UNIVERSIDADE DE SÃO PAULO
INSTITUTO BUTANTAN

Research project proposal

Beyond the textbook: a survey for breast cancer secondary features using machine learning methods

Students: Group 3 - Cássia Sampaio Sanctos, Fernanda Midori Abukawa, Juliana Correa Neiva Ferreira, Luiz Fernando de Camargo Rodrigues
Supervisor: Dr. Marcelo da Silva Reis

SÃO PAULO
2020

ABSTRACT

One of the most challenging tasks in the biological sciences is the study of cancer because of its tissue ubiquity, complexity and disease mortality. At the same time, human life expectancy is gradually increasing, mainly due to advances in the medical field, however older age also implies in higher cancer risk. Therefore, a better understanding on this type of disease becomes an increasingly pressing issue for scientists to tackle. Fortunately, scientific research has become more interdisciplinary, recent research in cancer can benefit from contributions from fields ranging from physics to computer science, which allows novel approaches and experimental techniques. An important new approach in medicine benefits from the increasing processing capacity of personal computers and large data repositories: it consists in the use of machine learning (ML) for regression and classification. It is especially useful, for example, in aiding physicians on deciding between different treatments for each patient, or searching for new features and defining characteristics of different types of cancer. The most frequent type of cancer in women is breast cancer, a highly heterogeneous disease. Public databases for breast cancer include clinical and molecular data, which may be related to protein expression, mRNA or genetic mutation, allowing for machine learning from different experimental observations. The earlier the diagnosis for breast cancer, the better the patient prognosis. The present proposal aims to apply ML to study breast cancer features regarding gene mutation, and its relation to clinical features, such as the widely used immunohistochemistry markers. This study also aims to search for the minimal set of features that define triple-negative breast cancer, which is a heterogeneous subtype of breast cancer with relapse growth patterns and that requires special treatment due to its specificities. We will use mainly the Random Forest algorithm due to its capacity to highlight underlying biological implications of our findings, but other methods may be necessary for data classification. We believe the project can benefit from the diverse backgrounds and interests of our members, and that it is a high risk/high gain project due to its potential impact on breast cancer detection, with possibility of better understanding the relationship between genetic and clinical features of patients.

INTRODUCTION AND JUSTIFICATION

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 (Spitale et al., 2009; Torre et al., 2015). Although survival rates for all age groups have increased within the past decades, studies have shown the association between survival and the diagnosis stage, related with tumor size and metastasis to the lymph nodes or beyond (Smigal et al., 2006; Tryggvadóttir et al., 2010; Rosso et al., 2010). According to the statistics of Cancer Research UK, the five years survival rate for breast cancer is almost 100% if detected at its earliest stage, but can be as low as 15% when detected at the latest stage (Yue et al., 2018).

This cancer is a highly heterogeneous disease, encompassing a number of biologically distinct entities with specific pathologic features and biological behaviors (Spitale et al., 2009; Tang et al., 2008). Clinically, breast cancer classification is done with tumor morphological characteristics into different types, and the histological approach promotes sub-classification (Tao et al., 2015). The recent development of new technologies and the increase of knowledge about the tumorigenesis progess, 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 (Dai et al., 2016; Tao et al., 2015; Nicolini et al., 2018).

Breast cancer is diagnosed according to standardized pathologic criteria. The most common breast cancer histology is invasive ductal carcinoma, which occurs in 50-75% of patients, followed by invasive lobular carcinoma (5-15% of patients), with mixed ductal/lobular carcinomas and other rarer histologies making up to the remaining of patients (Dillon et al., 2010).

Immunohistochemistry (IHC) markers, like estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2), along with clinicopathological variables, such as tumor size and grade, nodal involvement, histologic type and surgical margins, have been widely used for prognosis, prediction and treatment selection (Cheang et al., 2009; Vallejos et al., 2010). The IHC markers are known to mediate cell growth signaling but are also used for breast cancer subtyping (Fulford et al., 2006; Lippman et al., 1976; Moon et al., 1981). A summary of breast cancer subtypes can be found in Table 1.

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-) (Dai et al., 2016; Tao et al., 2015). In general, tumors with both ER and PR negativity have relatively poorer prognosis than tumors with either ER or PR positivity (Dai et al., 2016; Tao et al., 2015).

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 (Joshi and Press, 2018). The use of endocrine agents to downregulate ER signalling is the primary systemic therapy for ER+ ou PR+ breast cancers (Waks and Winer, 2019). 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 (Cameron et al., 2017).

Table 1: Summary of breast cancer tumor molecular subtypes (modified from Dai et al., 2018)

Subtype	**Alias**	**Biomarker Status**	**Grade**	**Outcome**	**Prevalence**
Luminal	Luminal A	[ER+ — PR+] HER2-KI67-	1 — 2	Good	23.7%
	Luminal B	[ER+ — PR+] HER2-KI67+	2 — 3	Intermediate	38.8%
		[ER+ — PR+] HER2+KI67-		Poor	14%
HER2 positive	HER2 over-expression	ER-PR-HER2+	2 — 3	Poor	11.2%
Triple negative	Basal	ER-PR-HER2-, basal marker+	3	Poor	10-25%
	Claudin-low	ER-PR-HER2-, EMT marker+, stem cell marker+, claudin-	3	Poor	7-14%
	Metaplastic breast cancer	ER-PR-HER2-, EMT marker+, stem cell marker+	3	Poor	1%
	Interferon-rich	ER-PR-HER2-, interferon regulated genes+	3	Intermadiate	10%
Molecular apocrine cancer (MAC)	MAC	ER-PR-AR+	2 — 3	Poor	13.2%

The utilization of machine learning (ML) approaches in the medical fields may be considered a good assistance in the decision making process of physicians (Agarap, 2018). ML techniques are playing a significant role in diagnosis and prognosis of breast cancer, it had been used for decades to classify tumors and other malignancies, predict sequences of genes responsible for cancer and determine the prognostic (Yue et al., 2018; Chakraborty et al., 2018). For instance, many efforts were made for developing computer-aided diagnosis/detection (CAD) algorithms that are based on the advances of digital image processing, pattern recognition and artifical intelligence. The CAD systems are expected to overcome the operator dependency, increase diagnosis rate, and reduce the expense of medical complementary modalities (Yassin et al., 2018).

Recent studies are focusing on applying ML algorithms on data of gene expression as a new approach to find genes differentialy expressed between breast cancer subtypes. ML was used to identify key genes of triple negative tumors (Naorem et al., 2019), prediction of estrogen receptor status (Alakwaa et al.,2018) and even treatment gene-related biomarkers (Tabl et al.,2019). 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 (Yue et al., 2018). The ML approach with medical open source databases can help clinicians to accurately classify breast cancer, allowing an early detection and hence increase the treatment chances and decrease mortality rate (Yassin et al., 2018; Yue et al., 2018).

OBJECTIVES

Our general objective is to search for breast cancer features to classify different variations and occurences. Such findings have the potential to reveal different metabolic pathways and relevant characteristics, which could be linked to this type of disease. The ML approach can serve as a useful tool in diagnosis, guiding future clinical and expression studies in this area.

Our specific objectives are to break down the study by answering two basic questions: a) what genes are important for predicting cancer receptor subtypes, and b) what is the minimal amount of features that can be used to predict triple-negative breast cancer. This subtype of cancer is of special interest to us because it is heterogenous and has the issue of not having the biomarkers receptors, which are common targets for most hormone therapies, therefore requiring different treatments.

MATERIALS AND METHODS

To tackle the problem, our group will use public datasets on breast cancer, mainly from *cBioPortal for Cancer Genomics*, and apply machine learning techniques. The process encompasses from data gathering and preprocessing to model optimization and validation.

I) Dataset

The machine learning algorithms will be trained to classify and predict breast cancer using the MSK, Cancer Cell 2018 dataset (Razavi et al., 2018), available in the *cBioPortal for Cancer Genomics*.

The dataset contains data of 1918 breast tumor specimens from 1756 patients. Characteristics of the patients are provided in Table 2. The histologic subtypes of breast cancer were determined based on clinical pathology reports of the primary tumors and metastatic sites. Pathology subtypes were classified as either invasive ductal carcinoma (IDC), invasive lobular carcinoma (ILC), mixed ductal and lobular carcinoma (mixed IDC/ILC), and several additional rare histologies (Table 1). The TNM classification (tumor size, lymph nodes affected and metastases) and overall tumor stage at diagnosis was defined per the American Joint Committee on Cancer (AJCC) 7th edition staging system.

The population on this dataset presents characteristics such as information regarding prior to and post treatment (mainly hormonal therapy followed by chemotherapy) tumor state both for primary and metastatic sites. There is also a larger percentage of patients with high-risk features, such as younger age and more advanced tumor status. These pieces of information, along with having the second largest N value for *cBioPortal* datasets on breast cancer, makes this dataset especially interesting for machine learning applications. Figure 1 contains a visual representation of dataset features.

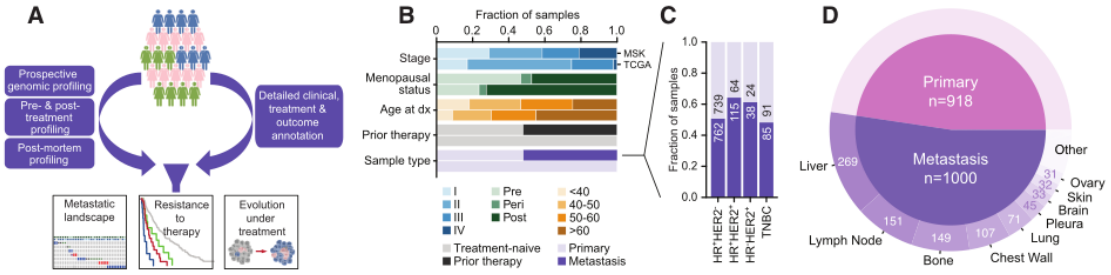


Figure 1: Features of the *cBioPortal* MSK Cancer Cell 2018 dataset. **(A)** Schematics on data collection. **(B)** Comparison of clinical features for patients of this dataset (MSK) compared to a contemporary study (TCGA). **(C)** Sample type composition regarding receptor subtypes. **(D)** Sample tissue origin distribution. (Razavi et al., 2018)

Patients defined as having metastatic disease were confirmed to have had a biopsy of a metastatic site that identified breast cancer. Tumor grade was defined based on the Nottingham combined histologic grade of the primary breast tumor (Bloom and Richardson, 1975; Elston and Ellis, 1991). The primary tumors with total tumor score of 3-5 were classified as G1 (well differentiated); 6-7: as G2 (moderately differentiated), and 8-9: as G3 (poorly differentiated). Patients were classified into breast cancer subtypes based on ER and PR IHC results and the HER2 IHC.

The mutation profiling was performed by the authors with a different method developed by their group (Memorial Sloan Kettering-integrated mutation profiling) to identify somatic mutations, DNA copy number alterations and select rearrangements in 468 cancer-associated genes (Cheng et al., 2015; Zehir et al., 2017). We will associate mutations in these genes, such as CDH1, TP53 and GATA3, with specific receptor types and histologies. The categorization of breast cancer through mutations and alterations in genes and metabolic pathways could influence clinical pratice and potential approachs for treatments.

II) Machine Learning

ML algorithms have been widely used in breast cancer diagnosis and prognosis to gain different insights from data samples. ML approaches deal with complex, large datasets with many dimensions, resulting in capacity of extracting key features that would be difficult to identify with the use of purely traditional statistics (Yue et al., 2018).

The ML approach involves data gathering, exploration, splitting into train and test samples, and preprocessing for use. After these steps, data serves as input for the machine learning model, which allows for predicting either a value or a data class.

II.1) Data Gathering, Exploration and Preprocessing

The first step of the process will be the merging of data. The dataset is a collection of Microsoft Excel Open XML Spreadsheet (XLSX) files that contain clinical as well as genomic data identified by patients and sample IDs. The XLSX spreadsheets will be merged according to patient and sample IDs (used as primary keys) in order to obtain a

Table 2: Clinical characteristics of the study cohort (modified from Razavi et al., 2018)

	HR+HER- n = 1364	**HR+HER2+ n = 166**	**HR-HER2+ n = 58**	**TNBC n = 168**	**p value**
Stage ad Diagnosis					<0.001
Stage I	429 (31.5)	31 (18.7)	14 (24.1)	47 (28)	
Stage II	401 (29.4)	44 (26.5)	9 (15.5)	54 (32.1)	
Stage III	271 (19.9)	36 (21.7)	14 (24.1)	35 (20.8)	
Stage IV	252 (18.5)	53 (31.9)	21 (36.2)	32 (19)	
Not available	11 (0.8)	2 (1.2)	0 (0)	0 (0)	
Histology					<0.001
Invasive Ductal	1007 (73.8)	136 (81.9)	51 (87.9)	147 (87.5)	
Invasive Lobular	253 (18.5)	15 (9)	3 (5.2)	3 (1.8)	
Mixed Ductal/ Lobular	73 (5.4)	8 (4.8)	1 (1.7)	1 (0.6)	
Carcinoma, NOS	18 (1.3)	6 (3.6)	2 (3.4)	3 (1.8)	
Other	13 (0.9)	1 (0.6)	1 (1.7)	14 (8.4)	
Histologic Grade (Primary Tumor)					<0.001
I - Well differentiated	87 (6.4)	2 (1.2)	1 (1.7)	0 (0)	
II - Moderately differentiated	382 (28)	26 (15.7)	6 (10.3)	9 (5.4)	
III - Poorly differentiated	747 (54.8)	120 (72.3)	47 (81)	156 (92.9)	
Not available	148 (10.9)	18 (10.8)	4 (6.9)	3 (1.8)	

consolidated data view. Then, as the second step, data exploration will be conducted, this will give information about the column types, percentage of missing values, data distribution and descriptive statistics, such as the mean, median, standard deviation and quantiles. This will determine the need for scaling, normalizing or centering data features.

Recent studies have shown that applying the appropriate preprocessing can improve classification results and data quality (Zeinab Sajjadnia, 2020) when using real world datasets. The third step will be to apply preprocessing methods such as data cleaning, error correction, resolving data inconsistency, noise removal, treating missing values with filling null values, and feature selection. Since there will be changes in the data, some exploration might be conducted after the preprocessing, going back and forth between steps 2 and 3.

II.2) Data Modeling

The fourth step is data modeling. For this, some supervised machine learning algorithms will be used such as *Random Forest*, *XGBoost*, *Support Vector Machine (SVM)*, *K Nearest Neighbors (KNN)* and there might also be some experimentation with *Naive Bayes* and *Neural Networks* classifiers.

II.2.1) Random Forest

Random Forest algorithm is an ensemble of decision trees in which each tree is fitted to a random sample of training data with replacement, and during the fitting, the trees aren't pruned. So each tree keeps splitting until it gets to the last data point. Since, this can be run in a parallell manner independently for each tree, Random Forest is also a bagging method. For a new data point, the prediction for each tree is obtained and then the best solution is

selected by means of voting. Figure 2 illustrates this learning model.

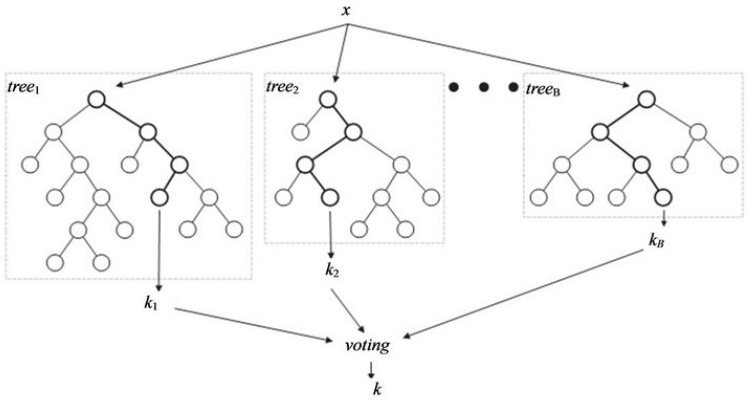


Figure 2: The Random Forest classifier makes use of trees, which allows better visualization of model decision making (Nguyen et al., 2013)

There are two stages in the Random Forest algorithm. The first one is the building of a classifier tree, and the second is making a prediction from the classifiers created during the first stage.

- For N samples in the learning set, select n cases at random - with replacement - from the original data.
- Use the resulting samples as the training set for growing the trees;
- Consider V as the total number of input variables, from that select randomly a variable v such that $v < V$ is specified at each node, pick the best split on v for splitting the tree node;
- Grow each tree to the largest possible extent, without pruning.

This model combines several trees classifiers, so it decreases variance and offers stability. Also, by having a number of decision trees participating in the process, it significantly reduces overfitting (Kabir and Ludwig, 2018).

When using a Rendom Forest model, we make use of Decision Trees. Trees are rule based models that offer high explainability and easy understanding of important data features relevant to the decision making process of the classification (Wang et al., 2020).

II.2.2) XGBoost

Extreme Gradient Boosting (XGBoost) is a tree based model originated from the Gradient Boosting Model (GBM) proposed by Friedman (Friedman, 2001). In the same way as Random Forest, it is an ensemble method. The algorithm selects random samples with replacement over and creates a tree, differently from Random Forest, this is done over a weighted training dataset, making XGBoost a boosting algorithm.

In boosting, each new tree is built sequentially. The first tree classifies the data and the next tree receives the data modified by weights for the samples that were previously misclassified. This keeps on going for each latter tree

created. After all sequential trees have been created, when a new data point is classified, its value is decided by taking the weighted average of each tree prediction. (Torlay et al., 2017).

XGBoost is a widely adopted algorithm due to its good results. Besides boosting the trees, it has other advantages such as parallel processing, built-in cross-validation and smoothing of the final weights (Ye et al., 2018). It also decreases variance, offers high stability and increasingly reduces error rate, although it can be prone to overfitting. The algorithm is illustrated in Figure 3.

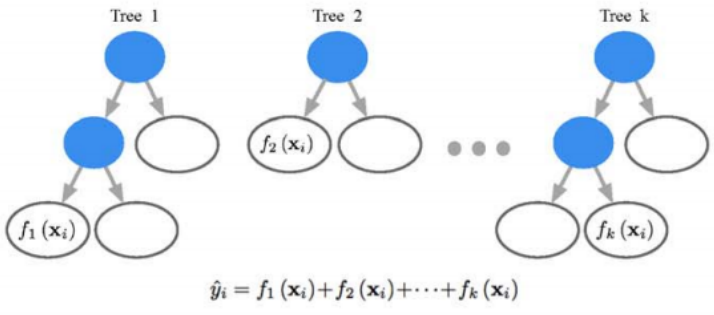


Figure 3: XGBoost classifier example, also having an approach to learning based on trees (Ye et al., 2018).

II.2.3) Support Vector Machine

The Support Vector Machine (SVM) model is very similar to the popular perceptron, but with elements of optimization. It represents data as points in space and looks for the hyperplane that divides data classes. The classification of new points is performed by verifying to which side of the hyperplane the unclassified point belongs to. There are infinite hyperplanes that can classify the data, but the optimum one presents the largest margin between data points, maximizing its distance on each side. This similarity to the basic perceptron is shown in Figure 4.

SVM usually can recognize subtle patterns in complex datasets (Aruna S, 2011) and it has been widely used in cancer classification since gene expression data became avaliable (Huang et al., 2018). The fact that SVM functions in a similar way than a perceptron allows the group to better understand it and to be able to visualize data classification.

II.2.4) K-Nearest Neighbors

K-Nearest Neighbors (KNN) is an algorithm that makes classification based on distances from data points, and it has a fundamental difference that it doesn't make any assumptions on data distribution (Islam et al., 2017). For new data points, KNN sorts which points are closest to it in terms of distance. This distance can be calculated by different metrics, such as the Euclidian distance, which is the most commonly used. After calculating the distance, each point is then classified as belonging to a class with the closest data points to it, and when the distribution of classes is skewed, the point may be classified by majority voting. Figure 5 presents an illustration of the KNN algorithm.

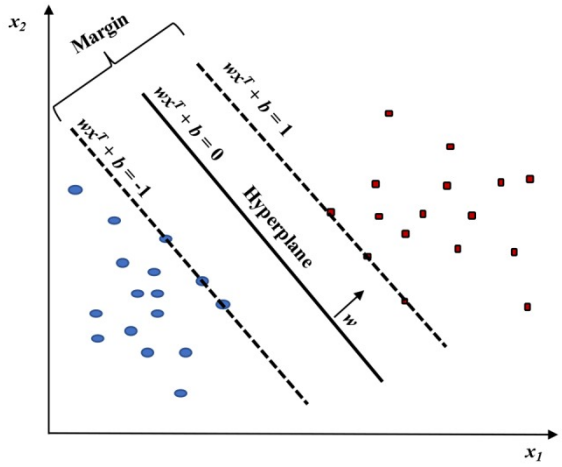


Figure 4: The SVM classifier, similar to the basic perceptron algorithm, uses a hyperplane to perform classification (Huang et al., 2018).

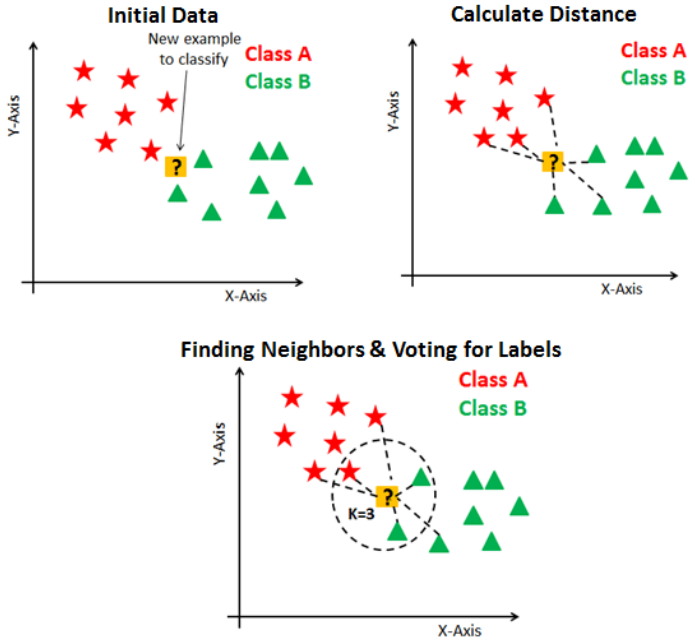


Figure 5: KNN classifier on two classes of data (A and B), which performs classification based on distance to nearest data points (Chauhan, 2019).

II.3) Evaluation metrics

The fifth step is to measure and analyze the model results. To better understand algorithm outputs and how much of the data was correctly or incorrectly classified, the confusion matrix can be employed to reveal the amount of false negatives/positives. Table 3 shows the structure of a confusion matrix, and the difference in the samples from the class and those that were predicted by the model can be visually and intuitively inspected.

To measure model results, metrics such as *precision*, *recall*, *F1-score* (which relates the previous two), and *accuracy*

Table 3: Structure of a confusion matrix, showing true and false positives/negatives for a model applied to a dataset. It is directly related to type I and type II classification errors for our proposal.

		Actual Class	
		Positive	Negative
Predicted Class	Positive	t_p	f_p
	Negative	f_n	t_n

will be used. These metrics give complementary information on results by relating entries of the confusion matrix and are given by

$$Precision = \frac{t_p}{t_p + f_p},$$

$$Recall = \frac{t_p}{t_p + f_n},$$

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n},$$

$$F1 = \frac{2}{recall^{-1} + precision^{-1}}.$$

The presented algorithms and metrics are implemented in the *scikit-learn* library for Python (Pedregosa et al., 2011). Having an already well-established software to be used allow us to save time in implementation and dedicate more time to critically assess the results from both biological and statistical learning points of view considering what has been taught in the IBI5031 course. The *F1-score*, *precision*, *recall* and *accuracy* metrics are part of the scikit-learn classification report, which also include the support column that shows how many samples were used for the classification. Table 4 shows an example of a report given by scikit-learn.

Table 4: Classification report example from the scikit-learn library for Python.

	Precision	Recall	F1-score	Support
Class 0	0.50	1.00	0.67	100
Class 1	0.00	0.00	0.00	100
Class 2	1.00	0.67	0.80	300
Accuracy			0.60	500
Macro avg	0.50	0.56	0.49	500
Weighted avg	0.70	0.60	0.61	500

II.4) Model Optimization and Validation

The sixth step is the model results optimization, and techniques such as *grid search* for hyperparameter search and *cross-validation* to enhance model generalization and prevent overfitting will be employed. The traditional way of performing hyperparameter optimization has been by using *grid search*: an exhaustive search method through a defined set of hyperparameters of the model being employed for learning. To prevent overfitting, a part of the dataset will be held out as a validation set.

By partitioning the available data into three sets, however, the number of samples is reduced, so the results may start to depend on the choice for the training and validation sets. To address this, one can use *cross-validation* (CV, illustrated in Figure 6), in which the test set is only used for final evaluation, with the training set being divided into smaller subsets.

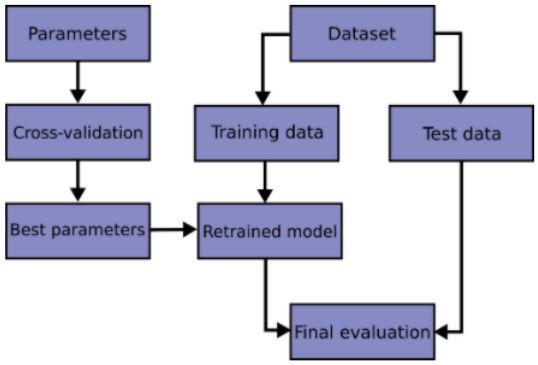


Figure 6: Flowchart for cross-validation usage, with partition of data into different sets (Pedregosa et al., 2011).

In K-Fold CV, an instance of Cross Validation, the training set is split into k smaller sets, and for each of the k fold the following steps are performed: i) the model is trained using $k - 1$ of the folds as training data; ii) the resulting model is validated on the remaining part of the data; iii) a chosen metric is computed for each data split; and iv) all metric values are averaged to constitute the k-fold performance measure.

This approach can be computationally expensive, but it optimizes data usage, which is an advantage when the number of samples is small (usually the case for biological data). Figure 7 illustrates the k-fold cross-validation process.

II.5) Further Models and Techniques

The focus of our work is to apply supervised learning, which is currently being studied in the IBI5031 course. Nevertheless, unsupervised learning (clustering) algorithms, such as *K-Means* and dimensionality reduction techniques like *Principal Component Analysis (PCA)* or *DBSCAN (Density-based spatial clustering)*, may also be used if needed

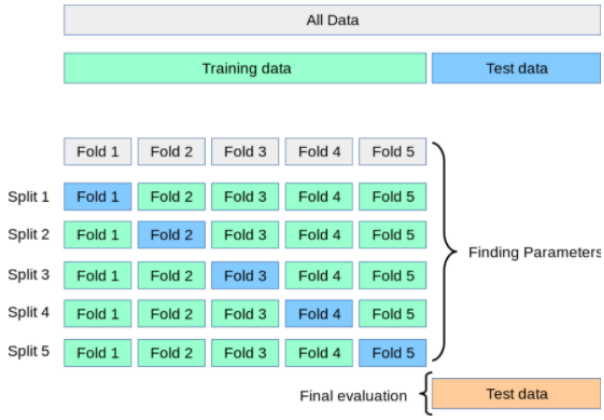


Figure 7: Example of k-Fold cross-validation (Pedregosa et al., 2011).

to visualize data and gather further insights. If those algorithms are used, the evaluation metrics employed will be *silhouette score* and *elbow method*.

II.6) Programming Language, Libraries and Packages

To accomplish the steps listed above we will use Python programming language, the Jupyter Notebook environment, the Scikit-learn library, Seaborn, Matplotlib, Scipy, Numpy and Pandas packages. We will also use the Time module in Python to understand on average how long our models take to run. If necessary, we might also use the SHAP (SHapley Additive exPlanations) library and its shap values to explain the output of the machine learning models.

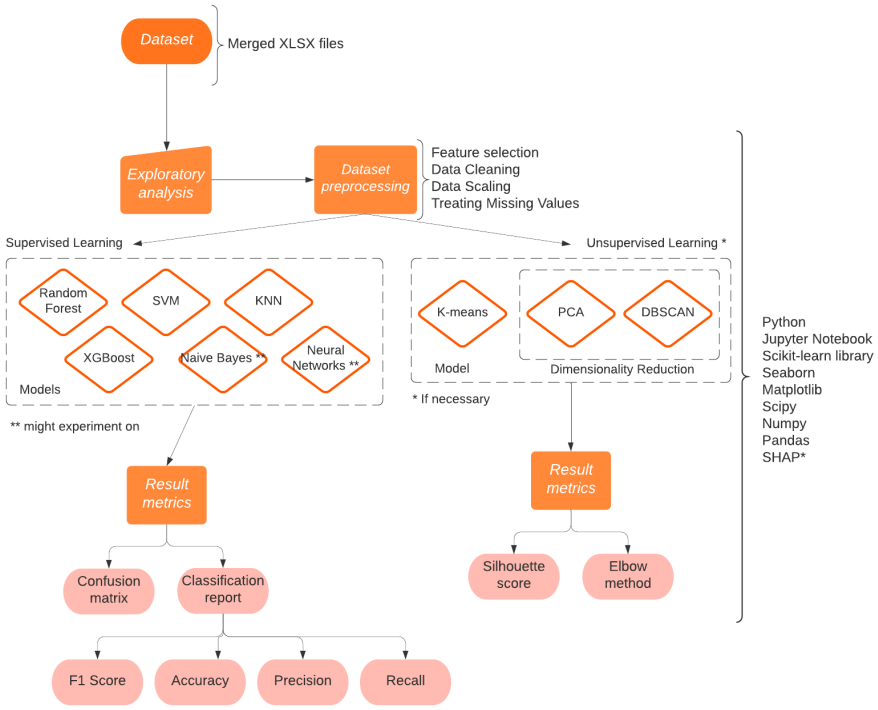


Figure 8: Workflow of the ML approach for the present project.

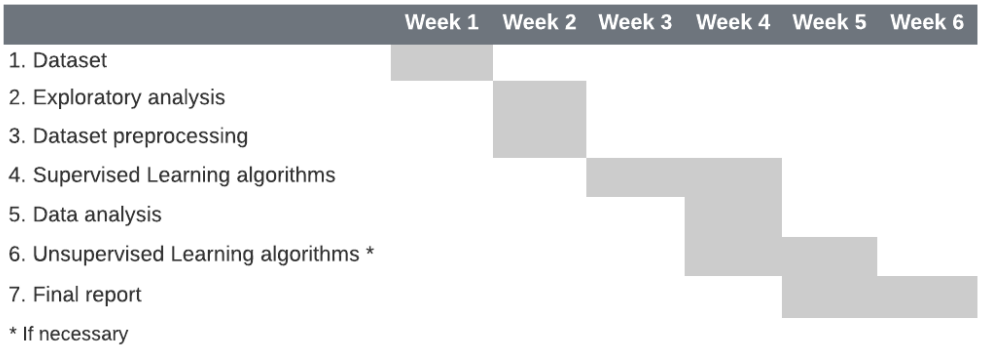
EXPECTED RESULTS

The main goal of this project is to apply ML methods to further study breast cancer subtypes merging biomarkers receptors expression data with other genetic information such as DNA copy number alterations for better subtype classification. We also hope that the result of this process may help guiding medical doctors decision making process on treatment by looking at mutation profiles, and expect to better elucidate some of the mutation features underlying breast cancer subtypes, which could guide potential targets for future *in vitro* studies.

We chose each of the models with specific purposes. By using tree based machine learning models we plan to get explainability on how the model is making decisions and with that further understand relations between the dataset features and the classification. By using KNN, we can understand which features are close to each other above the three dimensional space and by using support vector machines we can work with less data, maximize the split between the classified data points and find a best fitting hyperplane. So, with this setting we have explainability, closeness and margin maximization. Those are complementary ways that can help us better understand how the genomic features relate to breast cancer and be able to pinpoint which seem to be correlated to this disease, possibly even relating to different clinical conditions. For comparison, we will check our results over previous findings by other authors and existing pathways to check for biological soundness of this work's results.

WORK PLAN

We plan to conduct weekly meetings on wednesdays in which we can discuss results and ideas and also have more meetings if necessary. We have already begun to merge the different datasets to have a consolidated view. We have also done some data analysis. After that, we will start constructing the models, adjusting its hyperparamenters and optimizing results. While doing that we will also extract the most important features and the insights from each model. We will then reduce the number of model features to understand how they influence the results and make a final report. If necessary, we will also make use of dimensionality reduction and clustering algorithms to be able to visualize the shape of the data.



References

[1] A. F. M. Agarap. 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*, pages 5–9, 2018.

[2] F. M. Alakwaa, K. Chaudhary, and L. X. Garmire. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. *Journal of proteome research*, 17(1):337–347, 2018.

[3] R. S. Aruna S. A novel svm based cssffs feature selection algorithm for detecting breast cancer. *International Journal of Computer Applications*, 31(8):14–20, 2011.

[4] H. Bloom and W. Richardson. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. *British journal of cancer*, 11(3):359, 1957.

[5] D. Cameron, M. J. Piccart-Gebhart, R. D. Gelber, M. Procter, A. Goldhirsch, E. de Azambuja, G. Castro Jr, M. Untch, I. Smith, L. Gianni, 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.

[6] J. Chakraborty, A. Midya, and R. Rabidas. Computer-aided detection and diagnosis of mammographic masses using multi-resolution analysis of oriented tissue patterns. *Expert Systems with Applications*, 99:168–179, 2018.

[7] N. S. Chauhan. Classifying heart disease using k-nearest neighbors. `https://www.kdnuggets.com/2019/07/classifying-heart-disease-using-k-nearest-neighbors.html`, 2019. Online; Accessed: 18 October 2020.

[8] M. C. Cheang, S. K. Chia, D. Voduc, D. Gao, S. Leung, J. Snider, M. Watson, S. Davies, P. S. Bernard, J. S. Parker, 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.

[9] D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, 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.

[10] X. Dai, L. Xiang, T. Li, and Z. Bai. Cancer hallmarks, biomarkers and breast cancer molecular subtypes. *Journal of Cancer*, 7(10):1281, 2016.

[11] D. Dillon, A. Guidi, and S. Schnitt. Pathology of invasive breast cancer. *Diseases of the Breast*, 5:381–410, 2010.

[12] C. W. Elston and I. O. Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.

[13] J. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 11 2001. doi: 10.1214/aos/1013203451.

[14] L. Fulford, D. Easton, J. Reis-Filho, A. Sofronis, C. Gillett, S. Lakhani, and A. Hanby. Specific morphological features predictive for the basal phenotype in grade 3 invasive ductal carcinoma of breast. *Histopathology*, 49 (1):22–34, 2006.

[15] S. Huang, N. Cai, P. Pacheco, S. Narrandes, Y. Wang, and W. Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1):41–51, 2018.

[16] M. Islam, H. Iqbal, M. Haque, and M. Hasan. Prediction of breast cancer using support vector machine and k-nearest neighbors. 12 2017. doi: 10.1109/R10-HTC.2017.8288944.

[17] H. Joshi and M. F. Press. Molecular oncology of breast cancer. pages 282–307, 2018.

[18] M. F. Kabir and S. Ludwig. Classification of breast cancer risk factors using several resampling approaches. pages 1243–1248, 12 2018. doi: 10.1109/ICMLA.2018.00202.

[19] M. Lippman, G. Bolan, and K. Huff. The effects of glucocorticoids and progesterone on hormone-responsive human breast cancer in long-term tissue culture. *Cancer research*, 36(12):4602–4609, 1976.

[20] R. C. Moon, M. Pike, P. Siiteri, and C. Welsch. Influence of pregnancy and lactation on experimental mammary carcinogenesis. *Banbury report*, 8:353–364, 1981.

[21] L. D. Naorem, M. Muthaiyan, and A. Venkatesan. 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.

[22] C. Nguyen, Y. Wang, and H. Nguyen. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and Engineering*, 6:551–560, 2013. doi: 10.4236/jbise.2013.65070.

[23] A. Nicolini, P. Ferrari, and M. J. Duffy. Prognostic and predictive biomarkers in breast cancer: past, present and future. In *Seminars in cancer biology*, volume 52, pages 56–73. Elsevier, 2018.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] P. Razavi, M. T. Chang, G. Xu, C. Bandlamudi, D. S. Ross, N. Vasan, Y. Cai, C. M. Bielski, M. T. Donoghue, P. Jonsson, et al. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer cell*, 34(3):427–438, 2018.

[26] S. Rosso, A. Gondos, R. Zanetti, F. Bray, M. Zakelj, T. Zagar, G. Smailyte, A. Ponti, D. H. Brewster, A. C. Voogd, 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.

[27] C. Smigal, A. Jemal, E. Ward, V. Cokkinides, R. Smith, H. L. Howe, and M. Thun. Trends in breast cancer by race and ethnicity: update 2006. *CA: a cancer journal for clinicians*, 56(3):168–183, 2006.

[28] A. Spitale, P. Mazzola, D. Soldini, L. Mazzucchelli, and A. Bordoni. 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.

[29] A. A. Tabl, A. Alkhateeb, W. ElMaraghy, L. Rueda, and A. Ngom. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Frontiers in Genetics*, 10:256, 2019.

[30] P. Tang, J. Wang, and P. Bourne. Molecular classifications of breast carcinoma with similar terminology and different definitions: are they the same? *Human pathology*, 39(4):506–513, 2008.

[31] Z. Tao, A. Shi, C. Lu, T. Song, Z. Zhang, and J. Zhao. Breast cancer: epidemiology and etiology. *Cell biochemistry and biophysics*, 72(2):333–338, 2015.

[32] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu. Machine learning-xgboost analysis of language networks to classify patients with epilepsy. *Brain Informatics*, 4, 04 2017. doi: 10.1007/s40708-017-0065-7.

[33] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108, 2015.

[34] L. Tryggvadóttir, M. Gislum, F. Bray, Å. Klint, T. Hakulinen, H. H. Storm, and G. Engholm. 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.

[35] C. S. Vallejos, H. L. Gómez, W. R. Cruz, J. A. Pinto, R. R. Dyer, R. Velarde, J. F. Suazo, S. P. Neciosup, M. León, A. Miguel, 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.

[36] A. G. Waks and E. P. Winer. Breast cancer treatment: a review. *Jama*, 321(3):288–300, 2019.

[37] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin. An improved random forest-based rule extraction method for breast cancer diagnosis. *Applied Soft Computing*, 86:105941, 2020. ISSN 1568-4946. doi: 10.1016/j.asoc.2019.105941. URL http://www.sciencedirect.com/science/article/pii/S1568494619307227.

[38] N. I. Yassin, S. Omran, E. M. El Houby, and H. Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*, 156:25–45, 2018.

[39] L. Ye, T.-S. Lee, and R. Chi. 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, 01 2018.

[40] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu. Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2(2):13, 2018.

[41] A. Zehir, R. Benayed, R. H. Shah, A. Syed, S. Middha, H. R. Kim, P. Srinivasan, J. Gao, D. Chakravarty, S. M. Devlin, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine*, 23(6):703, 2017.

[42] M. R. M. Zeinab Sajjadnia, Raof Khayami. Preprocessing breast cancer data to improve the data quality, diagnosis procedure, and medical care services. *Sage*, 19:16, 2020.