RESEARCH

Classification of different breast cancer subtypes using supervised machine learning techniques

Carmen Toledo[1*], Jacqueline Boccacino[2], Letícia Braga[3] and Newton Silva[1]

*Correspondence: cmtoledo@usp.br
[1]Laboratory of CNS Disease Modeling, Microbiology Department, Institute of Biomedical Sciences, Universidade de São Paulo, São Paulo, BR
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background
Breast cancer is a leading cause of female death worldwide, being a heterogenous disease with distinct molecular subtypes. Early and accretive cancer classification is crucial for the effective treatment of breast carcinomas. One of the most promising ways to investigate the pathogenesis of breast cancer is undoubtedly research involving RNA sequencing (RNA-seq), as bioinformatics has been emerging as a powerful tool with increased essentiality for oncological advances.

Results
Mutual information was a more effective feature selector than Differential Expression Analysis, even it being more fast computationally. Both were beaten by boruta but a cost of much more computational power and more genes, chi-squared was the worse method

Random forest and XGBoost were more effective in their optimizer task, promoting greater efficiency in identifying breast cancer subtypes. And support vectors were not efficient in classifying breast cancer subtypes.

Conclusions
The evolution of science, clinical analyzes and clinical practices are favored by the implementation of machine learning and scientific data analysis routines, mainly on breast cancer. As well as the evolution in the search for quality of life, as well as the improvement in treatments, the diagnosis of patients with breast cancer.

Keywords: RNA-seq; machine learning; breast cancer; subtype identification

1 Background

Breast cancer is the leading cause of female mortality. Heterogeneity of histopathological manifestations is one of the main causes of this prevalent letality, also making diagnosis difficult [1, 2].

Accretive and early cancer classification is crucial for effective treatment of breast carcinomas. For playing an important role in reducing the mortality rate, due to the increased probability of preventing the occurrence of metastases, favoring an immediate and timely treatment of this pathology.

One of the most promising ways to investigate the pathogenesis of breast cancer is undoubtedly research involving RNA sequencing (RNA-seq) high-throughput

genomic RNA-seq allows for 'system-level analysis'. This fact offers the ability to measure the expression state of thousands of genes in parallel [3, 4].

RNA-Seq biological information processing is replacing the widespread use of microarrays. The RNA seq app was originally developed for the study of gene transcription. But today it is known that it helps in the mapping of gene transcription and expression [5].

Related research demonstrates the application of bio-informative RNA-seq aids in the development of more accurate techniques for classification of breast carcinoma. Studies conducted with data from 877 patients aiming to understand the relationship of RNA-seq, methyloma and miRNA with breast cancer initially 20 potential genes were obtained. Being two genes with the greatest carcinogenic potential [6].

Despite several existing computational methods and proposed for the identification of different existing breast cancer subtypes there is still no clear answer. The different genes involved in different subtypes of carcinomas are still not well understood. And this is one of the gaps that need to be filled for the advancement of cancer treatment involving breast cancer.

Research involving bioinformatics is proving to be more and more essential for oncological advances. Mainly aimed at the development of gene selection methods. Such advances have contributed a lot to the evolution of genomics and clinical analysis. And in the future it will provide a great improvement in the quality of life. [7].

Identifying the different types of genes associated with each breast cancer subtype is important for advancing a more humanized and personalized treatment for different patients.

2 Results

2.1 Feature Selection

Boruta feature selector [8] considered 833 genes as important for the classification and rejected 11850, for the other feature selectors used (Mutual information, Chi-squared and top differentialy expressed genes) we chose the top 100 genes. We compare this lists to verify whether this lists contain the same genes. Boruta list contained all genes in mutual information list, but only 41 of chi-squared list and 46 of differentialy expressed genes (DEGs) list. Otherwise mutual information and DEGs had 14 genes in common, mutual information and chi-squared had nine and there wasn't any gene in common between DEGs list and chi-squared (Figure 1).

2.2 XGboost

The XGboost model resulted in one of the best predictors of breast cancer subtypes. We implemented XGboost after performing distinct types of feature selection and, albeit this model had a great performance regardless of the kind of feature selection employed, the implementation of XGboost with genes chosen based on mutual information was the best predictor of breast cancer subtypes, as quantitatively measured by both ROC AUC and F1 scores (Table 1).

We further visualized the classification of tumor samples with XGboost and features selected by mutual information through a confusion matrix plot, which demonstrates that XGboost was able to classify breast cancer samples appropriately in

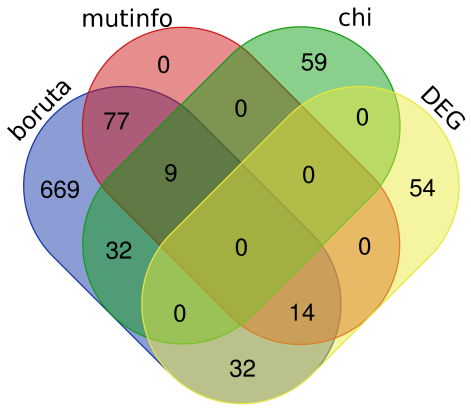

Figure 1 Venn diagram of common genes selected by each feature selection model

gene set	multinfo				chi			
model	XGBoost	Random	SVM	MLP	XGBoost	Random	SVM	MLP
AUC_score	0.978225	0.973492	0.500000	0.890841	0.942144	0.956097	0.50000	0.864612
F1_score	0.897632	0.839167	0.417926	0.807156	0.828302	0.845261	0.417926	0.702841
gene set	DEGs				Boruta			
model	XGBoost	Random	SVM	MLP	XGBoost	Random	SVM	MLP
AUC_score	0.951251	0.936947	0.500000	0.870438	0.977149	0.978106	0.500000	0.874820
F1_score	0.834283	0.806764	0.417926	0.752718	0.884179	0.866217	0.417926	0.799025

Table 1 AUC_score and F1_score from the best models chosen by randomized search. The models are XGBoost, Random Forest, Support Vector Machine and Multilayer Perceptron

the majority of cases (figure 2). Taken together, these results point out to XGboost as a promising machine learning approach to be used in breast cancer subtype identification.

2.3 Random forest

Random forest, on its turn, was also a good predictor of breast cancer subtypes, demonstrating elevated ROC AUC and F1 scores relative to the other models that were implemented (figure 3). In particular, the best approach with random forest was obtained upon the use of boruta as feature selection method. This highlights random forest as an alternative approach to the XGboost model, since both of them were shown to have a good performance in breast cancer subtype identification.

2.4 Support vector machines

Among all implemented machine learning approaches, support vector machines had the poorest outcome regarding the accurate classification of breast cancer subtypes for all feature selection methods employed (figure 4). It had the minor score possible in AUC showing that it wasn't any better than a coin toss. Maybe it could perform better under different hyperparameters set.

2.5 Multilayer perceptron

Multilayer perceptron showed a modest result, but was beaten by trees methods, showing that in this type of data trees methods tends to perform better than network data. Probably its performance could be enhanced by a better hyperparameter

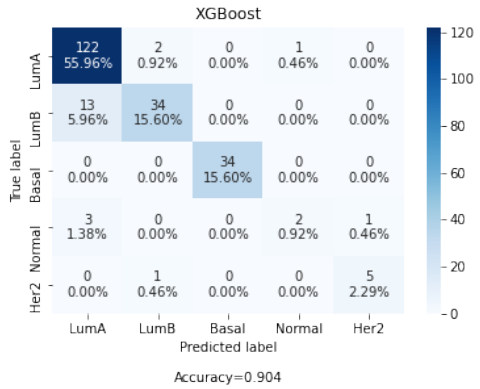

Figure 2 XGBoost confusion matrix using mutual information as gene set

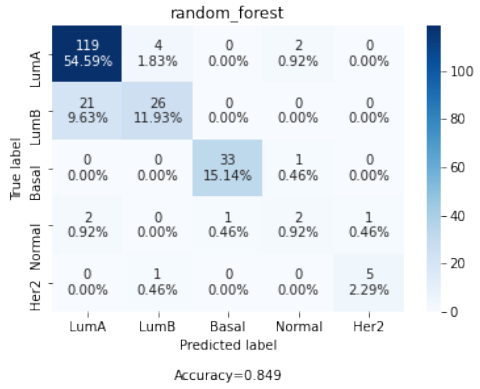

Figure 3 Random forest confusion matrix using mutual information as gene set

choice, showing that is a more complex model to fit than the trees methods used (figure 5).

2.6 Enrichment Analysis

Since mutual information and DEGs analysis were the best feature selectors we performed a KEGG pathway enrichment analysis [9] of than. Mutual information had three pathways enriched and DEGs had 5 (figure 6). Interestingly the same pathways enriched by genes selected by mutual information was also enriched by the genes differentialy expressed.

3 Discussion

Although boruta was the better model of feature selector it was more computationally expensive than the other methods and used a bigger number of genes. Since all genes in selected by mutual information was in the boruta we can consider that the gain of performance came from the information contained in the other genes.

Mutual information was the best feature selector, followed by DEGs analysis, still the analysis of mutual information is fastest of DEGs analysis showing that in this case a simpler statistic method is enough to select the best features. Chi-squared

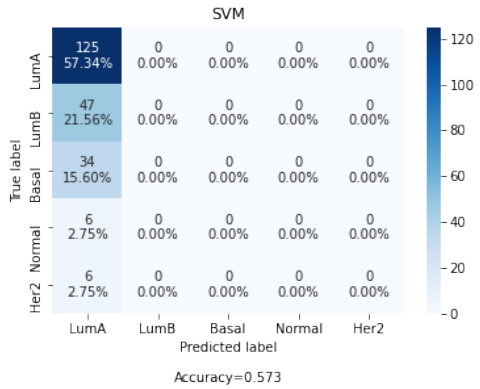

Figure 4 SVM confusion matrix using mutual information as gene set

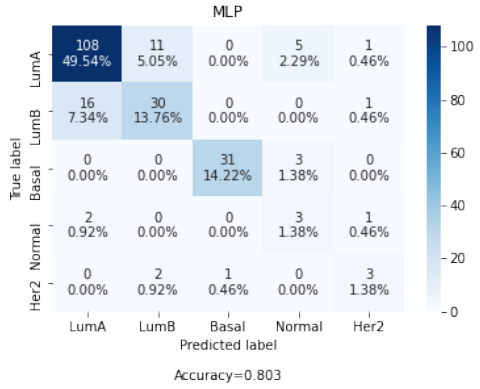

Figure 5 MLP confusion matrix using mutual information as gene set

had a slightly small result, but considering that it is the fastest method it can be considered a good model of feature selection.

Our DEGs analysis was made doing all cancers types versus normal cells types, maybe a better result could be achieved if instead was made each cancer type versus normal cells, but it would be 4 times more analysis.

Interestingly, although the three feature selectors models had a similar score it has low count of common genes. If we considered genes chosen by random the chance of getting n genes in common in two sets are given by a hypergeometric probability distribution:

$$P(n) = \frac{C_{100,n} C_{13212-100,100-n}}{C_{13212,100}} \tag{1}$$

Using this we can calculate the p-value of 14 genes in common ($< 10^{-13}$) and 9 genes in common ($< 10^{-7}$) so exist a correlation between the genes chosen by mutual information and those chosen by other feature selection methods.

Random forest was a great machine learning technique used. And it was remarkably efficient in solving the regression and classification problems that were targeted by the work.

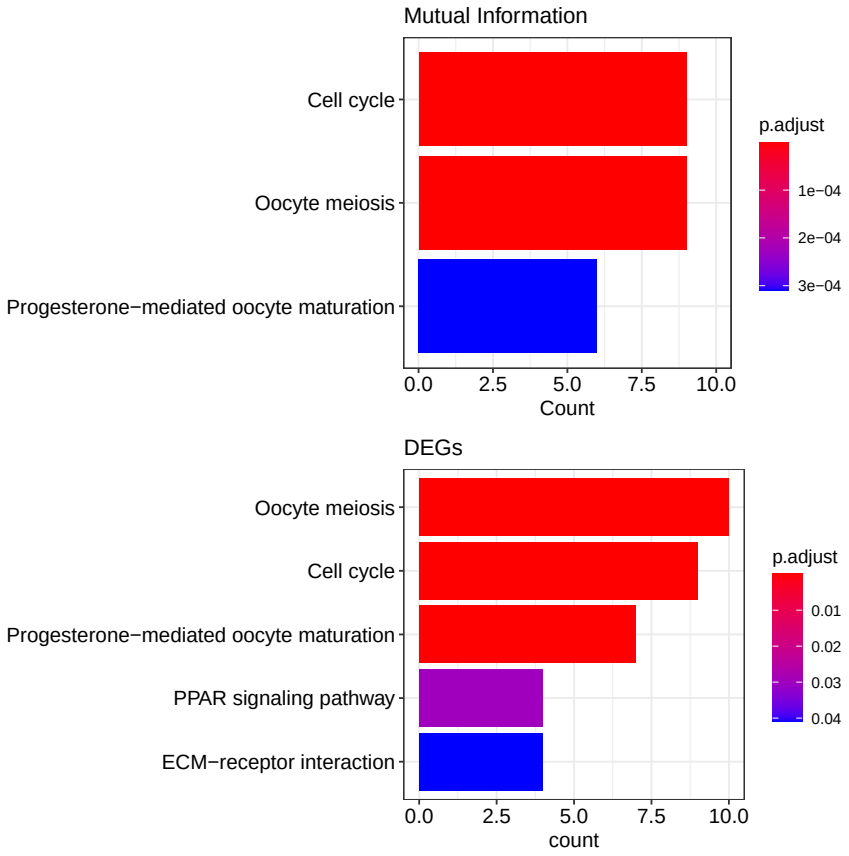

Figure 6 barplot of KEGG enrichment analysis performed in the two best feature selection methods

However, as shown in the results, XGBoost was more effective in its task as an optimizer, promoting greater efficiency in identifying breast cancer subtypes.

On the other hand, the machine learning method, support vectors was not demonstrated in the work of classifying them into breast cancer subtypes. Which makes it a less suitable method for solving this type of problem.

4 Conclusions

Mutual information statistics was proven the best feature selection considering its efficiency and velocity. DEGs analysis followed, showing that DEGs are not better predictors than genes with bigger mutual information. This is interesting considering that DEGs are though of having biological meaning while genes with bigger mutual information not necessarily.

Chi-squared could be a good feature selector if the necessity of velocity matters, it may be used as a pre-filter to lower the number of genes for further analysis. There was no advantage in using Boruta in this data.

Tree methods proven more useful to this kind of problem than those based in linear methods. There wasn't a significant difference between XGBoost and Random Tree, showing that both are very useful to this type of data. Besides they both had great results with all types of feature selection methods tested.

Methods

Data retrieval

The data used in our study come from The Cancer Genome Atlas (TCGA), a database where diverse genomics data are available for several cancer types. We obtained RNA sequencing (RNA-seq) data and corresponding clinical information from the TCGA breast cancer project, TCGA-BRCA, using the *R* package *TCGAbiolinks*. Out of the 1222 samples whose data were downloaded, 1102 corresponded to primary solid tumors, 7 to metastatic tumors, and 113 to solid normal tissues. In our analyses, we utilized only primary tumor samples.

Preliminary treatment of data

We filtered out 12 primary tumor samples which did not have a breast cancer subtype assignment, and 1090 samples remained to carry out downstream analyses. Thereafter, samples were divided into two distinct groups: a training group with 80% of the total number of samples; and a test group with 20% of this number.

Train and test raw count data were trimmed means of M-values (TMM)-normalized with *edgeR* [10, 11] in *R* in order to allow both within-sample and between-samples comparisons. As each sample had more than 56000 features - that is, genes whose expression was measured by RNA-seq -, we aimed to reduce this number to make our analyses more feasible. A pre-filtering was performed to remove genes that were not widely expressed across patients' cancer samples and whose counts were very evenly distributed among the different samples, being genes with low variance.

Feature selection

Subsequent to pre-filtering, different types of feature selection strategies were employed seeking the selection of the most predictive genes. We selected the most important features based on the following criteria:

mutual information

Mutual information measures the amount of information that a random variable tells about other variable. In our case we measures how much a gene expression tells about the cancer type. For this we used the feature selector implemented in scikit learn and select the top 100 more predictive genes

chi-squared method

We computed the chi-squared statistics between the features and the labels of our data using the feature selector implemented in scikit learn and selected the top 100 more predictive genes. One advantage of chi-squared is that this is the fastest feature selector used.

top differentially expressed genes between tumor samples and normal tissues

We used the R package *DESeq2* [12] to calculate the differential expression of genes between tumor samples and normal tissue, only the train data was used to the analysis to avoid data leakage. After that, we selected the top 100 genes most differentially expressed.

boruta
Boruta is a model of feature selection that tries to find all features relevant for prediction [8], we used boruta with 50 iterations to select all genes relevant for prediction.

Machine learning
As we aimed to create a classifier able to distinguish breast cancer subtypes, we implemented several machine learning algorithms to obtain a suitable model to predict to which subtype a breast tumor sample belongs. As machine learning models, we employed these models:

XGBoost
XGBoost is a model that uses gradient boosting algorithms to fit multiple trees models into data [13]. It is one of most used models in *Kaggle* competitions.

Random Forest
We used random forest model implemented in scikit-learn, this model uses an ensemble of decisions tree and select the classification most voted as result.

Support Vector Machine
We used the support vector machine implemented in scikit-learn, it was chosen due it high efficiency and ability to handle a big number a features

Multilayer Perceptron
For last, we implemented a deep neural network model known as multilayer perceptron, for this we also used scikit-learn implementation.

Randomized Search Cross Validation
To assure we used the best model we used 5-fold cross validation combined with random search to chose the best hyperparameters to each model, given a grid of hyperparams.

Measures
We calculated the *F1* and *AUC* scores of each model using each set of features selected and used this to compare models.

Enrichment Analysis
To understand if the features selected by our models had any biological meaning we performed a enrichment analysis using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in the genes selected by mutual information and by ones selected by differential expression analysis, we consider a cut-off of p-adjusted less than 0.05.

5 Declarations
5.1 Funding
This research was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, process number: $20/07450-5$), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, process number:).

5.2 Data and code availability

The data used in this work is publicly available for download through *TCGAbiolinks* in *R* or through Genomics Data Commons (GDC) Data Portal in the internet. All *R* and *Python* codes employed in our analyses may be found at https://github.com/cahmtoledo/IBI5031BreastCancerML

5.3 Competing interests

The authors declare that there were no competing interests throughout this work.

Author details

[1]Laboratory of CNS Disease Modeling, Microbiology Department, Institute of Biomedical Sciences, Universidade de São Paulo, São Paulo, BR. [2] Laboratory of Neurobiology and Stem Cells, Department of Cell and Developmental Biology, Institute of Biomedical Sciences, Universidade de São Paulo, São Paulo, Brasil. [3]Epilepsy research laboratory,Department of Neurosciences and Behavioral Sciences, Ribeirão Preto Medical School, Universidade de São Paulo, São Paulo, BR.

References

1. Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L., *et al.*: Sequence analysis of mutations and translocations across breast cancer subtypes. Nature **486**(7403), 405–409 (2012)
2. Yu, Z., Wang, Z., Yu, X., Zhang, Z.: Rna-seq-based breast cancer subtypes classification using machine learning approaches. Computational intelligence and neuroscience **2020** (2020)
3. Lal, S., Reed, A.E.M., de Luca, X.M., Simpson, P.T.: Molecular signatures in breast cancer. Methods **131**, 135–146 (2017)
4. Fu, C., Marczyk, M., Samuels, M., Trevarton, A.J., Qu, J., Lau, R., Du, L., Pappas, T., Sinn, B.V., Gould, R.E., *et al.*: Targeted rnaseq assay incorporating unique molecular identifiers for improved quantification of gene expression signatures and transcribed mutation fraction in fixed tumor samples. BMC cancer **21**(1), 1–10 (2021)
5. Castillo, D., Gálvez, J.M., Herrera, L.J., San Román, B., Rojas, F., Rojas, I.: Integration of rna-seq data with heterogeneous microarray data for breast cancer profiling. BMC bioinformatics **18**(1), 1–15 (2017)
6. Kothari, C., Osseni, M.A., Agbo, L., Ouellette, G., Déraspe, M., Laviolette, F., Corbeil, J., Lambert, J.-P., Diorio, C., Durocher, F.: Machine learning analysis identifies genes differentiating triple negative breast cancers. Scientific reports **10**(1), 1–15 (2020)
7. Koumakis, L., Kanterakis, A., Kartsaki, E., Chatzimina, M., Zervakis, M., Tsiknakis, M., Vassou, D., Kafetzopoulos, D., Marias, K., Moustakis, V., *et al.*: Minepath: mining for phenotype differential sub-paths in molecular pathways. PLoS computational biology **12**(11), 1005187 (2016)
8. Fay, M.P., Shaw, P.A.: Exact and asymptotic weighted logrank tests for interval censored data: the interval r package. Journal of statistical software **36**(2) (2010)
9. Kanehisa, M., Goto, S.: Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research **28**(1), 27–30 (2000)
10. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)
11. McCarthy, D.J., Chen, Y., Smyth, G.K.: Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. Nucleic acids research **40**(10), 4288–4297 (2012)
12. Love, M., Anders, S., Huber, W.: Deseq2 vignette. Genome Biol. doi **110** (2016)
13. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

Additional Files