Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers

Ertel Merouane[1]
Informatics and Applications Laboratory (IA), Faculty of Sciences
Moulay Ismail University
Meknes, Morocco

Amali Said[2]
Informatics and Applications Laboratory (IA), FSJES
Moulay Ismail University
Meknes, Morocco

El Faddouli Nour-eddine[3]
RIME Team, MASI Laboratory
E3S Research Center, EMI
Mohammed V University
Rabat, Morocco

***Abstract*—The volume and amount of data in cancerology is continuously increasing, yet the vast majority of this data is not being used to uncover useful and hidden insights. As a result, one of the key goals of physicians for therapeutic decision-making during multidisciplinary consultation meetings is to combine prediction tools based on data and best practices (MCM). The current study looked into using CRISP-DM machine learning algorithms to predict metastatic recurrence in patients with early-stage (non-metastatic) breast cancer so that treatment-appropriate medicine may be given to lower the likelihood of metastatic relapse. From 2014 to 2021, data from patients with localized breast cancer were collected at the Regional Oncology Center in Meknes, Morocco. There were 449 records in the dataset, 13 predictor variables and one outcome variable. To create predictive models, we used machine learning techniques such as Support Vector Machine (SVM), Nave Bayes (NB), K-Nearest Neighbors (KNN) and Logistic Regression (LR). The main objective of this article is to compare the performance of these four algorithms on our data in terms of sensitivity, specificity and precision. According to our results, the accuracies of SVM, kNN, LR and NB are 0.906, 0.861, 0.806 and 0.517 respectively. With the fewest errors and maximum accuracy, the SVM classification model predicts metastatic breast cancer relapse. The unbiased prediction accuracy of each model is assessed using a 10-fold cross-validation method.**

***Keywords*—Machine learning; classification; personalized medicine; CRISP-DM; metastasis; breast cancer**

I. Introduction

Breast cancer is a significant public health concern. According to data released by the World Cancer Observatory in 2018, 52,783 new cancer cases are reported in Morocco each year, with women accounting for 36.9% of these cases [1] , The key events linked to poor survival in breast cancer patients are disease progression and metastasis. Adjuvant chemotherapy (treatment given after surgery) combined with hormone therapy has been demonstrated in some trials to minimize the risk of recurrence and mortality from breast cancer [2], [3]. Due to the development of metastases and uncontrolled growth, various cases of female patients do not respond to therapeutic compounds in breast cancer in the same way [4].

Over the past two decades, personalized medicine has been defined in several ways. More broadly as a predictive, personalized, preventive and participatory health model ("P4 medicine") [5], and which also applies technologies to personalize and deliver care [6]. The use of personalized medicine or precision medicine in oncology aims to adapt treatments according to the characteristics of patients and their diseases by integrating all the biological and genetic, environmental, phenotypic and psychosocial knowledge found there clean [7]. Personalized medicine's ultimate goal is to provide the appropriate treatment to the appropriate person at the appropriate time [8].

The statistical method of machine learning techniques has shown to be a godsend for diagnostic, classification, prediction, and prognosis purposes in personalized medicine in cancer, given the amount of clinical data about each patient [9]–[13]. Various researchers are applying machine learning ideas to enhance cancer prediction and prognosis, this is done using a training data set whose variable assignments are already predetermined or known. Recently, researchers have focused more on decision trees, KNNs, SVMs and neural networks to predict cancer patient survival with high accuracy [14]–[16] . Web-based prediction models have been developed from cancer registry data to help determine the need for adjuvant therapy [17], [18]. PREDICT uses multivariate statistical analysis to calculate personalized survival probability based on the integration of clinical factors [19], [20]. However, the use of these models in clinical practice relies heavily on proof of the reliability of predictions and demonstration of acquired knowledge, moreover, the majority of them focus on overall survival rather than the risk of relapse. Given the paucity of predictive machine learning models that allow clinicians to identify patients at risk for metastatic relapse earlier by using a combination of various clinic-pathological characteristics, in particular Ki67 with tumor size, lymph node invasion and adjuvant therapy, we have seen fit to continue the current effort to resolve this problem.

In this study, our objective is to propose a supervised learning model, for predicting metastatic recurrence in individuals with early-stage breast cancer on an individual basis, which will guide the therapeutic decision in the multidisciplinary consultation meeting (MCM). Our model is fed by data including clinical, pathological, biological, therapeutic and prognostic characteristics. These data are collected from the files of patients with early-stage breast cancer, collected after the different stages of treatment

(diagnosis, relapse/progression, follow-up), offering a holistic view of previous successes and recommendations for good practices.

In the second part of this article, we will present the predictor variables introduced into the model, which predict the risk of metastatic relapse in patients before the start of adjuvant treatments (chemotherapy - Hormone therapy - Radiotherapy - Trastuzumab). The model proposal obtained according to the CRISP-DM process will be presented in the third section and in the last section we will analyze the results.

II. Related Work

In medical practice, the efficiency of breast cancer treatment is essentially determined on the ability to cancer prognosis, and cancer recurrence [21]. In recent years, with the use of machine learning technology in personalized medicine [6], modern oncology seeks to tailor treatments to expected results, through personalized predictive care models, based on patient characteristics patients and their pathologies by integrating all the biological and genetic, environmental, phenotypic and psychosocial knowledge that are specific to it. Tseng and Yi-Ju (2019) [22] propose an approach based on machine learning such as Random Forest (RF), Support Vector Machine (SVM), logistic regression (LR) and Naive Bayes (NB), to predict early breast cancer metastases using serum biomarkers and clinicopathologic data to reduce the risk of death. Tapak and Leili (2019) [23] proposed a model based on learning algorithms such as Naive Bayes (NB), Random Forest (RF), AdaBoost, Support Vector Machine (SVM), Least-squareSVM (LSSVM), Adabag, Logistic Regression (LR) and Linear Discriminant Analysis (LDA), for the prediction of breast cancer survival and metastasis.

In our research, we investigated four machine learning methods for predicting metastases in breast cancer patients: Support vector machine, Naive Bayes, K-Nearest Neighbour, and Logistic Regression. These algorithms are integrated into our proposed model according to the standard CRISP-DM process. The description of this model is the subject of the following section.

III. Materials and Methods

The phase of creating a predictive machine learning model is preceded by a preprocessing phase. In order to feed the model with clean data, the data may contain values that need to be transformed or eliminated, which can be useful for modeling. In our study, The CRISP-DM approach was employed (Cross Industry Standard Process for Data Mining) [24] which is considered to be an essential pillar for the success of a Machine Learning (ML) project. This method can help us find information and patterns hidden in a dataset with many features [25]. The CRISP method has six phases (see Fig. 1) that we will detail in the sections.

A. Data Understanding

1) Data source: Our predictive study included patients with localized breast cancer on all histological types of cancer collected at the regional oncology center of Meknes in Morocco, during the period 2014 to 2021, who had undergone surgery associated with adjuvant treatment during the years 2014 - 2016 (Chemotherapy and / or Hormonetherapy and / or Radiotherapy and / or Trastuzumab) with a follow-up of at least 48 months.

Our system's dataset contains 511 records and 14 variables. These variables provide demographic, clinical and therapeutic information about the patient, including the target variable (metastatic relapse). The data were collected from the computerized system which brings together the archives of patient files, which were then validated by experts (treating physicians).

2) Dataset features: The decision for adjuvant systemic treatment of breast cancer is based on clinical factors, such as (age) and histological (axillary lymph nodes, size, grade, vascular emboli), performance indicators, previous treatment methods, but also organic [26] Co-morbidities and of course the wishes of patients will continue to play an important role. To determine the adjuvant treatment of non-metastatic breast cancer [27]. Biologically, the expression of HR (Hormone Receptors) and the overexpression of HER2 (human epidermal growth factor receptor 2) are the main prognostic biomarkers and key predictors of the therapeutic effect [28], [29]. However, other biological parameters seem to have emerged recently, a study showed that the Ki67 index of cell proliferation in univariate and multivariate analyses of grade cancers, was the strongest predictor of overall and metastasis-free survival [30]. It is an important biomarker in the management of breast cancer, it can be used to guide clinical decisions regarding adjuvant chemotherapy [31].

The variables collected from the computer system of the regional oncology center of Meknes-Morocco, were the prognostic factors currently validated in breast cancer : The age of the patient, the size of the tumor, the pathological state of the lymph nodes lymphatic, grade, stage, histological type of tumor, estrogen receptor (ER) status, progesterone receptors (PR) grouped into hormone receptors (HR), HER2 overexpression, Ki67 status (cell proliferation), L ' surgical approach and types of adjuvant therapy.

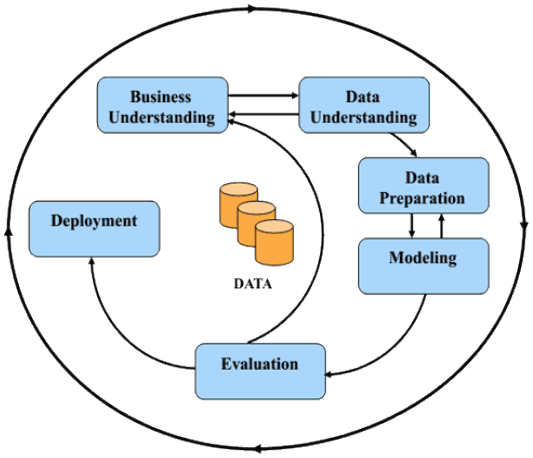

Fig. 1. Phases of the Current CRISP-DM Process Model for Data Mining

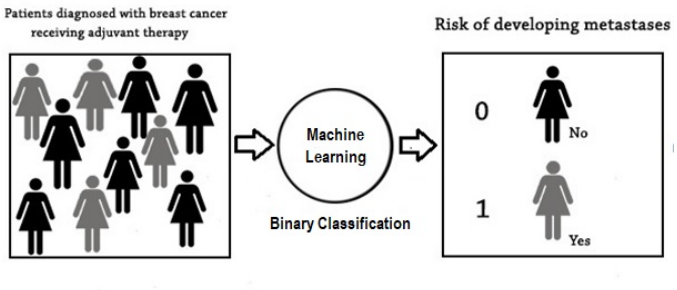

Fig. 2. Our Binary Classification Model for the Prediction of Metastatic Recurrence.

Our predictive model is based on these indicators to classify new patients with non-metastatic breast cancer into two classes: patients at low (0) or high (1) risk of metastatic relapse at 4 years (see Fig. 2).

B. Data Preparation

Data preparation is made up of several stages: Data cleaning, Data Transformation.

1) Data cleaning: The data collected from the information system of the Meknes Regional Oncology Center in Morocco is organized in the form of a database. This database has undergone a cleaning process to eliminate and reduce noise:

- Attribute noise is caused by input errors, missing variable values and redundant data.
- Class noise which is due to errors introduced when assigning instances to classes.

After removing the rows with substantial missing values, we checked for missing or null data points in the database using Python's pandas library (see Fig. 3).

The number of records kept is 449 records, each showing a different case of breast cancer with its own combination of treatments. Each of these cases is represented by 13 independent predictors / variables, plus 1 dependent / categorical variable that reflects metastatic relapse in breast cancer patients (No / Yes).

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 449 entries, 0 to 448
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age_diagnosis       449 non-null    int64
 1   Tumor_size          449 non-null    int64
 2   Lymph_Nodes         449 non-null    int64
 3   Tumor_stage         449 non-null    int64
 4   Cancer_Grade        449 non-null    int64
 5   HER2                449 non-null    object
 6   HR                  449 non-null    object
 7   Ki67                449 non-null    int64
 8   Surgery_Type        449 non-null    object
 9   Chemotherapy        449 non-null    object
 10  Trastuzumab         449 non-null    object
 11  Radiotherapy        449 non-null    object
 12  Hormonotherapy      449 non-null    object
 13  Metastatic_Relapse  449 non-null    object
dtypes: int64(6), object(8)
memory usage: 49.2+ KB
```

Fig. 3. Attribute Information of the Dataset.

2) Data transformation: The quality of the data and the amount of useful information are key factors that determine the learning ability of a machine learning algorithm. Therefore, it is absolutely essential to make sure that we encode categorical variables correctly, before using the data in a machine learning algorithm [32]. In this study we have 14 distinct attributes: 3 attributes represent numeric characteristics, 10 attributes represent object type variables, and the last attribute represents an object output variable, this means that our data contains object / categorical type variables, they must be coded by numbers before we can fit and evaluate our model.

For this, we used the technique (OneHotEncoder) from the Scikit-Learn library in the Pandas module of Python, to create a hot-encoding of integer-encoded values, which transforms the input categorical variables into numbers. This method increases the overall number of input characteristics, so this type of encoding creates a binary variable for each unique value of the nominal characteristic. The binary variable specifies (0) or (1) whether or not the category appears in observation (see Table I).

C. Modeling

The data preprocessing step is followed by a modeling process, which involves training the machine learning algorithms to predict the classes from the features. In this study, the presented entries are normalized so that all variables are on the same scale and distribution, in order to compare the performance of the models and evaluate them in the same way. We used the method (model_selection.KFold) of the SciKit-Learn library in Python, to train the model to create the cross-validation folds by 10. Indeed, this method is used to evaluate predictive models which divide the set original into a training sample that represents the training DataSet, and another set reserved for testing and evaluating the model. The result is a trained model that can be used for inference; making predictions on new data points (see Fig. 4).

TABLE I. CHARACTERISTICS OF PATIENTS WITH BREAST CANCER AND POSSIBLE VALUES

Attribute	***Type Attribute***	***Possible Value***
Age_diagnosis	*Numerical*	*20 - 80 (Years)*
Tumor_size	*Numerical*	*10 mm - 70 mm*
Lymph_Nodes	*Categorical*	*0 - 3*
Tumor_stage	*Categorical*	*0 - 3*
Cancer_Grade	*Categorical*	*1 - 3*
HER2	*Categorical*	*0 (Negative) - 1 (Positive)*
HR	*Categorical*	*0 (Negative) - 1 (Positive)*
Ki67	*Numerical*	*8 % - 60 %*
Surgery_Type	*Categorical*	*0 (Tumorectomy) - 1 (Mastectomy)*
Chemotherapy	*Categorical*	*0 (No) - 1 (Yes)*
Trastuzumab	*Categorical*	*0 (No) - 1 (Yes)*
Radiotherapy	*Categorical*	*0 (No) - 1 (Yes)*
Hormonotherapy	*Categorical*	*0 (No) - 1 (Yes)*
Metastatic_Relapse	*Categorical*	*0 (No) - 1 (Yes)*

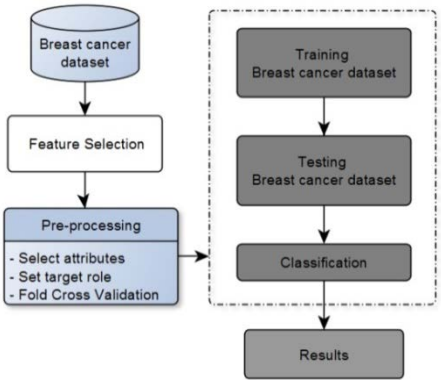

Fig. 4. Supervised Learning Workflow.

1) Classification methods: In the present study, four machine learning methods were used and compared to predict metastasis in breast cancer patients: Support Vector Machine, Naive Bayes, K-Nearest Neighbors, and Logistic Regression.

a) Support Vector Machine (SVM): SVM are a type of supervised learning algorithms for classification, regression analysis, and outlier identification that examine data [33]. It's a discriminating model described by a hyperplane; in our case, the hyperplane categorizes new instances into one of two classes: 0 or 1.

b) Naïve Bayes (NB): The influence of a variable value on a specific class is assumed to be independent of the values of other variables by NB classifiers. This is referred to as conditional class independence. When the training dataset is small, it is utilized to identify crucial classification parameters [33]. The NB classifier, which combines the Bayes probability model with a decision rule, is one of the most extensively used binary classification algorithms.

c) K-Nearest Neighbors (KNN): The non-assumption of the variable's distribution is one of the method's advantages. When comparing the two preceding techniques, this is a crucial consideration. To maximize classification and cope with the bias-variance trade-off, this approach must determine the optimal value of k, the number of neighbors. Optimal choices of k keep the bias-variance balance in check and, ideally, reduce both [34].

d) Logistic Regression (LR): LR is a classification algorithm generally used in binary classification problems [35], as is the case here with negative, 0 and positive response values, 1. It uses the maximum likelihood estimate for assess the probability of belonging to a class.

2) Performance measures: It is necessary to calculate the model's accuracy in order to test its capacity to anticipate occurrences in the proper class. The following procedures were employed.

a) Confusion Matrix: This is a statistic for evaluating a classification model's performance. It's also known as an error matrix since it may be used to figure out where the model is off in its predictions. The confusion matrix analyzes the number of accurate and wrong predictions after the prediction. On the basis of these factors, classifier comparisons are made (see Fig. 5).

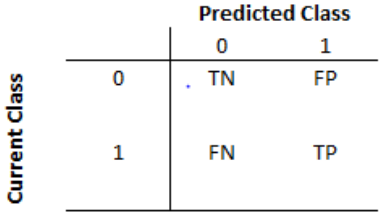

Fig. 5. Confusion Matrix.

We may designate one class as positive and one as negative per row and true or false per column in binary classification, giving us:

- TP: correct relapse expected.
- TN: correction of the expected non-relapse.
- PF: incorrectly predicted relapse.
- FN: incorrect non-relapse prediction.

b) Classification report: A classification report is used to assess the classification model's quality.

The proportion of right guesses in the overall number of correct forecasts is known as accuracy. It is calculated by (1), where TP and TN indicate the number of properly categorized positive and negative cases, respectively, and FN and FP represent the number of incorrectly classified negative and positive examples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

The ratio of true positives to all positives is called precision (2). This would be the measurement of individuals accurately identified as having a risk of metastatic recurrence among all patients truly at risk for our issue statement.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Equation (3) defines the true negative rate (specificity). Among all negative data points, the false positive rate is the fraction of negative data points that are correctly classified as negative.

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

The real positive rate, calculated by equation (4), is the recall (sensitivity). Out of all positive data points, this rate represents the proportion of positive data points that are accurately classified as positive.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

The Roc curve and AUC: When the decision threshold is changed, a Receiver Operating Characteristic (ROC) curve displays the rate of true positives (sensitivity) versus the rate of false positives (1 - specificity) [36]. The area under the curve (AUC) is a measure of the likelihood that the model would rate a positive random example higher than a negative random example. Its values range from 0 to 1. The AUC of a model with 100% incorrect predictions is 0. Its AUC is 1 if all of its predictions are right.

The comparison of the performance of learning algorithms, discussed in the next section, is based on these indicators (Accuracy; Precision; specificity; recall; AUC).

IV. Results and Discussion

We utilized Jupyter Notebook, Python modules (pandas, matplotlib, bumpy), and the scikit-learn framework to process ML algorithms for our analysis. To predict metastasis in breast cancer patients, the following approaches were tested: (SVM, NB, k-NN, and LR).

First, we performed training for 70% of the dataset (314 random records), applying the cross-validation method checking all the metrics mentioned previously. Then we ran a test of the remaining 30% of the data set. Table II illustrates the prediction results by successfully classified and wrongly categorized examples for the methods (SVM, NB, kNN, and LR).

Then we compared the difference between the precision results found in the test and the total, this comparison is based on the indicators Accuracy, Precision, sensitivity, specificity, Roc curve and AUC, to measure the performance of these algorithms based on the Confusion Matrix entries. The findings are shown in Table III.

The best classification performance is obtained with SVM, as shown in Table II, which correctly predicts 123 instances out of 135 (94 instances 0 which are in fact 0 and 29 instances 1 which are in fact 1), and 12 badly predicted instances (03 instances of class 0 predicted as 1 and 09 instances of class 1 predicted as 0). We also notice that NB has the lowest value of correctly classified instances and the highest value of misclassified instances (36 badly predicted instances) compared to the other classifiers (12 incorrect instances for kNN and LR).

In Table III, the results of the performance measurements of the four classification algorithms clearly show that the SVM and kNN achieved the highest precision (91.1%). kNN has reached the highest sensitivity (Recall), which is 81.6%. And NB the worst specificity (51.7%). We can also notice that SVM surpasses the other classifiers in terms of Precision (90.6%), Specificity (96.9%), AUC (92.9%). This is why, with a score of (91.1%) and a smaller error, the SVM outperforms the other classification approaches utilized in our study.

TABLE II. Confusion Matrix of Classification Techniques BLE

Classifiers	***Predicted***		***Test Size = 0.30***	
	0	***1***		
SVM	*94*	*3*	***0***	***Current***
	9	*29*	***1***	
kNN	*92*	*5*	***0***	
	7	*31*	***1***	
LR	*90*	*7*	***0***	
	9	*29*	***1***	
NB	*69*	*28*	***0***	
	8	*30*	***1***	

TABLE III. Classifiers Performance

Classifiers	***Accuracy (%)***	***Precision***	***Specificity***	***Recall***	***AUC***
SVM	*91,1*	*0,906*	*0,969*	*0,763*	*0,929*
kNN	*91,1*	*0,861*	*0,948*	*0,816*	*0,882*
LR	*88,1*	*0,806*	*0,928*	*0,763*	*0,914*
NB	*73,3*	*0,517*	*0,711*	*0,789*	*0,750*

The ROC curve, on the other hand, offers for a better grasp of a machine learning algorithm's capability. Fig. 6 shows the ROC curves displayed for the fitted test models in our investigation.

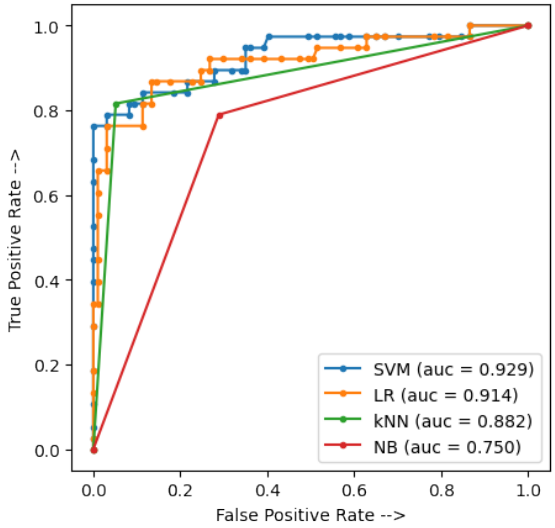

Fig. 6. Roc Curve (AUC) Models.

SVM is the best classifier, as seen in Fig. 6, since the curve is squished towards the top left edge, then travels towards the upper right corner (90.6 % sensitive and 96.9% specific), followed by the other algorithms: LR, kNN, and NB.

The results obtained for various performance indicators show the effectiveness of the SVM model in predicting metastatic relapse in patients with early-stage breast cancer with the highest precision value of 91.1% and AUC score of 92.9%.

V. Conclusion

In this article, we proposed a model that could be used during the multidisciplinary consultation meeting (MCM), as a personalized prediction tool for the systematic management of patients with early breast cancer. Our model predicts the risk of metastatic relapse after four years for breast cancer patients likely to receive adjuvant therapy. This prediction can help decision-making in order to improve therapeutic management and increase the overall survival and quality of life of patients. We also presented a comparative study on the efficiency and effectiveness of the SVM, NB, k-NN and LR algorithms in terms of accuracy, precision, and sensitivity to find the best classification precision. The results obtained show that SVM has proven its efficiency and achieves the best performance in terms of precision and low error rate.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[2] Q. W. Lopez, "Évaluation de la réponse aux traitements et détermination de facteurs prédictifs et pronostiques dans le cancer du sein luminal (récepteurs hormonaux positifs/HER2-)," p. 222.

[3] E. Deluche and J.-Y. Pierga, "Chimiothérapie et femme jeune dans le cancer du sein : quelle prise en charge ?," Bulletin du Cancer, vol. 106, no. 12, pp. S19–S23, Dec. 2019, doi: 10.1016/S0007-4551(20)30043-6.

[4] A. Mailliez, C. Decanter, and J. Bonneterre, "Chimiothérapie adjuvante de cancer du sein et fertilité : estimation de l'impact, options de préservation et place de l'oncologue," Bulletin du Cancer, vol. 98, no. 7, pp. 741–751, Jul. 2011, doi: 10.1684/bdc.2011.1391.

[5] L. Hood and M. Flores, "A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory," New Biotechnology, vol. 29, no. 6, Art. no. 6, Sep. 2012, doi: 10.1016/j.nbt.2012.03.004.

[6] R. Snyderman, "Personalized health care: From theory to practice," Biotechnology Journal, vol. 7, no. 8, Art. no. 8, Aug. 2012, doi: 10.1002/biot.201100297.

[7] I. Greenwalt, N. Zaza, S. Das, and B. D. Li, "Precision Medicine and Targeted Therapies in Breast Cancer," Surgical Oncology Clinics of North America, vol. 29, no. 1, Art. no. 1, Jan. 2020, doi: 10.1016/j.soc.2019.08.004.

[8] J. C. O'Donnell, "Personalized Medicine and the Role of Health Economics and Outcomes Research: Issues, Applications, Emerging Trends, and Future Research," Value in Health, vol. 16, no. 6, Art. no. 6, Sep. 2013, doi: 10.1016/j.jval.2013.06.004.

[9] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Computer Science, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[10] P. Jayapaul, A. Balasundaram, K. P. D. Seturamalingam, and K. Sekar, "Performance analysis of machine learning techniques for the prediction of breast cancer in big data environment," Erode, India, 2020, p. 140006. doi: 10.1063/5.0011116.

[11] M. M. Ibrahim, D. Ahmed, and R. Ahmed, "Deep Learning Hybrid with Binary Dragonfly Feature Selection for the Wisconsin Breast Cancer Dataset," IJACSA, vol. 12, no. 3, 2021, doi: 10.14569/IJACSA.2021.0120314.

[12] K. Rajendran, M. Jayabalan, and V. Thiruchelvam, "Predicting Breast Cancer via Supervised Machine Learning Methods on Class Imbalanced Data," IJACSA, vol. 11, no. 8, 2020, doi: 10.14569/IJACSA.2020.0110808.

[13] T. A. Khan, K. A., S. Nasim, M. Alam, Z. Shahid, and M. S. Mazliham, "Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease Classification," IJACSA, vol. 11, no. 11, 2020, doi: 10.14569/IJACSA.2020.0111170.

[14] G. Battineni, N. Chintalapudi, and F. Amenta, "Performance analysis of different machine learning algorithms in breast cancer predictions," EAI Endorsed Transactions on Pervasive Health and Technology, vol. 6, no. 23, p. 166010, Sep. 2020, doi: 10.4108/eai.28-5-2020.166010.

[15] W. Kim, K. S. Kim, and R. W. Park, "Nomogram of Naive Bayesian Model for Recurrence Prediction of Breast Cancer," Healthc Inform Res, vol. 22, no. 2, p. 89, 2016, doi: 10.4258/hir.2016.22.2.89.

[16] "Performance of Support Vector Machine Kernels (SVM-K) on Breast Cancer (BC) Dataset," ijrte, vol. 8, no. 2S7, pp. 412–417, Sep. 2019, doi: 10.35940/ijrte.B1076.0782S719.

[17] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," Cancer Inform, vol. 2, p. 117693510600200, Jan. 2006, doi: 10.1177/117693510600200030.

[18] W. Kim et al., "Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine," J Breast Cancer, vol. 15, no. 2, p. 230, 2012, doi: 10.4048/jbc.2012.15.2.230.

[19] S. Mook et al., "Calibration and discriminatory accuracy of prognosis calculation for breast cancer with the online Adjuvant! program: a hospital-based retrospective cohort study," The Lancet Oncology, vol. 10, no. 11, pp. 1070–1076, Nov. 2009, doi: 10.1016/S1470-2045(09)70254-2.

[20] G. C. Wishart et al., "RPeRseaErcDh aIrCticlTe : a new UK prognostic model that predicts survival following surgery for invasive breast cancer," Breast Cancer Research, p. 10, 2010.

[21] M. Fekih et al., "Utilisation de référentiels et hétérogénéité décisionnelle des indications de chimiothérapie adjuvante dans les cancers du sein exprimant les récepteurs hormonaux, HER2-négatifs: résultats d'un sondage national en France," Bulletin du Cancer, vol. 101, no. 10, pp. 918–924, Nov. 2014, doi: 10.1684/bdc.2014.2030.

[22] Y.-J. Tseng et al., "Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies," International Journal of Medical Informatics, vol. 128, pp. 79–86, Aug. 2019, doi: 10.1016/j.ijmedinf.2019.05.003.

[23] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," Clinical Epidemiology and Global Health, vol. 7, no. 3, Art. no. 3, Sep. 2019, doi: 10.1016/j.cegh.2018.10.003.

[24] P. Chapman et al., "Step-by-step data mining guide," p. 76.

[25] H. L. Afshar, M. Ahmadi, M. Roudbari, and F. Sadoughi, "Prediction of Breast Cancer Survival Through Knowledge Discovery in Databases," Global Journal of Health Science, vol. 7, no. 4, p. 7, 2015.

[26] A. Gonçalves, J. Moretta, F. Eisinger, and F. Bertucci, "Médecine personnalisée et cancer du sein : médecine anticipatoire, évaluation pronostique et ciblage thérapeutique," Bulletin du Cancer, vol. 100, no. 12, Art. no. 12, Dec. 2013, doi: 10.1684/bdc.2013.1856.

[27] C. Georges-Tarragano, F. Tapié de Celeyran, J. Platon, and J.-L. Misset, "Décider en cancérologie dans les situations médicosociales complexes: les réunions de concertation pluriprofessionnelles médicosociales et éthiques à l'hôpital Saint-Louis de Paris," Oncologie, vol. 16, no. 1, pp. 55–62, Jan. 2014, doi: 10.1007/s10269-014-2368-5.

[28] M. Scimeca et al., "Novel insights into breast cancer progression and metastasis: A multidisciplinary opportunity to transition from biology to clinical oncology," Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, vol. 1872, no. 1, pp. 138–148, Aug. 2019, doi: 10.1016/j.bbcan.2019.07.002.

[29] H. Rizki, C. Hillyar, O. Abbassi, and S. Miles-Dua, "The Utility of Oncotype DX for Adjuvant Chemotherapy Treatment Decisions in Estrogen Receptor-positive, Human Epidermal Growth Factor Receptor 2-negative, Node-negative Breast Cancer," Cureus, Mar. 2020, doi: 10.7759/cureus.7269.

[30] F. Penault-Llorca and N. Radosevic-Robin, "Ki67 assessment in breast cancer: an update," Pathology, vol. 49, no. 2, Art. no. 2, Feb. 2017, doi: 10.1016/j.pathol.2016.11.006.

[31] C. Criscitiello et al., "High Ki-67 score is indicative of a greater benefit from adjuvant chemotherapy when added to endocrine therapy in Luminal B HER2 negative and node-positive breast cancer," The Breast, vol. 23, no. 1, pp. 69–75, Feb. 2014, doi: 10.1016/j.breast.2013.11.007.

[32] P. Cerda, "Similarity encoding for learning with dirty categorical variables," Mach Learn, p. 18, 2018.

[33] M. N. Murty and V. S. Devi, Pattern recognition: an algorithmic approach. London: Springer, 2012.

[34] G. James, D. Witten, T. Hastie, and R. Tibshirani, Eds., An introduction to statistical learning: with applications in R. New York: Springer, 2013.

[35] F. C. Pampel, Logistic regression: a primer. Thousand Oaks, Calif: Sage Publications, 2000.

[36] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.