SIMULATING BREAST CANCER TREATMENT EFFICACY: A COMPUTATIONAL APPROACH TO OPTIMIZING PATIENT CARE

MEROUANE ERTEL[1] , AZIZ MENGAD[2] ,SAMIRA FADILI[3] ,YOUNES BOUFERMA[4] OUSSAMA RHARIB[5] ,MERYEM CHAKKOUCH[6] ,SAID AMALI[7]

[1,6] Informatics and Applications Laboratory (IA), Faculty of Sciences, Moulay Ismail University, Morocco

[2]Centre for Doctoral Studies "Life and Health Sciences"-Drug Sciences Formation, Laboratory of Pharmacology and Toxicology (LPTR), Faculty of Medicine and Pharmacy of Rabat (FMPh), Impasse Souissi Rabat, Morocco

[3]Laboratory of Education, Culture, Arts and Teaching of French Language and Literature (ECADLLF), Faculty of Sciences of Education, Mohammed V University in Rabat, Morocco

[4]Faculty of Legal, Economic and Social Sciences, Moulay Ismail University, Meknes, Morocco

[5]Sociology and Psychology Laboratory, Faculty of Letters and Human Sciences Dhar El Mahraz, University Sidi Mohammed Ben Abdellah of FES, Morocco

[7]Informatics and Applications Laboratory (IA), FSJES, Moulay Ismail University, Morocco

E-mail: [1]m.ertel@edu.umi.ac.ma, [2]mengad.aziz@hotmail.com, [3]sfadili705@gmail.com, [4]bahaebik@gmail.com, [5]Oussama.rharib@usmba.ac.ma, [6]chakkchou.m@gmail.com, [7]s_amali@yahoo.com

ABSTRACT

Treatment of breast cancer with chemotherapy is common, but its effectiveness can vary significantly depending on the individual characteristics of the patient and the type of cancer. In this context, computer simulation based on machine learning can constitute a solution to optimize the treatment strategy of patients suffering from this disease. This study uses a dataset of 490 breast cancer patients, to feed a machine learning model and uses simulation techniques to simulate different treatment strategies. Machine learning algorithms, such as Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and artificial neural networks (ANN), have been evaluated for their performance. The results indicate that the RF algorithm achieved the highest accuracy rate of 76.9%, while the NB algorithm recorded the lowest accuracy rate of 66.5%.The study demonstrates that machine learning-based computer simulation can help identify breast cancer patients at high risk of metastatic relapse and predict an individualized therapeutic combination to reduce this risk.

Keywords: *Computer Simulation, Machine Learning; Modeling, Personalized Medicine; Combination Therapy; Prediction of Therapeutic Response; Breast Cancer.*

1. INTRODUCTION

The most frequent malignancy in women is breast cancer, with approximately 2.3 million new cases diagnosed each year worldwide. According to World Health Organization (WHO) statistics from 2020, In Morocco, breast cancer is the main reason why women pass away, accounting for 19.4% of all cancer-related deaths. Breast cancer accounts for approximately 30% of all cancers diagnosed in women in Morocco [1]. Chemotherapy is one of the most common treatments used to fight breast cancer. It consists of administering anti-cancer drugs intravenously or orally. The drugs work by destroying cancer cells by disrupting their cycle of growth and division. Chemotherapy can be given before or after surgery to reduce the size of the tumor and prevent it from spreading[2]. Once the tumor is removed, it is crucial for a breast cancer patient to choose an adjuvant therapy that can eradicate tiny foci of cancerous cells which, if ignored, could spread and develop into cancer metastatic[3]. Clinicians use a combination of therapies depending on the characteristics to achieve the best results. The combination of treatments may include surgery, chemotherapy , radiation therapy, hormone therapy, and immunotherapy [4], [5], [6].

In this context, Computer simulation based on machine learning can be used to predict the therapeutic combination in breast cancer based on various individual factors such as disease stage, genetic and molecular characteristics of the tumor, and potential response to the different treatments available [7], [8], [9].Computer simulation uses mathematical models and algorithms to simulate tumor behavior and potential treatments. These models can be based on real clinical data, such as biopsy data, radiology images, and genetic test results[10], [11], [12], [13], [14]. Using these models, clinicians can predict a patient's response to different treatments and design a personalized therapy combination that can deliver the best results. Computer simulation can also be used to optimize drug doses and treatment regimens, which could mitigate adverse effects and improve patients' quality of life.

Recently, several studies have used computer simulation to predict the therapeutic combination in breast cancer. A study published in the journal Breast Cancer Research and Treatment in 2018 used computer simulation algorithms to predict the response of metastatic breast cancer patients to different treatment combinations [15]. The results showed that the personalized treatment combinations led to a significant improvement in progression-free survival and overall survival. Another study published in the journal Nature Communications in 2019 used computer simulation models to identify optimal treatment regimens for patients with triple-negative breast cancer, an aggressive subtype of breast cancer[16]. The results showed that the combination of chemotherapy and immunotherapy could improve the survival of patients with this subtype of breast cancer. A study published in the journal Clinical Cancer Research in 2020 used computer simulation to assess the effectiveness of different treatments for patients with HER2-positive breast cancer [17]. The results showed that the combination of targeted therapy and chemotherapy might offer the best results for patients with this subtype of breast cancer. These studies show that computer simulation can be a valuable tool for predicting therapeutic combination in breast cancer and improving patient outcomes. However, it is important to note that computational models still need to be validated on a large scale before being widely used in clinical practice.

In this paper, we present a computer simulation model utilizing different machine learning algorithms including Naive Bayes (NB), Decision Tree (DT), Random Forest (RF) and artificial neurons networks (ANN), we evaluated the performance of the models. This model based on a variety of parameters, such as tumor grade, size, hormone receptors, as well as different therapies, in particular, this model can help identify patients who have a high risk of metastatic relapse and recommend treatments, Individualized measures that can reduce this risk. This research could have important implications for improving the quality of care and outcomes for breast cancer patients.

The predictor variables utilized in the model will be presented in the second section of this study. In the third section, various data processing (collecting, pre-processing, cleaning, and transformation) for the clinical, biological and pathological data set will be described. A description of the recommended multinomial regression methods is given in the fourth section. The last section will give the analysis of the data utilized to determine the ideal treatment for those with breast cancer to prevent metastatic recurrence.

2. RELATED WORK

The prediction of optimal therapeutic strategies for breast cancer, particularly those aimed at reducing metastatic risks, has been the subject of extensive research. Given the global significance of breast cancer as a health concern, numerous studies have delved into diverse methodologies to enhance treatment outcomes and diminish the likelihood of metastatic recurrence.

In recent times, researchers have focused on the development of web-based models leveraging extensive data from cancer registries. These models aim to ascertain the most suitable therapeutic approach for early-stage breast cancer patients [13], [18], [19], [20], [21], [22], [23]. Noteworthy among these is the PREDICT tool, which calculates individualized survival probabilities by integrating clinical variables through multivariate statistical analysis [24]. This tool is highly recommended for treatment planning.

Additionally, tools like Adjutorium assess the necessity of prescribing adjuvant therapies, such as chemotherapy and hormonal therapy, in conjunction with surgery [17].

In 2022, Jonathan M. Ji and Wen H. Shen designed a web application to predict breast cancer survival rates, providing valuable insights for medical decision-making and treatment guidance. The study compared eight classical models (KNN, Logistic Regression, Decision Tree, Random Forest, Extra Trees, AdaBoost, SVC, and XG Boost) [25] .

Aligned with these advancements in predictive machine learning models, which empower clinicians to accurately anticipate the efficacy of various therapeutic combinations, our study aims to utilize easily accessible clinical information. Binary variables representing the treatment protocols (Surgery, chemotherapy, hormonal therapy, Herceptin, and radiotherapy) will be employed to elucidate the contribution of each treatment to relapse-free survival rates. This approach will enable us to predict effective treatment combinations for individual breast cancer patients, thereby preventing instances of undertreatment or overtreatment.

3. MATERIALS AND METHODS

3.1 Data Understanding

3.1.1 Data Source

At the Regional Oncology Center of Meknes in Morocco, from (2014 – 2021), patients with breast cancer localized, were treated with one or a combination of the following treatments (surgery, chemotherapy, hormone therapy, radiotherapy, therapies) between 2014 and 2016 were included in this predictive study.

Our model dataset has 14 variables and 490 inputs. These variables, including the target variable, provide information on the patient's demographics, clinical status, and therapy. Data was collected from a hospital information system.

3.1.2 Dataset features

The attending physicians entered in the system the features of the tumor factors, patient follow-up, and treatment results. The following data were taken from each patient: age, size of the first tumor (TS), age at menopause, histological classification, marker of cell proliferation (Ki67), number of axillary lymph nodes involved, epidermal growth factor receptor 2 (Her2) status, estrogen (ER) and progesterone (PR) status, as well as a variety of treatment protocols (types of surgery, chemotherapy, radiation therapy hormone therapy, and Herceptin). Table 1 lists dataset.

Table 1 : Information about patients' demographics, cancer, and treatment.

#	Variable	Definition	Non-Null	Dtype
0	Age_Diagnosis	20- 34, 34 to 45, 46-55, 56 ≥ years old	325	Int64
1	Post_menopausal	0 ≤50 1 = >50	325	Int64
2	Tumor_Size	≤3 4-6 ≥7	325	Float64
3	Lymph_Nodes	0 = "No" 1 = "1–4" 2 = "5–9" 3 = ">9"	325	Int64
4	Tumour Grade	“1”- “2” or “3”	325	Int64
5	HER2	“0” = Negative, “1” = Positive	325	Int64
6	ER	“0” = Negative, “1” = Positive	325	Int64
7	PR	“0” = Negative, “1” = Positive	325	Int64
8	KI 67	≤ 14 % , 15-24 % , 25-29 % , ≥30	325	Int64
9	Surgery	“No Mastectomy “ “Yes Mastectomy”	325	Object
10	Chimiotherapy	No Yes	325	Object
11	Herceptine	No Yes	325	Object
12	Radiotherapy	No Yes	325	Object
13	Hormonotherapy	No Yes	325	Object

4. METHODOLOGY

3.1. Simulation model

We also used simulation techniques to simulate different treatment strategies for breast cancer patients using the results of machine learning algorithms[10], [26]. We used these simulations to assess the effectiveness of different treatment strategies and identify the most effective strategies for breast cancer patients.

This research presents a simulation computer model based on a combination of clinical and biological information and recommendations for adjuvant therapy. We constructed our prediction model based on the five most common treatment approaches to produce projected results (Surgery, chemotherapy, Hormonotherapy, Herceptin and Radiotherapy) (Fig 1). We studied the methods of cancer treatment employed at the regional oncology center of Meknes. This model forecasts therapy approaches that successfully lower the probability of metastatic recurrence.

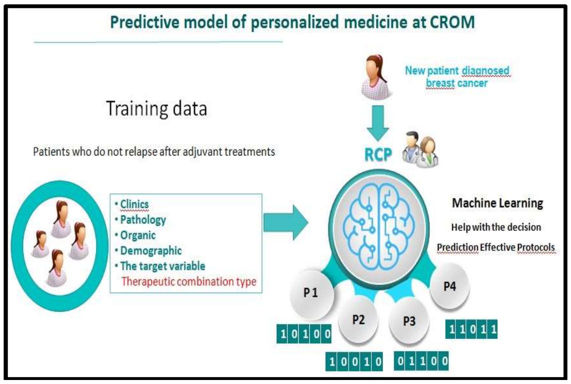

Predictive model of personalized medicine at CROM
New patient diagnosed breast cancer
Training data
Patients who do not relapse after adjuvant treatments
RCP
• Clinics
• Pathology
• Organic
• Demographic
• The target variable
Therapeutic combination type
Machine Learning
Help with the decision
Prediction Effective Protocols
P1
P2
P3
P4
1 0 1 0 0
1 0 0 1 0
0 1 1 0 0
1 1 0 1 1

Fig.1 Our predictive model of adjuvant treatment combinations

The processes involved in developing the multi-class prediction model for our investigation are described in detail in the section that follows.

3.2. Data Preparation

3.2.1. Data cleaning

The database extracted from the system included 490 patients who received treatment for breast cancer. To remove and reduce noise, the database has undergone a cleaning procedure. 120 people having tumor recurrences throughout the course of the research were disregarded, and 45 registries with insufficient data were also eliminated. The result was a data collection of 325 records, each representing a distinct instance of breast cancer treated with a particular therapy plan (see Table 1).

Each instance in Table 1 can be represented by one of 14 alternative attributes, with attributes ranging from 0 to 8 representing various clinicopathologic traits (independent variables) and attributes between 09 to 13 representing various treatment options or dependent variables/ categorical.

3.2.2. Multi-label classification encoding:

Multi-label classification is a type of machine learning problem where an instance may belong to more than one class at the same time. Binary relevance is one approach to solving this problem, where a separate binary classification model is trained for each class, treating the problem as multiple binary classification problems.

In binary relevance, the problem is transformed into a set of binary classification problems, where each problem predicts the presence or absence of a single class. To train the models, the original dataset is converted into multiple datasets, each containing the same instances but with the target variable (i.e., the classes) modified to reflect the presence or absence of a single class.

Once the binary classification models are trained, they can be used to predict the presence or absence of each class for a given instance. The final prediction for an instance is a set of binary labels, one for each class. In this study, in this study , the target variable had categorical values, represented by strings such as "Surgery", "Chemotherapy", "Herceptin", "Radiotherapy", and "Hormone therapy". In order to use these values as the target variable in a machine learning model, they needed to be converted into numerical values.
The Python library "convert objects (convert numeric=True)" was used to convert the objects. This parameter indicates that the function convert the categorical values into numerical values.

It is common practice in machine learning to represent categorical variables as numerical values in order to use them in models. This is because most machine learning algorithms operate on numerical data (see Table 2).

Table 2. Binary relevance coding method used in this study for combinations of therapeutic strategies.

Treatment Protocol		Encoding	Example of protocol combination code
Surgery	No	0	1
	Yes	1	
Chimiotherapy	No	0	0
	Yes	1	
Herceptin	No	0	0
	Yes	1	
Radiotherapy	No	0	1
	Yes	1	
Hormonotherapy	No	0	0
	Yes	1	

3.2.3. Conversion to Categorical Data:

In machine learning, conversion to categorical data refers to the process of transforming a numerical or text-based feature into a categorical feature, where the values represent different categories or classes. This transformation is necessary when working with algorithms that require categorical input, such as decision trees, random forests, and some neural networks. There are several methods for converting numerical or text-based data to categorical data[27].

Pandas are a popular data manipulation library in Python that provides powerful data structures for working with tabular data, such as data frames. In this study, the concatenation method in Pandas is used to combine five data frames by appending rows from one to another or by concatenating columns from one data frame to another, des variables de traitement "Surgery", "Chemotherapy", "Herceptin", "Radiotherapy", and "Hormone therapy" encodées ont été fusionnées à l'aide de la méthode de concaténation disponible dans la bibliothèque Pandas en Python. The Combination Therapy Binary Code variable is the name of this new one with 05 values, column (N°09) (see Table 3).

Table 3. Target Combination Variables And Dataframe Details Following Encoding

#	Variable_Name	Non-Null	Dtype
0	Age_Diagnosis	325	Int64
1	Postmenopausal	325	Int64
2	Tumor_Size	325	Float64
3	Lymph_Nodes	325	Int64
4	Tumour_Grade	325	Int64
5	HER2	325	Int64
6	ER	325	Int64
7	PR	325	Int64
8	KI67	325	Int64
9	Combination_Therapy Binary Code	325	Category

Table 3 shows a novel dependent variable that integrates therapies for breast cancer patients. Thus, a mixture of five-digit categories reflecting various types of protocols would make up the predicted result.

3.3. Modiling

After the data preprocessing stage, the modeling stage follows. In this stage, various machine learning algorithms are trained on the preprocessed data to predict the classes or values of the target variable. The modeling stage involves selecting an appropriate algorithm, defining the model architecture, and training the model on the preprocessed data [28].

During the training process, the model is presented with input data, and it adjusts its parameters to minimize the difference between its predictions and the true labels of the data. Once the model is trained, it can be used to make predictions on new, unseen data. We applied the following machine learning techniques in this study: (NB), (DT), (RF) and (ANN). These are the most often used algorithms for classification problems of this kind. In this study, individual medication combinations that minimize the risk of metastatic relapse in patients with early-stage breast cancer were categorized using these classification models based on their initial clinical and demographic data. We employed Python scikit-learn to examine the data. Fig. 2 illustrates the whole course of the experiment.

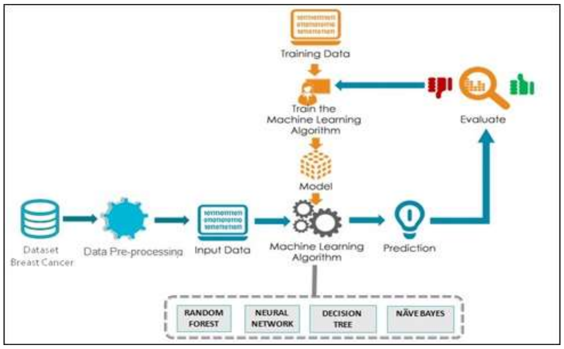


Fig.2.Model Machine Learning Use

We used a 10-fold cross-validation test, a technique used to assess the performance of predictive models. This involves dividing the original dataset into ten equal-sized subsets. The model is then trained on nine subgroups and tested on the rest of the subgroup, this process being repeated 10 times, so that each subset is used once as a test set[29].

Using this method, the effectiveness and efficiency of the model can be assessed by examining its performance on each of ten test sets. The results from each fold can then be averaged to provide an overall estimate of model performance.

3.3.1. Machine Learning Algorithms

3.3.1.1. Decision tree*:* is a machine learning algorithm used for classification and regression tasks. It is a hierarchical structure that represents a sequence of decisions and their possible consequences[30].

In a decision tree, each internal node represents a decision based on a feature value, and each leaf node represents a class label or a numerical value. The goal of building a decision tree is to create a model that can predict the value of a target variable based on several input variables[31].

The decision tree algorithm works by recursively splitting the data into subsets based on the value of a selected feature. The feature with the highest information gain is selected as the splitting criterion at each node. Information gain is a measure of the reduction in entropy or impurity of the data after splitting.

3.3.1.2. Naïve Bayes (NB*):* is a probabilistic machine learning algorithm used for classification tasks. It is based on Bayes' theorem, which describes the probability of an event given some prior knowledge. The algorithm is considered "naïve" because it makes the simplifying assumption that all features are independent of one another, which is not always the case in real-world datasets[32].

The basic idea behind the Naïve Bayes algorithm is to calculate the probability of each class given a set of input features. It does this by first calculating the prior probability of each class, which is the proportion of examples in the training data that belong to each class. Then, for each input feature, it calculates the likelihood of that feature given each class. Finally, it combines the prior probabilities and likelihoods to calculate the posterior probability of each class given the input feature [32].

3.3.1.3.Artificial neural network (ANN): are a type of machine learning model that is inspired by the structure and function of the human brain. ANNs consist of a large number of interconnected nodes or "neurons," which work together to process and analyze complex data.

Each neuron receives input from other neurons, processes this input, and then produces an output signal. The connections between neurons can be adjusted or "trained" through a process called backpropagation, which involves adjusting the weights of the connections between neurons to improve the accuracy of the model's predictions [33].

3.3.1.4. Random forest (RF): is a machine learning algorithm that is used for both classification and regression tasks. It belongs to the family of ensemble learning methods, which means it combines multiple models to improve the accuracy of predictions.

The algorithm works by building a large number of decision trees (known as the forest) and combining their results to make a final prediction. Each decision tree is constructed using a random subset of the training data and a random subset of the input features. This helps to reduce overfitting and improve the generalization ability of the model [34].

When making a prediction, each decision tree in the forest independently produces a prediction, and the final prediction is obtained by averaging or taking the majority vote of all the predictions from the trees. This approach tends to reduce the variance of the predictions, which helps to improve the accuracy of the model.

In the paragraph that follows, we will compare how well each classifier performs in terms of accuracy, speed at which the model may be built, and the proportion of cases that are successfully and wrongly categorized.

3.4. Performance indicators

It is important to evaluate the performance of a machine learning model in order to understand its effectiveness and identify areas where it could be improved[35].

In this study we used the performance indicators (Accuracy, Recall, F-measure, AUC-ROC, Confusion matrix) to evaluate the performance of our machine learning model and make informed decisions to improve its accuracy and predictive ability[19].

3.4.1. Confusion Matrix

A confusion matrix is a performance evaluation metric used in machine learning and statistics to evaluate the accuracy of a classification algorithm. It is a table that summarizes the performance of a classification algorithm by comparing the predicted labels with the true labels.

For a multi-class classification problem, the confusion matrix is a square matrix that shows the counts of true positives, true negatives, false positives, and false negatives for each class. The rows of the matrix represent the true classes, and the columns represent the predicted classes. Fig.3 shows the confusion matrix for a multi-label model [28].

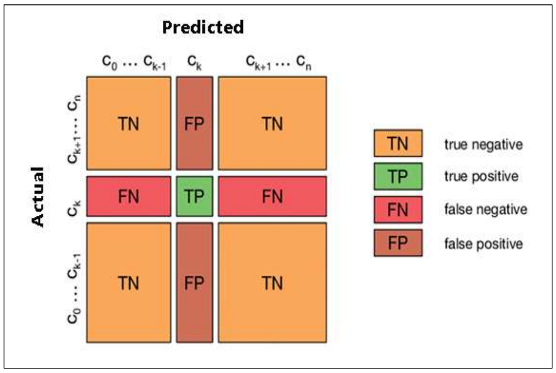


Fig.3. Multi-Label Classification -Confusion Matrix

4.2. Classification report:

A categorization report is a tool used to evaluate how well a categorization model performs. It provides a summary of the model's performance by calculating various metrics such as precision, recall, F1 score, and accuracy. The classification report typically includes the following information:

Overall accuracy is a measure of the proportion of correctly classified instances in a dataset. It is commonly used as a performance metric for classification models.

Mathematically (1), overall accuracy is computed by dividing the total number of cases in the dataset by the instances that were properly classified:

$$OverallAccuracy = \frac{\sum_{i=1}^{N} c_{i,i}}{\sum_{i=1}^{N}\sum_{j=1}^{N} c_{i,j}} \quad (1)$$

Precision (2) measures the proportion of true positives (correctly identified instances of a class) among all instances classified as that class. A high precision score means that the model's positive predictions are mostly correct.

$$Precision_{class} = \frac{TP_{class}}{TP_{class} + FP_{class}} \quad (2)$$

Equation defines the specificity of the actual negative rate (3). The percentage of negative data points out of all negative data points that are mistakenly interpreted as negative is known as the false positive rate.

$$Specificity\ _{class} = \frac{TN_{class}}{FP_{class} + TN_{class}} (3)$$

The Recall (Sensitivity) measures the proportion of true positives that are correctly identified by the model among all instances that belong to that class. A high recall score means that the model can correctly identify most instances of a class.

$$Recall\ _{class} = \frac{TP_{class}}{TP_{class} + FN_{class}} (4)$$

Sensitivity and specificity are two measures used to evaluate the performance of diagnostic tests, as they are important for diagnostic tests because a test with high sensitivity but low specificity can generate many false positive results, while a testing with high specificity but low sensitivity can generate many false negative results. Therefore, a good diagnostic test must have both high sensitivity and specificity.

The F1 score is the harmonic mean of precision and recall, and it provides a balance between precision and recall. A high F1 score means that the model can balance both precision and recall well [36].

$$F\,1 - Score = \frac{2 * TP_{class}}{2 * TP_{class} + FN_{class} + FP_{class}} (5)$$

The ROC curve and AUC are useful for comparing different models and selecting the best one for a particular application. They can also help identify the optimal classification threshold for a given problem, based on the trade-off between TPR and FPR. [37].

5. RESULTS AND DISCUSSION

4.1. Analysis of Result

By using classification techniques and a confusion matrix, this study assesses the simulation model's quality, in order to improve performance of (ROC surface and accuracy) without compromising accuracy, we used the variables (age, (TS), postmenopausal, histological grade, (Ki67), number of involved axillary lymph nodes and type of surgery). The characteristics HER2, (PR and ER) are also included in our dataset from the breast cancer registry. The outcomes of various classification methods are displayed in Table 4 below, with all results being 10-fold cross-validation results and each being the method's ideal result (A-B-C-D).

Table 4. Confusion Matrix Of The Target Variable For Each Classifier

A: Confusion matrix of target variable for each combination of categorical values by Random Forest

Random Forest - Classifier		Predicted				
		01100	10010	10100	11011	∑
Current	01100	48	5	11	1	65
	10010	5	53	1	1	60
	10100	13	7	47	17	84
	11011	0	2	12	102	116
∑		66	67	71	121	325

B: Confusion matrix of target variable for each combination of categorical values by Naïve Bayes

Naive Bayes - Classifier		Predicted				
		01100	10010	10100	11011	∑
Current	01100	33	14	17	1	65
	10010	3	48	6	3	60
	10100	14	9	41	20	84
	11011	0	6	16	94	116
∑		50	77	80	118	325

C: Confusion matrix of target variable for each combination of categorical values by ANN

ANN - Classifier		Predicted				
		01100	10010	10100	11011	∑
Current	01100	46	9	9	1	65
	10010	9	36	6	9	60
	10100	12	5	47	20	84
	11011	0	1	15	100	116
∑		67	51	77	130	325

D: Confusion matrix of target variable for each combination of categorical values by Decision Tree

DecisionTree -Classifier		Predicted				
		01100	10010	10100	11011	∑
Current	01100	46	3	15	1	65
	10010	4	53	3	0	60
	10100	18	8	45	13	84
	11011	3	3	14	96	116
∑		71	67	77	110	325

The findings demonstrated that the Random Forest classifier was more accurate in identifying promising therapeutic combinations. The therapy combination 10100 had the largest number of false positive predictions (17), which was then followed by Decision Tree, Neural Network, and Naive Bayes. On the other side, we see that the Naive Bayes model has the most erroneous predictions overall, at 109.

Furthermore, the therapy combination codes (11011 - 10010 - 01100- 10100) for which Random Forest achieved the highest AUC are important because they represent the therapy regimens that are most effective in reducing the risk of recurrence. By using Random Forest to predict the combination of adjuvant therapies, clinicians can select the most effective regimen for each patient, which can lead to improved patient outcomes.

4.2.Performance Evaluation

To compare the effectiveness of the four methods, classification metrics were computed. According to Table 5, the Random Forest algorithm produced the greatest outcomes in terms of accuracy (76.9%), specificity (76.3%), sensitivity (76.9%), and f1 measure (76.5%). When AUC was taken into account, Random Forest likewise obtained the best specificity (92.7%).

Table 5. Analyzing The Four Machine Learning Algorithms That Are Used

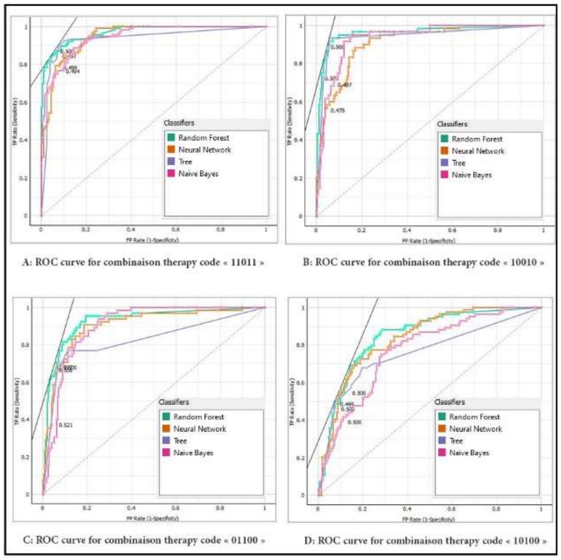
Model	AUC	CA	F1	Precision	Recall
Random Forest	0.927	0.769	0.765	0.763	0.769
Decision Tree	0.867	0.738	0.737	0.738	0.738
Neural Network	0.902	0.705	0.700	0.700	0.705
Naive Bayes	0.888	0.665	0.660	0.664	0.665

Roc and AUC curve:

The study found that all machine learning classifiers achieved an accuracy level of more than 66% for classifying the combination of therapies for breast cancer patients, indicating excellent performance in predicting therapy combinations. The ROC curve, which is based on the true positive rate (TPR) and false positive rate (FPR), is an important measure of the classification results. The ROC curve helps to evaluate the performance of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

classifiers and determine which one achieves the highest AUC (area under the curve) for ROC. In this study, the Random Forest classifier achieved the highest AUC for ROC in the therapy combination codes (11011 - 10010 - 01100- 10100) see Fig 4. This suggests that the Random Forest classifier may be the best option for predicting the combination of therapies for breast cancer patients.



A: ROC curve for combinaison therapy code « 11011 »

B: ROC curve for combinaison therapy code « 10010 »

C: ROC curve for combinaison therapy code « 01100 »

D: ROC curve for combinaison therapy code « 10100 »

Fig.4. Graphic Representation Of The ROC Curve Of Four Variables Predicted By The Classifications Used In This Study

We observe in Fig.4 that the ROC and AUC curves showing similar patterns in the upper left corner suggests that the classifiers are performing well at identifying the positive cases (i.e., the therapeutic combination codes) with high accuracy. The order of the codes mentioned in the ROC curves (11011, 10010, 01100, and 10100) may reflect their prevalence or frequency in the test dataset, with 11011 being the most common and 10100 being the least common.

Overall, it seems that the classifiers are able to correctly predict the therapeutic combination codes with good accuracy, as indicated by the high AUC values and the patterns observed in the ROC curves.

This work has certain limitations, despite the fact that machine learning has shown robust results in predicting a variety of effective therapy for breast cancer. Due to lack of information, some patients were excluded, which may cause selection bias. In addition, due to the retrospective data, our study was not able to more accurately predict the ideal adjuvant therapy group for some postoperative breast cancer subgroups, such as those with breast cancer associated with other malignancies, this may have limited the applicability of the study results. Future research on this topic needs to include more studies.

Computer simulation based on machine learning can be a valuable tool for breast cancer treatment decision-making. This method can help customize treatments for individual patients based on their characteristics and response to treatment. Additionally, the use of computer simulation can help optimize treatment outcomes while minimizing side effects. However, it's important to note that these techniques should be used in conjunction with clinical expertise and patient-specific information to ensure the best possible treatment outcomes. Additionally, further research may be needed to validate the findings of this study and determine the applicability of machine learning techniques in clinical practice.

6. CONCLUSION

The proposed computer simulation model in this study is based on machine learning, which can analyze large amounts of data and generate predictions based on patterns and trends in the data. The study compared four different machine learning algorithms to determine which one was most effective in predicting optimal treatments for breast cancer patients. The algorithms compared were Random Forest, Decision Tree, Artificial Neural Network, and Naive Bayes.

The results of the study showed that the Random Forest algorithm had the highest accuracy and the lowest error rate in predicting adjuvant therapy treatment protocols. This means that the Random Forest algorithm was the most effective at identifying the best treatment plan for breast cancer patients based on the patient's individual characteristics and medical history.

Overall, Computer simulation based on machine learning is a promising technique for the optimization of chemotherapy for breast cancer, as well as for other applications in medicine and health. By using machine learning algorithms to analyze patient data, healthcare professionals can deliver personalized and effective care, improve research into new treatments, accelerate the development of targeted therapies, improve the quality of clinical trials, and contribute precision medicine and the prevention and early detection of breast cancer.

REFERENCES

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, Art. no. 2, Mar. 2011, doi: 10.3322/caac.20107.

[2] 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.

[3] A. M. Brewster *et al.*, "Residual Risk of Breast Cancer Recurrence 5 Years After Adjuvant Therapy," *JNCI: Journal of the National Cancer Institute*, vol. 100, no. 16, Art. no. 16, Aug. 2008, doi: 10.1093/jnci/djn233.

[4] J. Beaudet, Québec (Province), and Direction de la lutte contre cancer, *Guide d'utilisation du trastuzumab (Herceptin MC) dans le traitement adjuvant du cancer du sein: mise à jour*. 2008. Accessed: Sep. 17, 2021. [Online]. Available: https://www.deslibris.ca/ID/220526

[5] A. Sarradon-Eck and I. Pellegrini, "Le traitement adjuvant du cancer du sein par tamoxifène: Entre risques et bénéfices thérapeutiques," *Sciences sociales et santé*, vol. 30, no. 1, p. 47, 2012, doi: 10.3917/sss.301.0047.

[6] M. Mimouni, W. Chaouki, H. Errihani, and N. Benjaafar, "Analyse des délais de traitement du cancer du sein : expérience d'un centre de référence tertiaire au Maroc," *Bulletin du Cancer*, vol. 105, no. 9, pp. 755–762, Sep. 2018, doi: 10.1016/j.bulcan.2018.05.010.

[7] A. Prodan, L. Liyanage, and J. A. Ginige, "Exploring Cannulation Process in Chemotherapy through a Computer Simulation".

[8] X. Lai *et al.*, "Toward Personalized Computer Simulation of Breast Cancer Treatment: A Multiscale Pharmacokinetic and Pharmacodynamic Model Informed by Multitype Patient Data," *Cancer Research*, vol. 79, no. 16, pp. 4293–4304, Aug. 2019, doi: 10.1158/0008-5472.CAN-18-1804.

[9] G. Dinstag *et al.*, "Clinically oriented prediction of patient response to targeted and immunotherapies from the tumor transcriptome".

[10] K. Chakradeo, S. Vyawahare, and P. Pawar, "Breast Cancer Recurrence Prediction using Machine Learning," in *2019 IEEE Conference on Information and Communication Technology*, Allahabad, India: IEEE, Dec. 2019, pp. 1–7. doi: 10.1109/CICT48419.2019.9066248.

[11] E. Merouane, A. Said, and E. F. Nour-eddine, "Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers," *IJACSA*, vol. 13, no. 2, 2022, doi: 10.14569/IJACSA.2022.0130222.

[12] M. Aa and A. Um, "Application of Machine Learning Techniques in Predicting of Breast Cancer Metastases Using Decision Tree Algorithm, in Sokoto Northwestern Nigeria," p. 5.

[13] M. Ertel, S. Azeddine, A. Said, and E. F. Nour-eddine, "PREDICTION OF THE MOST EFFECTIVE ADJUVANT THERAPEUTIC COMBINATIONS FOR BREAST CANCER PATIENTS USING MULTINOMIAL CLASSIFICATION," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 23, Dec. 2022, [Online]. Available: http://www.jatit.org/volumes/onehundred23.php

[14] M. Chakkouch, M. Ertel, A. Mengad, and S. Amali, "A Comparative Study of Machine Learning Techniques to Predict Types of Breast Cancer Recurrence," *IJACSA*, vol. 14, no. 5, 2023, doi: 10.14569/IJACSA.2023.0140531.

[15] S. P. Somashekhar *et al.*, "Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board," *Annals of Oncology*, vol. 29, no. 2, pp. 418–423, Feb. 2018, doi: 10.1093/annonc/mdx781.

[16] G. D. Farmer, M. Pearson, W. J. Skylark, A. L. J. Freeman, and D. J. Spiegelhalter, "Redevelopment of the Predict: Breast Cancer website and recommendations for developing interfaces to support decision-making," *Cancer Med*, vol. 10, no. 15, pp. 5141–5153, Aug. 2021, doi: 10.1002/cam4.4072.

[17] A. M. Alaa, D. Gurdasani, A. L. Harris, J. Rashbass, and M. van der Schaar, "Machine learning to guide the use of adjuvant therapies for breast cancer," *Nat Mach Intell*, vol. 3, no. 8, pp. 716–726, Aug. 2021, doi: 10.1038/s42256-021-00353-8.

[18] M. A. Naji, S. E. Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms

For Breast Cancer Prediction And Diagnosis," *Procedia Computer Science*, vol. 191, pp. 487–492, 2021, doi: 10.1016/j.procs.2021.07.062.

[19] M. Ertel and S. Amali, "'Artificial Intelligence (AI) in oncology: Predicting Treatment Response in Women with Breast Cancer'. (2021). 1st International Meeting on science at the service of health in the context of public-private partnership. May 27 - 28, Meknes, Morocco.," 2021.

[20] M. A.-E. Zeid, K. El-Bahnasy, and S. E. Abu-Youssef, "AN EFFICIENT OPTIMIZED FRAMEWORK FOR ANALYZING THE PERFORMANCE OF BREAST CANCER USING MACHINE LEARNING ALGORITHMS," . *Vol.*, no. 14, 2022.

[21] T. E. Mathew, "AN OPTIMIZED EXTREMELY RANDOMIZED TREE MODEL FOR BREAST CANCER CLASSIFICATION," . *Vol.*, no. 16, 2022.

[22] N. M. Swelam, A. E. Khedr, and H. Auda, "BREAST CANCER DIAGNOSIS AND PROGNOSIS USING STACKING ENSEMBLE TECHNIQUE," . *Vol.*, no. 14, 2022.

[23] S.-J. Sammut *et al.*, "Multi-omic machine learning predictor of breast cancer therapy response," *Nature*, vol. 601, no. 7894, pp. 623–629, Jan. 2022, doi: 10.1038/s41586-021-04278-5.

[24] F. J. Candido Dos Reis *et al.*, "An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation," *Breast Cancer Res*, vol. 19, no. 1, p. 58, Dec. 2017, doi: 10.1186/s13058-017-0852-3.

[25] J. M. Ji and W. H. Shen, "A Novel Machine Learning Systematic Framework and Web Tool for Breast Cancer Survival Rate Assessment," Oncology, preprint, Sep. 2022. doi: 10.1101/2022.09.16.22280052.

[26] S. N. de Oliveira, A. Massaroli, J. G. Martini, and J. Rodrigues, "From theory to practice, operating the clinical simulation in Nursing teaching," *Rev. Bras. Enferm.*, vol. 71, no. suppl 4, pp. 1791–1798, 2018, doi: 10.1590/0034-7167-2017-0180.

[27] P. Cerda, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *Mach Learn*, vol. 107, no. 8–10, pp. 1477–1494, Sep. 2018, doi: 10.1007/s10994-018-5724-2.

[28] H. Akhmouch, H. Bouanani, G. Dias, and J. G. Moreno, "Stratégie Multitâche pour la Classification Multiclasse," p. 10.

[29] G. C. Wishart *et al.*, "PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2," *Br J Cancer*, vol. 107, no. 5, pp. 800–807, Aug. 2012, doi: 10.1038/bjc.2012.338.

[30] S. Marne, S. Churi, and M. Marne, "Predicting Breast Cancer using effective Classification with Decision Tree and K Means Clustering technique," in *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India: IEEE, Mar. 2020, pp. 39–42. doi: 10.1109/ESCI48226.2020.9167544.

[31] M. Takada *et al.*, "Prediction of axillary lymph node metastasis in primary breast cancer patients using a decision tree-based model," *BMC Med Inform Decis Mak*, vol. 12, no. 1, p. 54, Dec. 2012, doi: 10.1186/1472-6947-12-54.

[32] A. Jamain and D. J. Hand, "The Naive Bayes Mystery: A classification detective story," *Pattern Recognition Letters*, vol. 26, no. 11, pp. 1752–1760, Aug. 2005, doi: 10.1016/j.patrec.2005.02.001.

[33] R. Daoudi-Dabladji, "Classification du cancer du sein par des approches basées sur les systèmes immunitaires artificiels," p. 150.

[34] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[35] B. Alain, D. Marcel, F. P. Emmanuel, and R. Caroline, "107 Development of quality indicators for lung cancer surgery from the national database EPITHOR," *BMJ Quality & Safety*, vol. 19, no. Suppl 1, pp. A58–A59, Apr. 2010, doi: 10.1136/qshc.2010.041624.12.

[36] L. Sazonova, G. Osipov, and M. Godovnikov, "Intelligent system for fish stock prediction and allowable catch evaluation," *Environmental Modelling & Software*, vol. 14, no. 5, pp. 391–399, Mar. 1999, doi: 10.1016/S1364-8152(98)00100-5.

[37] N. Massart *et al.*, "Correction to: Characteristics and prognosis of bloodstream infection in patients with COVID-19 admitted in the ICU: an ancillary study of the COVID-ICU study," *Ann. Intensive Care*, vol. 12, no. 1, p. 4, Dec. 2022, doi: 10.1186/s13613-022-00979-w.