Journal of the Brazilian Computer Society, 2022, 28:1, doi: 10.5753/jbcs.2022.2209

Classification of breast cancer subtypes: A study based on representative genes

Rayol Mendonca-Neto [**Federal University of Amazonas** | ***rayol@icomp.ufam.edu.br***]
João Reis [**Federal University of Amazonas** | ***joao.reis@icomp.ufam.edu.br***]
Leandro Okimoto [**Federal University of Amazonas** | ***okimoto@icomp.ufam.edu.br***]
David Fenyö [**New York University** | ***david@fenyolab.org***]
Claudio Silva [**New York University** | ***csilva@nyu.edu***]
Fabíola Nakamura [**Federal University of Amazonas** | ***fabiola@icomp.ufam.edu.br***]
Eduardo Nakamura [**Federal University of Amazonas** | ***nakamura@icomp.ufam.edu.br***]

Federal University of Amazonas, Av. Gen. Rodrigo Octávio, 6200, Setor Norte do Campus Universitário, Coroado, Manaus, AM, 69080-900, Brazil.

Received: 15 September 2021 • **Accepted:** 05 November 2022 • **Published:** 29 December 2022

Abstract Breast cancer is the second most common cancer type and is the leading cause of cancer-related deaths worldwide. Since it is a heterogeneous disease, subtyping breast cancer plays an important role in performing a specific treatment. In this work, we propose an evaluation framework that uses different machine learning techniques for a broader analysis of the PAM50 list in the classification of breast cancer subtypes. The experiments show that the best method to be used in the classification of breast cancer subtypes is the SVM with linear kernel, which presented an F1 score of 0.98 for the Basal subtype and 0.90 for the Her 2 subtype, the two subtypes with worse prognosis, respectively. We also presented a gene analysis for the classification methods using SHAP values, where we found which genes are important for the classification of each subtype.

Keywords: Breast Cancer, Gene Expression, Subtypes Classification

1 Introduction

Breast cancer is the second most common type of cancer and is the leading cause of disease-related death worldwide [Bray *et al.*, 2018]. As a highly heterogeneous disease, breast cancer shows distinct genetic variations, clinical outcomes, and treatment strategies among its subtypes [Chen *et al.*, 2016]. Breast cancer has four main molecular subtypes: Basal, Her 2, Luminal A, and Luminal B. Basal and Her 2 are the subtypes with the worst prognosis (they have the highest fatality rate), respectively, while Luminal A and Luminal B are linked to a better prognosis, as there are effective targeted therapies for them [Dwivedi *et al.*, 2019].

Classification methods are commonly used to identify cancer subtypes because they allow efficient, accurate, and objective diagnosis. Diagnosis of a tumor based on its biological (or “intrinsic”) subtype provides significant prognostic and predictive information for patients with breast cancer [Parker *et al.*, 2009]. Thus correct classification into its subtypes is critical for the effective treatment of patients [Mendoncaneto *et al.*, 2021]. Nowadays, for patient prognosis and management, classical immunohistochemistry markers (e.g., ER, PR and HER2), along with traditional clinicopathological variables (e.g., tumor size, tumor grade, and nodal involvement), are commonly used [Dai *et al.*, 2015].

Recent advances in DNA *microarray* technology have allowed us to monitor the expression levels of thousands of genes simultaneously during important biological processes [Jiang *et al.*, 2004], resulting in gene expression data. Presenting a viable alternative to employ in cancer classification, these gene expression data pose a challenge for analysis, as there are usually thousands of genes for a few hundred samples.

Despite the wealth of these data, genetic mapping of breast cancer and its subtypes is still far from complete. Currently, there is a list called PAM50 [Chia *et al.*, 2012], which includes fifty genes accepted as representative for the characterization of breast cancer and is considered the referential set of genes to differentiate the subtypes. However, there is still a need to investigate these subtypes further, as there is no precise way to distinguish them using the PAM50 gene list.

This shows that although the study of gene expression is already a reality, there is still no definitive understanding of all genes related to breast cancer and, especially, a definitive understanding of the interactions between these genes. Thus, an important contribution to accurate diagnosis is identifying a subset of genes capable of characterizing the subtypes and differentiating them from each other.

In this context, we propose an evaluation framework that uses different machine learning techniques to classify breast cancer subtypes and investigate the features. Given the particularities of gene expression data, mainly caused by the sensitivity of different technologies for its acquisition, it is not a simple application of machine learning methods and packages from a computational point of view. Since depending on the technologies used (e.g., cDNA microarray [Schena *et al.*, 1995] or oligonucleotide arrays Lockhart *et al.* [1996]) to quantify the gene expression data. Results are presented differently. For example, RNA-Seq is a more recent and advanced technique. It is capable of investigating at high resolution all the RNAs present in a sample, characterizing their sequences, and quantifying their abundances at the same time. Therefore, combining results from two distinct technologies is challenging.

Accordingly, there is a clear need to investigate existing techniques to treat these input data with different characteristics and reliability. Based on this, the main contributions of this work are: (i) Study of different methods in the task of classifying breast cancer subtypes, (ii) analysis of the PAM50 list in the classification of breast cancer subtypes and (iii) identify a list of genes that are important for the classification in each subtype.

2 Motivation

Studying the characteristics of thousands of genes simultaneously offered a deep insight into cancer classification problem. The gene expression profiles available in cancer datasets introduced an abundant amount of data ready to be explored [Tarek *et al.*, 2017] and laid the problem to a high-dimensional data problem.

Despite the large data, as there are usually thousands of genes for a few hundred samples, there is a movement to employ gene expression data in the cancer classification [Yip *et al.*, 2011]. One way to simplify the high-dimensional data classification problem is selecting genes [Guyon *et al.*, 2002]. The majority of the cancer classification approaches are employed for binary classification (cancer/not cancer). Classification of multiclass cancer problem such as breast cancer subtypes based on gene expression profiles is less common but has more practical implications for prognosis as well as the potential to further improve our understanding of gene expression of various cancer problems [Nguyen and Rocke, 2002].

Thus, employing a feasible breast cancer subtype classification using representative genes is an important task, which needs the best methods to narrow the gene selection and improve the classification [Shukla *et al.*, 2018]. Additionally, this is the best way to simplify machine learning methods, speed up the classification task and significantly improve the performance of the classifier [Gatto *et al.*, 2021; Alanni *et al.*, 2019; Díaz-Uriarte and De Andres, 2006].

3 Related Work

When employing gene expression data in cancer classification context, authors try to reduce the uneven number of samples versus genes, by selecting only relevant genes, this approach is known as gene filtering or gene selection. Selecting only important genes, can improve the classification performance and also reduce the computational effort.

Graudenzi *et al.* [2017] proposed a classification framework based on Support Vector Machines (SVMs) with a feature selection strategy based on the concept of pathway activity. They identified and analyzed a list of enriched pathways in four different breast cancer subtypes, and used this information to perform the feature selection method in the classifier implementation. In terms of overall accuracy, the proposed classifier presents an accuracy around 85.00%, using 400 genes from the feature selection method.

Lee *et al.* [2020] used a pathway-based approach for feature selection, and applied a deep learning model with attention mechanism and network propagation for cancer classification. They used five TCGA[1] cancer datasets. The average F1 score of their method was 93.74% for urothelial bladder Carcinoma (BLCA), 85.52% for breast invasive carcinoma (BRCA), 87.01% for colon adenocarcinoma (COAD), 89.62% for prostate adenocarcinoma (PRAD), and 91.49% for (Stomach adenocarcinoma) STAD. They selected a total of $5,515$ genes for the classification task.

Mostavi *et al.* [2020] proposed three distinct Convolutional Neural Networks (CNN) for cancer classification task. Regarding the prediction of breast cancer subtypes, the 1D-CNN model was employed. The authors used poor statistics methods for the feature selection step, such as standard deviation and mean. After selecting $7,091$ genes, they used their model for the classification task and achieved an average accuracy of 88.42% among five subtypes.

In the work of Li *et al.* [2017], the authors divided the process into two stages. First, a genetic algorithm is used as a gene selection mechanism, and the KNN (k-nearest neighbors) algorithm as a classification method. The dataset contains 31 tumor types. For the sorting task using KNN, k was set to 5 with a majority voting rule. The results show that the classification accuracy was greater than 90% for 28 of the 31 types of cancer.

Lyu and Haque [2018] incorporated gene expression data into 2-D images and used a Convolutional Neural Network (CNN) to classify 33 distinct types of cancer. The authors transform cancer classification based on the gene expression problem into an image problem. The main problem is that gene expression data is highly dimensional, whereas most deep learning architectures are for 2-D imaging. As a result, the authors achieved a mean F1 across cancer types of 95.43% using $20,531$ genes.

Table 1 synthesize the related works. In summary, research that explores the problem of multiclass (subtypes) classification encounters difficulty concerning performance. The works of Li *et al.* [2017] and Lyu and Haque [2018] dealt with the binary classification, even though they used more than 30 types of tumour, all the samples were classified as cancer or non cancer, therefore achieving high results.

When dealing with the multiclass classification, these works use hundreds of genes and do not reach such expressive results, not reaching 90% of *accuracy*. Different subtypes can share important genes for their identification, so classifying a type of cancer among the subtypes makes it a much more complex task. In this context, we will work with the PAM50 list as it is considered representative for breast cancer subtypes and has only fifty genes to generate better results using fewer genes.

[1] The Cancer Genome Atlas Program

Table 1. Summary of the related work.

Author	# of genes	Features	Classes	Classifier	Evaluation Metrics
[Graudenzi *et al.*, 2017]	400	Basedo em pathways	Multiclass	SVM	Precision, recall and accuracy
[Mostavi *et al.*, 2020]	7,091	Standard Deviation and mean	Multiclass	1D-CNN	Precision, recall and F1
[Lee *et al.*, 2020]	5,015	Baseado em Pathways	Multiclass	GCN+MAE	F1
[Li *et al.*, 2017]	-	Genetic algorithm	Biclass	KNN	Accuracy
[Lyu and Haque, 2018]	20,531	Variance	Biclass	GA/KNN	Precision, recall, accuracy e F1

4 Evaluation Framework

The evaluation framework in this work consists of the following steps. (i) collection of databases that have gene expression; (ii) data pre-processing to select only the genes involved in the study; (iii) classification of samples among breast cancer subtypes; and (iv) analysis of the performance of classifiers with different evaluation metrics.

4.1 Dataset and Pre-processing

This framework starts by choosing the dataset, where the data can be extracted from genomic data repositories containing gene expression data. After the data collection phase, pre-processing is necessary to identify if the database has all 50 genes from the PAM50 list. Thus, among all the genes that exist in the chosen dataset, only the 50 genes from the PAM50 list are selected, if the 50 genes are not in the dataset, another dataset needs to be used to validate the work. In this step, we understand that employing the PAM50 genes in the classification task is already a way of feature selection since it reduces our scope from thousands of genes to only 50.

4.2 Classification

After selecting only the PAM50 genes for the training and testing basis, we classify them with different classifiers. This step aims to understand how different classification methods are able to distinguish breast cancer subtypes using gene expression data.

4.3 Data Analysis

To measure the performance of the methods, we apply traditional metrics such as *precision*, *recall*, *F-measure*, and *accuracy*. Since biological data usually have a sparse dataset [Chicco, 2017], we also measure the performance of the methods using the Matthews correlation coefficient (*MCC*) [Baldi *et al.*, 2000] and *precision vs. recall* curve (*AUPRC*). Both metrics were selected because they are suitable for unbalanced databases. While *MCC* is more appropriate for binary classification, the *precision vs. recall* curve is a more reliable and informative indicator of statistical performance in multiclass problems [Chicco, 2017].

We also calculate *specificity* as this measure is used to see how correctly we can classify an individual into the correct cancer subtype [Parikh *et al.*, 2008]. Metrics are defined as:

$$Precision = \frac{TP}{TP + FP}, \quad (1) \qquad Recall = \frac{TP}{TP + FN}, \quad (2)$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}, \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}}, \quad (5)$$

$$Specificity = \frac{TN}{TN + FP}, \quad (6)$$

in which *TP* are true positives, *TN* are true negatives, *FP* are false positives, and *FN* are false negatives. The higher the value of these metrics, the better the result.

After performing the classification and evaluating the performance of each classifier, we apply the Shapley Values (SHAP) [Lundberg and Lee, 2017] to evaluate the feature importance for each of the classifiers. SHAP is a game theory based approach to describe the performance of a machine learning model. SHAP can provide explanations for local and global models, estimating feature contributions to the output of the model. To produce an interpretable model, SHAP uses an additive feature attribution method. According to Bi *et al.* [2020], the SHAP values can be calculated as follows:

$$\phi_i = \sum_{S \subseteq F, \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{s\cup\{i\}} \left(x_{s\cup\{i\}} \right) - f_s \left(x_s \right) \right], \quad (7)$$

where F represents the set of all features and S represents all feature subsets obtained from F after removing the i^{th} feature. Then, two models, $f_{s\cup\{i\}}$ and f_S, are trained again, and predictions of these models are compared to the current input $\left[f_{s\cup\{i\}} \left(x_{s\cup\{i\}} \right) - f_s \left(x_s \right) \right]$, where x_S represents the values of the input features in the set S. To estimate ϕ_i from $2^{|F|}$ differences, the SHAP approach approximates the Shapley value by either performing Shapley sampling or Shapley quantitative influence. It is interesting to note that, SHAP estimates the feature importance (magnitude of the contribution) as well as the sign (positive or negative).

5 Evaluation of the proposed Framework

This section presents an analysis of the different classifiers used to classify breast cancer subtypes using the PAM50 gene set. Additionally, we detail the methodology used to apply the proposed approach.

5.1 Methodology

In this subsection, we describe the evaluation methodology used in this work. We detail the characteristics of the datasets used in the experiments. We present the parameters chosen for the machine learning methods and also explain the evaluation metrics used.

Dataset

We used two distinct gene expression datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [Edwards *et al.*, 2015] to validate the methods. The datasets present the gene expression of patients with breast cancer. These specimens are divided into four intrinsic breast cancer subtypes (see Section 1).

To evaluate the performance of the classifiers, we employ a 10-fold cross-validation classification. We combined both datasets using ComBat [Johnson *et al.*, 2007], a widely used tool for correcting technical biases in gene expression data. Therefore, obtaining a larger dataset with 194 samples. Then, we separated the resulting merged dataset into 70% for training and 30% for testing. The training set was used for the 10-fold cross-validation.

Table 2 summarizes the characteristics of the databases used in the experiments:

Table 2. Dataset description.

Dataset	**# of genes**	**Subtypes**	**# of samples**	**Total # of samples**
Cptac 2C	23122	Basal	29	117
		Her 2	14	
		Luminal A	57	
		Luminal B	17	
Cptac 2D	16525	Basal	18	77
		Her 2	12	
		Luminal A	23	
		Luminal B	24	

Classifiers

To evaluate the performance of the PAM50 list for classifying breast cancer subtypes, we employed five distinct methods. The Grid search [Bergstra and Bengio, 2012] was used to optimize the parameters for each classifier, the parameters set for the Grid Search were chosen empirically. Table 3 presents the chosen classifiers and parameters. The remaining parameters have been set to the scikit-learn[2] default configuration.

Table 3. Classifier parameters.

Method	Parameters
SVM(Linear)	$C = 0.1$
SVM(RBF)	$C = 1.1$
KNN	$p = 1, n\,neighbors = 5, weights = uniform$
Random Forest	$bootstrap = False, min\,samples\,split = 6, n\,estimators = 28$
XGBoost	$gamma = 0.04\,, learning\,rate = 0.07$

To calculate the evaluation metrics, we used the scikit-learn and pandas-ml[3] libraries. We compare the different classifiers to see which method performs better overall and for each subtype separately. We use the evaluation metrics presented in Section 4.

5.2 Results

Different classifiers take into account distinct ways of classifying samples into different classes. Some use spacing between classes to differentiate them (SVM), while others check which class predominates among the elements closest to the analyzed sample (KNN). Some use decision trees (Random Forest) to perform the classification, and others start from a primary hypothesis and try to improve it to reach a better result (XGboost). Therefore, we expect that there will be different results for the tested classifiers, even if they are submitted to the same test conditions.

Classification Analysis

In the first experiment, we compared the results obtained by all methods to classify the four subtypes (Figure 1). Figure 1a illustrates performance in terms of precision, recall, and F1. Figure 1b illustrates performance in terms of accuracy, MCC, and specificity. The X axis presents the precision (Figure 1a) and the accuracy (Figure 1b). The Y axis presents the recall (Figure 1a) and the specificity ((Figure 1b). The larger the circle, the higher the F1 in Figure (Figure 1a) and the specificity in (Figure 1b). The color of the circle indicates the method.

Comparing the performance obtained by the methods, we see that SVM(Linear) outperformed all other methods in the six analyzed macro metrics. This performance can be explained by the fact that it is the classifier that best managed to separate the samples from Luminal A from Luminal B.

Figure 2 shows the average confusion matrices obtained by each method. Each row represents the instances of an actual class, and each column represents the instances of a predicted class. We can see that the results obtained with the *SVM(Linear)* (Figure 2a presented better results than the other classifiers, where 7,69% of the Basal samples were wrongly classified as Her 2, 19,09% of Her 2 samples and 0,59% of Luminal A misclassified as Luminal B. In the case

[2] https://scikit-learn.org/stable/

[3] https://pypi.org/project/pandas-ml/

Table 4. Classification results using *precision*, *recall* and *F1* micro metrics. Best results in bold.

Method	Precision				Recall				F1			
	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB
SVM(Linear)	**0,99**	0,91	**0,87**	**0,86**	0,96	**0,88**	**0,93**	**0,81**	**0,98**	0,90	**0,90**	**0,83**
SVM(RBF)	**0,99**	0,91	0,83	0,84	0,96	0,84	0,89	0,79	**0,98**	0,88	0,86	0,82
KNN	0,96	**0,97**	0,83	0,84	**0,99**	0,86	0,89	0,79	0,97	**0,91**	0,86	0,81
Random Forest	0,96	0,85	0,80	0,80	0,93	0,78	0,86	0,74	0,95	0,81	0,83	0,77
XGBoost	0,97	0,90	0,80	0,85	0,91	0,84	0,85	0,79	0,94	0,87	0,82	0,82

Table 5. Classification results using *MCC*, *AUPRC* and *Specificity* micro metrics. Best results in bold.

Method	*MCC*				*AUPRC*				*Specificity*			
	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB	Basal	Her2	LumA	LumB
SVM(Linear)	**0,95**	0,79	**0,80**	**0,66**	0,96	**0,88**	**0,93**	**0,81**	**1,00**	0,97	**0,86**	**0,95**
SVM(RBF)	**0,95**	0,75	0,72	0,63	0,96	0,84	0,89	0,79	**1,00**	0,98	0,82	0,94
KNN	**0,95**	**0,82**	0,72	0,63	**0,99**	0,86	0,89	0,79	0,97	**1,00**	0,82	0,94
Random Forest	0,90	0,62	0,65	0,54	0,93	0,78	0,86	0,74	0,99	0,97	0,77	0,93
XGBoost	0,87	0,73	0,65	0,63	0,91	0,84	0,85	0,79	**1,00**	0,98	0,77	**0,95**

of *SVM(RBF)* (Figure 2b, the classifier had difficulty separating the Her 2 samples, where 14,55% of the samples were incorrectly classified as Luminal A and 15,45% incorrectly classified as Luminal B.

The *KNN* classifier (Figure 2c) had the same classification problem as the *SVM(RBF)*, where 35,56% of the Luminal B samples were classified as Luminal A. In addition, 15,45% of the Her 2 samples were incorrectly classified as Luminal B. The KNN was the only method to correctly classify 100% of the Basal subtype. Tests with the *Random Forest* (Figure 2d) present that 39,44% of Luminal B were erroneously classified as Luminal A. Finally, *XGBoost* (Figure 2e) misclassified 27,69% of the Basal samples and confused 35,56% of the Luminal B samples with Luminal A subtype.

In summary, the *SVM(Linear)* classifier provides the best performance, with the least amount of wrongly classified samples. It can also be noted that the subtype with the worst prognosis, Basal, had the few characterization problems, regardless of classifier, thus being the most characteristic subtype among the four. In contrast, the Luminal B subtype is the subtype where the classifiers have greater difficulty in classifying the samples. It also happens that all classifiers have classified at least 10% of the Her 2 subtype as Luminal B.

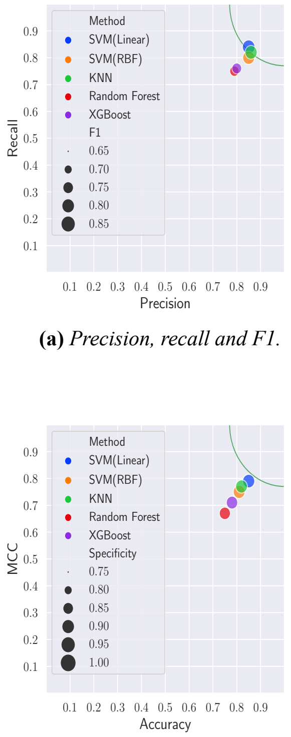
Analyzing Table 4, containing the results obtained by the five classifiers tested, we can identify that the Basal subtype obtained a precision score of 99% in both SVM classifiers. The Luminal A subtype had the highest precision score of 87% with SVM(Linear), plus a recall above 85% in four of the five classification methods used. Observing the Her 2 subtype, it can be seen that it manages to obtain a score of 97% of precision with the KNN classifier, with a recall near 90% only with SVM(Linear) classifier. The Luminal B subtype had a maximum precision of 86% and a maximum recall of 81%.

When we analyze the F1 score, the SVM(Linear) classifier holds three highest scores in the four subtypes, The Basal subtype had a score of 98%, while the Luminal A subtype had 90% and the Luminal B subtype had 83%. For the Her 2 subtype, the KNN classifier had a score of 91%.

Table 5 contains the results for the metrics *MCC*, *AUPRC* and *Specificity*. Examining the data obtained with the metric *MCC*, we notice that the classifier SVM(Linear) along with SVM(RBF) and KNN have the highest scores for the subtypes Basal with 95%. Only SVM(Linear) achieves a *MCC* of 80% in Luminal A. In contrast, the Her 2 subtype had a

Table 6. Classification results using macro metrics compared with the related work. The best results for each metric are bold.

Method	# of genes	F1 Macro	ACC	MCC	AVG AUPRC	AVG Specificity
SVM(Linear)	50	**0,85**	**0,84**	**0,79**	**0,89**	**0,94**
SVM(RBF)	50	0,83	0,81	0,75	0,87	0,93
KNN	50	0,84	0,82	0,77	0,88	0,93
Random Forest	50	0,76	0,75	0,67	0,83	0,91
XGBoost	50	0,80	0,77	0,71	0,85	0,92
1D-CNN (Mostavi)	50	0,63	0,73	-	-	-



(a) *Precision, recall and F1.*

(b) *Accuracy, MCC and specificity.*

Figure 1. Performance obtained by the methods using macro metrics.

maximum score of 82% with the *KNN* classifier.

In the *AUPRC* metric, the SVM(Linear) classifier obtained scores of 88% for the Her 2 subtype, 93% for the Luminal A subtype, and 81% for the Luminal B subtype. The KNN classifier obtained an AUPRC of 99% for the Basal subtype. Analyzing the *Specificity*, we notice that the Basal subtype got 100% with the classifiers SVM(Linear), SVM(RBF), and XGBoost. The Her 2 subtype obtained a maximum of 100% using the KNN classifier. The Luminal A subtype obtained 86% with the SVM(Linear) classifier, and the Luminal B subtype obtained 95% of *Specificity* using the SVM(Linear) and XGBoost.

Analyzing the data from the Table 6, with the results of the macro metrics, and the results of the related works. We can conclude that SVM(Linear) outperforms all of the other classifiers. The work of [Mostavi *et al.*, 2020] uses a CNN but only achieves a F1 score of 63%. The SVM(Linear) has the best score in the following employed evaluation metrics: 85% for *F1*, 84% for *Accuracy*, 79% for *MCC*, 89% for *AUPRC* and 94% for *specificity*.

In general, we notice that the Basal, the subtype with the worst prognosis, is the most characteristic among all since the classifiers obtained better results in this subtype. Concerning Her 2 (subtype with the second-worst prognosis), we noticed that it obtained the second best result among the subtypes. While Luminal B had the worst result, thus being the most difficult to be classified. Finally, the results showed that the evaluation framework combined reveals that the PAM50 gene list has good results when classifying the breast cancer subtypes, and the SVM(Linear) is the best classifier to employ.

Gene Analysis

In this step, we analyze which features (genes) are more important for each of the methods to classify the breast cancer subtypes using SHAP values (Section 4). Although to compute the SHAP values, we face an exponential computational complexity [Messalas *et al.*, 2019], in the scope of our project, we were able to apply it to all samples since our merged dataset has 194 samples.

Figure 3 shows the features SHAP values obtained by each method. The larger the bar, the more critical the gene is for the classification of the subtype. We can see that distinct classifiers present distinct gene importance. For the SVM(Linear) (Figure 3a, the larger bars belong to the Basal subtype, therefore showing why this subtype has the best classification score.

For the SVM(RBF) (Figure 3b, the genes are also important in classifying the Luminal B subtype, explaining why these method only looses to SVM(Linear) when classifying the Luminal B subtype. The KNN (Figure 3c also presents the genes as important for the Basal subtype, explaining the performance in the classification of this subtype. For the other subtypes, this method presents similar results when compared with SVM(RBF) and Random Forest.

The Random Forest (Figure 3d) presents genes very important for the basal subtype classification and also for Her 2 and Luminal A. Finally, the XGBoost (Figure 3e) presents the worst result among all the methods.

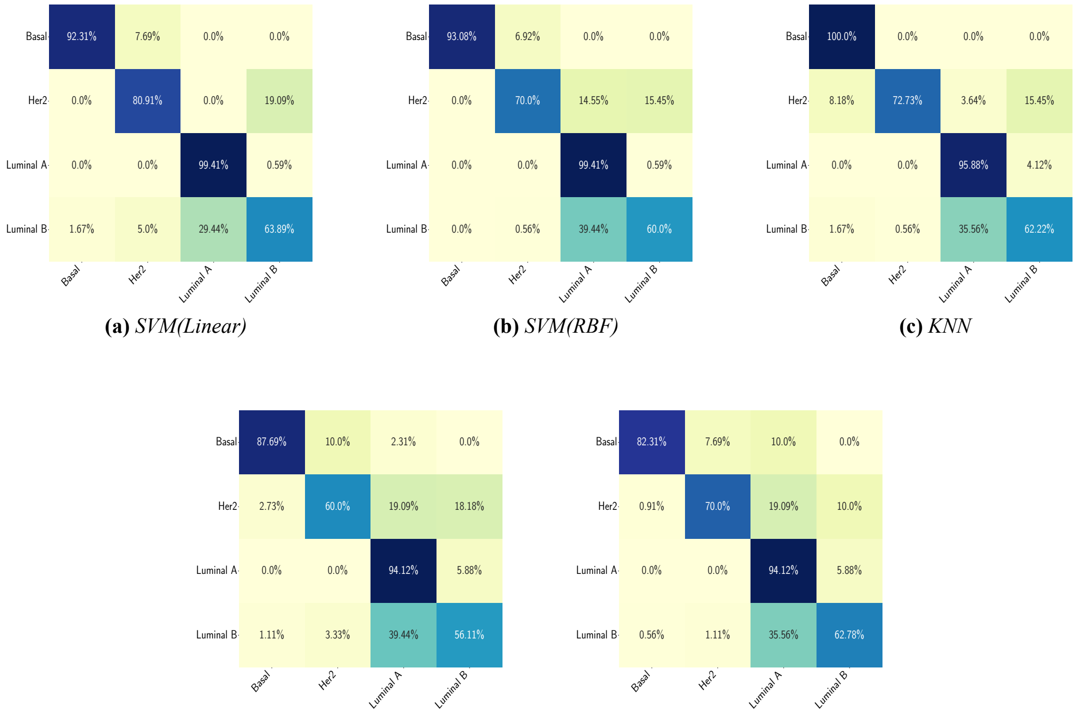
Summarizing the results, we can see that the ESR1 gene has almost the same importance for each classifier and is the most important gene to classify the Her 2 subtype. This is because mutations on this gene are acquired frequently in metastatic hormone receptor-positive breast cancer [Turner *et al.*, 2020].

When we look for the PGR gene, we can see that it is more important for the Luminal A subtype in all the methods. This is because the expression of PGR is a potent prognostic indicator for evaluating the long-term prognosis of Luminal A [Kurozumi *et al.*, 2017]. The FOXA1 and MLPH genes are the more critical genes to classify the basal subtype. Both are implicated in the development of breast cancer [He *et al.*, 2015]. FOXA1 segregates with genes that characterize the luminal subtypes in DNA microarray analyses [Badve *et al.*, 2007].

In our next experiment (Figure 4), we analyze the SHAP values individually for each class (subtype). We chose the SVM(*Linear*) for this analysis since it presented the best classification performance among all the methods tested (Subsection 5.2). This summary plot shows the importance of features and how their SHAP values are spread across the data. The plot uses SHAP values to show the distribution of each feature's impacts on the model output. The dots represent each sample in the test dataset.

For each subtype (Figures 4a, 4b, 4c and 4d), we can see which features are most influential in the model's output, the importance of the features are ranked in ascending order. For example, in Figure 4a, the ESR1 gene is the more important gene for the Basal subtype, while the CCNE1 gene is the 15th more important gene.

The horizontal location of the samples (the dots across the



(a) *SVM(Linear)* **(b)** *SVM(RBF)* **(c)** *KNN*

(d) *Random Forest* **(e)** *XGBoost*

Figure 2. Confusion matrices obtained by each method.

plot) shows whether the effect of that value is associated with a higher or lower prediction, and the color shows whether that variable is high (in red) or low (in blue) for that observation. As can be seen in Figures 4a, 4b, 4c and 4d, the more important gene for each subtype is more spread across the plots.

We can see that the importance of genes is different for each subtype. It is interesting to note that ESR1 is the most important gene for Basal, Her 2, and Luminal B, while for Luminal A is the second most important gene. We can also see that the distribution of the samples varies depending on the subtype.

For example, while for the Basal subtype (Figure 4a) most samples have negative SHAP values and a high correlation with the gene expression, for the Luminal B subtype (Figure 4d), most samples have positive SHAP values. These results complement the experiments presented in Figure 3a, as they demonstrate the behavior of the SHAP values in each of the samples for each of the subtypes.

6 Conclusion and Future Directions

This paper presents an evaluation framework for classifying breast cancer subtypes based on the PAM50 gene list. We employed distinct classification methods, each with different characteristics, to analyze whether there is a difference between them when classifying the breast cancer subtypes. Seven evaluation metrics were employed to evaluate the methods to get an overview of how the methods perform.

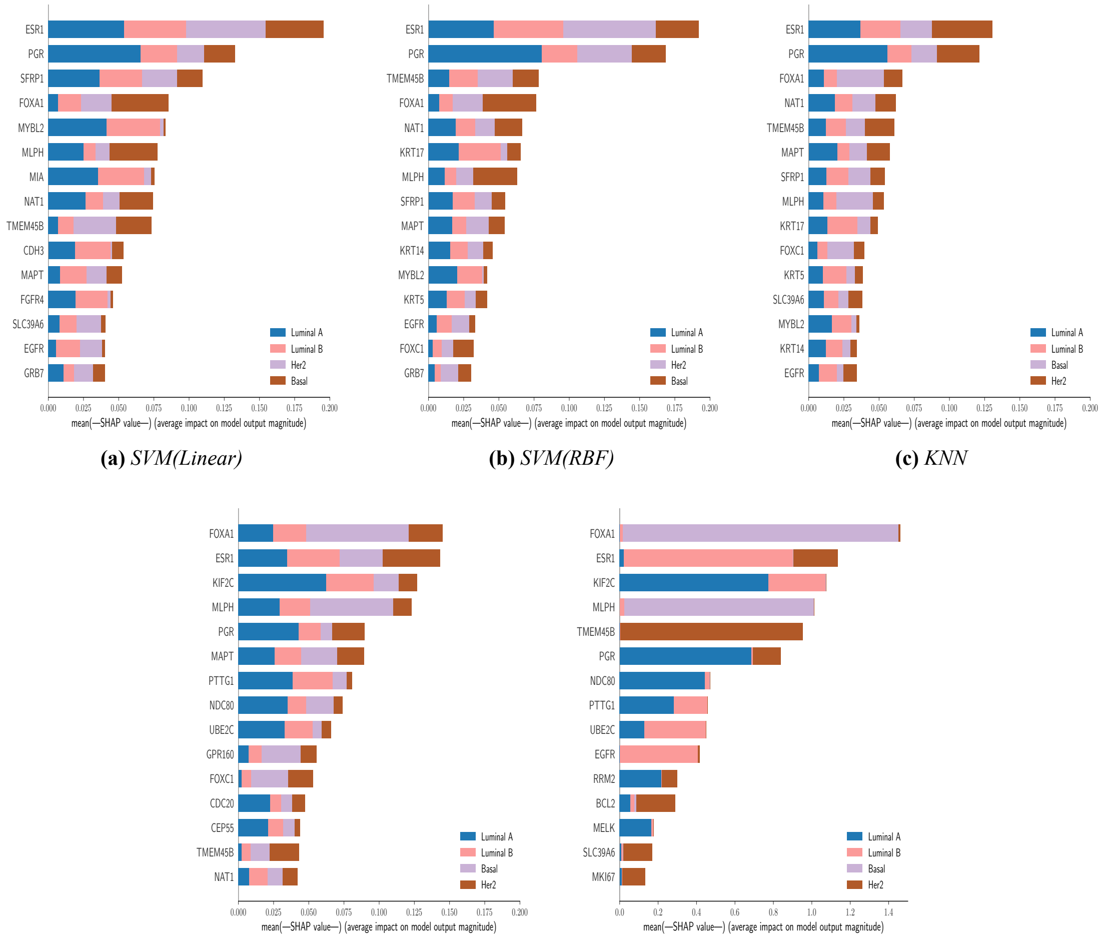
As a result, we noticed that the SVM(Linear) obtained better macro results than the others. We also verified that the Basal subtype (the one with the worst prognosis and the most characteristic), the classifier KNN outperformed all the other methods, reaching an F1 score of 100%. In addition, the other classifiers remained with a score above 80% for this subtype.

It is noticed that Her 2, the subtype with the second-worst prognosis, has the third best results in the classification. It reaches a maximum F1 score of 80%, achieving a minimum of 60% with the classifier Random Forest, in which the Her 2 samples are confused with all other subtypes.

Among the Luminal A and Luminal B subtypes, there is confusion between the samples, given that they are highly correlated. Although the PAM50 has only 50 genes, this is a good set for ranking as it scored in four of the five classifiers an F1 Macro score above 75%. At the micro-level, SVM(Linear) also managed to maintain an F1 score above 82% for all subtypes.

When analyzing the features, we can see that SHAP values identify the more important genes for the classification of each subtype and when we study those genes, we understand how they are related to the classified subtypes.

As future work, we intend to extend the analysis to a multilevel classification, in which we will employ a hierarchical classifier to perform the classification of breast cancer subtypes. Therefore, we will isolate the subtypes and investigate if there are classifiers that present better performances for the analyzed subtype, using the genes from the PAM50 list.



(a) *SVM(Linear)* **(b)** *SVM(RBF)* **(c)** *KNN*

(d) *Random Forest* **(e)** *XGBoost*

Figure 3. Feature SHAP values for each method.

Declarations

Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. This work was developed with the support of the Amazonas State Government through the Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM), with the granting of a scholarship. This research, according to Article 48 of Decree nº 6.008/2006, was funded by Samsung Electronics of Amazonia Ltda, under the terms of Federal Law nº 8.387/1991, through agreement nº 003, signed with ICOMP/UFAM.

Authors' Contributions

Rayol Mendonca-Neto: Formal analysis, Methodology, Writing - Review Editing. João Reis: Software, Writing - Original Draft. Leandro Okimoto: Formal analysis, Software. David Fenyö: Conceptualization, Resources, Supervision. Claudio Silva: Visualization. Fabíola Nakamura: Supervision. Eduardo Nakamura: Conceptualization, Formal analysis, Supervision.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Data can be made available upon request.

References

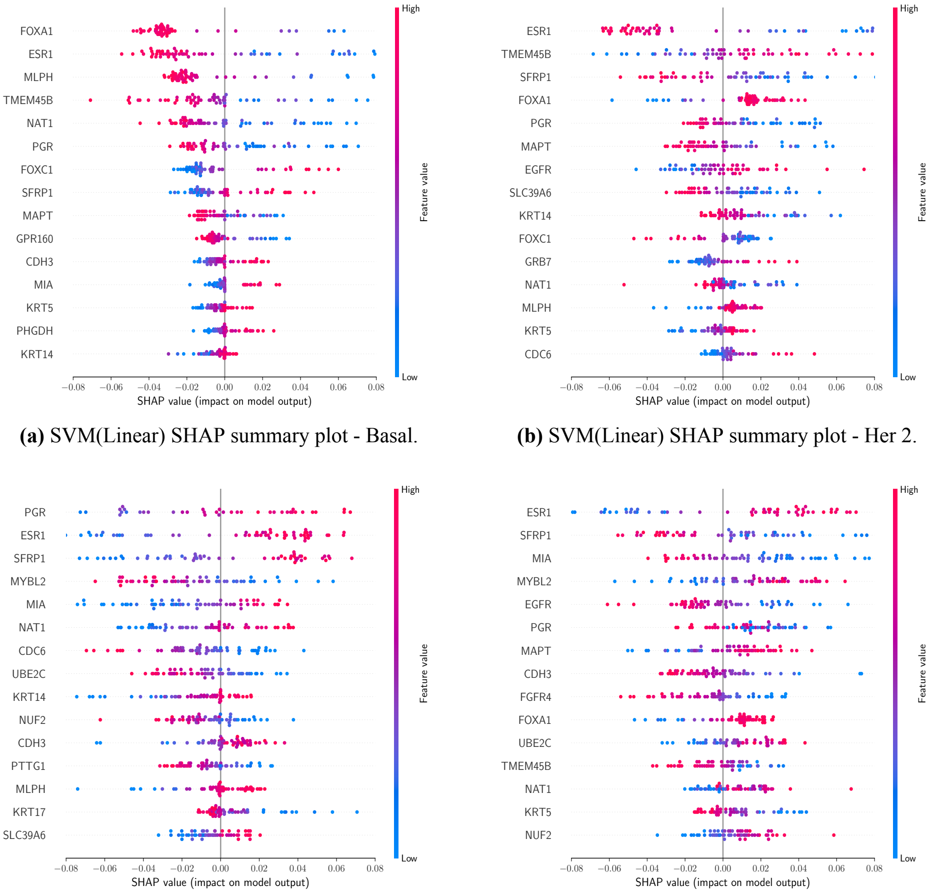
Alanni, R., Hou, J., Azzawi, H., and Xiang, Y. (2019). Deep gene selection method to select genes from microarray datasets for cancer classification. *BMC bioinformatics*, 20(608):1–15.

Badve, S., Turbin, D., Thorat, M. A., Morimiya, A., Nielsen, T. O., Perou, C. M., Dunn, S., Huntsman, D. G., and Nakshatri, H. (2007). Foxa1 expression in breast cancer—correlation with luminal subtype a and survival. *Clinical cancer research*, 13(15):4415–4421.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305.

Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., and Song, J. (2020). An interpretable prediction model for identify-



(a) SVM(Linear) SHAP summary plot - Basal.

(b) SVM(Linear) SHAP summary plot - Her 2.

(c) SVM(Linear) SHAP summary plot - Luminal A.

(d) SVM(Linear) SHAP summary plot - Luminal B.

Figure 4. SVM(Linear) SHAP summary plot for each subtype.

ing n7-methylguanosine sites based on xgboost and shap. *Molecular Therapy-Nucleic Acids*, 22:362–372.

Bray, F., Ferlay, J., Soerjomataram, I., L. Siegel, R., Torre, L., and Jemal, A. (2018). Global cancer statistics 2018. *CA: A Cancer Journal for Clinicians*, 68:394–424. DOI: 10.3322/caac.21492.

Chen, X., Hu, H., He, L., Yu, X., Liu, X., Zhong, R., and Shu, M. (2016). A novel subtype classification and risk of breast cancer by histone modification profiling. *Breast cancer research and treatment*, 157(2):267–279.

Chia, S. K., Bramwell, V. H., Tu, D., *et al.* (2012). A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clinical cancer research*, 18(16):4465–4472.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):1–17.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research*, 5(10):2929.

Díaz-Uriarte, R. and De Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(3):13.

Dwivedi, S., Purohit, P., Misra, R., Lingeswaran, M., *et al.* (2019). Application of single-cell omics in breast cancer. In *Single-Cell Omics*, volume 2, pages 69–103. Elsevier.

Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., Madhavan, S., and Ketchum, K. A. (2015). The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6):2707–2713.

Gatto, B. B., Santos, E. M. d., Koerich, A. L., Fukui, K., and Junior, W. S. (2021). Tensor analysis with n-mode generalized difference subspace. *Expert Systems with Applications*, 171:1–11.

Graudenzi, A., Cava, C., Bertoli, G., Fromm, B., *et al.* (2017). Pathway-based classification of breast cancer subtypes. *Front Biosci*, 22(10):1697–1712.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422.

He, J., Yang, J., Chen, W., Wu, H., Yuan, Z., Wang, K., Li, G., Sun, J., and Yu, L. (2015). Molecular features of triple negative breast cancer: microarray evidence and further integrated analysis. *PloS one*, 10(6):e0129842.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge & Data Engineering*, 16(11):1370–1386.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Kurozumi, S., Matsumoto, H., Hayashi, Y., Tozuka, K., Inoue, K., Horiguchi, J., Takeyoshi, I., Oyama, T., and Kurosumi, M. (2017). Power of pgr expression as a prognostic factor for er-positive/her2-negative breast cancer patients at intermediate risk classified by the ki67 labeling index. *BMC cancer*, 17(1):1–9.

Lee, S., Lim, S., Lee, T., Sung, I., and Kim, S. (2020). Cancer subtype classification and modeling by pathway attention and propagation. *Bioinformatics*, 36(12):3818–3824.

Li, Y., Kang, K., Krahn, J. M., Croutwater, N., *et al.* (2017). A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC genomics*, 18(1):508.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., *et al.* (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675–1680.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

Lyu, B. and Haque, A. (2018). Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 89–96. ACM.

Mendoncaneto, R., Fenyo, D., Li, Z., Nakamura, E. F., Nakamura, F. G., and Silva, C. T. (2021). A gene selection method based on outliers for breast cancer subtype classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Messalas, A., Kanellopoulos, Y., and Makris, C. (2019). Model-agnostic interpretability with shapley values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–7. IEEE.

Mostavi, M., Chiu, Y.-C., *et al.* (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, 13(44):1–13.

Nguyen, D. V. and Rocke, D. M. (2002). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18(9):1216–1226.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45–50.

Parker, J. S., Mullins, M., Cheang, M. C., *et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470.

Shukla, A. K., Singh, P., and Vardhan, M. (2018). A hybrid gene selection method for microarray recognition. *Biocybernetics and Biomedical Engineering*, 38(4):975–991.

Tarek, S., Elwahab, R. A., and Shoman, M. (2017). Gene expression based cancer classification. *Egyptian Informatics Journal*, 18(3):151–159.

Turner, N. C., Swift, C., Kilburn, L., Fribbens, C., Beaney, M., Garcia-Murillas, I., Budzar, A. U., Robertson, J. F., Gradishar, W., Piccart, M., *et al.* (2020). Esr1 mutations and overall survival on fulvestrant versus exemestane in advanced hormone receptor–positive breast cancer: A combined analysis of the phase iii sofea and efect trials. *Clinical Cancer Research*, 26(19):5172–5177.

Yip, W.-K., Amin, S. B., and Li, C. (2011). A survey of classification techniques for microarray data analysis. In *Handbook of Statistical Bioinformatics*, pages 193–223. Springer.