Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review

PEDRO HENRIQUES ABREU and MIRIAM SEOANE SANTOS, CISUC, Department of Informatics Engineering, Faculty of Sciences and Technology of Coimbra University, Portugal
MIGUEL HENRIQUES ABREU, Portuguese Institute of Oncology of Porto, Portugal
BRUNO ANDRADE, CISUC, Department of Informatics Engineering, Faculty of Sciences and Technology of Coimbra University, Portugal
DANIEL CASTRO SILVA, LIACC, Department of Informatics Engineering, Faculty of Engineering of Porto University, Portugal

Background: Recurrence is an important cornerstone in breast cancer behavior, intrinsically related to mortality. In spite of its relevance, it is rarely recorded in the majority of breast cancer datasets, which makes research in its prediction more difficult. Objectives: To evaluate the performance of machine learning techniques applied to the prediction of breast cancer recurrence. Material and Methods: Revision of published works that used machine learning techniques in local and open source databases between 1997 and 2014. Results: The revision showed that it is difficult to obtain a representative dataset for breast cancer recurrence and there is no consensus on the best set of predictors for this disease. High accuracy results are often achieved, yet compromising sensitivity. The missing data and class imbalance problems are rarely addressed and most often the chosen performance metrics are inappropriate for the context. Discussion and Conclusions: Although different techniques have been used, prediction of breast cancer recurrence is still an open problem. The combination of different machine learning techniques, along with the definition of standard predictors for breast cancer recurrence seem to be the main future directions to obtain better results.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Machine learning approaches**; • **Applied computing** → *Bioinformatics*;

Additional Key Words and Phrases: Breast cancer recurrence, pattern recognition, clinical decision-making

ACM Reference Format:
Pedro Henriques Abreu, Miriam Seoane Santos, Miguel Henriques Abreu, Bruno Andrade, and Daniel Castro Silva. 2016. Predicting breast cancer recurrence using machine learning techniques: A systematic review. ACM Comput. Surv. 49, 3, Article 52 (October 2016), 40 pages.
DOI: http://dx.doi.org/10.1145/2988544

1. INTRODUCTION

Breast cancer (BC) figures among the major causes of concern worldwide. According to the latest GLOBOCAN statistics [World Health Organization 2012], it was the second most frequently diagnosed cancer and the fifth cause of cancer mortality worldwide, responsible for 6.4% of all deaths.

The mortality associated to this pathology is mostly related to metastization [Moody et al. 2005], the spread of cancer to other parts of the body remote from the breast,

Authors' addresses: P. H. Abreu, M. S. Santos, and B. Andrade, Polo II, Pinhal de Marrocos 3030-290 Coimbra; emails: pha@dei.uc.pt, miriams@student.dei.uc.pt, bandrade@student.dei.uc.pt; M. H. Abreu, Rua Dr. António Bernardino de Almeida, 4200-072 PORTO; email: antonio.m.abreu@ipoporto.min-saude.pt; D. C. Silva, Rua Dr. Roberto Frias s/n 4200-465 Porto; email: dcs@fe.up.pt.

DOI: http://dx.doi.org/10.1145/2988544

Table I. A Comparison between KDD, SEMMA, and CRISP-DM Knowledge Discovery Processes [Azevedo and Santos 2008]

KDD	SEMMA	CRISP-DM
Pre-KDD	————	Business understanding
Selection	Sample	Data understanding
Preprocessing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	————	Deployment

and recurrence (or relapse), which describes cancer that reappears after treatment [Mendonza 2013]. Being documented in 10%–15% of all BC patients [Van den Hurk 2011], recurrence assumes a pivotal importance in their prognosis. However, it is not as well studied as BC itself. Searching in Thomson Reuters [2015] platform for research works with the expression "breast cancer" in the title yields more than 330,000 results. A similar search focused on recurrence yields only around 20,000 results (approximately 6%), obtained when the search terms are extended to "recurrence(s)," "relapse(s)," and "metastasis (es)" (individually or in combination). These results can be partially explained by the fact that, for instance, none of the three major American cancer registries reports cancer recurrence information [In et al. 2014].

Besides the obvious implications of recurrence in mortality, BC patients also face serious treatment-related complications, which increases their risk of death from causes unrelated to BC itself [Farr et al. 2013]. In this scenario, accurate prediction of BC behavior assumes an important role, since it aids clinicians in their decision-making process, enabling a more personalised treatment for patients. Some of the studies regarding cancer recurrence involve the use of statistical methodologies, or machine learning algorithms, which have a long history in cancer research [Kononenko 2001; Cruz and Wishart 2006; Kouroua et al. 2015]. This research work attempts to provide an overview of the prediction of BC recurrence using machine learning techniques. The challenge is to accurately predict recurrence events, within a binary outcome (yes/no). This challenge encompasses not only the choice of a good dataset (containing quality data) but also the selection of the most appropriate features, as well as the most advantageous algorithm.

The remainder of the article is organized as follows: Section 2 covers the steps used by different authors to predict BC recurrence, highlighting the datasets, variables included in the reviewed studies, data mining algorithms, sampling strategies, and evaluation metrics used. Section 3 depicts the analyzed works in more detail and Section 4 presents a discussion on the different works. Finally, some conclusions and future directions are discussed in Section 5.

2. PREDICTING BC RECURRENCE PHASES AND TASKS

The most common processes to develop a data mining approach are Knowledge Discovery in Databases (KDD) [Fayyad et al. 1996]; Sample, Explore, Modify, Model, and Assess (SEMMA) [SAS Institute 2015]; and Cross-Industry Standard Process for Data Mining (CRISP DM) [Chapman et al. 2000]. The first two are composed of five steps each; despite the different designations, their steps are generally equivalent [Azevedo and Santos 2008]. The third strategy, CRISP DM, presents two novel steps that consist in business understanding, where, after the evaluation phase, the results are interpreted from a business perspective; and the deployment step, where the final process achievements are somehow incorporated in a product/service (more related to a business perspective). Table I highlights the difference between these processes. As all BC recurrence studies analyzed use a KDD strategy, its steps are used to highlight the

methodology followed by each of the works. In the following subsections, we explore some of the steps in the KDD approach and how they were addressed in the reviewed works.

2.1. Selection

This step consists in the selection of a dataset and an appropriate set of features for knowledge extraction. The datasets can be publicly available (e.g., online) or they may result from a collaboration between institutions and research teams, not available for the general public. Feature selection may be performed manually, or using variable selection algorithms. In particular, for BC recurrence, 7/17 of the studies used manual selection (five with the help of medical experts), while some of the others took advantage of well-known feature selection algorithms; for instance, Jonsdottir et al. [2008].

The datasets and the number of patients used in the analyzsed studies are summarized in Table II. From the 17 reviewed works, the majority uses available datasets (nine works). Among those, four works use the Wisconsin prognostic breast cancer (WPBC) dataset and three use the BC dataset, both available from the University of California, Irvine machine learning repository (UCI Repository) [Lichman 2015]. The remaining two datasets are available from van 't Veer's study [van 't Veer et al. 2002] and the widely known Surveillance, Epidemiology, and End Results (SEER) database (U.S. National Cancer Institute). The unavailable databases are collaborations with specialized BC centers, registers, or teaching hospitals, in several different countries (Sweden, Spain, California, Iceland, South Korea, and Ljubljana). The Institute of Oncology in Ljubljana was the greatest contributor for BC recurrence studies, providing the data for five of the reviewed works.

Although the endpoint for predicting BC recurrence is not defined for some cases, most of the datasets are associated with a specific time period for recurrence prediction (e.g., 4–5 years after the diagnosis, 10 years after surgery). Moreover, the great majority of the datasets suffered from a considerable class imbalance, with uneven cases of "recurrence" versus "no-recurrence," following a 30%–70% distribution. The most affected works are Mani et al. [1997] and Razavi et al. [2007], with a class distribution (recurrence/no-recurrence) of 10%–90% and 20%–80%, respectively. Going against this trend, three works (Sun et al. [2007], Trumbelj et al. [2010], and Tomczak [2013]) perform their experiments on balanced datasets with approximately 50%–50% class distribution. The distribution is unknown for two works, Razavi et al. [2005] and Jerez-Aragonés et al. [2003].

Data imbalance occurs when one class (w_i) is represented by a larger number of examples than the others. Considering a two-classification problem, a dataset is said to be imbalanced if there exists an underrepresented class (the minority class) in comparison to the other class (the majority class) [Chawla 2010; Longadge and Dongre 2013]. For instance, consider the case of women with BC, where w_1 represents "recurrent" tumors and w_2 represents "nonrecurrent" tumors in a total population of 100 women. If 50 women suffer recurrence while the other 50 do not, then clearly there is no class imbalance. On the contrary, if 90 women do not suffer from recurrence, then this class is more represented than the "recurrence" class. Imbalance data is known to deteriorate the performance of a classifier, since there is a "concept," a "class," that is underrepresented, meaning that the classifier does not "learn" this class as well as the other: it tunes its predictions for the larger class, while developing "blind spots" toward the minority one [Kotsiantis et al. 2006; Garcia et al. 2007]. In many machine learning problems, dealing with class imbalance problems remains a knotty subject for classification algorithms, due to their tendency to overlook small, underrepresented concepts, in favor of better represented, more clear, larger concepts. However, this is especially problematic when there is a higher cost of misclassification of the minority examples [Ganganwar 2012], which is the case of BC recurrence. A false positive

Table II. Datasets and Feature Selection used in the Analyzed Studies (NU - Not Used, ER - Estrogen Receptor, PR - Progesterone Receptor, ICD - International Classification of Diseases)

Publications	Dataset (recurrence/no recurrence)	Patient Characteristics	Tumor Characteristics	Treatments
Mani et al. [1997]	Breast Cancer Center, Orange County, California (85/802)	Lymphedema	Tumor presence and its invasive nature, size, lymph nodes involvement, stage	NU
Jerez-Aragonés et al. [2003]	Hospital Clínico Universitario Malaga, Spain (1035 patients, not effective, distribution unknown)	Age, menarchy age, menopausal age, first pregnancy age/pregnancies number, number of miscarriages	Size, grade, lymph nodes involvement, expression of ER, PR, p53 accumulation, ploidy, S-phase	NU
Razavi et al. [2005, 2007]	Swedish Regional BC Register (**2005**: 3949 patients, distribution unknown; **2007**: 3699, 664/3035)	Age	Size, perigland growth, lymph nodes involvement, expression of ER, PR, S-phase	NU
Sun et al. [2007]	Publicly available microarray data (van 't Veer et al. [2002]) (46/51)	Age	Size, vascular invasion, lymphocytic infiltration, expression of ER, PR, 70-gene profile	NU
Ryu et al. [2007a]	Breast Cancer Dataset (available from UCI Repository) (85/201)	Age, menopausal status	Location, size, grade, lymph nodes involvement	Radiotherapy
Jonsdottir et al. [2008]	Rose dataset University Hospital in Iceland (73/184)	Age, comorbidities, carcinoembryonic antigen/cancer antigen 15-3 (CEA/CA-15-3) values	Clinical detectable, histological type, size, local invasion, inflammatory component, lymphovascular invasion, lymph nodes involvement, metastasis (lung, bone), expression of ER, PR, S-phase	Chemotherapy (pre- and postsurgery), hormonotherapy, radiotherapy
Fan et al. [2010]	SEER Public-Use Data 2005 (46,996,113 patients, not effective, distribution unknown)	Race, age, marital status	Behavior defined by ICD-03, location (laterality, breast regions), size, grade, local invasion, lymph nodes involvement	Surgery, radiotherapy
Belciug et al. [2010]	Wisconsin Prognostic BC (available from UCI Repository) (47/151)	NU	Area, perimeter, compactness, texture, concavity, concave points, size, lymph nodes involvement	NU
Trumbelj et al. [2010]	oncology BCR (Institute of Oncology, Ljubljana) (449/432)	Age, menopausal status, personal or familiar previous malignancies	Histological type, size, grade, local invasion, lymphovascular invasion, lymph nodes involvement, expression of ER, PR	Chemotherapy, hormonotherapy

(Continued)

Table II. Continued

Publications	Dataset (recurrence/no recurrence)	Patient Characteristics	Tumor Characteristics	Treatments
Kim et al. [2012]	Tertiary Teaching Hospital, South Korea (195/484)	NU	Number of tumors, size, grade, local invasion, lymphovascular invasion, lymph nodes involvement, expression of ER	NU
Salama et al. [2012]	Wisconsin Prognostic BC (available from UCI Repository) (47/151)	NU	Area, perimeter, compactness, texture, concavity, concave points, size, lymph nodes involvement	NU
Murti [2012]	Breast Cancer Dataset (available from UCI Repository) (81/196)	Age, menopausal status	Location, size, grade, lymph nodes involvement	Radiotherapy
Tomczak [2013]	Institute of Oncology, Ljubljana (followup from Strumbelj et al.) (949 patients, distribution unknown but assumed the same as the one from Strumbelj's: 51%/49%)	Age, menopausal status, personal or familiar previous malignancies	Histological type, size, grade, local invasion, lymph/vascular invasion, lymph nodes involvement, stage, expression of ER, PR	Chemotherapy, hormonotherapy
Pawlovsky and Nagahashi [2014]	Wisconsin Prognostic BC (available from UCI Repository) (46/148)	NU	Area, perimeter, compactness, texture, concavity, concave points, size, lymph nodes involvement	NU
Beheshti et al. [2014]	Wisconsin Prognostic BC (available from UCI Repository) (47/151)	NU	Area, perimeter, compactness, texture, concavity, concave points, size, lymph nodes involvement	NU
Chaurasia and Pal [2014]	Breast Cancer Dataset (available from UCI Repository) (85/201)	Age, menopausal status	Location, size, grade, lymph nodes involvement	Radiotherapy

(classifying a "no-recurrence" case as "recurrence") has a strong impact on the physical and mental state of the patient; but more importantly, a false negative (classifying a "recurrence" case as "no-recurrence") is more costly because it could potentially turn into a life-threatening situation, since a diseased patient believes everything is going perfectly with her treatment, when in fact the cancer is back. For that reason, dealing with the problem of class imbalance is crucial, whether by using classification algorithms that can handle these characteristic issues of imbalanced data (Section 2.3), by finding an appropriate sampling strategy to change the original class distribution (Section 2.4), or alternatively, by selecting evaluation metrics capable of accurately translating the classifier's behavior in terms of predicting the positive and negative examples in particular (Section 2.5) [Jo and Japkowicz 2004; Chawla et al. 2004; He and Garcia 2009].

Concerning the feature selection—and for some years—many factors were linked to BC recurrence, namely, age at diagnosis, size, stage and grade of tumor, involvement of lymph nodes, menopausal status, Estrogen (ER) and Progesterone Receptors (PR), and HER2 pattern (Human Epidermal Growth Factor Receptor 2) [Mendonza

2013]. Frequently, some of them are associated, given that tumors in younger patients (premenopausal) tend to be high grade, with a triple negative phenotype: without expression of ER, PR, and also HER2. Variables used in the BC recurrence prediction in the previously analyzed studies were compared using three groups: patient characteristics, tumor characteristics, and treatments.

From the analysis of Table II, it is important to highlight that there are many datasets with different origins (local and open source) used to deal with this problematic. Also, different authors used dissimilar varieties with a weak attention for the treatment followed by the patient (only 7/17 studies focused on this factor). Attending to patient characteristics, the majority of studies (10/17) identified age as an important predictor, followed by menopausal status (5/17). This last factor could not be totally independent from age as very young patients are also premenopausal and very old ones are always postmenopausal. All the studies considered size as the main predictor in the tumor characteristics group, followed by lymph nodes involvement (16/17).

2.2. Data Cleaning and Preprocessing

Data cleaning and preprocessing tasks are performed to reduce noise and increase the consistency of data. The preprocessing steps most addressed in the reviewed research works were normalization/standardization of data and missing data handling. Two simple ways of data preprocessing are normalization (Min-Max transformation) and standardization (Z-Score transformation) [Suarez-Alvarez et al. 2012]. Normalization refers to the feature scaling between its minimum and maximum values, while standardization rescales the features so that they follow a standard normal distribution (zero mean and unitary standard deviation). The objective of normalization/standardization is to make features with different scales and ranges of measurement (e.g., age, hemoglobin values) comparable, so that none has more influence than the others on classification task [Shalabi and Shaaban 2006].

Missing Data (MD) can result from a huge variety of events and represents a common challenge in healthcare contexts [Abreu et al. 2013b; García-Laencina et al. 2015].

In brief, MD can be produced At Random (MAR), completely At Random (MCAR), or completely Not At Random (MNAR) [Little and Rubin 2002]. Over the years, several strategies have been studied to handle this issue. The most simple one is Listwise Deletion, where records with MD are simply discarded. This approach may be inappropriate, especially in environments like health care, where most often patients are characterized with a large number of variables with high probability of missing observations. According to the literature, imputation is a more appropriate strategy to deal with MD: using the available complete data, the MD are estimated and filled with plausible values [García-Laencina et al. 2010; Cismondi et al. 2013]. Mean/Mode is one of the simplest imputation strategies, where continuous variables are imputed according to their mean, and categorial variables using their mode. Despite its simplicity, this strategy causes the data to lose some variability, which constitutes its major drawback.

A more sophisticated strategy is using mixture models trained with the Expectation-Maximization (EM) approach, which consists of two steps [Dempster et al. 1977]: the expectation step (E-step) and the maximization step (M-step). Basically, the EM algorithm is based on finding the maximum likelihood of data in order to find the best estimates for missing observations. For the algorithm to start, the E-step makes an initial guess of the model parameters. Using those parameters, and according to the observed (complete) data, it produces estimates for missing observations. The M-step is then responsible for computing new model parameters using the current MD estimations. This process continues repeatedly until the algorithm converges. More specific details can be found in Dempster et al. [1977], Tsikriktsis [2005], Bishop [2006], and Zio et al. [2007].

Multiple Imputation (MI) substitutes every missing observation M times ($M > 1$), using M different estimators (e.g., EM, Markov Chain Monte Carlo methods) [Rubin 2004]. As the name implies, multiple complete datasets are generated, each with different estimates for the absent observations. Then, the M complete datasets have to be analyzed using standard methods—for instance, classification models—in order to combine the different estimates and obtain a single set of results (a discussion on combination rules is given in Little and Rubin [2002]). On one hand, MI is able to reflect the data variability due to missing values. On the other hand, it is computationally expensive, given the generation of different MD estimations, and the required time to further analyze its results.

2.3. Machine Learning Methods

Throughout the years, many Machine Learning (ML) algorithms have been used to predict BC recurrence. A possible taxonomy for the categorization of these methods consists in dividing them into "black-box" and "white-box" methods [Larose 2005]. Black-box algorithms work on the basis of "input stimulus" and "output reactions," without any knowledge of their internal procedures. From the user's perspective, this type of algorithms raises a wide range of questions that will always remain unclear, such as how the results are generated or how they can be explained by the internal methods, given a specific input, among others. This issue becomes especially critical when the user (e.g., a clinician) considers interpretability as a key requirement, in order to use these kinds of approaches and benefit from them in his daily decision-making activities. Contrarily to black-box algorithms, white-box algorithms allow the inspection and explanation of their internal rules, that is, the results of a white-box algorithm may be analytically (mathematically) derived from a given set of inputs [Larose 2005]. This section presents a review on the algorithms used to predict BC recurrence in the studied research works, starting with the white-box algorithms followed by the back-box algorithms.

2.3.1. Decision Trees. Decision Trees (DTs) are defined by recursively partitioning the input space from a root node to multiple branch nodes [Quinlan 1993; Mitchell 1997]. The root node is the "first division" of a DT, from which outgoing edges create several other nodes. Nodes with outgoing edges (with the exception of the root node) are known as internal (or test) nodes, while the remaining (that only have incoming edges) are called leaves, each one assigned to a class. The test nodes divide the input feature space into $p \geq 2$ subspaces according to a condition test of the input features values. Typically, a single feature is considered in each test node and the feature space is divided according to that feature's values. For continuous features, each outgoing edge represents a certain range. An input vector is classified in the DT by sorting it from the root to a leaf, according to the results of the conditions tested along the path.

Some of the most popular DT approaches are the Iterative Dichotomiser (ID3) algorithm [Quinlan 1986], its successors C4.5 and C5.0 [Quinlan 1993; Kantardzic 2011], and Classification And Regression Tree (CART) [Breiman et al. 1984], which use entropy-based measures as splitting criterion during the tree construction process. Some frequently used measures are information gain, which at each node of the tree determines the attribute that provides the most information if used as splitting criteria, or the gain ratio, which is an improvement over information gain that takes into account the number and size of branches when choosing the attribute to be used as splitting criteria [Mitchell 1997; Witten and Frank 2005].

C5.0 is an extension of C4.5. Its DT induction is essentially the same, but it offers some improvements over C4.5 in terms of speed, memory usage, pruning, and weighting schemes [Patel and Rana 2014].

DTs are computationally efficient, can easily handle mixed variables (continuous and discrete), and the rules generated by them are relatively easy to interpret and understand, particularly in health care contexts [Chen et al. 2005]. However, noise and MD can contribute to drastically decrease the accuracy of these algorithms [Liu et al. 2005; Zhang et al. 2005; Atla et al. 2011].

To understand how DTs work in the context of BC recurrence, consider 10 patients and two classes: "recurrence" and "no-recurrence," Each sample (patient) could be characterized by several attributes, such as age, tumor size, Hormonal Receptor (HR) status, and so on. To build a classification tree, DT algorithms start by determining which attribute should be the tree's root, according to some attribute selection measure. Say that "tumor size" is chosen as the root node, and includes two possible values (branches): "$<2cm$" and "$\geqslant 2cm$." The training instances are therefore divided according to their 'tumor size" values, and this attribute is no longer available to be used again. For each branch of "tumor size" the most informative attribute must be found following the same logic. For instance, for "$\geqslant 2cm$," the "hormonal receptor status" with branches *"positive"* and *"negative"* could follow. This process is repeated until (i) all samples belong to the same class or (ii) there are no more attributes to proceed with the division. As an example, consider that "tumor size $< 2cm$" includes three patients, all "no-recurrence" cases and "tumor size $\geqslant 2cm$" has seven patients, divided into four patients with "positive HR status" and three patients with "negative HR status." Furthermore, the HR− patients are all "no-recurrence" cases, but the HR+ patients have three cases of "recurrence" and one of "no-recurrence." Imagine that the remaining attribute (age) adds no new information to this division, proving to be irrelevant. The final tree is therefore complete, and when samples do not belong to the same class while in the same branch, a majority voting scheme is used: the leaf represents the most common class. After the DT is constructed, the test examples left out while building the tree are used for evaluation. To classify those instances, the tree's path is followed from the root up to its final leaf, according to the instance's values. For instance, a patient with "tumor size = 2cm" and "hormonal receptor status = positive" would be classified as a "recurrence" event.

2.3.2. Naive Bayes. Naive Bayes (NB) classifier takes into account the probability distribution of the patterns in each class to make a decision, assuming that there is a probabilistic relationship between predictors (features) and the output (class) [Luttrell 1994]. Bayesian classification determines the probability of a given pattern represented by $\mathbf{x}$ to belong to class ω_i, $P(\omega_i \mid x)$, called *posteriori* probability. Considering a binary classification problem, where two *posteriori* probabilities exist, $P(\omega_1 \mid x)$ and $P(\omega_2 \mid x)$, NB decision rule considers that

—If $P(\omega_1 \mid x) > P(\omega_2 \mid x)$, then $\mathbf{x}$ belongs to ω_1;
—If $P(\omega_1 \mid x) < P(\omega_2 \mid x)$, then $\mathbf{x}$ belongs to ω_2.

Alternatively, if $P(\omega_1 \mid x) = P(\omega_2 \mid x)$, then the choice is arbitrary. The *posteriori* probabilities are calculated according to the well-known Bayes' law (Equation (1)).

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)\, P(\omega_i)}{p(\mathbf{x})}, \tag{1}$$

where $P(\omega_i)$ is the *prior* probability of class ω_i, that is, an estimate of the probability of pattern $\mathbf{x}$ to belong to ω_i; $p(\mathbf{x} \mid \omega_i)$ is the *likelihood* of $\mathbf{x}$, that can be estimated through the Probability Density Function (*pdf*) of $\mathbf{x}$; and $p(\mathbf{x})$ is the total probability of $\mathbf{x}$, which

can be determined using Equation (2):

$$p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x} \mid \omega_i) P(\omega_i). \tag{2}$$

Due to the fact that NB uses probability rules, it inherits the strengths of statistics. Also, another advantage of this method is allowing the researcher to include his domain experience in the modeling process of NB classifiers. Moreover, being a white-box method, it can be more easily understood, for instance, by clinicians. However, its computational complexity, especially when a large dataset is used, constitutes its main drawback [Lee and Abbott 2003].

Consider two classes, "recurrence" and "no-recurrence," and only one attribute, "hormone receptor status," with values "HR+" and "HR−," and a test instance (a test patient) with "HR+." To determine if this patient is going to suffer recurrence or not, NB would have to base its predictions on the values of $P(\omega_i)$ and $p(\mathbf{x} \mid \omega_i)$. For $i = 2$ classes, $P(\omega_i)$ are simply the probabilities of "recurrence" and "no-recurrence" found in the training data: $P(recurrence)$ and $P(no\text{-}recurrence)$. The values $p(\mathbf{x} \mid \omega_i)$ represent the conditional class probabilities of an "HR+" case, given that "recurrence" and "no-recurrence" events were observed in the training data: $p(HR+ \mid recurrence)$ and $p(HR+ \mid no\text{-}recurrence)$. The final *posteriori* probabilities $P(recurrence \mid HR+)$ and $P(no\text{-}recurrence \mid HR+)$ result from the direct multiplication of the priors and likelihoods since the $p(\mathbf{x})$ term in Equation (2) can be ignored, because due to the sum expression it is the same for both classes. Thus, NB decides on the classification of the test patient based on the *posteriori* probabilities' values; for example, if $P(recurrence \mid HR+) > P(no\text{-}recurrence \mid HR+)$, then the patient is classified as a "recurrence case"; otherwise, as a "no-recurrence" one.

2.3.3. Logistic Regression. Logistic Regression (LR) is a mathematical method that aims to describe the relation between a group of independent variables and a dichotomous dependent variable. To achieve that, LR tries to estimate a set of unknown parameters using a maximum likelihood method [Kleinbaum et al. 2002]. The term "Logistic Regression" may be slightly misleading, since regression is mostly used to build models where the target feature is continuous. However, LR is used for classification, not regression. In brief, LR involves a probabilistic view of classification. It maps a point of a multidimensional feature space to a value in the range [0,1], using a logistic function. The logistic model can therefore be interpreted as a probability of class membership by applying a certain threshold to such probability. In conclusion, LR gives the class probability for each considered feature vector. The class assignment depends on the chosen threshold. One of the main advantages of this method is that it clearly illustrates how the inputs justify the outputs through the final generated equation. However, its performance drops when the dataset contains MD.

To explain the application of LR to BC recurrence scenarios, let us consider the occurrence of recurrence events in a group of women as a function of their age at menarche (for instance, 9–17), as training data. Recurrence events are coded as 1 and no-recurrence events are coded as 0. To determine whether a given patient (test data) will suffer recurrence, LR fits a logistic function to the training data, defining a probabilistic function that maps an input (age at menarche) to a probability of recurrence. The probability of a recurrence event (as all probabilities) will lie on the interval [0,1], but can be translated into a binary class assignment ("no-recurrence" = 0 and "recurrence" = 1), by applying a decision threshold. For instance, if we consider a threshold of 0.7, a patient whose probability of recurrence is 0.85 will belong to the class "recurrence."

2.3.4. K-Means Algorithm. K-Means is one of the most well-known clustering algorithms, due to its easy implementation, efficiency, and success over a wide range of pattern recognition applications [Jain and Dubes 1988]. K-Means is a partitional clustering algorithm, which means that it does not impose a hierarchical structure and finds clusters through the recursive partitioning of data, according to a similarity criteria between data points [Jain 2010]. In brief, the K-Means algorithm works as follows. First, the desired number of clusters, k, needs to be specified. Then, k randomly chosen *centroids* (which are simply pseudodata points with the same dimensionality as the ones intended to cluster) are initialized. The distances (e.g., Euclidean distance) of each point to those k centroids is calculated, and each point is assigned to its closer centroid. The initial partitions of data are defined at this point. However, the objective of K-Means is to find a partition such as the sum of the squared error over all k clusters is minimized [Jain and Dubes 1988]. For that reason, the *centroids* of each partition are updated at each iteration of the algorithm. The new *centroids* are given by the mean vectors of the points belonging to each cluster. Again, new distances are calculated for each point, now considering the new centroids. This is repeated until there are no changes in cluster membership (none of the points changes cluster).

It is worth mentioning that, unlike the previously discussed algorithms, K-Means is an unsupervised learning algorithm. Moreover, despite its simplicity and low computational cost, K-Means has some drawbacks that relate to the number of clusters, the initialization of centroids, and the presence of noisy data [Jain and Dubes 1988; Marques de Sá 2001]. The number of centroids k needs to be specified *a priori*, which sometimes is not trivial, especially without sufficient domain knowledge. Also, since the initialization of centroids is random, different runs of the algorithm may return different results. Finally, K-Means is not robust against noisy data, which may skew the update of cluster centers in some cases.

Imagine that we intended to cluster a group of women that includes "recurrence" and "no-recurrence" cases. K-means clustering works on the basis of finding similar points (similar patients) to group in the same cluster. In theory, "recurrence" and "no-recurrence" patients should map onto two groups with a high intrasimilarity (between elements of the same cluster) and low intersimilarity (between elements of distinct clusters). Note that in practice this is not so linear, given the existent heterogeneity between patients of the same class, as is natural in a disease with such a biological variability as BC [Polyak 2011]. However, to define two distinct groups among these patients, K-means would specify $k = 2$ random centroids, and a distance metric, so that the distance between each pair (i, j) of patients can be computed and similar patients grouped together. The algorithm would iterate, as the centroids adjust to provide a solution that maximizes the distance between the two required groups [Jain and Dubes 1988]. In the particular case of BC recurrence, the idea is that patients with similar values of tumor size and location, age at menarche, hormonal status, and so on are placed in the same cluster. When the K-means algorithm stabilizes, we would end up with two groups. However, both these groups can contain "recurrence" and "no-recurrence" patients (due to the tumor variability, as mentioned).

If the objective is solely to characterize the determined groups, no more procedures are necessary: the analysis would resume to the study of the attributes' values in each group. However, if K-means is to be used for classification, a majority voting scheme is necessary: the most common class in each group defines the class of the group. Then, the centroids of both groups are saved, to allow the classification of new patients: the distance of a new patient to each group's centroid is calculated and the patient assumes the class of the closest group [Zheng et al. 2014].

2.3.5. Bagging. Created by Breiman, Bootstrap Aggregating (Bagging) uses several bootstrap samples to train different classifiers, that are afterwards combined to achieve the final classification results. Each bootstrap sample is created by randomly selecting examples (with replacement) from a training set of size m. From n bootstrap samples, n classifiers $C_1, C_2, \ldots, C_n$ are built [Efron and Tibshirani 1994], each one using a different training set. The final classifier C_b is built from all the C_n classifiers, by combining them through majority voting, where ties are broken arbitrarily (for more details, please refer to Breiman [1996, 1998]).

The application of Bagging to the problem of BC recurrence is straightforward: from a dataset comprising "recurrence" and "no-recurrence" cases, several bootstrap samples are taken and each serves as training data for a particular classifier (e.g., $C_1 = Alg_1$, $C_2 = Alg_2$, $C_3 = Alg_3$). The patients left out for testing are then classified by the chosen group of classifiers. The outputs of each classifier are compared, so that the most frequent class for a given patient is established as its final class assignment: for instance, a patient x_i that achieves a classification of $C_1 =$ *no-recurrence*, $C_2 =$ *recurrence*, and $C_3 =$ *no-recurrence* belongs to class "no-recurrence."

2.3.6. Boosting. Boosting was created to improve the accuracy of a specific algorithms' family called "weak learning algorithms," which are typically slightly correlated to the true classification. On the contrary, strong learners are algorithms well correlated with the true labels, providing good classification results. One of the advantages of weak learners is that they are usually much faster than strong ones. The first Boosting procedure was introduced by Schapire [1990] and worked somehow similarly to bagging. A subset of n examples ($n <$ the total number of training examples N) was taken randomly without replacement from an initial training set (considered to be the training subset Z_1). Z_1 was then used to train a weak classifier C_1. Afterwards, a training subset Z_2 (with $n < N$) was built, containing half the samples misclassified by C_1, and another weak classifier C_2 was trained. Finally, all the samples of the initial training set N for whose C_1 and C_2 predictions disagreed were trained against a third weak classifier, C_3. The final classifier was obtained by a voting scheme of C_1, C_2, and C_3 [Kumar 2012] .

In 1995, Freund and Schapire [1995] introduced the most well-known boosting algorithm, called Adaptive Boosting (AdaBoost). In AdaBoost, the idea is to consider a weighting scheme to select the training subsets. This algorithm starts by considering a maximum number of classifiers M and weighting each training example equally. The misclassified examples get their weights increased for the next classification stages, while the correctly classified examples get their weights decreased. These "weights" simply determine their probability of being chosen for the training set in the next stages and therefore, wrongly labeled examples have a higher probability of being used (learned) again. Moreover, each classifier will get a specific weight attending to its performance in the training set and the final classifier is defined by a linear combination of all the considered M classifiers, each one contributing with its associated weight [Kumar 2012].

Formulated by Friedman et al., LogitBoost is a probabilistic interpretation of AdaBoost. It fits an additive LR model using Newton steps to find estimates for its parameters via the maximum likelihood [Friedman et al. 2000; Kumar 2012]. Instead of using an exponential loss function (which AdaBoost does), LogitBoost minimizes the logistic loss, which makes it less sensitive to outliers (known to be a bad feature of AdaBoost).

The application of Boosting for BC recurrence scenarios is very similar to Bagging's. The only differences are that the classifiers chosen for the ensemble $C_b = \{C_1, C_2, \ldots, C_n\}$ need to be weak classifiers and the training data varies for each classifier: half of the wrongly classified patients from C_1 will be included in C_2's training data

and so forth. At the end of this training step, the test patients are classified according to the same procedure as explained for Bagging.

2.3.7. Linear Discriminant Analysis. Linear Discriminant Analysis (LDA) is a linear transformation technique, generally used to reduce the dimensionality of a dataset in the preprocessing phase, in order to decrease the computational cost of classification and avoid overfitting. Nevertheless, it is also used for classification alone, based on the concept of searching for a linear combination of features that allows the maximization of between-class variance, while minimizing the within-class variance. In other words, the optimization criterion of LDA is to maximize the ratio of between-class and within-class scatter. LDA, also called Fisher Discriminant Analysis (FDA), was first developed by Fisher to deal with only two classes [Fisher 1936]. However, after more than 10 years, this method was extended to deal with multiple classes [Rao 1948].

For simplicity, let us consider the case where a BC recurrence dataset includes a group of patients (training data) described by two attributes ("age at diagnosis" and "number of lymph nodes involved"), and two possible outcomes (classes "recurrence" and "no-recurrence"). LDA would find an optimal linear model that best separates the two classes, which in this example, with two-dimensional data, would simply be a straight line. To understand what is meant by "maximizing the between-class variance" and "minimizing the within-class variance", consider that "number of lymph nodes" is plotted as a function of "age at diagnosis." Additionally, "recurrence" patients are represented as red dots and "no-recurrence" patients as blue dots. The LDA discriminant function, represented by the straight line, is defined in such a way that the red dots and blue dots are close among themselves (minimum within-class variance), but as far apart as possible (maximum between-class variance). After this discriminant function is determined, classifying a new (test) patient is straightforward: the values of age at diagnosis and number of lymph nodes involved are replaced in the function's equation, enabling the computation of the class assignment. For this example, the class assignment can even be determined visually, by plotting the test patient in the same graph where the decision boundary (straight line) is represented and observing in which region the point falls onto.

2.3.8. Support Vector Machines. Support Vector Machines (SVMs) were first introduced by Vladimir Vapnik for two-class classification [Vapnik 1999]. Basically, this algorithm tries to find the optimal decision hyperplane that maximizes the separation margin between data points of distinct classes [Boser et al. 1992]. The middle of the separation margin defines the decision boundary (optimal hyperplane) and the data points that are closest to it are the support vectors. SVMs belong to the general category of kernel methods. Kernel methods can operate in high-dimensional spaces, since they depend on the data only through dot-products. This has two main advantages: it allows the generation of nonlinear decision boundaries and enables the classification of data that has no obvious fixed-dimensional vector space representation [Scholkopf and Smola 2002; Shawe-Taylor and Cristianini 2004]. SVMs are known for excellent classification performance, since they can handle high-dimensionality problems and have a good generalization behavior. They balance the model's complexity against its success at fitting the data, which translates into a successful trade-off between the model's flexibility and the error in training data [Scholkopf and Smola 2002]. However, and despite being a white-box algorithm, it requires a comprehensive understanding of how it works. When training a SVM, researchers have to face several decisions concerning the preprocessing stages of the input data and the SVM's hyperparameters (e.g., kernel function, regularization constant).

Consider a two-dimensional and two-class ("recurrence" and "no-recurrence") linearly separable recurrence problem, where patients are characterized by "size of tumor"

and "age at diagnosis." There are several decision boundaries that could be defined to separate both classes, such as the ones resulting from LDA or Multilayer Perceptrons. However, the optimization criteria behind SVMs guarantees that the margin of the decision boundary is maximized, and therefore, the decision boundary will be as far away from both "recurrence" and "no-recurrence" cases as possible. In a scenario where the recurrence problem is nonlinear, SVMs use a kernel function to transform the data points to a higher dimensional space, making classification easier. For instance, imagine that in the original input space, the patients' data $x_i = \{v_{1(tumor\ size)}, v_{2(age)}\}$ of class "recurrent" form a filled circle in the input space (e.g., red dots), while the patients' data from class "no-recurrent" (e.g., blue dots) surround that circle. A linear boundary in that input space would be impossible to achieve. SVMs solves this problem by mapping the data points onto a higher dimensional (three-dimensional, in this case) space, by adding a new feature $v_3 = {v_1}^2 + {v_2}^2$. Since "recurrent" data points have lower values of v_1 and v_2, their values of v_3 will be lower as well. A projection onto an input space defined by v_1 and v_3, for instance, would show a linearly separable problem (in fact, v_3 alone is enough to separate both classes).

2.3.9. k-Nearest Neighbors. k-Nearest Neighbors (KNN) is a supervised classification algorithm in which the k nearest neighbors of a point are chosen, found by minimizing a similarity measure (e.g., Euclidean distance, Mahalanobis distance) [Altman 1992]. To determine the class of an unlabeled example, KNN computes its distance to the remaining (labeled) examples, and determines its k-nearest neighbors and respective labels. The unlabeled object is then classified either by majority voting—the predominant class in the neighborhood—or by a weighted majority, where a greater weight is given to points closer to the unlabeled object. The major drawback of KNN is related to the fact that it is a lazy learning algorithm. That means that there is no "model": the training data is not used to perform any generalization. Therefore, whenever KNN searches for each instance's nearest neighbors, it needs to go through the entire dataset, which is especially problematic for large databases. Another issue is finding the optimal number of neighbors (k) and the most appropriate distance metric to use. This requires a careful study of the dataset and the development of several KNN models, in order to achieve the best results.

For BC recurrence problems, KNN can be used in a similar way to the k-means algorithm, except that no groups are formed. To determine the class assignment of a test patient, its distance to all the training patients included in BC recurrence dataset is computed. A majority voting or weighted scheme is used to choose the class the patient belongs to, according to the class assignment of its k closest patients. Imagine a patient x with three nearest neighbors y (recurrent), w (recurrent), and z (no-recurrent). A 3-Nearest Neighbor scheme with majority voting would assign patient x to class "recurrent." The value of k has to be defined a priori, or set to a range of values (say, $k = 1, \ldots, 30$) to be evaluated one by one. The k value that maximizes the KNN algorithm performance (see Section 2.5 for more information on how to measure performance) is chosen to be the best k for the given dataset [Medjahed et al. 2013; García-Laencina et al. 2015].

2.3.10. Association Rule Learning. Association Rule Learning allows one to unveil the relationship among variables in a dataset. Proposed by Agrawal et al. [1993], this method assumes that all variables are categorical and because of that it is not a good algorithm to deal with numerical data. Each identified association rule follows two main concepts: support and confidence. Support identifies the percentage of the population that follows a specific rule. Confidence is the measure of certainty associated with each discovered rule. In a simple manner, association rules can be perceived as "if-then" rules that describe relations between the data. They are extremely advantageous due

to their exhaustive exploration of the data [Molina et al. 2013]. Moreover, the final rules returned by this algorithm are usually simple enough to be understood by users. Nevertheless, some of their inconveniences are that they are affected by noisy data and have a slight tendency to overfit the data.

Association Rule Learning may be used to discover interesting relationships between attributes related to BC recurrence. Imagine a BC recurrence dataset comprising the following attributes: tumor size, age, HR status, race, and family history of BC. In association rule mining, the objective is to find frequent cooccurrences of items, in this case, attribute's values that seem to frequently appear together. An an example, a relationship between race and hormonal status can be found, if several patients show associated values for those two attributes—for instance, $\{Caucasian, HR+\}$ or $\{Asian, HR-\}$. Similarly, associations between larger itemsets are also possible. For instance, an association rule including race, family history of BC, and tumor size could be $\{Caucasian, Yes \Longrightarrow \geqslant 2cm\}$ meaning that "*if* the patient's $race = Caucasian$ and the *family history of breast cancer* $= Yes$, *then* the patient will have a tumor size $= \geqslant 2cm$." This rule will have associated measures of Support and Confidence that allow researchers and clinicians to assess its relevance with respect to the context and objectives of the study.

2.3.11. Isotonic Separation. Isotonic Separation, developed by Chandrasekaran et al. [2005], is a linear programming model that follows the principles of isotonic consistency. The isotonic consistency constraint assumes an ordering relation of data points in the feature space, given by $S = \{(i, j) : a_i \geq a_j\}$, where a_i and a_j are coordinate vectors that represent the attribute values of i and j, in all d dimensions (for more details please consult Ryu et al. [2007b]). Therefore, S consists of (i, j) pairs of ordered data points such that, considering a two-classification problem:

—If i is classified as belonging to class ω_1, then j must be classified as belonging to class ω_1 and conversely.
—If j is classified as belonging to class ω_2, then i must be classified as belonging to class ω_2.

As an example application of this scheme to BC recurrence, consider, for instance, that certain patients have registered values of age, size of tumor, expression of estrogen receptor, and so on, that cause them to be classified as "recurrent"; then all patients registering the same (or greater) values are also considered "recurrent." Isotonic separation also takes into account misclassification costs, where each misclassified data point receives a penalty, for instance, $\alpha > 0$ for each "recurrent" patient classified as "nonrecurrent" and $\beta > 0$ for each "nonrecurrent" patient classified as "recurrent." Isotonic Separation minimizes the total cost of misclassification, $\alpha n_i + \beta n_j$, where n_i is the number of wrongly classified recurrent patients and n_j is the number of wrongly classified nonrecurrent patients.

2.3.12. Random Forests. Created by Breiman [2001] Random Forests (RFs) instantly became a commonly used method, mainly due to its simplicity (in terms of training and tuning) and performance [Trevor et al. 2009]. Similar to the bagging algorithm (in the sense that it uses individual DTs as individual classifiers), RFs construct correlated trees. However, in this case, for each tree node, v features out of the total V input features are randomly selected (considering $v \ll$ V) and the best split of v features divides the node. Finally, the forest picks the most voted class, over all the trees in the forest, either by considering the mode of the classes of the individual trees (classification) or their mean (regression). By designing a multitude of DTs and later combining their predictions, RFs decrease the risk of overfitting, usually associated with individual DTs [Verikas et al. 2011].

As for BC recurrence scenarios in particular, RFs work very similarly to individual DTs. The differences are that (i) the selection of training data and input features is random (this is why they are called "Random"); and (ii) the final class of a patient is based on the classification results obtained by several individual trees (creating a "Forest"), which resembles Bagging procedure as well. In the training phase, several individual C_n DT models are generated, using a random partition of the patients' dataset. Then, to determine the class of a given test patient, her input data is tested by all the considered C_n individual trees (as explained in Section 2.3.1). The final class assignment is determined by majority voting.

2.3.13. Neural Networks. Artificial Neural Networks (ANNs) [McCulloch and Pitts 1943] are mathematical-computational models inspired by neuronal cells' functioning, simulating human reasoning. A generic ANN model is composed by three layers: the input, output, and processing layer (or hidden layer) [Minsky and Papert 1969]. The input layer receives the data, while the output layer communicates the result. The hidden layer is responsible for data processing and results' calculation. ANN analyzes existing patterns in the information they receive and derive associations between input and output variables. These associations are used to produce the most correct output for each input, which is then compared to the correct output and, based on this comparison, the algorithm resets the associations between the input data and the previously determined output. This process continues iteratively until the correct result is determined or the maximum of iterations is achieved. Then, the system memorizes the model of such association between inputs and outputs in order to classify new cases.

A Multilayer Perceptron (MLP) is a modification of the standard linear perceptron and can distinguish nonlinearly separable data [García-Laencina et al. 2013]. In its basic form, it is simply a type of feed-forward ANN. It consists of multiple nodes interconnected in a directed graph, where the input layer passes the input vectors to the network and the output layer communicates the response. A MLP can have one or more hidden layers, composed by neurons with nonlinear activation functions (e.g., sigmoid, tangential), responsible for the computation of results.

The major advantage of ANN models is that they avoid the construction of "if-then" rules, and its definition by experts. They also do not need a very large set of data to produce estimates, though the larger the training set, the more accurate the results. On the other hand, the training phase can be time-consuming. However, the main disadvantage of this type of algorithms is their model's interpretation—these algorithms are black-box models, since the associations between data are complex and difficult to explain.

Consider a BC recurrence problem where patients are described by two features: "tumor size" and "age at diagnosis," and the objective is to determine the outcome of such patients, their "recurrence state" (whether a patient suffers "recurrence" or "no-recurrence"). An ANN approach based on this scenario could consider two inputs in the input layer and two outputs on the output layer. It is also necessary to specify the hidden layers configuration (number of layers and number of neurons each of them contains), their activation function, and the initial weights of synapses. When training the artificial neural network, the hidden layers can iteratively learn the weights of the synapses that achieve the most accurate classification results. This means the ANN can achieve the best matrix of weights that, multiplied by the inputs "tumor size" and "age at diagnosis," achieve the most accurate values of "recurrence state" (1/0), coded as "recurrence" = [1 0] and "no-recurrence" = [0 1]. It is important to notice that for classification problems, ANNs do not output a single class assignment, like most of the described algorithms. For each patient, they will output a "probable answer for each class,", for instance, patient's x_i output could be [0.84 0.16], which means patient x_i has

an 84% chance of "recurrence" versus a 16% chance of "no-recurrence." This result can be easily transformed to a final assignment of "recurrence state" = 1, so that the results are comparable with other ML algorithms. When the training stage is complete, the ANN model (the final matrix of weights) is used to classify the test patients, following the same logic.

2.3.14. Self-Organizing Maps. Self-Organizing Maps (SOMs) are a type of artificial networks that use a form of unsupervised learning (competitive learning) to represent the input data in a low-dimensional space (a map), typically with one or two dimensions [Kohonen 1995]. A SOM network is built from a grid of neurons ("nodes"), where each node has a specific position in the grid and is completely connected to the input layer. Furthermore, each node is associated with a weight vector, which has the same dimension as the input feature space: feature vectors of d dimensions will origin nodes with weight vectors of size n. Like most ANNs, SOMs perform training and testing, or in this case, training and mapping. In the training phase, SOMs build the map using the input examples, by placing each one next to the node with the most similar weights, known as Best Matching Unit (BMU) [García-Laencina et al. 2010]. The BMU's and adjacent nodes' weights iteratively adapt every time a new training input is given to the SOM. In the mapping phase, the test input vectors are classified according to their distance to the existing nodes in the map, constructed in the training phase.

In general (supervised learning) ANNs, the target is required to guide the update of the network's parameters (e.g., weights). Therefore, when training ANNs with a group of patients $x_n = \{v_{1(tumorsize)}, v_{2(age)}\}$, their class assignment is fundamental: without it, the model is unable to adjust its parameters. However, for SOMs, the training stage is absolutely independent of the class assignment. SOMs are completely data-driven, and thus only the patients data ("tumor size" and "age at diagnosis" in this example) is required to build the SOM map. The map's weights are adjusted until each patient is placed next to its most similar neighbor, in a way, clustering the input data. After the map is complete and its weights are defined, the training class most frequently assigned to a neuron in the map becomes its label (majority voting). This labeling allows for the test patients to be assigned according to their distance to the existing "clusters": the nodes that better represent patients with similar characteristics to theirs (most similar in its weight structure) [Li and Eastman 2006].

2.3.15. Classification Restricted Boltzmann Machines. Restricted Boltzmann Machines (RBMs) are a variant of Boltzmann machines, with the constraint that there can be only interlayer connections, that is, there cannot be connections of nodes within a layer, only between layers. RBMs can be seen as stochastic neural networks, given their neuron-like units whose activation has a probabilistic element [Fischer and Igel 2014]. They typically have one layer of visible units (inputs), and one layer of hidden units. They may or may not have a bias unit. Each visible unit is connected to all the hidden units, and the connections are symmetric, meaning that each hidden unit is also connected to all the visible units. As mentioned previously, no hidden unit is connected to another hidden unit, and no visible unit is connected to another visible unit.

Despite being mostly used as unsupervised learners, RBMs can also be used as supervised, black-box algorithms for classification [Larochelle and Bengio 2008]. Classification Restricted Boltzmann Machines (ClassRBMs) are a variant of RBMs oriented to classification. In classification tasks, RBMs are treated as parametric models (considering that the number of hidden layers is fixed) of the joint distribution between the layer of hidden units (neurones) and the visible layer of inputs. Based on this joint probability, ClassRBMs can compute the distribution $p(y \mid x)$, that is, the probability of x belonging to y, which allows the determination of the most probable class label (see Larochelle et al. [2012] for more details).

Consider a two-class BC recurrence scenario ("recurrence" versus "no-recurrence"), where patients are characterized by n features (e.g., "hormone status," "age at diagnosis," "tumor size," "lymph nodes involvement"). Based on these features, a RBM will try to discover the latent factors that may explain their expression. In a BC recurrence problem, a RBM should be able to learn two latent units (hidden units) underlying a patient's probability of recurrence, that is, two natural groups in the patients' training set, one that corresponds to "recurrence" events, and another to "no-recurrence" events. During the training phase, the states of the visible units (patients' input features) are set and the hidden units are updated accordingly until the network converges (see Fischer and Igel [2012] for further details). When classifying a test patient, the RBM takes her feature values $\{v_{1(hormone\text{-}status)}, v_{2(age\text{-}diagnosis)}, v_{3(tumor\text{-}size)}, v_{4(lymph\text{-}nodes)}\}$ and determine which of the hidden units was activated by her characteristics. However, like for most artificial networks, there is not a specific hidden unit that is completely activated. All units are activated with associated probabilities. Therefore, if patient x_i has characteristics of a recurrent cancer, the hidden unit representing "recurrence" events will be activated with a higher probability that the "no-recurrence," which is how her class assignment is determined.

2.3.16. Genetic Algorithms. Genetic Algorithms (GAs) are inspired in Darwin's evolutionary theory which explains the evolution of species through natural selection. As species evolve in order to adapt to their environment, a GA also uses a "survival of the fittest" philosophy in order to obtain the result that best fits the data from a population of individual potential solutions [Mitchell 1996]. A fitness function determines which solutions should be kept and which should be eliminated. At each generation, a new population is generated and the fitness values of all individuals are evaluated based on their performance in the problem domain. Three main genetic operators can actuate over each selected population, so as to generate the next-generation population—copy, crossover, and mutation. These mechanisms are repeated, and the population continues evolving, until the optimal solution (fitness value) is produced, or a stopping condition is reached (e.g., a maximum number of generations).

In the prediction of a BC recurrence problem, GAs can be used as an optimization method for other algorithms, such as ANNs. Imagine an initial population of 40 randomly generated ANNs (each ANN will have its own configuration, i.e., a different number of hidden layers and number of nodes per layer), and a fitness function determined by the performance of the ANN as a classifier for the BC recurrence problem. At each generation, each ANN is evaluated, and the ones with the best fitness values (best classification performance) move on to the next generation; also, new ANNs are generated by combining different individuals with good results (for instance, using the configuration of the first hidden layer from one ANN and the configuration of another hidden layer from another ANN; mutations can also occur, for instance changing the number of nodes in one specific layer) so that the number of individuals in the population remains stable. This process is repeated until an optimal ANN is found or the maximum number of generations is reached (at this point, the best ANN is selected).

2.4. Sampling Strategies

To evaluate a classifier, researchers need to find its true error rate, that is, the classifier's error rate in the entire population. However, in real-world applications, it is not possible to access the entire population. Only a finite set of examples is available, from which an estimation of the true error rate must be calculated. A naive approach to this issue of finite datasets would be to use all the available data to train and test the models. However, this would return an overly optimistic error estimation and serious overfitting [Marques de Sá 2001]. For that reason, another approach must be pursued:

the division of the available (and labeled) examples into training and test sets. In the context of BC recurrence problem, this approach means that the patients records will be split into two sets: one (training set) used to build the model and another (test set) used to evaluate the performance of the model. The techniques to divide data into training and test sets are called sampling strategies and in this section we will review the ones used in the reviewed works: holdout method, random subsampling, k-fold cross validation, and leave-one-out [Marques de Sá 2001; Arlot and Celisse 2010; Han et al. 2011; Duda et al. 2012].

2.4.1. Holdout Method. The holdout method simply divides the available examples into two disjoint sets, according to some percentage. Traditionally, train and test sets are divided in a 50%–50% partitioning scheme [Marques de Sá 2001], although most authors consider a train/test division of 70%–30% or 80%–20%. Holdout is the simplest of the sampling strategies, but along with its simplicity come some limitations. In general ML problems, one of the holdout method's shortcomings is that it may be subjected to "unfortunate splits," and therefore the training data may not be representative of the population, leading to biased results [Marques de Sá 2001; Bishop 2006]. This issue is even more evident with scarce and imbalanced datasets that arise in BC recurrence contexts, where (i) setting aside an even smaller part of the dataset for testing is not acceptable and (ii) the probability of unfortunate splits is higher, since "recurrence" and "no-recurrence" classes are not equally represented [Srivastava 2013; Menardi and Torelli 2014]. To overcome these limitations, several other sampling methods were proposed, namely, random subsampling, k-fold cross validation, and leave-one-out, which we describe herein.

2.4.2. Random Subsampling. In random subsampling, we consider several p experiments (runs or splits). Each split considers a fixed number of random training and test examples, selected without replacement. Then, for each split, training and testing is performed individually. The individual error rates e_i (determined using each test set) are averaged to form the final error estimate, according to Equation (3). When choosing the samples for each split, the samples cannot be repeated. However, between splits, the same samples may be selected. Therefore, random subsampling does not guarantee that all samples are used for training and testing, which constitutes its major drawback.

$$error = \frac{1}{p}\sum_{i=1}^{p} e_i. \tag{3}$$

2.4.3. k-Fold Cross Validation. k-fold cross validation divides the data into k subsets (folds) that rotate, in order to consider all folds for both training and testing. More specifically, k-fold cross validation considers $k > 1$ distinct folds, where $k - 1$ folds are used to train a classifier and the left-out fold is used for validation. This is performed k times, so that every fold is considered in both training and test design. Similarly to random subsampling, the true error rate is estimated by averaging the each error e_i, obtained from each fold.

The choice of k influences the bias-variance trade-off in performance estimation through k-fold cross validation. For small values of k, the bias increases, although the variance is low; for higher values of k, the error estimate is closer to the true error (low bias), but the variance increases.

2.4.4. Leave-One-Out. Leave-One-Out (LOO) is a particular case of k-fold cross validation, when $k = N$, the total number of available examples. Therefore, in LOO, $N - 1$ examples are used for training, while the held out example is used to test the classifier.

Table III. Confusion Matrix

		Actual Class	
		Negative	Positive
Predicted Class	Negative	True negative (TN)	False negative (FN)
	Positive	False positive (FP)	True positive (TP)

Thus, there are N error estimates that need to be averaged to determine the final estimate of the error rate.

In LOO, only one sample is used for testing, which leads to a high variance in error estimation. On the other hand, since all $N-1$ are used in the training design, the bias is low. For that reason, the averaged test set error is a good estimate of the performance error. When the sample size is low, LOO is the best approach to provide an accurate estimate of the true error [Bishop 2006; Markov and Larose 2006; Santos et al. 2015].

2.5. Evaluation Metrics Background

The performance evaluation of a classifier is normally based on a confusion matrix (Table III). This matrix illustrates the actual versus the predicted class in classification problems, where each column of the matrix represents the instances in an actual class and the rows represent the instances in a predicted class.

True positives (TP) and true negatives (TN) represent the number of examples correctly classified in the positive and negative classes, while false positives (FP) and false negatives (FN) represent the number of misclassified positive and negative examples, respectively. Accuracy (or inversely, the error rate, which is 1-accuracy), are the two most widely used measures to evaluate the performance of a classifier. Accuracy represents how many predictions of the classifier were in fact correct (Equation (4)), whereas the error rate is the percentage of misclassified examples in total.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4}$$

Nevertheless, the accuracy/error might not be appropriate performance measures for imbalanced datasets where the class priors are very different, because they will be strongly biased toward the majority class [He and Garcia 2009]. As an example, consider the diagnosis of BC recurrence where 98 out of 100 women have nonrecurrent tumors (negative class), whereas the remaining two have recurrent tumors (positive class). A classifier that labels all the examples as negative would have an accuracy of 98% and an error of only 2%, which may seem a good classification result, but in fact, the classifier has failed all the examples of the class of interest ("recurrence" class). For that reason, an imbalanced scenario requires alternative evaluation metrics that consider the performance of both positive and negative class independently [He and Garcia 2009; Longadge and Dongre 2013].

Recall or sensitivity represents how many positive examples the classifier was able to correctly identify (in the BC recurrence problem this is the percentage of patients with recurrence identified as such) (Equation (5)).

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{5}$$

Specificity represents how accurately the classifier behaves in terms of predicting the negative class (in the BC recurrence problem this is the percentage of patients without recurrence identified as such) (Equation (6)).

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{6}$$

Precision shows the proportion of the correctly predicted positive cases relative to all the predicted positive ones (in the BC recurrence problem this is the percentage of patients identified as having recurrence that actually recur) (Equation (7)).

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{7}$$

Mean Square Error (MSE) of a classifier represents the difference between a vector of n predictions ($\hat{Y}_i$) and the true observable vector (Y_i) for all n examples (Equation (8)).

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2. \tag{8}$$

Other measures that can also be used are Area Under the Curve (AUC), F-measure, and Cohen's kappa. The AUC values measure how well a classification model can distinguish between two classes by representing the trade-off between TP and FP values. AUC is often used when a representative measure of discrimination is needed and it can even replace accuracy as a performance measure [Huang 2005]. In its turn, the F-measure is defined as the harmonic mean of precision and recall, providing a balance between both performance metrics (Equation (9)), that better reflects the performance of a classifier in the presence of an underrepresented class [Chawla 2010].

$$\text{F-measure} = \frac{2 * precision * recall}{precision + recall}. \tag{9}$$

Nevertheless, F-measure does not take the TN cases into account, and therefore measures such as the Cohen's kappa (kappa values) [Cohen 1960] may be used to complement the classifier's evaluation. The kappa values measure the agreement between the classifier's predictions (predicted class) and the real outcomes (actual class), as follows:

$$K = \frac{P(a) - P(e)}{1 - P(e)} \qquad \text{with} \qquad \begin{cases} P(a) = \frac{TP+TN}{N} \\ P(e) = \frac{(TP+FN)\cdot(TP+FP)+(TN+FP)\cdot(TN+FN)}{N^2} \end{cases}. \tag{10}$$

N represents the total number of cases, $P(a)$ is the percentage of agreement between the actual and predicted classes (and is equal to accuracy), whereas $P(e)$ stands for the chance agreement, the hypothetical probability that the predicted classification outcomes match the actual outcomes [McHugh 2012]. K can range from -1 to 1: a perfect agreement will achieve $K = 1$, while a $K = -1$ represents a perfect disagreement. Overall, kappa values lower than 0 indicate no agreement, since $K = 0$ itself relates to the agreement expected from chance [Eugenio and Glass 2004].

Note, however, that in the study of imbalanced datasets, researchers are most often interested in achieving the high TP values and the lowest FN values as possible, since the positive class is more rare, as explained in Section 2.1. This does not mean that the TN rate should be overlooked, but explains why kappa values are not commonly reported as a measure of classifier performance in most ML studies on imbalanced data in general [Chawla 2010; Kotsiantis et al. 2006; Ferri et al. 2009; He and Garcia 2009; Ganganwar 2012]. This is also true for BC recurrence studies in particular, which are predominantly imbalanced. In fact, only one of the reviewed studies (Jonsdottir's study) has included kappa values as an evaluation metric [Jonsdottir et al. 2008].

3. APPLICATION OF BC RECURRENCE

In 1997, Mani et al. [1997] compared the performance of rule-based classifiers (DTs and Association Rules) with a well-known probabilistic classifier (Naive Bayes), in the identification of tumor features associated with BC recurrence. The data was collected from a BC center in California, where 887 patients were characterized by demographics and tumor-specific information, including diagnostic and treatment features. From the initial set of features, six were handpicked by a medical expert to proceed with the study. Since only 10% of the collected patients suffered from a recurrence event (85 patients), the majority class (no-recurrence) was randomly divided into six datasets, in order to follow a 60%–40% class distribution (no-recurrence/recurrence) in each dataset. NB proved to be the best approach with an average accuracy of 68.3%, overcoming all others in the majority of the tested datasets. A particular type of association rule, the First Order Combined Learner (FOCL) has also stood out with an average accuracy of 66.4%.

This work raises a common controversial topic within the bioinformatics community: the trade-off between classification results and interpretability. The use of rule-based classifiers is generally encouraged in medical contexts [Mani et al. 1997; Intrator and Intrator 2001; Zhou and Jiang 2003], due to the additional information they provide. However, as shown by this work, they did not offer leading results in terms of classification accuracy. Moreover, most of the generated rules reflected a somewhat obvious knowledge domain, which does not constitute a meaningful contribution to medical experts. From a technical perspective, there are some points to be further discussed in this work, namely, the feature selection phase, the sampling phase, and the evaluation metrics used. Although some works make use of clinical guidance to select relevant variables to study, they should be clearly stated to allow a proper comparison with related works. Furthermore, the explanation of the division of patients into the considered six datasets is too vague. As is stated, it seems that the division of the majority class (no-recurrence) followed a random subsampling method, which does not guarantee that all samples are used to model/test the classifiers. Therefore, important information might be unused due to the sampling phase design. Finally, in spite of the authors' efforts to design datasets with more appropriate class distributions, they still suffer from a considerable imbalance, which requires the evaluation of algorithms to go beyond the traditional accuracy measure, more efficiently applied to balanced classes. However, accuracy was the only metric used, which may hint on misleading conclusions.

In 2003, Jerez-Aragonés et al. [2003] employed a hybrid model, combining ANNs and DTs, to a database from a hospital in Malaga, Spain, in order to determine whether a patient will suffer a postsurgical relapse at any period during follow-up time, considering an endpoint of 5 years. Out of 85 available features (including demographics, postsurgical and treatment information), a subset of 14 features was selected by medical experts as the most relevant for predicting outcome. The hybrid model was then used to predict BC recurrence for seven different time intervals from the surgical intervention. The first six intervals are equally spaced (10-month periods), while the remaining one considered a period of over 60 months. Initially, the dataset was constituted by 1,035 patients, but records with MD were discarded, resulting in a decrease of the sample size: from 845 patients for the first interval to 466 patients for the last interval. Using a holdout method (80% train/20% test), the performance of the proposed hybrid approach was compared to a Cox statistical technique, commonly used by medical experts. In terms of accuracy, the proposed approach outperformed the Cox model in all intervals (with results ranging from 93.4% to 96%), except for the last interval ($>$60 months), lagging behind by just 0.3%. To complement the accuracy analysis of the

proposed method, recall, precision, and specificity measures are also included. Recall varied between 78.7% and 88.7%, while precision ranged from 64.8% to 77.2% and specificity between 94.5% and 97.2%.

The combination of ANNs with DTs is an interesting approach, since it does not discard the advantages of one in favor of the other. DTs provide useful information for selecting the most relevant prognostic factors for each considered interval, while ANNs are able to use that information to make an accurate prediction, using personalized topologies for different time intervals. By not choosing one algorithm over the other, the authors take advantage of each one's potential, achieving accurate, yet interpretable results, which is an improvement from the previous illustrated work. Also, in this work, the authors have in mind that accuracy is not always the best classification metric, and complement this information with additional metrics (such as sensitivity and specificity), allowing for a better evaluation of the power of the proposed method. Nevertheless, some topics remain for discussion. Although the authors mention that their approach is appropriate for data with a considerable number of features with missing values, this is not supported by the work itself, since the MD perspective is ignored. Furthermore, its application for a high number of features is not discussed yet, since a medical team performed the feature selection phase beforehand.

Amir Razavi et al. followed the idea of combining DTs with other algorithms to improve the prediction of BC relapse during the first 5 years after diagnosis [Razavi et al. 2005]. In 2005, they applied Canonical Correlation Analysis (CCA) as a preprocessing step, prior to classification, to study the influence of dimensionality reduction in prediction performance. They used a dataset obtained from a Swedish regional center, with 3,949 patients characterized by more than 150 features. Following the same methodology as the previous discussed works, the feature selection phase was performed by a team of medical experts, resulting in a decrease of the feature space to only 17 predictors. However, unlike previous works, values for MD fields were imputed using the EM algorithm [Razavi et al. 2005] (explained at the end of Section 2.2).

A 10-fold cross validation procedure was used to evaluate the performance of three different predictive models: (i) DT coupled with CCA, (ii) DT without any preprocessing step, and (iii) DT with MD imputation as the only preprocessing step. The results showed that DT coupled with CCA overcame the other two approaches in terms of accuracy (67%) and specificity (63%); yet lagging behind both in terms of sensitivity, which is generally not a good indicator. However, it is important to state that this solution yields trees with only 10% the size of those without preprocessing, resulting in a simpler system, and improving interpretability. Still, it would be interesting to make a comparison between rule-based models and other types of classifiers, and their behavior when coupled with the mentioned preprocessing strategies. In 2007, the same authors applied the previously developed combined model (CCA + DT) to predict BC recurrence within 4 years after diagnosis [Razavi et al. 2007]. The used dataset consisted of 3,699 patients (with absent observations), where 664 (18%) suffered recurrence in the first 4 years of follow-up. MD imputation was performed using Multiple Imputation (MI), and 10-fold cross validation was used to estimate performance error. A hundred cases were previously separated from the initial dataset (by stratified random sampling) to validate the developed model against the predictions of two medical experts (Oncologists 1 and 2).

Although the comparison of the Area Under the Curve (AUC) values between the three approaches (DT and two oncologists) did not significantly differ, a more detailed analysis on the performance results is required. The authors present the confusion matrix for the validation set of 100 patients (81 without and 19 with recurrence) but at no point do they make a critical comparison of the results. In fact, despite a higher accuracy (82%) and precision (57.1%), DT is overpowered by one of the two medical

doctors collaborating in the study (Oncologist 2) in terms of AUC values, specificity, and F-measure. With a training and validation set presenting such a high imbalance (80% without recurrence, 20% with recurrence), the sensitivity should assume a pivotal importance. Sensitivity results are poor for DT (21.1%); however, Oncologist 1 had an even lower rate of sensitivity (5.56%), despite having a higher percentage of specificity (97.5%).

The poor results of DTs in this work might be explained by the dominance of "no-recurrence" cases in the training model. Although the authors discuss the usage of sampling strategies to balance the dataset, they do not apply them, presenting arguments such as the small sample size or the lack of representativeness they would generate. However, this should have been considered since it is not clear that the proposed approach is the most suitable to predict BC recurrence: Oncologist 2 had a sensitivity rate of 57.8% without severely compromising precision (50%), by analyzing patient records with MD.

In the same year, Sun et al. [2007] combined clinical and genetic information to create a "hybrid signature," capable of predicting BC recurrence in the first 5 years after diagnosis. This work makes use of microarray data, publicly available in the Nature website [Nature Publishing Group 2015]. The dataset includes 97 patients, 46 of which suffered recurrence, while 51 remained recurrence-free. According to previous works using this dataset [Guo-Zheng 2011], it also contains MD, although this perspective is not addressed throughout the work. Preprocessing steps before classification include data standardization (Min-Max) and a feature selection algorithm (I-RELIEF), developed earlier by Sun [2007]. Four different approaches are tested: one using only genetic markers; another using only clinical markers; a hybrid signature (including genetic and clinical information); and St. Gallen's criterion [Harbeck et al. 2013], a consensus criterion to determine recurrence used in oncology guidelines. To compare the performance between the approaches, the authors specified a threshold for each one, in a way the sensitivity is 90% for all. Then, the comparison was done by analyzing the corresponding specificity values: 47%, 48%, 67%, and 12% for genetic-only, clinical-only, hybrid signature, and St. Gallen approaches, respectively. The Receiver Operating Characteristic (ROC) curves for the first three approaches were also compared, with the hybrid signature outperforming the other two (which in turn showed a similar behavior).

It must be noted that this is the work that includes the smallest number of patients. With a small sample size, there is a higher danger of overfitting the training data. To avoid this problem, a nested Leave-One-Out Cross Validation (LOOCV) is adopted [Sun et al. 2007]. In a nested cross validation, an inner loop is responsible for the selection of the optimal classification parameters (for I-RELIEF in this case), while in the outer loop the classification of the held-out sample is performed. Linear Discriminant Analysis (LDA) is used in the classification task (outer loop), since it does not require the estimation of hyperparameters, and thus makes the experiments computationally less expensive.

Before Sun, some authors had previously attempted to combine genetic and clinical information, but rather unsuccessfully [Dettling and Buhlmann 2004; Gevaert et al. 2006], which reinforces the achievements of this work. On one hand, one may argue that 67% is far from an optimal specificity result. On the other hand, it must be noted that the proposed hybrid signature improved the specificity of the remaining approaches by nearly 20% to 60%. This proves that the combination between genetic and clinical information is a suitable approach to determine the prognosis of BC patients, even though combination strategies are difficult to design.

Also in 2007, Ryu et al. [2007a] used the Ljubljana BC dataset (available in UCI Repository [Lichman 2015]) to compare several methods to predict BC recurrence in the

first 5 years after removal of tumor: isotonic separation, robust Linear Programming (LP), and three variants of DT (C4.5, OC1, and QUEST), SVM, AdaBoost, and learning vector quantization. The dataset contains 286 patients, where 201 (70.3%) did not suffer recurrence and the remaining 85 (29.7%) had recurrence events. Each patient is characterized by nine features, and there are some missing values. In particular, "node-caps" and "breast-quad" are responsible for the nine missing observations present in this dataset. From those nine missing values, five belong to the "no-recurrence" class, while the other four belong to the "recurrence." All the methods were evaluated according to a holdout method (70% for train and 30% for test), except for QUEST decision three, which has its own sampling scheme (threefold cross validaton). The results are presented in terms of error rate; however, to make them comparable with the remaining works, we translated them to accuracy values ($1 - error$). Isotonic separation outperformed all others with 80% of accuracy. Only the accuracy (conversely, the error rate) was determined for each classifier, which may be considered a limitation, since the dataset is known to have a 29.7%/70.3% of recurrence/no-recurrence distribution [Lichman 2015]. Moreover, a backward sequential elimination process for feature selection showed that age, menopause status, node capsules, tumor grade, and irradiation were the most relevant features for recurrence.

The most comprehensive study in terms of tested classifiers, feature selection algorithms, and performance measures was performed by Jonsdottir et al. [2008], in 2008. They implemented 17 classification algorithms, including NB, several variants of DT and other rule base classifiers (OneR, PART, Jrip), LR, and some metaclassifiers, including boosting, bagging, and ensemble schemes. Moreover, this work also uses a wide range of feature selection algorithms, such as OneR, Correlation-based Feature Selection (CFS) method, Las Vegas Filter (LVF) algorithm, RELIEF, information gain, and C4.5 decision tree. Furthermore, the existing knowledge domain (from previously published works in BC, medical experts, and authors' experience) is also explored to select the most relevant features. However, the results from these feature selection methods are not discussed. The authors do not state if one returned better results than the other, or, alternatively, if each one's results were combined to select a final subset of features. The classification results were evaluated in terms of accuracy, kappa values, AUC, sensitivity, and specificity. The algorithms were run on a relatively small dataset (257 patients) with high dimensionality (400 features), obtained from the University Hospital of Iceland (Rose dataset). Jonsdottir's study focused on two main goals: (i) predicting BC recurrence within 5 years after diagnosis and (ii) predicting recurrence risk (low, intermediate, or high) within the same time period. The latter is out of the scope of this review, although the recurrence risk, not as outcome but as predictor, was also included in (i), in order to determine if the inclusion of this feature had any influence in predicting recurrence events. In order to fulfill objective (i), the authors conducted a feature selection phase that resulted in three different datasets:

(1) Base-DS, including 98 features selected according to the experience of a medical expert and the results of the feature selection methods;
(2) Med-DS, with 22 features selected from Base-DS by a medical doctor; and
(3) Small-DS, where only five features were manually selected from Base-DS.

The distribution of recurrence/no-recurrence events was 28.4%/71.6%, exhibiting a considerable imbalance between classes, which the authors have counteracted by random subsampling of patients. The MD perspective was not directly addressed (there is no information on absent observations in the data); however, all the used classification algorithms can handle MD directly.

A 10-fold cross validation scheme was used across all classifiers for evaluation. Only the results for the algorithms with the best results were discussed in this work, namely,

NB, DTs, and PART. However, the incorporation of results for the remaining ones would have been interesting to allow for a cross-sectional evaluation among different works. In terms of accuracy, and considering the "risk of recurrence" as an extra feature, DT overcomes the other two approaches in Base-DS (76%) and Med-DS (77%), lagging behind PART in Small-DS by just 1% (PART achieves 80% of accuracy). When the "risk of recurrence" is added, DT still maintains its superiority in both Base-DS (75%) and Med-DS (76%), although lagging behind NB in Small-DS by just 1% (NB obtains a 78% accuracy). Regarding AUC values, NB outperforms the other two approaches with and without considering "recurrence risk" as an extra feature, although its superiority is not highly pronounced. DT obtains sensitivities of 48%, 45%, and 37% for Base-DS, Med-DS, and Small-DS, rivaling with PART, whose results were 48%, 33%, and 37% for the same datasets, without considering "recurrence risk." When the extra feature is added, both approaches still obtain similar results: 48%, 40%, and 30% for DT versus the 51%, 40%, and 32% of PART. Finally, in terms of specificity, all three approaches obtain very similar (and high) results—from 87%–96% for NB, 86%–96% for DT, and 78%–97% for PART—which makes it harder to assess the best one. The authors are not very conclusive in assessing the best approach, suggesting either NB or C4.5 DT are both suitable to be selected as the best classifier. It can be discussed that DT achieves the best performance for Base-DS and Med-DS, while Small-DS benefits the most by using PART. However, in a general view, DT seems to be the best approach, since it always achieves higher sensitivity results, and higher or comparable specificity. Furthermore, they have the advantage of producing interpretable rules, and as the authors mention, may be clearly visualized when the dataset is small. In addition, the risk of recurrence did not improve the classification results; nor did the presence of a high number of features: the results are similar across all datasets. Therefore, it can be discussed that small dimensional spaces are suitable to address the BC recurrence context, having as main advantage the reduced complexity of the classification models.

The study developed by Fan et al. [2010] in 2010 targeted the internationally available SEER dataset [SEER Research 2015], where 46,996,113 patients are described by a set of 117 features. The SEER Public-Use Data used in this work includes patients diagnosed with BC from 1973 to 2005; however, the endpoint for determining recurrence is not specified. Records with MD were ignored, but the number of patients kept in the final dataset is not mentioned. The feature selection phase was performed according to the validation of medical experts, with 13 features being selected as final inputs. A holdout method (80% train/20% test) was used to evaluate the performance of five different algorithms, namely, ANN and other four variants of DT. The results show that all the DT variants outperformed ANN in terms of accuracy, with the best accuracy results being achieved by C5.0 algorithm (71.17%). However, ANN had the highest precision rate for recurrence events (77.79%), while CART had the highest precision rate for no-recurrence events (73.75%).

Although nearly all accuracy and precision results are above 70%, a more detailed discussion should have been presented. Since the final number of patients included in the study (and of those, how many did or did not recur) is unknown, the readers do not have the necessary information to evaluate the results. If the class distribution of recurrence and no-recurrence is not balanced, the accuracy results are not reliable. Furthermore, although precision results are a common finding in most ML studies, they are so in combination with recall results, which translate the sensitivity of the classifiers. This information is not presented in the paper, and is of much importance when the objective is to accurately predict a particular class of interest, in this case, the "recurrence" class. Furthermore, no single classifier overcomes all others in the three considered metrics (accuracy, precision for "recurrence," and precision

for "no-recurrence"). Therefore, it is not possible to determine which is the most suitable approach.

Belciug et al. [2010] compared the performance of k-means, SOM, and cluster network in the detection of BC recurrence events, using the WPBC dataset [Lichman 2015]. Each record of WPBC dataset represents the follow-up data for one patient. This dataset includes invasive BC cases with no evidence of distance metastases at the time of diagnosis, from 1984 to 1995. Therefore, prior to any classification study, the dataset should first be filtered to translate a defined endpoint (e.g., recurrence within 2 years after diagnosis). The authors do not specify such in their research, and thus the true BC problematic cannot be identified. The WPBC dataset is available from UCI Repository, and contains 198 (47 recurrent and 151 nonrecurrent) patients characterized by 34 features. These features describe the characteristics of the cell nuclei observable in an image of the patient's breast mass. According to the dataset description [Lichman 2015], it contains absent observations. Specifically, "lymph node status" is missing in four cases (three "nonrecurrent" and one "recurrent"). However, the MD perspective is never mentioned in this work. From the 34 features in the original dataset, the authors chose 12 to be used, discarding redundant information unnecessary to the clustering algorithms. The feature selection process is not mentioned: the features are selected according to the authors' assumption on their relevance to the study, without further elaboration on the subject. The authors compare both training and testing performance between the used algorithms, except for k-means, where only the training performance is assessed. For SOM and cluster network, a 10-fold cross validation procedure is used, and the accuracy results are then averaged to achieve the final classification results. Cluster network obtained the highest accuracy results in both training (83%) and testing (78%), versus the 72% and 67% obtained for SOM and the 62% (training) for k-means. On one hand, cluster network has shown an efficient behavior in predicting BC relapse, achieving accuracy results higher than the majority of the discussed works. On the other hand, this paper fails by disregarding the clear class imbalance between "recurrence" and "no-recurrence" events (76.3%/23.7%). In such cases, as previously mentioned, accuracy is not an appropriate metric, and other metrics, such as sensitivity and specificity, should be presented.

Furthermore, the fact that the MD perspective is not addressed is quite intriguing, since clustering methods cannot generally analyze data points with MD, without further constraints. Usually, if no specifications are provided to the algorithm, data points with MD are discarded, and thus not clustered. Therefore, if MD has not been handled in any way, the results may be somewhat biased. Also, the 10-fold cross validation procedure could have been extended to k-means as well, for testing purposes. Using a LOO approach, each sample would be held out while the remaining were used for training. Then, the held-out sample would be assigned to a class according to its proximity to the cluster centers, by majority voting, for instance. This would provide the testing performance for k-means, making it comparable with the remaining approaches. Furthermore, the authors raise a question that is not further discussed in the paper: "how many clusters are needed so that the clustering process is optimal?" [Belciug et al. 2010]. SOM and cluster networks do not require an a priori specification of the number of clusters, and thus the data points are labeled according to a majority voting of the points belonging to the cluster they are assigned. This could also be achieved by k-means, where different numbers of clusters had to be tested. The fact that k-means was predefined for $k = 2$ may help explain its poor results, when compared to the other two algorithms, that do not have that limitation. In conclusion, despite proving good results, this work raises some technical questions.

In the same year, Trumbelj et al. [2010] addressed the problem of BC recurrence in two directions: as a classification problem, predicting recurrence/no-recurrence events

within 10 years after surgery, and as a regression problem, determining how many years would it take until cancer reappears. The latter is out of scope of this review.

Regarding the prediction of recurrence events, Strumbelj compared the performance of several well-known classifiers, namely, NB, DTs, SVMs, RFs, and Multilayer Perceptron (MLP) with the evaluation of two oncologists. A bagging procedure coupled with NB was also considered. The initial data was provided by Ljubljana Institute of Oncology (not to be confused with the Breast Cancer Dataset), consisting of 1,035 patients characterized by 32 features. After removing some features (due to their redundancy) and some patients (whose follow-up was inferior to 10 years), the final dataset included 881 patients and 13 features, all categorical. Although the authors state that some of the collected features are redundant, and that not all features are considered for classification, the criteria to select the most relevant features is not depicted. Furthermore, no information was given regarding MD.

The distribution of recurrence/no-recurrence events is 51%/49%, thus class imbalance is not a constraint and accuracy is considered an appropriate evaluation metric. A 10-fold cross validation scheme was used to access the performance of the chosen ML algorithms, where DT, NB (both as a standard formulation and coupled with a bagging scheme), and RF performed similarly with accuracies ranging from 67.4% to 68%, outperforming SVM (59.9%) and MLP (60.8%). Therefore, the best approaches (DT, NB, and RF) were further compared with the predictions of two medical experts (two oncologists), using a validation test of 100 randomly chosen patients. NB classifier achieved the best results (both standard and considering bagging) with an accuracy of 70%. However, overall, the accuracy results were very similar, with DT and RF obtaining accuracies of 67% and 68%. Both oncologists lagged slightly behind the ML algorithms, with accuracies of 63% and 65%. In fact, ML results did not prove to be significantly higher than the predictions of medical experts. The fact that the final dataset contains only categorical variables is a topic for discussion in this work. The authors state that their "preliminary analysis" has not shown significant differences between numerical or discretized versions of some features in prediction results, although these results are not presented in the work. This may explain the poor results achieved by SVM and MLP, which generally tend to deal better with continuous variables [Kotsiantis 2007; Irshad et al. 2014]. However, it has to be stated that the main objective of this work was not to achieve optimal classification results. More than building a successful model to predict BC recurrence, the aim of this work is to improve the interpretability of ML models and develop a method to assess their reliability. This topic is also out of scope of this study; however, it highlights the increasing interest of ML experts in developing accurate, yet still easy to use and interpretable strategies to be used by non-ML experts, in particular, medical doctors, when dealing with a health care context.

In 2012, Kim et al. [2012] studied the application of SVMs, ANNs, and the Cox-regression model to the prediction of BC recurrence within 5 years after surgery. To assess the performance of the proposed approaches, three well-known BC prognostic models were also selected: St. Gallen's guidelines [Harbeck et al. 2013], Nottingham Prognostic Index (NPI), [Galea et al. 1992], and Adjuvant! Online [Olivotto et al. 2005]. The initial dataset was composed by 1,541 patients from a tertiary hospital in South Korea. However, after discarding patients with incomplete follow-up, late-stage, and male BC patients, as well as patients suffering from other types of cancers (besides BC), the study population consisted of 679 patients, with 195 recurrence cases (28.7%) and 484 no-recurrence cases (71.3%). Out of 193 available features, seven were chosen to be included in the prediction models, namely, histological grade, tumor size, number of metastatic lymph nodes, ER status, Lymphovascular Invasion (LVI), local invasion of tumor, and number of tumors [Kim et al. 2012]. They were selected beforehand by the authors in collaboration with medical experts, and further refined based

on Kaplan-Meier and Cox-regression analysis. The results were evaluated in terms of accuracy, sensitivity, specificity, precision, AUC, and Negative Predictive Value (NPV), using a holdout method (70%–30%). Regarding the computational models, SVMs and ANNs performed similarly, outperforming the Cox model except in terms of specificity: 73%, 52%, and 94% for SVMs, ANNs, and Cox model, respectively. ANNs achieved the best sensitivity (95%) and precision (80%) results, followed by SVMs with 89% and 75%, respectively. However, SVMs proved to be the best approach, outperforming the others in terms of NPV (89%), accuracy (84.58%), and AUC (0.85). The authors further compared the performance of SVMs with the previously mentioned prognostic models: St. Gallen's, NPI, and Adjuvant!. St. Gallen's achieved the highest sensitivity and NPV (100%); however, it had poor results in the remaining metrics. Similarly, Adjuvant! also returned high sensitivity and NPV results (95% and 83%), although its superiority was not verified for the other metrics. The same may be said of the Cox model, which had the highest specificity (94%), but failed to keep its advantage over the other performance metrics. Thus, SVM proved a superior performance over the "classic" models for the prognosis of BC recurrence. The authors highlight that although ML algorithms generally achieve higher performances, their use in clinical practice is still very limited "because they cannot be easily calculated with a traditional calculator." In our opinion, they are right in that ML are currently not used in practice, despite their undoubtedly higher performance. However, we do not agree on the reason. The real reason boils down to the interpretability again. Even if tools to calculate ML predictions are made available, medical doctors will not "trust" models they cannot fully understand and interpret. Another point mentioned by the authors is that ML algorithms can be adjusted to data. For instance, SVM hyperparameters may be adjusted to different subject populations. This may bring an important advantage over traditional prognostic models that impose a universal prediction model for all races or countries.

In that same year, Salama et al. [2012] compared the performance of DTs, MLP, SVMs, NB, and KNN in the prediction of BC recurrence using the WPBC dataset (similarly to Belciug et al., the endpoint is not defined). The fusion between classifiers was also explored, to assess if a multiclassifier approach could bring some benefit in terms of classification performance. The comparison between classifiers was performed using a 10-fold cross validation sampling scheme, and evaluating their accuracy. Among the five considered classifiers, SVM and DT outperformed all others, with an accuracy of 76.3%, followed by MLP, KNN, and NB with 66.5%, 64.4%, and 50.5%, respectively. Moreover, a fusion analysis of two, three, and four classifiers was conducted. The first fusion considered SVM coupled with the remaining: SVM-NB, SVM-MLP, SVM-DT, and SVM-KNN. All the combinations have achieved the same accuracy results: 76.3%. The fusion of three classifiers considered SVM and DT coupled with the remaining: SVM-DT-NB, SVM-DT-MLP, and SVM-DT-KNN. Once again, all the combinations showed the exact same accuracy: 76.3%. Finally, the third fusion considered the coupling of SVM, DT, and MLP with the remaining: SVM-DT-MLP-KNN and SVM-DT-MLP-NB. The combination of SVM-DT-MLP-KNN resulted in an improvement of accuracy, 77.3%, while SVM-DT-MLP-NB did not improve the previous results, achieving an accuracy of 74.2%. In conclusion, the fusion of SVM, DT, MLP, and KNN proved to be superior when compared to the remaining combinations of classifiers and the other setups of standalone classifiers. This work shows that the combination of classifiers may be beneficial to the classification performance. However, the authors do not mention what type of combination was used (using probability results, majority voting, or combination rules, for instance). Also, and as previously mentioned, WPBC is an imbalanced dataset, and therefore more appropriate performance metrics would be required, namely, sensitivity and specificity. Finally, the MD perspective is also ignored in this work, which constitutes another of its limitations.

Also in 2012, Murti [2012] used three rule-based classifiers to predict BC recurrence within 5 years after surgery, namely, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Decision Table, and Decision Table with Naive Bayes (DTNB). To conduct the experiments, the database from the Oncology Institute of Ljubljana (Breast Cancer Dataset) was used. The initial dataset was preprocessed to remove missing values, and although the final number of patients included in the study is not mentioned, we assume that all nine records with MD are eliminated, thus resulting in a final dataset of $(286 - 9) = 277$ patients. The algorithms are compared in terms of precision, recall, F-measure, and AUC, for both "no-recurrence" and "recurrence" events. As previously mentioned for the case of Fan et al., when the objective is to predict recurrence (and considering that this class always has a lower number of cases), it should be defined as the positive class, and the metrics should be analyzed having that in mind. Only such analysis would provide a meaningful and suitable comparison with the other studied works. For that reason, only the performance results of the "recurrence" class are analyzed and compared within this review. Accordingly, RIPPER obtained 72.3%, 36.5%, 84.6%, 50%, 0.4, and 0.58 in terms of accuracy, sensitivity, specificity, precision, F-measure, and AUC, respectively. In turn, Decision Table achieved 72.7%, 23.5%, 91%, 53%, 0.33, and 0.64 for the same metrics. Finally, outperforming these approaches, DTNB returned an accuracy of 75.2%, sensitivity, specificity, and precision of 36.5%, 89.6%, and 59.6%, respectively, while achieving an F-measure of 0.45 and AUC of 0.68. Although DTNB outperformed the other approaches, achieving a good accuracy (75.2%) and specificity (89.6%), the sensitivity results are very poor, being among the worst approaches reviewed. Similarly, the F-measure and particularly the AUC results show that this is not a feasible approach to predict BC recurrence, being only slightly better than random guessing.

In 2013, Tomczak [2013] used the Classification Restricted Boltzmann Machine (ClassRBM) to predict BC recurrence within 10 years after surgery and determine input features (symptoms) relevant for disease reappearance. Several methods for learning ClassRBM are discussed, namely, DropOut, Drop Connect, and DropPart [Tomczak 2013]. These algorithms are compared to classical approaches such as NB, SVM, RF, and CART (coupled with AdaBoost, Bagging, and LogitBoost). This work also counted with the collaboration of two oncologists in order to provide a comparison of ML techniques with the opinion of medical experts. A holdout method (70% train-30% test) was used across all computational methods, while predictions from oncologists were obtained using 100 cases of the test set. Overall, the computational approaches achieved better accuracy results than the medical experts, except for SVM, which performed poorly. ClassRBM and ensemble approaches had very similar results; however, from all the considered algorithms, the ensemble LogitBoost + CART outperformed all others with an accuracy of 75%. This work is somewhat of a follow-up of Strumbelj's study, using the same dataset provided by the Oncology Institute of Ljubljana (not Breast Cancer Dataset). However, in Tomczak's study, the final dataset is composed of 949 patients (there are more patients with a minimum follow-up of 10 years) and 15 features (the feature selection process is not discussed). These 15 features include all 13 used by Strumbelj and two more, regarding the application of two different types of therapy (cTherapy and hTherapy). All input features were binarized, resulting in a dataset composed of 55 binary features. As discussed in Strumbelj's study, the binarization of all input features could explain the poor performance of SVMs. The distribution of "recurrence"/"no-recurrence" events is not depicted; however, we assume it is very similar to Strumbelj's estimates (51%–49%). The authors do not perform a thorough discussion on the best approach to predict BC recurrence. Nevertheless, they highlight the ClassRBM's ability to retrieve relevant information regarding the most important input features while also achieving a high classification performance.

Pawlovsky and Nagahashi [2014] proposed a method based on scoring to select the best configuration to be used in KNN classification of WPBC dataset (the endpoint was not defined). In their approach, patients with MD are removed from the study (four patients), and only 32 features are kept. After discussing the effects of different combinations of training size and number of neighbors and runs considered, the authors present their scoring scheme and perform its validation by addressing the BC recurrence problem. The best classification setting is chosen according to the preprocessing method used (raw data, standardization, or normalization), number of k neighbors, number of runs, sample size for classification, and average, maximum, minimum, and standard deviation of the accuracy results. Their strategy provided the best results for a configuration using raw data, 19 neighbors, 80% of samples in classification, and 100 simulation runs. These configurations achieved a mean accuracy of 76%, and minimum and maximum values of 62% and 90%, respectively. It is also important to note that, overall, the preprocessing method used does not significantly affect the final classification results.

Although discussing an interesting topic, this work is more focused on finding a strategy to select appropriate KNN configurations than addressing the particular problem of BC recurrence: no feature selection is performed and the class imbalance problem is not addressed (again, only accuracy results are presented). Nevertheless, it takes into account the existence of absent observations by removing them. The generalization of this work could possibly be a topic for further research, and its extension to include sensitivity/specificity results could possibly be a more suitable approach to the BC recurrence problem. However, as discussed in Section 2.3, it must be noted that KNN is a lazy learner, as it makes local approximations, without further generalization, and thus the classification task for this algorithm is very time-consuming. With a considerable amount of data, given the number of different combinations to be tested, it could become infeasible for real-time applications.

In the same year, Beheshti et al. [2014] tackled the principles of Genetic Programming, by comparing the performance of several genetic approaches when coupled with MLP: Centripetal Accelerated Particle Swarm Optimization (CAPSO), Particle Swarm Optimization (PSO), Gravitational Search Algorithm (GSA), and Imperialist Competitive Algorithm (ICA). These four hybrid approaches (CAPSO-MLP, PSO-MLP, GSA-MLP, and ICA-MLP) were applied to nine medical datasets targeting different diseases. Among them is the WPBC dataset [Lichman 2015], previously presented. Before running the simulations, the dataset was normalized and absent observations were handled with mean imputation. All approaches were evaluated in terms of MSE, AUC, accuracy, sensitivity, and specificity, following a holdout scheme (80% train-20% test). GSA-MLP achieved 0.167 of MSE, 0.55 of AUC, and 79.3%, 7.86%, and 80.23% of accuracy, sensitivity, and specificity. In turn, ICA-MLP and PSO-MLP obtained MSE results of 0.177 and 0.173 and AUC results of 0.57 and 0.6, respectively. In terms of accuracy, sensitivity, and specificity, these approaches have returned the same results: 78.3%, 43%, and 83%. CAPSO-MLP achieved an MSE of 0.170 and an AUC of 0.63 while returning accuracy, sensitivity, and specificity results of 80.3%, 52.3%, and 83.4%, clearly outperforming all others and being considered the most suitable approach for unseen data.

According to the authors, the adjustment of the parameters of the PSO algorithm is time-consuming. The CAPSO approach was created to solve this problem, by using less a priori parameters, resulting in a simplified tuning process (more automated). The inclusion of the original PSO approach was an important step, to evaluate if the new technique (CAPSO) improves the results by comparison. However, applying only this type of algorithms does not provide a real assessment of their performance. There should have been a setup including a more traditional approach as a baseline for

comparison, for example, backpropagation, in accordance with other authors cited in this work [Chau 2007; Socha and Blum 2007; Ozkan et al. 2011; Ahmadi et al. 2013; Mahmoudi et al. 2013]. This would present the opportunity to compare the two different methodologies, verifying whether the proposed algorithms generate better results. Moreover, the chosen algorithms do not agree with the literature review of this article: GSA is not referred to in any of the cited articles, while others were inexplicably left out (e.g., Artificial Immune System, Ant Colony Optimization, and Artificial Bee Colony). ICA and GSA happen to be the two most recent approaches mentioned, but there was no explicit indication of the reason to choose them. Nevertheless, the used algorithms are thoroughly explained, which is especially important in modern techniques.

Still in 2014, Chaurasia and Pal [2014] investigated the performance of DDTs, ANNs, and LR in BC recurrence within 5 years after surgical intervention, using the previously described Breast Cancer Dataset to conduct their experiments.

The results were evaluated through a 10-fold cross validation procedure, by determining the accuracy, true positive rate, false positive rate, precision, and recall for both "recurrence" and "no-recurrence" classes. As discussed in previous works (Fan et al., Murti), the class of interest is "recurrence." For that reason, and to allow an appropriate comparison between all the research works, we restrict our analysis to accuracy, sensitivity, specificity, and precision results considering "recurrence" as the positive class. In terms of accuracy, specificity, and precision, LR performed the best, with 74.5%, 92.5%, and 64.3%, respectively, versus the 71.3%, 92%, and 54.3% of DT and 73.8%, 88.6%, and 58.9% of ANN. Regarding sensitivity, ANN was the best approach, with 38.8%, over the 31.8% and 22.4% obtained by LR and DT, respectively. Although ANN achieves the best sensitivity results (is the best classifier in identifying recurrence events), LR is overall the best approach, outperforming all others in terms of accuracy, specificity, and precision. The authors have also analyzed the impact of the chosen feature to recurrence prediction, which revealed that tumor grade is the most explanatory feature, followed by lymph nodes involvement, node capsules, tumor size, irradiation, age, breast quad, breast, and menopause. As many of the works use WPBC dataset, the fact that this work neglects the MD perspective is its main weakness.

Table IV presents a résumé of the ML algorithms used in each research work and the performance results of the best approach (highlighted in bold). The results are measured in terms of accuracy (Acc), sensitivity (Sen), specificity (Spe), and AUC values. The strategies used for data sampling and handling MD are also depicted.

Based on the 17 analyzed revised works, and despite the fact that a direct comparison between these works needs to be performed with due reservations (as they use different algorithms and approaches as well as distinct datasets), there seems to be a slight advantage of approaches using combined ML methods over approaches using a single ML method. Combined approaches present maximum results of 96%, 90%, and 97.2% for accuracy, sensitivity, and specificity, while single approaches reach maximum results of 84.58%, 89%, and 96%, respectively (the AUC metric was not compared since only five of the 17 revised works present such information). This tendency (combined approaches achieving better results than single ones) is in fact stated in the literature for other BC classification problems, as illustrated in the works developed by Abreu et al. [2013a] and Srinivas and Mohan [2015], or as a general ML approach [Stefanowski 2005].

4. DISCUSSION

Predicting BC recurrence is a very important challenge for oncological clinicians because it has direct influence in their daily practice, for example, in choosing the most beneficial treatment for a patient. Over the past decade, several works have tried to propose suitable approaches to model BC behavior; however, after performing this

Table IV. Comparison of ML Algorithms (Bold Indicates the One That Presented the Best Performance), Achieved Results and Sampling Strategies used in the Analyzed Studies

Publications	Algorithms	Acc	Sen	Spe	AUC	MD	Sampling Strategy
Mani et al. [1997]	**NB**, CART, C4.5, C4.5 rules, FOCL	68.30%	–	–	–	–	Stratified Random Subsampling Construction of six datasets (40%–60%)
Jerez-Aragonés et al. [2003]	**ANN+DT**	93,4%–96%	78,7%–88.7%	94,5%–97.2%	–	Removed	10-fold Cross validation
Razavi et al. [2005]	**C4.5+CCA**, C4.5 + EM, C4.5	67%	80%	63%	–	EM	10-fold Cross validation
Razavi et al. [2007]	**C4.5+CCA**, Two oncologists	82%	21.10%	96.30%	0.76	MI	10-fold Cross validation Stratified Random Sampling for validation
Sun et al. [2007]	**LDA+hybrid**, LDA+genetic, LDA+clinical, St.Gallen criterion	–	90%	67%	–	Unknown	LOO
Ryu et al. [2007a]	**Isotonic Separation**, Robust Linear Programming, DT (C4.5, OC1, QUEST), SVM, AdaBoost, Learning Vector Quantization	80%	–	–	–	–	Holdout (70%–30%)
Jonsdottir et al. [2008]	**C4.5**, NB, LMT, REP tree, RF, SVM, Log, Slog, MetaClass1, MetaClass2, MetaClass3, Bag +REP tree, DT, OneR, PART, Jrip, VFI	79% (Small-DS)	48% (Base-DS)	96% (Small-DS)	0,70 (Base-DS)	Not mentioned, but handled by algorithms	10-fold Cross validation
Fan et al. [2010]	**C5.0**, ANN, CHAID, CART, QUEST	71.20%	–	–	–	Removed	Holdout (80%–20%)
Belciug et al. [2010]	**Cluster network**, k-means, SOM	78%	–	–	–	Unknown	10-fold Cross validation
Trumbelj et al. [2010]	**NB**, **NB+Bagging** DT, SVM, RF	70%	–	–	–	Considered as a separate feature value.	10-fold Cross validation
Kim et al. [2012]	**SVM**, ANN, Cox model, St. Gallen, NPI, Adjuvant!	84.58%	89%	73%	0.85	Removed	Holdout (70%–30%)
Salama et al. [2012]	**SVM-DT-MLP-KNN**, DT, MLP, SVM, NB, KNN, SVM-NB, SVM-MLP, SVM-DT, SVM-KNN, SVM-DT-NB, SVM-DT-MLP, SVM-DT-KNN, SVM-DT-MLP-NB	77.30%	–	–	–	Unknown	10-fold Cross validation

(Continued)

Table IV. Continued

Publications	Algorithms	Acc	Sen	Spe	AUC	MD	Sampling Strategy
Murti [2012]	**DTNB**, RIPPER, Decision Table	75.17%	37%	90%	0.676	Removed	Unknown
Tomczak [2013]	**CART+LogitBoost**, ClassRBM, ClassRBM+DropOut, Class-RBM+DropConnect, Class-RBM+DropConnect, CART+Bagging, CART+AdaBoost, NB, SVM, RF, Two oncologists	75%	–	–	–	Unknown	Holdout (70%–30%)
Pawlovsky and Nagahashi [2014]	**KNN**	76%	–	–	–	Removed	n.a.
Beheshti et al. [2014]	**CAPSO-MLP**, PSO-MLP, GSA-MLP, ICA-MLP	80.25%	52.33%	83.38%	0.63	Mean	Holdout (80%–20%)
Chaurasia and Pal [2014]	**LR**, C4.5, ANN	74.50%	31.80%	92.50%	–	Unknown	10-fold Cross validation

revision, it is clear that this is still an open problem. This observation is based on five problems detected in the reviewed works (RWs):

Lack of Data. The majority of RWs used local datasets (datasets that contain only data from a local/regional center), which complicates the replication and further comparison of results by other researchers. Also, the number of patients enrolled in most of these studies can be considered small (less than 1,000 patients), especially for a common pathology like BC. The reduced size of the datasets becomes even more critical when most of the works do not deal with MD, either at all (more than 80%) or with proper thoroughness. Only three research works have addressed this issue (Razavi et al. [2005, 2007] and Beheshti et al. [2014]).

Imbalanced Binary Decision Problem. The second problem, as mentioned in the Introduction, is that the prediction of BC recurrence is a binary classification problem where the goal is to accurately predict whether a BC patient will or will not recur. To achieve that, these two classes should be balanced (have similar proportions in the dataset); otherwise, the algorithms could predict one class better than the other. From our analysis, it can be noted that the majority of the RW presented imbalanced datasets, which will somehow degrade the performance of ML techniques. This point could be easily overcome by using appropriate sampling strategies to balance data, such as Synthetic Minority Oversampling Technique (SMOTE) [Chawla et al. 2002].

Feature Selection. The third problem concerns feature selection. Only a small number of works used computational feature selection techniques. Most of the RW use a manual feature selection process, in which medical doctors are consulted to select the variables to use in the prediction studies. However, this process has one great disadvantage: the information that the algorithms are able to find is exactly what they were expected to find: the doctors select the variables using previously established knowledge or informed intuition, which may prevent potentially useful variables from being used in the models, and new relationships between variables and recurrence

to be found. Also, it is important to note that only one study tried to mix clinical markers with genetic information [Sun et al. 2007]. Many of the variables selected in the RWs were not mutually exclusive (e.g., BC stage is the conjunction of tumor size, lymph nodes involved, and the presence or absence of metastasis) and some were not routinely described as important recurrence factors (e.g., tumor location included breast regions and laterality), which could return somewhat misleading conclusions. Moreover, important factors that must be present in daily clinical practice are missing, such as HER2 expression. The determination of HER2 expression is mandatory for the definition of intrinsic subgroups that define BC behavior. Even in the most frequent BC subgroup (that express HRs), patients with HER2 enriched tumors are associated with a high rate of brain, liver, and lung metastasis [Beca et al. 2014]. These tumors also display different patterns of relapse and metastatic spread depending on HR status, with a median relapse-free survival of 19.5 months after surgery in HR negative patients compared to 32.0 months in HR positive patients. Patients with HER2+/HR− disease have significantly increased hazard of early (0–2 and 2–5 years), but not late death (>5 years), when compared to HER2+/HR+ [Van den Hurk 2011].

The nonstandardization of such set also hampers direct comparisons between studies and compromises future investigations in the field.

A consensus in the definition of important variables to study and its validation over appropriate datasets is still a current challenge.

Interpretability. The fourth problem is interpretability, an important concept in the health care area. If the expert/clinician cannot validate the approach, it will never be accepted by the community as a valid one. This sometimes leads to a scenario where researchers try to find a trade-off between interpretability and performance for their approaches. Accordingly, it is not surprising that 13 out of the 17 RW used ML techniques that are well known for their interpretability, like DTs. However, other techniques that are in the opposite side (traditionally achieving higher performance, although less interpretable) have not been neglected, such as ANNs. The comparison between these different methods is impossible due to a number of factors: the used datasets are different, the selected set of features and algorithms do not always match, and finally the evaluation metrics used are not always the same. Some studies even use clinicians to validate their approaches. However, hybrid or combination algorithms generally seem to be among the best approaches.

Evaluation Metrics. Finally, regarding the metrics used in the evaluation phase, it is quite surprising how eight of the 17 RWs only used accuracy to measure classification performance, especially considering the class imbalance present in the associated datasets. Accuracy is not the most appropriate metric for imbalanced datasets, since it does not properly identify the true positive and true negative rates (i.e., sensitivity and specificity). When considering other studies that present both accuracy and sensitivity results, it can be noted that it is easier to achieve a good accuracy performance than sensitivity results (only three of the RWs presented good sensitivity). This may also be explained by the imbalanced distribution between "recurrence" and "no-recurrence" cases.

5. CONCLUSIONS AND FUTURE WORK CHALLENGES

As discussed in this survey article, predicting recurrence is a key point in the BC context. However, and in spite of the fact that researchers have tried to address this topic in the past decade, it remains an open challenge. Based on the analyzed RWs, the works using a combination of ML techniques seem to have a slight advantage over the ones that used a single approach (e.g., the one proposed by Jerez-Aragonés et al. [2003]), which falls in line with the literature review works [Stefanowski 2005]. Also, another

important aspect not yet fully addressed in the RWs is related to dealing with MD. This is a crucial problem and imputation strategies are only possible if the used dataset has a sufficient number of patients, also benefiting from balanced datasets. To achieve that, the clinicians community must establish a standard characterization for such patients (that will be used as predictors) which will lead to the creation of datasets with large patients records.

Finally, the development of new ML algorithms or the exploration of ML algorithms that have never been used in this context may also constitute a valid future perspective.

REFERENCES

P. H. Abreu, H. Amaro, D. C. Silva, P. Machado, and M. H. Abreu. 2013b. Personalizing breast cancer patients with heterogeneous data. In *Proceedings of the IFMBE International Conference on Health Informatics*. 39–42.

P. H. Abreu, H. Amaro, D. C. Silva, P. Machado, M. H. Abreu, N. Afonso, and A. Dourado. 2013a. Overall survival prediction for women breast cancer using ensemble methods and incomplete clinical data. In *Proceedings of the Mediterranean Conference on Medical and Biological Engineering and Computing*. 1366–1369.

R. Agrawal, T. Imielinski, and A. N. Swami. 1993. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. 207–216.

M. A. Ahmadi, M. R. Ahmadi, and S. R. Shadizadeh. 2013. Evolving artificial neural network and imperialist competitive algorithm for prediction permeability of the reservoir. *Neural Computing and Applications* 13, 2 (2013), 1–9.

N. S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.

S. Arlot and A. Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4 (2010), 40–79.

A. Atla, R. Tada, V. Sheng, and N. Singireddy. 2011. Sensitivity of different machine learning algorithms to noise. *Journal of Computing Sciences in Colleges* 26, 5 (2011), 96–103.

A. Azevedo and M. F. Santos. 2008. KDD, SEMMA and CRISP-DM: A parallel overview. In *Proceedings of Informatics and Data Mining*. 182–185.

F. Beca, R. Santos, D. Vieira, L. Zeferino, R. Dufloth, and F. Schmitt. 2014. Primary relapse site pattern in women with triple-negative breast cancer. *Pathology - Research and Practice* 210, 9 (2014), 571–575.

Z. Beheshti, S. M. H. Shamsuddin, E. Beheshti, and S. S. Yuhaniz. 2014. Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis. *Soft Computing* 18, 11 (2014), 2253–2270.

S. Belciug, F. Gorunescu, A. B. Salem, and M. Gorunescu. 2010. Clustering-based approach for detecting breast cancer recurrence. In *Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA)*. 533–538.

C. M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

B. Boser, I. Guyon, and V. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Annual Workshop on Computational Learning Theory*. 144–152.

L. Breiman. 1996. Bagging predictors. In *Machine Learning* 24, 2 (1996), 123–140.

L. Breiman. 1998. Arcing classifiers. *The Annals of Statistics Journal* 26, 3 (1998), 801–849.

L. Breiman. 2001. Random forests. *Machine Learning Journal* 45 (2001), 5–32.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Wadsworth & Brooks, Monterey, CA.

R. Chandrasekaran, Y. U. Ryu, V. S. Jacob, and S. Hong. 2005. Isotonic separation. *INFORMS Journal on Computing* 17, 4 (2005), 462–474.

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. 2000. *CRISP-DM 1.0: Step-by-Step Data Mining Guide.* SPSS.

K. W. Chau. 2007. Application of a PSO-based neural network in analysis of outcomes of construction claims. *Automation in Construction* 16, 5 (2007), 642–646.

V. Chaurasia and S. Pal. 2014. Data mining techniques: To predict and resolve breast cancer survivability. *International Journal of Computer Science and Mobile Computing* 3, 1 (2014), 10–22.

N. V. Chawla. 2010. *Data Mining and Knowledge Discovery Handbook* (2nd ed.). Springer US. 875–886.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research* 16, 2002 (2002), 321–357.

N. V. Chawla, N. Japkowicz, and A. Kotcz. 2004. Editorial: Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter* 6, 1 (2004), 1–6.

H. Chen, S. S. Fuller, C. Friedman, and W. Hersh (Eds.). 2005. *Medical Informatics—Knowledge Management and Data Mining in Biomedicine*. Vol. 8. Springer-Verlag US.

F. Cismondi, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. Sousa, and S. N. Finkelstein. 2013. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine* 58, 1 (2013), 63–72.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

J. A. Cruz and D. S. Wishart. 2006. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics* 2 (2006), 59–77.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39, 1 (1977), 1–38.

M. Dettling and P. Buhlmann. 2004. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* 90, 1 (2004), 106–131.

R. O. Duda, P. E. Hart, and D. G. Stork. 2012. *Pattern Classification* (2nd ed.). John Wiley & Sons.

B. Efron and R. Tibshirani. 1994. *An Introduction to the Bootstrap* (1st ed.). Chapman and Hall/CRC.

B. D. Eugenio and M. Glass. 2004. The kappa statistic: A second look. *Computational Linguistics* 30, 1 (2004), 95–101.

Q. Fan, C. J. Zhu, and L. Yin. 2010. Predicting breast cancer recurrence using data mining techniques. In *Proceedings of International Conference on Bioinformatics and Biomedical Technology*. 310–311.

A. Farr, R. Wuerstlein, A. Heiduschka, C. F. Singer, and N. Harbeck. 2013. Modern risk assessment for individualizing treatment concepts in early-stage breast cancer. *Reviews in Obstetrics and Gynecology* 6, 3 (2013), 165–173.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. 1996. From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine* 17, 3 (1996), 37–54.

C. Ferri, J. Hernández-Orallo, and R. Modroiu. 2009. An experimental comparison of performance measures for classification. *Pattern Recognition Letters* 30, 1 (2009), 27–38.

A. Fischer and C. Igel. 2012. An introduction to restricted Boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 14–36.

A. Fischer and C. Igel. 2014. Training restricted Boltzmann machines: An introduction. *Pattern Recognition* 47, 1 (2014), 25–39.

R. A. Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 2 (1936), 179–188.

Y. Freund and R. E. Schapire. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the 2nd European Conference on Computational Learning Theory*. 23–37.

J. Friedman, T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: A statistical view of boosting. *Annals of Statistics* 28, 2 (2000), 337–407.

M. H. Galea, R. W. Blamey, C. E. Elston, and I. O. Ellis. 1992. The Nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment* 22, 3 (1992), 207–219.

V. Ganganwar. 2012. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering* 2, 4 (2012), 42–47.

P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonso. 2015. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in Biology and Medicine* 59, 2015 (2015), 125–133.

P. J. García-Laencina, J. L. Sancho-Gómez, and A. Figueiras-Vidal. 2010. Pattern classification with missing data: A review. *Neural Computing & Applications* 19, 2010 (2010), 263–282.

P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal. 2013. Classifying patterns with missing values using multi-task learning perceptrons. *Expert Systems with Applications* 40, 4 (2013), 1333–1341.

V. Garcia, R. A. Mollineda, R. Alejo, and J. M. Sotoca. 2007. The class imbalance problem in pattern classification and learning. In *II Congreso Español de Informática (CEDI'07)*. 978–84.

O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor. 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22, 14 (2006), 184–190.

L. Guo-Zheng. 2011. Machine learning for clinical data processing. In *Machine Learning: Concepts, Methodologies, Tools and Applications*. IGI Global, 875–897.

J. Han, M. Kamber, and J. Pei. 2011. *Data Mining: Concepts and Techniques: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

N. Harbeck, C. Thomssen, and M. Gnant. 2013. St. Gallen 2013: Brief preliminary summary of the consensus discussion. *Breast Care* 8, 2 (2013), 102–109.

H. He and E. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21, 9 (2009), 1263–1284.

J. Huang. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 17, 3 (2005), 290–310.

H. In, K. Y. Bilimoria, A. K. Stewart, K. E. Wroblewski, M. C. Posner, M. S. Talamonti, and D. P. Winchester. 2014. Cancer recurrence: An important but missing variable in national cancer registries. *Annals of Surgical Oncology* 21 (2014), 1520–1529.

O. Intrator and N. Intrator. 2001. Interpreting neural-network results: A simulation study. *Computational Statistics and Data Analysis* 37, 3 (2001), 373–393.

H. Irshad, A. Gouaillardd, L. Rouxa, and D. Racoceanub. 2014. Multispectral band selection and spatial characterization: Application to mitosis detection in breast cancer histopathology. *Computerized Medical Imaging and Graphics* 38, 5 (2014), 390–402.

A. Jain and R. Dubes. 1988. *Algorithms for Clustering Data*. Prentice-Hall, Inc, Upper Saddle River, NJ.

A. K. Jain. 2010. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters Journal* 31, 8 (2010), 651–666.

J. M. Jerez-Aragonés, J. A. Gomez-Ruiz, G. Ramos-Jimenez, J. Munoz-Perez, and E. Alba-Conejo. 2003. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine* 27, 1 (2003), 45–63.

T. Jo and N. Japkowicz. 2004. Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter* 6, 1 (2004), 40–49.

T. Jonsdottir, E. T. Hvannberg, H. Sigurdsson, and S. Sigurdsson. 2008. The feasibility of constructing a predictive outcome model for breast cancer using the tools of data mining. *Expert Systems with Applications* 34, 1 (2008), 108–118.

M. Kantardzic. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms* (2nd ed.). Wiley-IEEE Press.

W. Kim, K. S. Kim, J. E. Lee, D. Y. Noh, S. W. Kim, Y. S. Jung, M. Y. Park, and R. W. Park. 2012. Development of novel breast cancer recurrence prediction model using support vector machines. *Journal of Breast Cancer* 15, 2 (2012), 230–238.

D. Kleinbaum, M. Klein, and E. Pryor. 2002. *Logistic Regression: A Self-Learning Text. Statistics for Biology and Health Series*. Springer-Verlag.

T. Kohonen. 1995. *Self-Organizing Maps*. Springer, Berlin.

I. Kononenko. 2001. Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine* 23 (2001), 89–109.

S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 1 (2006), 25–36.

S. B. Kotsiantis. 2007. Supervised machine learning: A review of classification techniques. *Informatica* 31 (2007), 249–268.

K. Kouroua, T. P. Exarchosa, K. P. Exarchosa, M. V. Karamouzisc, and D. I. Fotiadisa. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13 (2015), 8–17.

B. S. Kumar. 2012. Boosting techniques on rarity mining. *International Journal of Advanced Research in Computer Science and Software Engineering* 2, 10 (2012), 27–35.

H. Larochelle and Y. Bengio. 2008. Classification using discriminative restricted Boltzmann machines. In *Proceedings of the International Conference on Machine Learning*. 536–543.

H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio. 2012. Learning algorithms for the classification restricted Boltzmann machine. *Journal of Machine Learning Research* 13, 1 (2012), 643–669.

D. T. Larose. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley.

S. Lee and P. A. Abbott. 2003. Bayesian networks for knowledge discovery in large datasets: Basics for nurse researchers. *Journal of Biomedical Informatics* 36, 2003 (2003), 389–399.

Z. Li and J. R. Eastman. 2006. The nature and classification of unlabelled neurons in the use of Kohonen's self-organizing map for supervised classification. *Transactions in GIS* 10, 4 (2006), 599–613.

M. Lichman. 2015. UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml.

R. J. A. Little and D. B. Rubin. 2002. *Statistical Analysis with Missing Data* (2nd ed.). Wiley.

P. Liu, L. Lei, and N. Wu. 2005. A quantitative study of the effect of missing data in classifiers. In *Proceedings of the International Conference on Computer and Information Technology*. 28–33.

R. Longadge and S. Dongre. 2013. Class imbalance problem in data mining review. *International Journal of Computer Science and Network* 1, 2 (2013), 83–87.

S. P. Luttrell. 1994. Partitioned mixture distribution: An adaptive Bayesian network for low-level image processing. *IEE Proc Vision, Image Signal Process* 141, 4 (1994), 251–260.

M. T. Mahmoudi, F. Taghiyareh, N. Forouzideh, and C. Lucas. 2013. Evolving artificial neural network structure using grammar encoding and colonial competitive algorithm. *Neural Computing and Applications* 22, 1 (2013), 1–16.

S. Mani, M. J. Pazzani, and J. West. 1997. Knowledge discovery from a breast cancer database. *Artificial Intelligence in Medicine* 1211 (1997), 130–133.

Z. Markov and D. T. Larose. 2006. *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*. John Wiley & Sons, Inc.

J. P. Marques de Sá. 2001. *Pattern Recognition: Concepts, Methods and Applications*. Springer-Verlag.

W. McCulloch and W. Pitts. 1943. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5, 4 (1943), 115–133.

M. L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22, 3 (2012), 276–282.

S. A. Medjahed, T. A. Saadi, and A. Benyettou. 2013. Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. *International Journal of Computer Applications* 62, 1 (2013), 1–5.

G. Menardi and N. Torelli. 2014. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery* 28, 1 (2014), 92–122.

E. Mendonza. 2013. Predictors of early distant metastases in women with breast cancer. *Journal of Cancer Research and Clinical Oncology* 139, 4 (2013), 645–652.

M. Minsky and S. Papert. 1969. *An Introduction to Computational Geometry*. MIT Press.

M. Mitchell. 1996. *An Introduction to Genetic Algorithms*. MIT Press.

T. M. Mitchell. 1997. *Machine Learning* (1st ed.). McGraw-Hill, Inc.

C. Molina, B. Prados-Suarez, D. R. M. Prados, and Y. C. Pena. 2013. Improving hospital decision making with interpretable associations over datacubes. *Studies in Health Technology and Informatics* 197 (2013), 91–95.

S. E. Moody, D. Perez, T. C. Pan, C. J. Sarkisian, C. P. Portocarrero, C. J. Sterner, K. L. Notorfrancesco, R. D. Cardiff, and L. A. Chodosh. 2005. The transcriptional repressor snail promotes mammary tumor recurrence. *Cancer Cell* 8, 3 (2005), 197–209.

M. S. Murti. 2012. Using rule based classifiers for the predictive analysis of breast cancer recurrence. *Journal of Information Engineering and Applications* 2, 2 (2012), 12–19.

Nature Publishing Group. 2015. Nature International Weekly Journal of Science. (2015). http://www.nature.com/nature.

I. A. Olivotto, C. D. Bajdik, P. M. Ravdin, C. H. Speers, A. J. Coldman, B. D. Norris, and K. A. Gelmon. 2005. Population-based validation of the prognostic model adjuvant! for early breast cancer. *Journal of Clinical Oncology* 23, 22 (2005), 2716–2735.

C. Ozkan, O. Kisi, and B. Akay. 2011. Neural networks with artificial bee colony algorithm for modeling daily reference evapotranspiration. *Irrigation Science* 29, 6 (2011), 431–441.

B. R. Patel and K. K. Rana. 2014. A survey on decision tree algorithm for classification. *Journal of Engineering Development and Research* 2, 1 (2014), 5 pages.

A. P. Pawlovsky and M. Nagahashi. 2014. A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis. In *Proceedings of IEEE-EMBS International Conference on Biomedical and Health Informatics*. 189–192.

K. Polyak. 2011. Heterogeneity in breast cancer. *Journal of Clinical Investigation* 121, 10 (2011), 3786–3788.

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (1986), 81–106.

J. R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

C. R. Rao. 1948. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)* 10, 2 (1948), 159–203.

A. R. Razavi, H. Gill, H. Ahlfeldt, and N. Shahsavar. 2005. A data pre-processing method to increase efficiency and accuracy in data mining. In *Artificial Intelligence in Medicine*. Vol. 3581. Springer, Berlin, 434–443.

A. R. Razavi, H. Gill, H. Ahlfeldt, and N. Shahsavar. 2007. Predicting metastasis in breast cancer: Comparing a decision tree with domain experts. *Journal of Medical Systems* 31, 4 (2007), 263–273.

D. B. Rubin. 2004. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Y. U. Ryu, R. Chandrasekaran, and V. S. Jacob. 2007a. Breast cancer prediction using the isotonic separation technique. *European Journal of Operational Research* 181, 2 (2007), 842–854.

Y. U. Ryu, R. Chandrasekaran, and V. S. Jacob. 2007b. Data classification using the isotonic separation technique: Application to breast cancer prediction. *European Journal of Operational Research* 181 (2007), 1–30.

G. I. Salama, M. B. Abdelhalim, and M. A. E. Zeid. 2012. Experimental comparison of classifiers for breast cancer diagnosis. In *Proceedings of International Conference on Computer Engineering and Systems (ICCES)*. 180–185.

M. S. Santos, P. H. Abreu, P. J. García-Laencina, A. Simão, and A. Carvalho. 2015. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of Biomedical Informatics* 58 (2015), 49–59.

SAS Institute. 2015. SAS Enterprise Miner - SEMMA. Retrieved from https://web.archive.org/web/20120308165638/http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html/.

R. E. Schapire. 1990. The strength of weak learnability. *Machine Learning* 5, 2 (1990), 197–227.

B. Scholkopf and A. Smola. 2002. *Learning with Kernels*. MIT Press, Cambridge, MA.

SEER Research. 2015. Surveillance, Epidemiology, and End Results (SEER) Program. Retrieved from http://seer.cancer.gov/data/access.html.

L. A. Shalabi and Z. Shaaban. 2006. Normalization as a preprocessing engine for data mining and the approach of preference matrix. In *Proceedings of the International Conference on Dependability of Computer Systems*. 207–214.

J. Shawe-Taylor and N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

K. Socha and C. Blum. 2007. An ant colony optimization algorithm for continuous optimization: Application to feed-forward neural network training. *Neural Computing and Applications* 16, 3 (2007), 235–247.

M. Srinivas and C. K. Mohan. 2015. Multi-level classification: A generic classification method for medical data sets. In *Proceedings of the IEEE International Conference on E-Health Networking, Application Service*. 6 pages.

R. Srivastava. 2013. *Research Developments in Computer Vision and Image Processing: Methodologies and Applications*. IGI Global.

J. Stefanowski. 2005. An experimental study of methods combining multiple classifiers—Diversified both by feature selection and bootstrap sampling. In *Issues in the Representation and Processing of Uncertain and Imprecise Information*. 337–354.

M. M. Suarez-Alvarez, D. Pham, Y. Mikhail, and Y. I. Prostov. 2012. Statistical approach to normalization of feature vectors and clustering of mixed datasets. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 468, 2145 (2012), 2630–2652.

Y. Sun. 2007. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 6 (2007), 1035–1051.

Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie. 2007. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* 23, 1 (2007), 30–37.

Thomson Reuters. 2015. Web of Science. (2015). http://thomsonreuters.com/thomson-reuters-web-of-science/.

J. M. Tomczak. 2013. Prediction of breast cancer recurrence using classification restricted Boltzmann machine with dropping. *CoRR* abs/1308.6324 (2013), 9 pages.

H. Trevor, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics.

E. Trumbelj, Z. Bosnic, I. Kononenko, B. Zakotnik, and C. Kuhar. 2010. Explanation and reliability of prediction models: The case of breast cancer recurrence. *Knowledge Information System* 24 (2010), 305–324.

N. Tsikriktsis. 2005. A review of techniques for treating missing data in OM survey research. *Journal of Operations Management* 24, 1 (2005), 53–62.

C. Van den Hurk. 2011. Unfavourable pattern of metastases in M0 breast cancer patients during 1978–2008: A population-based analysis of the Munich cancer registry. *Breast Cancer Research and Treatment* 128, 3 (2011), 795–805.

L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415 (2002), 530–536.

V. Vapnik. 1999. *The Nature of Statistical Learning Theory (Information Science and Statistics)* (2nd ed.). Springer.

A. Verikas, A. Gelzinis, and M. Bacauskiene. 2011. Mining data with random forests: A survey and results of new tests. *Pattern Recognition* 44, 2 (2011), 330–349.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

World Health Organization. 2012. GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012. Retrieved from http://globocan.iarc.fr.

S. Zhang, Z. Qin, C. X. Ling, and S. Sheng. 2005. "Missing is useful": Missing values in cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering* 17, 12 (2005), 1689–1693.

B. Zheng, S. Yoon, and S. S. Lam. 2014. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications* 41, 4, Part 1 (2014), 1476–1482.

Z. H. Zhou and Y. Jiang. 2003. Medical diagnosis with c4.5 rule preceded by artificial neural network ensemble. *IEEE Transactions on Information Technology in Biomedicine* 7, 1 (2003), 37–42.

M. D. Zio, U. Guarnera, and O. Luzi. 2007. Imputation through finite Gaussian mixture models. *Computational Statistics and Data Analysis* 51, 11 (2007), 5305–5316.

Received July 2015; revised May 2016; accepted August 2016