic data has provided researchers with critical insights into the molecular mechanisms potentially driving its development.

This study demonstrated the effectiveness of machine learning models in classifying breast cancer transcriptomic data. Among the classifiers evaluated, the Random Forest and RBF SVM models achieved the highest accuracy—exceeding 99%—and consistently outperformed others in terms of stability and generalization. Their low standard deviations and

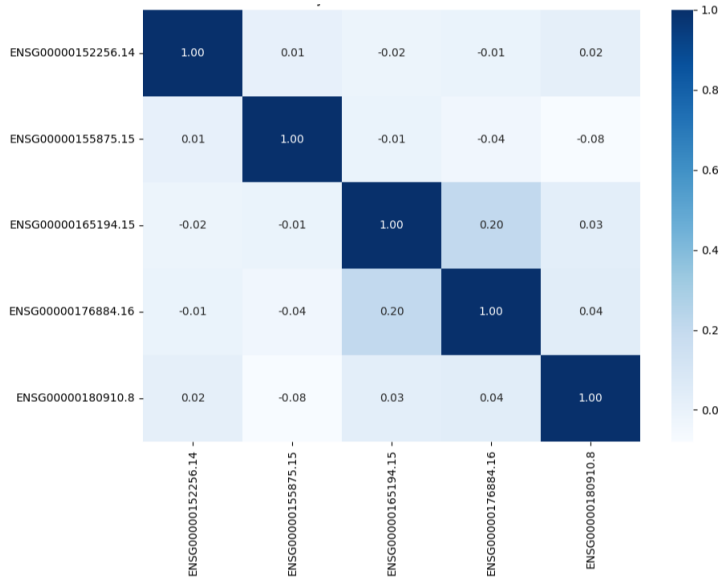

Fig. 2: Heatmap graph of the 5 most relevant features found with feature selection algorithm using the Decision Tree classification model

Table 3: Description of Genes and their Types

Gene	Descrição do Ensembl	Gene Type
ENSG00000152256.14	Gene PDK1, Pyruvate Dehydrogenase Kinase 1	Protein coding
ENSG00000155875.15	Gene SAXO1, Axonemal Microtubule Stabilizer 1	Protein coding
ENSG00000165194.15	Gene PCDH19, Protocadherin 19	Protein coding
ENSG00000176884.16	Gene GRIN1, Ionotropic Glutamate Receptor NMDA Type Subunit 1	Protein coding
ENSG00000180910.8	Gene TTTY11, Expressed Testis Transcription, Linked to Y 11	Unprocessed pseudogene transcript

balanced precision and recall scores further confirmed their robustness and reliability.

The analysis of confusion matrices revealed that these top-performing models effectively minimized both false positives and false negatives, making them particularly well-suited for clinical applications that demand high diagnostic accuracy. In contrast, classifiers such as the Gaussian Process and Neural Network exhibited significant limitations, particularly in terms of specificity, which may reduce their practical utility in medical settings.

The application of five-fold cross-validation enhanced the reliability of performance estimates, ensuring that the results were not overly dependent on any single data partition. However, the underlying class imbalance in the dataset remained a potential challenge that could influence predictive outcomes.

The incorporation of feature selection methods enhanced interpretability by identifying biologically meaningful genes. Among the five most relevant, *PDK1* is known for regulating cancer cell metabolism, *SAXO1* may influence cytoskeletal remodeling, *PCDH19* plays a role in cell adhesion and tumor invasiveness, *GRIN1* is involved in glutamatergic signaling linked to cell proliferation, and *TTTY11*, a pseudogene, might regulate gene networks through ceRNA mechanisms. These insights reinforce the potential of these genes as biomarkers or therapeutic targets.

In summary, this study provides a reliable, interpretable, and high-performance framework for transcriptomic analysis in breast cancer, supporting both the discovery of biomarkers and the advancement of precision oncology.

Additional Information

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

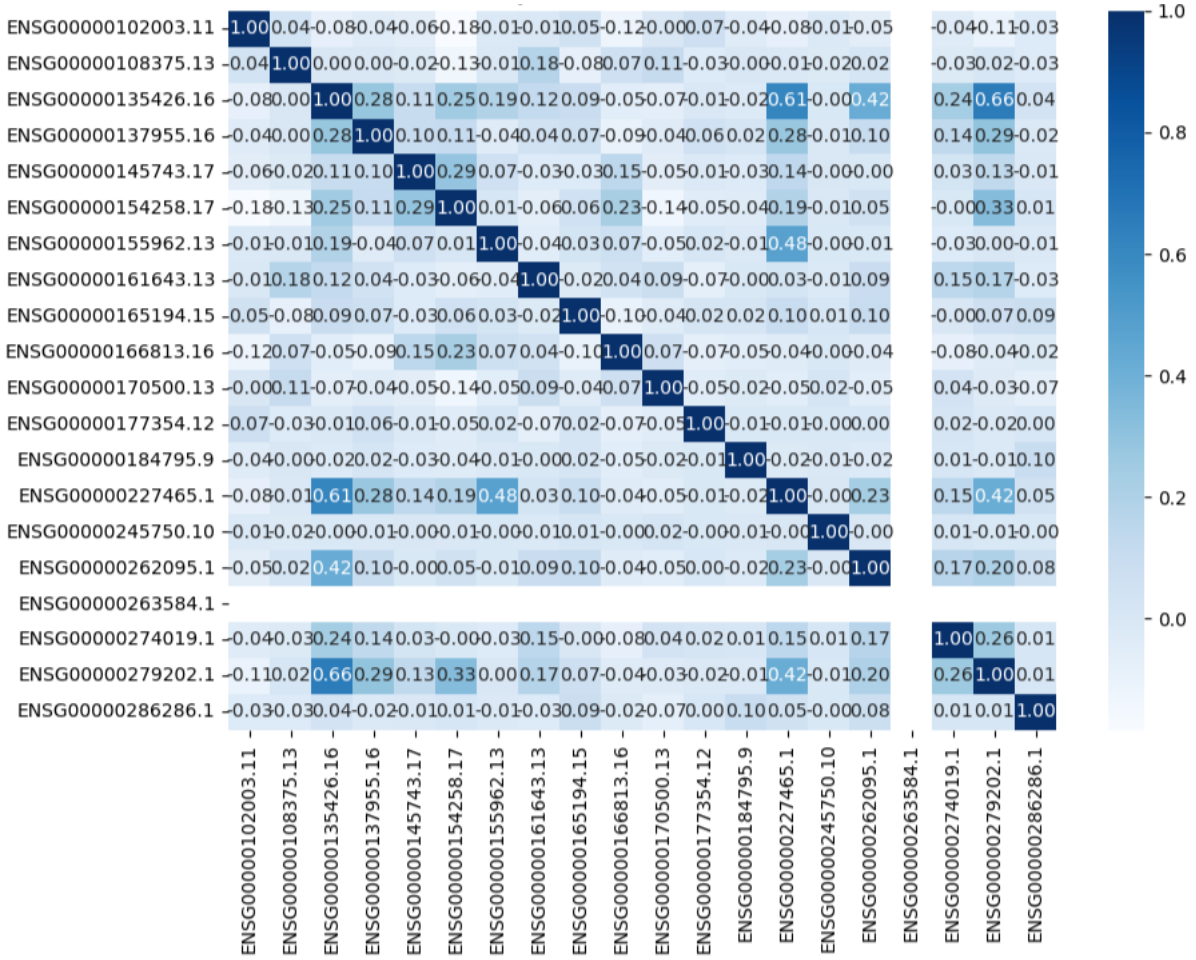

Fig. 3: Heatmap graph of the 20 most relevant features found with the SelectKBeast statistical method

Acknowledgement

The first author acknowledges the financial support from the Coordination for the Improvement of Higher Education Personnel (CAPES). The authors also thank the Federal University of Technology – Paraná (UTFPR) and the Graduate Program in Bioinformatics for institutional support.

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] L. Breiman. Classification and regression trees. Routledge, (2017).

[2] G.V. Batista, J.A. Moreira, A.L. Leite, C.I.H. Moreira. Câncer de mama: fatores de risco e métodos de prevenção. Research, Society and Development, **9**(12), e15191211077 (2020).

[3] J. Liñares-Blanco, A. Pazos, C. Fernandez-Lozano. Machine learning analysis of TCGA cancer data. PeerJ Computer Science, **7**, e584 (2021).

[4] T.C. Silva, A. Colaprico, C. Olsen, et al. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. F1000Research, **5** (2016).

[5] A. Colaprico, T.C. Silva, C. Olsen, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research, **44**(8), e71 (2016).

[6] M. de Oliveira Santos, F.C.S. de Lima, L.F.L. Martins, et al. Estimativa de incidência de câncer no Brasil, 2023-2025. Revista Brasileira de Cancerologia, **69**(1) (2023).

[7] A.A. Ionkina, G. Balderrama-Gutierrez, K.J. Ibanez, et al. Transcriptome analysis of heterogeneity in mouse model of metastatic breast cancer. Breast Cancer Research, **23**, 1-16 (2021).

[8] K. Tomczak, P. Czerwińska, M. Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, **2015**(1), 68-77 (2015).

[9] What Is Data Mining. Data mining: Concepts and techniques. Morgan Kaufmann, **10**(559-569), 4 (2006).

[10] E. Carrizosa, C. Molero-Río, D. Romero Morales. Mathematical optimization in classification and regression trees. Top, **29**(1), 5-33 (2021).
[11] T. Beckers. An introduction to Gaussian process models. arXiv preprint arXiv:2102.05497 (2021).
[12] M.M. Bejani, M. Ghatee. A systematic review on overfitting control in shallow and deep neural networks. Artificial Intelligence Review, **54**(8), 6391-6438 (2021).
[13] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor. An introduction to statistical learning: With applications in Python. Springer Nature, (2023).
[14] J. Brownlee. How to choose a feature selection method for machine learning. Machine Learning Mastery, **10**, 1-7 (2019).
[15] K.L. Howe, P. Achuthan, J. Allen, et al. Ensembl 2021. Nucleic Acids Research, **49**(D1), D884-D891 (2021).
[16] E.Y. Boateng, J. Otoo, 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, **8**(4), 341-357 (2020).
[17] R. Couronné, P. Probst, A.L. Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, **19**, 1-14 (2018).
[18] J. Brownlee. How to choose a feature selection method for machine learning. Machine Learning Mastery, **10**, 1-7 (2019).
[19] Zhang, S. L., et al. "PDK1 and the Warburg Effect in Cancer." *Cancer & Metabolism* 6.1 (2018): 1-12.
[20] Kwon, Y., et al. "Cytoskeletal Dynamics and SAXO1 in Cancer Cell Motility." *Journal of Cell Science* 134.2 (2021): jcs256338.
[21] Stelzer, Y., et al. "Protocadherins and Their Emerging Role in Cancer." *Cell Adhesion & Migration* 10.1-2 (2016): 41–45.
[22] Zhou, Y., et al. "Glutamate Signaling in Cancer Cells." *Cancer Research* 79.4 (2019): 889–896.
[23] Chen, Y., et al. "Pseudogenes and Their Emerging Roles in Tumor Biology." *Seminars in Cancer Biology* 67 (2020): 179–191.

Natalia Padre Holds a degree in Biotechnology from the Federal University of Bahia and recently obtained a Master's degree in Bioinformatics from the Federal University of Technology – Paraná (UTFPR), in the Graduate Program in Bioinformatics. Their research interests include genomic and transcriptomic analysis, data integration in cancer, and the application of machine learning techniques in bioinformatics.

Monique Borges Seixas is an undergraduating student og Bioprocess Engineering at UTFPR (Universidade Tecnológica Federal do Paraná). Her work to focus on deep learning and image segmentation in medical imaging.

Paulo Victor dos Santos is a researcher at Albert Einstei Hospital - Brazil, with expertise in areas like deep learning, machine learning, and computer vision. He has contributed to studies involving medical imaging and optimization techniques.

Glaucia Maria Bressan is an associate professor and researcher of Universidade Tecnológica Federal do Paraná (UTFPR) in Brazil and in PPGBIOINFO Graduate Program. Her work spans areas like optimization, machine learning, and fuzzy systems.

Heron Oliveira dos Santos Lima is a full professor and researcher of Universidade Tecnológica Federal do Paraná (UTFPR) in Brazil. He has worked in areas such as Chemistry, Biotechnology and Biology, machine learning, and optimization.

Marcella Scoczynski is a researcher and associate professor at UTFPR and PPGBIOINFO Graduate Program with expertise in fields like machine learning, combinatorial optimization, and evolutionary computation.