RESEARCH ARTICLE

Beyond the textbook: a survey for breast cancer secondary features using Machine Learning methods

Fernanda Midori Abukawa[1*], Juliana N.C. Ferreira[2], Luiz Fernando C. Rodrigues[3] and Cássia S. Sanctos[1]

[*]Correspondence: fernanda.abukawa@butantan.gov.br
[1]Laboratório Especial de Toxinologia Aplicada - LETA, Instituto Butantan, São Paulo, Brazil
Full list of author information is available at the end of the article

Abstract

Background: Breast cancer is the leading cause of death from cancer among women, and the high mortality rate of this type of cancer is largely due to the complexity of invasive breast cancer and its significantly varied clinical outcomes. For this reason, the improvement of accuracy in the prediction of breast cancer has important significance and becomes one of the major research areas in the medical and biomedical fields.

Results: Performance analysis of supervised Machine Learning models indicates a better performance of Random Forest and Support Vector Machine models for classification of the triple-negative breast cancer subtype.

Conclusion: This study demonstrate that clinical pathology report and mutation profiling can be used with supervised Machine Learning methods to classify the triple-negative breast cancer subtype. The classification models implemented in this study could provide a reasonable prediction of this subtype of breast cancer based on the mutation profile of patients.

Keywords: Machine learning; Breast Cancer prediction; Genomic data

Introduction

Breast cancer is the most frequently occurring cancer in women, accounting for one-third of newly diagnosed malignancies, and one of the main causes of cancer deaths around the world [1] [2]. Survival rates for all age groups have increased within the past decades, with the five years survival rate being almost 100% if detected at its earliest stage, but as low as 15% when detected at the latest stage, according to the statistics of Cancer Research UK [3]. Previous studies have shown the association between survival and the diagnosis stage, related with tumor size and metastasis to the lymph nodes or beyond [4] [5] [6].

Clinically, breast cancer classification is done with tumor morphological characteristics into different types, and the histological approach promotes sub-classification [7]. The recent development of new technologies and the increase of knowledge about the tumorigenesis progress, resulted in the identification of new biomarkers and novel subtypes, helping in more accurate disease management, but making the understanding of breast cancer heterogeneity even more complex [8] [7] [9].

The breast cancer is a highly heterogeneous disease, encompassing a number of biologically distinct entities with specific pathological features and biological behaviors [1] [10]. This cancer is diagnosed according to standardized pathological criteria, and the immunohistochemistry (IHC) markers, like estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2), along with clinicopathological variables have been widely used for prognosis, prediction and treatment selection [11] [12]. The IHC markers are known to mediate cell growth signaling but are also used for breast cancer subtyping [13] [14] [15].

The estrogen receptor alpha (ER) is expressed in approximately 70% of breast cancers. ER is a steroid hormone receptor and a transcription factor that when activated by estrogen, activates oncogenic growth pathways in cancer cells. Expression of PR is also a marker of ER signalling [16]. The use of endocrine agents to downregulate ER signalling is the primary systemic therapy for ER+ or PR+ breast cancers [17]. The HER2 is a transmembrane receptor tyrosine kinase in the epidermal growth factor receptor family that can be amplified or overexpressed in approximately 20% of breast cancer [18].

The IHC markers identifies breast tumors into four basic subgroups: i) tumors with either ER or PR positivity, and HER2 negativity ([ER+—PR+]HER2-); ii) tumors with either ER or PR positivity, and HER2 positivity ([ER+—PR+]HER2+); iii) tumors with ER and PR negativity, and HER2 positivity, also named HER2 positive (ER-PR-HER2+); iv) tumors with ER, PR, HER2 negativity, also named triple negative (ER-PR-HER2-) [8] [7]. In general, tumors with both ER and PR negativity have relatively poorer prognosis than tumors with either ER or PR positivity [8] [7]. Triple-negative breast cancer accounts for 10-20% of all breast cancer, and are characterized by lack of the ER, PR and HER2 receptors, as previously described [19]. This is the most aggressive form of breast cancer and is very heterogeneous [20]. The complexity of this subtype is due to its high risk of relapse and overall survival [21].

The utilization of machine learning (ML) approaches in the medical fields may be considered a good assistance in the decision making process of physicians [22]. ML techniques are playing a significant role in diagnosis and prognosis of breast cancer, and it had been used for decades to classify tumors and other malignancies, predict sequences of genes responsible for cancer and to determine the prognostic [3] [23]. Recent studies are focusing on applying ML algorithms on data of gene expression as a new approach to find genes deferentially expressed between breast cancer subtypes. ML was used to identify key genes of triple negative tumors [24], prediction of estrogen receptor status [25] and even treatment gene-related biomarkers [26]. Although a great number of algorithms have achieved very high accuracy for breast cancer diagnosis, the development of different improved ML algorithms is still necessary for providing more alternative solutions to breast cancer data [3].

In the present study we investigate breast cancer features to classify the most aggressive subtype of breast cancer. Such findings have the potential to reveal different metabolic pathways and relevant features which could be linked to this disease. The aims of this study were to assess the genes that are important for predicting the triple-negative cancer and the minimal amount of features that can be used to predict triple-negative breast cancer. This subtype of breast cancer is of special interest to us because it is heterogeneous and has the issue of not having the biomarkers receptors that are common for most hormone therapies, therefore requiring different treatments. The ML approach can serve as a useful tool in diagnosis, guiding future clinical and expression studies in this area.

Materials and Methods

Dataset

The dataset contains data of 1918 breast tumor specimens from MSK, Cancer Cell 2018 dataset [27], available in the *cBioPortal for Cancer Genomics*. The histologic subtypes of breast cancer were determined based on clinical pathology reports of the primary tumors and metastatic sites. The mutation profiling was performed by the authors to identify somatic mutations, DNA copy number alterations and select rearrangements in 468 cancer-associated genes [28] [29]. We associated mutations in these genes with specific receptor types and histologies.

Data Preprocessing and Feature Selection

The first step of the process were to merge the data. The dataset was a collection of Microsoft Excel Open XML Spreadsheet (XLSX) files that contained clinical as well as genomic data identified by patients and sample IDs. The XLSX spreadsheets were merged according to patient and sample IDs (used as primary keys) in order to obtain a consolidated data view.

The resulting spreadsheet contained only female patient IDs, receptor status and genomic information from patients with mutation in at least one gene, consisting of 1907 samples of 423 genes. This proportion is difficult for learning, so for dimensionality reduction the Sequential Forward Selection (SFS) procedure was performed by means of the mlxtend Python package [30]. This was chosen due to it being faster and simpler than more sophisticated algorithms such as floating methods [31]. This is a deterministic search algorithm to feature selection, which starts from an empty subset of genes and sequentially selects genes, one at a time, until no further improvement is achieved in the evaluation function value. We fixed a specific number of features estimated from the clustering method with the K-means algorithm, with the *elbow method* as the evaluation metric employed, and set as a maximum number of features to be selected by the SFS method.

The final dataset was characterized in two types, according to the mutation profile. We transformed the gene content information in 0 when the patient had no mutation in the gene, and 1 otherwise. This was called the treated dataset, while the untreated had the original gene information values from [27], which present specific values for different gene mutations. Both datasets were used for results comparison.

Models used and selection

Different models were tested for data classification in an attempt to balance classification accuracy and resulting biological insight. Random Forest, for example, allows for visualization of features in an hierarchical way to understand what are the most relevant factors for classification, while others, such as SVM, do not have this ability. The Random Forest model combines several tree classifiers, decreasing variance and offering stability. Also, by having a number of decision trees participating in the process it significantly reduces overfitting [32]. SVM usually can recognize subtle patterns in complex datasets [33], and has been widely used in cancer classification since gene expression data became available [34].

The Extreme Gradient Boosting (XGBoost), in the same way as Random Forest, is a ensemble method, but differently it selects random samples with replacement and creates a tree over a weighted training dataset. This algorithm is widely adopted due to its good results, and also decreases variance, offers high stability and increasingly reduces error rate, although it can be prone to overfitting [35] [36]. Finally, the K-nearest Neighbors (KNN) was used to classify based on distances from data points without making assumptions on data distribution [37]. This algorithm sorts which points are closest to it in terms of distance, then each point is classified as belonging to a class with the closest data points to it, and when the distribution of classes is skewed the point might be classified by majority voting. GridSearch was used to find the optimized model parameters for each algorithm with a 5-fold cross-validation, which is an approach to evaluate the performance of a ML model on a validation set. Then, the data was trained with the best parameters found and the predictor model was obtained.

To better understand algorithms outputs and how much of the data was correctly or incorrectly classified, the confusion matrix was employed to reveal the amount of false negatives and positives. To measure model results the classification report was implemented with metrics such as *F1-score*, *precision*, *recall* and *accuracy*. The F1-score is of special interest because the dataset has only about 9% of triple-negative breast cancer samples, so it provided more specific information regarding this part of the data than the accuracy and was used as the main criterion for model selection.

All of the analysis were performed with intense use of the Scikit Learn library [38].

Results and Discussion

The SFS method applied to each algorithm and dataset resulted in the selection of distinct numbers of features that were used to compare the models. Tables 1 and 2 show the weighted average of accuracy, precision, F1-score, recall and the minimal number of features for different classification techniques, while Tables 3 and 4 display the confusion matrix of classification results of each model. The clustering method with the K-means resulted in 42 clusters that were used as the number of features to select in the implementation of the SFS, where this number of features needs to be less than the full feature set.

All methods have their own limitations and strengths specific to the type of application. KNN, XGBoost and SVM (best results with the RVB kernel) models resulted in high, and similar, values of F1-score for the treated (0.89, 0.88, 0.90, respectively) and untreated (0.89, 0.89, 0.90, respectively) datasets. This pattern was different for the Random Forest algorithm since this score was 0.91 for the treated dataset and 0.43 for the untreated. In general, the Random Forest model was a better classifier for the treated dataset and SVM for the untreated, according to the F1-score, which is defined as the mean between precision and recall, and it is used as a statistical measure to rate performance.

The other parameters in the classification report, such as precision, recall and accuracy demonstrated a similar pattern as described by the F1-score. However, the Random Forest had the precision value (0.91) similar to the other models (KNN: 0.90, XGBoost: 0.90, and SVM: 0.92) in the untreated dataset, despite the recall and accuracy showing the smallest values. As for the minimal number of features for each ML model, the Random Forest had similar values for both datasets (treated:

42, and untreated: 41). On the other hand, the feature selection for the KNN had a higher outcome with the treated dataset (30) then the untreated (12). For the XGBoost and SVM the opposite was observed, the highest values were with the untreated dataset (27 and 37, respectively) and the smallest with the treated (6 and 29, respectively). In summary, the XGBoost has the smallest minimal number of features in the treated dataset, as for the untreated the smallest minimal number is for the KNN model. For both datasets, Random Forest has the biggest minimal number of features.

Comparison of the confusion matrix between the treated and untreated datasets indicated that the KNN, XGBoost and SVM models had similar values for the true/false positives and negatives. The XGBoost had the smallest value of true positives (19) in the treated dataset, while the Random Forest had the highest (51). As for the true negatives, XGBoost and SVM had the highest values, both with 1722, while the RF had the smallest (1710). Concerning the untreated dataset, Random Forest had the highest true positive value (170) and XGBoost the smallest (24), while for the true negative values the SVM model had the highest (1722) and Random Forest the smallest (509). Even though the Random Forest had a better performance classifying the triple-negative for the untreated dataset, this model had the smallest values of F1-score and accuracy for this particular dataset, so the results must be interpreted carefully.

Feature selection for each model resulted in the identification of some genes in common between most or even all algorithms for the treated dataset (CCND3 and ETV6) and the untreated (CCND2 and CCND3). These genes could be linked with specific metabolic pathways of triple-negative breast cancer, helping the classification of this subtype. There are already studies suggesting a connection between these genes and breast cancer [39] [40], so our findings regarding feature selection seem to be of biological significance.

There were some limitations in the current study. The proportion of samples with triple negative breast cancer in the dataset were around 9.2% while samples without triple negative were 90.8%, representing an imbalance in the classification of these two classes. The learning objective of these models is to construct a discriminative classifier, which can be understood as finding the true hypothesis from the entire possible hypothesis space. Based on total accuracy, it was shown that all classification methods performed almost similarly for classifying triple-negative breast cancer in both types of dataset, with exception for the Random Forest with the untreated data.

Many studies have compared the performance of classifiers to predict an outcome of interest. Different classification techniques were developed for breast cancer diagnosis, and the accuracy of many of them was evaluated using the dataset taken from Wisconsin breast cancer database [41]. For example, in [42] the Naive Bayes and KNN algorithms were used for breast cancer classification, with the KNN achieving a higher efficiency (97.51%). In the study of [37], the breast cancer classification was employed with SVM and KNN models, resulting in 99.68% of accuracy with the SVM in the training phase.

Similarly with the present study, [43] presented a comparative analysis of four widely used ML techniques: Multi-Layer Perceptron, SVM, KNN and Decision Tree, with the dataset from the Wisconsin breast cancer database previously mentioned. This study showed a better performance of the Multi-Layer Perceptron (98.12%) as compared to the other models (Decision Tree: 96.05%, SVM: 96.19%, and KNN: 94.29%), with also performing better then the others when the cross-validation metric was used.

Brenton et al. [44], indicates that, given the complexity and heterogeneity of breast cancer, a strategical approach is to use both clinical data and gene profiling. This strategy was used by [45], with the combination of clinical information, omics data, ML and bioinformatics to identify candidate genes that could be linked with triple-negative breast cancer classification. Another study [46] demonstrated that clinically distinguishable breast cancer subtypes can be identified solely based on somatic mutation profile data. Their classification model can be used to predict unknown subtypes of breast cancer, given the somatic mutation profile of a patient. Our methodology also allows the classification of breast cancer subtypes, but in our case we restrict to only the triple-negative subtype as a first step. This was due to the fact that its one of the most aggressive form of breast cancer, with the highest mortality due to high rate of relapse, resistance and lack of an effective treatment [45]. The categorization of breast cancer through mutations and alterations in genes and metabolic pathways could influence clinical pratice and potential novel approaches for treatments.

The used ML techniques in this study are non-parametric and provide efficient solutions for classification problems without considering any special assumption regarding the distribution of data. They also deal with non linearity and high order interactions. However, the performance of a method is data dependent and, in general, there is no method that always performs as the best technique in classification problems.

To continually evaluate this methodology for breast cancer classification, these models could be trained with different datasets. This approach can also be applied to classify and predict other breast cancer subtypes than the triple-negative, and might even be adequate to provide candidate genes that could be further pursued for the classification of this type of cancer, like [45]. This study used ML algorithms to analyse breast cancer data and have identified two potential genes that could be successfully differentiate triple-negative subtype from non triple-negative, despite of their heterogeneity. This is important since these genes could serve as biomarkers for this aggressive subtype of breast cancer, having the potential to be developed as therapeutics targets.

Conclusion

Breast cancers are highly heterogeneous diseases. Therefore, accurate classification of this type of cancer is an important step towards making accurate treatment decisions. The focus of this study was to evaluate the performance of different classifiers in detecting the triple-negative subtype of breast cancer patients. This study demonstrate that clinical pathology report and mutation profiling can be used with supervised ML methods to classify this subtype of breast cancer, and the classification models implemented could provide a reasonable prediction based on the mutation profile of patients. ML techniques have been applied extensively in various medical applications. Breast cancer classification based on gene expression data remains a challenging task, especially to identify potential points for therapeutics intervention, understand tumour behaviour and facilitate drug development. Although a great number of algorithms have achieved very high accuracy with breast cancer classification, the development of different and improved ML algorithms is still necessary for more alternative solutions to the complex breast cancer data. These algorithms can help to build healthcare systems to assist physicians as a second opinion for their final decision, improving decision accuracies.

Acknowledgements
We thank professor Marcelo Reis for his helpful suggestions to improve this research and our colleagues from IBI5031.

Funding
This work was supported by Fundação Butantan of Instituto Butantan; CNPq (88887.369943/2019-00) and CAPES (88882.332864/2019-01)

Abbreviations
ML: Machine Learning; RF: Random Forest; KNN: K-nearest neighbors; SVM: Support Vector Machines;

Availability of data and materials
Code and datasets are freely available at project home page and cBioPortal for Cancer Genomics

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors performed the experiments and analyses, designed the research and wrote the manuscript.

Author details
[1]Laboratório Especial de Toxinologia Aplicada - LETA, Instituto Butantan, São Paulo, Brazil. [2]Oceanographic Institute, University of São Paulo, São Paulo, Brazil. [3]Institute of Physics, University of São Paulo, São Paulo, Brazil.

References

1. Spitale, A., Mazzola, P., Soldini, D., Mazzucchelli, L., Bordoni, A.: Breast cancer classification according to immunohistochemical markers: clinicopathologic features and short-term survival analysis in a population-based study from the south of switzerland. Annals of oncology **20**(4), 628–635 (2009)
2. Torre, L.A., Bray, F., Siegel, R.L., Ferlay, J., Lortet-Tieulent, J., Jemal, A.: Global cancer statistics, 2012. CA: a cancer journal for clinicians **65**(2), 87–108 (2015)
3. Yue, W., Wang, Z., Chen, H., Payne, A., Liu, X.: Machine learning with applications in breast cancer diagnosis and prognosis. Designs **2**(2), 13 (2018)
4. Smigal, C., Jemal, A., Ward, E., Cokkinides, V., Smith, R., Howe, H.L., Thun, M.: Trends in breast cancer by race and ethnicity: update 2006. CA: a cancer journal for clinicians **56**(3), 168–183 (2006)
5. Tryggvadóttir, L., Gislum, M., Bray, F., Klint, Å., Hakulinen, T., Storm, H.H., Engholm, G.: Trends in the survival of patients diagnosed with breast cancer in the nordic countries 1964–2003 followed up to the end of 2006. Acta Oncologica **49**(5), 624–631 (2010)
6. Rosso, S., Gondos, A., Zanetti, R., Bray, F., Zakelj, M., Zagar, T., Smailyte, G., Ponti, A., Brewster, D.H., Voogd, A.C., *et al.*: Up-to-date estimates of breast cancer survival for the years 2000–2004 in 11 european countries: the role of screening and a comparison with data from the united states. European Journal of Cancer **46**(18), 3351–3357 (2010)
7. Tao, Z., Shi, A., Lu, C., Song, T., Zhang, Z., Zhao, J.: Breast cancer: epidemiology and etiology. Cell biochemistry and biophysics **72**(2), 333–338 (2015)
8. Dai, X., Xiang, L., Li, T., Bai, Z.: Cancer hallmarks, biomarkers and breast cancer molecular subtypes. Journal of Cancer **7**(10), 1281 (2016)
9. Nicolini, A., Ferrari, P., Duffy, M.J.: Prognostic and predictive biomarkers in breast cancer: past, present and future. In: Seminars in Cancer Biology, vol. 52, pp. 56–73 (2018). Elsevier
10. Tang, P., Wang, J., Bourne, P.: Molecular classifications of breast carcinoma with similar terminology and different definitions: are they the same? Human pathology **39**(4), 506–513 (2008)
11. Cheang, M.C., Chia, S.K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P.S., Parker, J.S., *et al.*: Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. JNCI: Journal of the National Cancer Institute **101**(10), 736–750 (2009)
12. Vallejos, C.S., Gómez, H.L., Cruz, W.R., Pinto, J.A., Dyer, R.R., Velarde, R., Suazo, J.F., Neciosup, S.P., León, M., Miguel, A., *et al.*: Breast cancer classification according to immunohistochemistry markers: subtypes and association with clinicopathologic variables in a peruvian hospital database. Clinical breast cancer **10**(4), 294–300 (2010)
13. Fulford, L., Easton, D., Reis-Filho, J., Sofronis, A., Gillett, C., Lakhani, S., Hanby, A.: Specific morphological features predictive for the basal phenotype in grade 3 invasive ductal carcinoma of breast. Histopathology **49**(1), 22–34 (2006)

14. Lippman, M., Bolan, G., Huff, K.: The effects of glucocorticoids and progesterone on hormone-responsive human breast cancer in long-term tissue culture. Cancer research **36**(12), 4602–4609 (1976)
15. Moon, R.C., Pike, M., Siiteri, P., Welsch, C.: Influence of pregnancy and lactation on experimental mammary carcinogenesis. Banbury report **8**, 353–364 (1981)
16. Joshi, H., Press, M.F.: Molecular oncology of breast cancer, 282–307 (2018)
17. Waks, A.G., Winer, E.P.: Breast cancer treatment: a review. Jama **321**(3), 288–300 (2019)
18. Cameron, D., Piccart-Gebhart, M.J., Gelber, R.D., Procter, M., Goldhirsch, A., de Azambuja, E., Castro Jr, G., Untch, M., Smith, I., Gianni, L., *et al.*: 11 years' follow-up of trastuzumab after adjuvant chemotherapy in her2-positive early breast cancer: final analysis of the herceptin adjuvant (hera) trial. The Lancet **389**(10075), 1195–1205 (2017)
19. Podo, F., Buydens, L.M., Degani, H., Hilhorst, R., Klipp, E., Gribbestad, I.S., Van Huffel, S., van Laarhoven, H.W., Luts, J., Monleon, D., *et al.*: Triple-negative breast cancer: present challenges and new perspectives. Molecular oncology **4**(3), 209–229 (2010)
20. Karaayvaz, M., Cristea, S., Gillespie, S.M., Patel, A.P., Mylvaganam, R., Luo, C.C., Specht, M.C., Bernstein, B.E., Michor, F., Ellisen, L.W.: Unravelling subclonal heterogeneity and aggressive disease states in tnbc through single-cell rna-seq. Nature communications **9**(1), 1–10 (2018)
21. Mustacchi, G., De Laurentiis, M.: The role of taxanes in triple-negative breast cancer: literature review. Drug design, development and therapy **9**, 4303 (2015)
22. Agarap, A.F.M.: On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In: Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, pp. 5–9 (2018)
23. Chakraborty, J., Midya, A., Rabidas, R.: Computer-aided detection and diagnosis of mammographic masses using multi-resolution analysis of oriented tissue patterns. Expert Systems with Applications **99**, 168–179 (2018)
24. Naorem, L.D., Muthaiyan, M., Venkatesan, A.: Integrated network analysis and machine learning approach for the identification of key genes of triple-negative breast cancer. Journal of cellular biochemistry **120**(4), 6154–6167 (2019)
25. Alakwaa, F.M., Chaudhary, K., Garmire, L.X.: Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. Journal of proteome research **17**(1), 337–347 (2018)
26. Tabl, A.A., Alkhateeb, A., ElMaraghy, W., Rueda, L., Ngom, A.: A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. Frontiers in Genetics **10**, 256 (2019)
27. Razavi, P., Chang, M.T., Xu, G., Bandlamudi, C., Ross, D.S., Vasan, N., Cai, Y., Bielski, C.M., Donoghue, M.T., Jonsson, P., *et al.*: The genomic landscape of endocrine-resistant advanced breast cancers. Cancer cell **34**(3), 427–438 (2018)
28. Cheng, D.T., Mitchell, T.N., Zehir, A., Shah, R.H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z.Y., Won, H.H., Scott, S.N., *et al.*: Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (msk-impact): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. The Journal of molecular diagnostics **17**(3), 251–264 (2015)
29. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., *et al.*: Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nature medicine **23**(6), 703 (2017)
30. Raschka, S.: Mlxtend: providing machine learning and data science utilities and extensions to python's scientific computing stack. Journal of open source software **3**(24), 638 (2018)
31. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern recognition letters **15**(11), 1119–1125 (1994)
32. Kabir, M.F., Ludwig, S.: Classification of breast cancer risk factors using several resampling approaches, pp. 1243–1248 (2018). doi:10.1109/ICMLA.2018.00202
33. Aruna S, R.S.: A novel svm based cssffs feature selection algorithm for detecting breast cancer. International Journal of Computer Applications **31**(8), 14–20 (2011)
34. Huang, S., Cai, N., Pacheco, P., Narrandes, S., Wang, Y., Xu, W.: Applications of support vector machine (svm) learning in cancer genomics. Cancer genomics & proteomics **15**(1), 41–51 (2018)
35. Ye, L., Lee, T.-S., Chi, R.: A hybrid machine learning scheme to analyze the risk factors of breast cancer outcome in patients with diabetes mellitus. Journal of Universal Computer Science **24**, 665–681 (2018)
36. Friedman, J.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29** (2001). doi:10.1214/aos/1013203451
37. Islam, M.M., Iqbal, H., Haque, M.R., Hasan, M.K.: Prediction of breast cancer using support vector machine and k-nearest neighbors. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 226–229 (2017). IEEE
38. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
39. Justenhoven, C., Pierl, C.B., Haas, S., Fischer, H.-P., Hamann, U., Baisch, C., Harth, V., Spickenheuer, A., Rabstein, S., Vollmert, C., *et al.*: Polymorphic loci of e2f2, ccnd1 and ccnd3 are associated with her2 status of breast tumors. International journal of cancer **124**(9), 2077–2081 (2009)
40. Letessier, A., Ginestier, C., Charafe-Jauffret, E., Cervera, N., Adélaïde, J., Gelsi-Boyer, V., Ahomadegbe, J.-C., Benard, J., Jacquemier, J., Birnbaum, D., *et al.*: Etv6 gene rearrangements in invasive breast carcinoma. Genes, Chromosomes and Cancer **44**(1), 103–108 (2005)
41. Zhou, Z.-H., Jiang, Y., Yang, Y.-B., Chen, S.-F.: Lung cancer cell identification based on artificial neural network ensembles. Artificial Intelligence in Medicine **24**(1), 25–36 (2002)
42. Amrane, M., Oukid, S., Gagaoua, I., Ensarİ, T.: Breast cancer classification using machine learning. In: 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), pp. 1–4 (2018). IEEE
43. Gupta, M., Gupta, B.: A comparative study of breast cancer diagnosis using supervised machine learning

techniques. In: 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), pp. 997–1002 (2018). IEEE

44. Brenton, J.D., Carey, L.A., Ahmed, A.A., Caldas, C.: Molecular classification and molecular forecasting of breast cancer: ready for clinical application? Journal of clinical oncology **23**(29), 7350–7360 (2005)
45. 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)
46. Vural, S., Wang, X., Guda, C.: Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. BMC systems biology **10**(3), 62 (2016)

Tables

Table 1 Classification report and minimal number of features comparison between ML algorithms for treated data (*SVM kernel: RBF).

	Random Forest	KNN	XGBoost	SVM*
Precision	0.91	0.91	0.89	0.91
Recall	0.92	0.92	0.91	0.92
F1-score	0.91	0.89	0.88	0.90
Accuracy	0.92	0.92	0.91	0.92
n features	42	30	6	29

Table 2 Classification report and minimal number of features comparison between ML algorithms for untreated data (*SVM kernel: RBF).

	Random Forest	KNN	XGBoost	SVM*
Precision	0.91	0.90	0.90	0.92
Recall	0.35	0.92	0.91	0.92
F1-score	0.43	0.89	0.89	0.90
Accuracy	0.35	0.92	0.91	0.92
n features	41	12	27	37

Table 3 Confusion matrix comparison betwween ML algorithms treated data (*SVM kernel: RBF).

	Random Forest	KNN	XGBoost	SVM*
True Positive	51	31	19	34
False Positive	21	10	9	9
True Negative	1710	1721	1722	1722
False Negative	125	145	157	142

Table 4 Confusion matrix comparison betwween ML algorithms untreated data (*SVM kernel: RBF).

	Random Forest	KNN	XGBoost	SVM*
True Positive	170	33	24	41
False Positive	1222	17	11	9
True Negative	509	1714	1720	1722
False Negative	6	143	152	135