STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **12**, May 2024, pp 617–629.
Published online in International Academic Press (www.IAPress.org)

A hybrid sampling combining Smote and Random Forest algorithm for cancer chemotherapy regimen Classification: A Case of Moroccan breast cancer patients

Houda AIT BRAHIM*, Oumayma BANOUAR, Salah EL-HADAJ, Abdelmoutalib METRANE

Laboratory of Computer and Systems Engineering, Faculty of Science and Technology, Cadi Ayad University, Marrakech, Morocco

Abstract Breast cancer causes the highest number of deaths weekly. It is the most common type of cancer and the leading cause of death in women in the world. Deep learning and machine learning are a way to predict perfect therapeutic protocols for patients. This paper presents different models to perform the prediction of breast cancer chemotherapy protocol and the number of cycles of cure using deep learning and machine learning algorithms. The database of therapeutic protocols related to 600 patients with breast cancer pathology was constructed at the University Hospital Center of Marrakesh, Morocco, it was treated following three distinct procedures before being used. In the first procedure, even though it is imbalanced, the database was left as it was. In the second procedure, the database was divided into three equal classes of 200 samples (Three balanced classes). As for the third, a database augmentation was realized, so that the minority classes have the same number of samples as the majority classes to have a balanced database. The main objective is to determine with our restricted database that we have collected since 2018, the best algorithm for the prediction of the therapeutic protocol and number of cycles, in accuracy terms, balanced accuracy, time, and precision. The results show that machine learning algorithms, especially Random Forest and XGBoost gave the best scores when the database was amplified.

Keywords Machine learning, Deep learning, Breast cancer, database Augmentation, Therapeutic protocol

DOI: 10.19139/soic-2310-5070-1941

1. Introduction

Precision medicine is a new approach to health care that takes into account individual differences in genes, environment, and lifestyle. This approach is in contrast to a one-size-fits-all approach that does not take into account differences between individuals. Precision medicine has the potential to improve healthcare by allowing doctors and researchers to more accurately predict which treatments and prevention strategies will be effective for specific diseases in specific groups of people. Other tests look for genetic changes or variants (sometimes called mutations) in cancer cells. This will help determine whether treatment is needed and which treatment will be most effective. For example, breast tumor cells can be tested to see if they produce large amounts of a protein called HER2. Patients with HER2-positive breast cancer are more likely to respond to drugs that target this protein. Some genetic tests can reveal whether the body activates (activates) certain drugs, which can help determine the best treatment. In 2021, the cancer treatment market has reached 185 billion USD, a rise of 12% compared to 2020. It could more likely reach 270 billion USD in 2025, and 307 billion USD (+66 %) in 2026 [1]. Between 2010 and 2020, the cost of care for cancer in the US increased by 27%, due to an aging population and increasingly complex therapies. However, alongside this rapid and wide growth, persistent delays are being reported in chemotherapy [1].
According to the National Medicines Safety Agency, ruptures or risks of ruptures concerned 2,160 drug references in 2021, compared to 871 in 2018. In 2022, the phenomenon accelerated, since in mid-August, supply disruptions

*Correspondence to: Houda AIT BRAHIM (Email: houda.aitbrahim@ced.uca.ma)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

concerned 12.5% of references, compared to 6.5% in January [2].
In the oncology field, the data related to patients with breast cancer is too limited and imbalanced, which affects the accuracy of the prediction of therapeutic protocols. This difficulty can cause a disturbance in the supply of chemotherapeutic molecules and consequently cause the hospital department to lose lots of money.
This article describes a scientific experiment that compares and measures the effectiveness of data augmentation on machine learning (ML) and deep learning (DL) algorithms in predicting treatment protocols for breast cancer patients.
The remainder of this article is as follows, section 2 presents the background and motivation, section 3 dresses the related works, section 4 discusses the materials and methods, and the last section is the conclusion and future work which provides a synthesis and an opening to future work.

2. Background and motivation

Breast cancer is the most common cancer in women, accounting for one in eight cancer diagnoses worldwide [3]. In 2020, 2.3 million new infections were reported, particularly in countries in transition [4]. In terms of economic costs, breast cancer is one of the most expensive cancers in the world, at $4.4449 trillion according to the OECD [5]. Another study using the FIGO staging system [6] showed that the cost of this disease increases with the advancing stage. Cumulative treatment costs in 2015 were $29,724 for Stage I, $39,322 for Stage II, $57,827 for Stage III, and $62,108 for Stage IV. In percentage terms, these treatments were 32%, 95%, and 109% higher in stages II, III, and IV than in stage I. This shows how important early diagnosis is. Other than that, the consequences of this disease can affect both the psychological and physical health of the patients. Many organs can be affected such as the lungs and liver. Sometimes, the cancer can spread through the bloodstream to other parts of the body. As for mortality, on a global scale, 685,000 deaths occurred because of breast cancer in 2020, with the largest numbers in transitioning countries [4]. To minimize these effects and reduce the social and economic burden of breast cancer, many techniques have been developed to allow better diagnosis and implement more appropriate treatment protocols. For example, radiotherapy is often used to predict tumor response [7]. Studies have also shown that gene expression profiling can be developed into a diagnostic tool for pathological complete response (PCR) to neoadjuvant chemotherapy (NAC) in breast cancer [8]. Artificial intelligence (AI) also plays a role in this area. It is a broad field of computer science that builds intelligent machines that can perform tasks that typically require human intelligence [9].ML and DL are the two most important concepts for realizing AI.
Breast cancer detection and classification can be performed using various ML and DL technologies [10]. For example, diseases can be classified using multimodal medical imaging. In this study, we divided ML into five techniques: Support Vector Machine (VSM), Decision Tree (DT), Nearest Neighbor, Naive Bayesian Network, and ANN. Additionally, some ML methods using quantitative MRI data can be used to predict the final treatment response of breast cancer patients after one cycle of neoadjuvant chemotherapy (NAC). By the time a significant number of patients are found to be unresponsive to treatment protocols, consider that the disease may no longer be surgically respectable. Therefore, it is important to develop techniques to predict tumor response early in treatment planning [11]. Deep Learning technology is also used in predicting breast cancer response to chemotherapy across multiple datasets.
First, the tumor volume is reliably segmented by deep neural networks, and then the response is automatically predicted thanks to multiple databases provided by several international organizations [12]. Additionally, ML can also be used to predict breast cancer prognosis and metastatic outcomes, even in complex structures with a large number of variables [13]. Various comparisons between four algorithms were considered to analyze medical data. Early studies compared SVM, logistic regression, random forests, and K-NN (K nearest neighbors) using different datasets. The author [14] (2020) COMPARED SVM, Naive Bayes (NB), k-NN, and DT models to perform predictions on the Wisconsin breast cancer dataset [15]. As a result, SVM outperforms all other algorithms in terms of efficiency and effectiveness based on accuracy, precision, sensitivity, and specificity, proving its efficiency in predicting and diagnosing breast cancer.
Although ML and DL approaches have demonstrated the ability to detect and classify tumors and respond to treatment, few studies have demonstrated their use in predicting treatment protocols for breast cancer. Predicting the

treatment outcome is of great importance to prevent treatment complications and side effects, allowing physicians to choose the optimal treatment protocol according to the diagnosis and patient [13].

3. Related works

Recently, attention has been increasing on using machine learning and deep learning algorithms to predict the best chemotherapy regimens for patients suffering from breast cancer. Machine learning and deep learning algorithms can analyze a patient's medical history and make better predictions about the best chemotherapy regimens. This can help reduce the risk of treatment failure and improve the overall effectiveness of the treatment.
Unfortunately, there are still relatively few studies that examine the use of machine learning and deep learning algorithms to predict the best chemotherapy regimens for breast cancer patients. This is because these algorithms require a large amount of real data to be effective. Additionally, developing these algorithms is a costly and time-consuming process.
Some studies did run some protocols to investigate the best models (Machine learning, deep learning) for predicting drug response of the different chemotherapy regimens on breast cancer patients.
This paper [16] presents a novel approach to predict the therapeutic responses of cancer patients to anti-cancer drugs using Deep Learning Neural Networks (DLNNs). Through the combination of Association Rule Mining and DLNNs, a large data set of molecular profiles from 1001 cancer cell lines was used to generate cancer-specific signatures, which were then used to predict pharmacological responses to a wide variety of anti-cancer drugs. The proposed algorithm was able to outperform existing state-of-the-art drug-response prediction methods. Furthermore, this work introduced a strategy for identifying potential therapeutic targets and drug combinations with high therapeutic potential. The successful application of DLNNs to predict therapeutic responsiveness demonstrated a major milestone in the field of precision medicine.
Authors in [17] presented BRISK, a deep learning model for predicting malignancy in patients with BI-RADS 4 mammograms. The BRISK model combines ultrasound images, mammography reports, patient demographic variables, and pathology results to accurately predict patient outcomes. They evaluated the model using a dataset of 5,147 patients from Houston Methodist Hospital, and the results showed that the BRISK model had 100% sensitivity and 74% specificity, with a total accuracy of 81%. They compared Their model to seven other machine learning methods and found that BRISK had a higher area under the curve than all of the other methods.
A study in [18] showed that machine learning algorithms trained on patient-reported preoperative and clinical data can be used to accurately predict economic toxicities associated with breast cancer treatment. The algorithm was developed and tested using data from her 611 patients at the University of Texas MD Anderson Cancer Center and had an overall accuracy of 82%, sensitivity of 85%, and specificity of 81%. Receipt of neoadjuvant therapy, use of autologous breast reconstruction, and low credit score were identified as important clinical factors associated with financial burden. These results suggest that machine learning algorithms have the potential to stratify patients into high-risk populations for economic toxicity.
A study in [19] proposed a mechanistic model to predict postoperative metastatic recurrence in early breast cancer patients using data available only at the time of diagnosis. The model achieved a C-index of 0.65, similar to standard predictive models and machine learning algorithms. This model is based on the biology of the metastatic process, provides insights not achievable through statistical analysis alone, and allows testing of whether covariance is associated with proliferation and/or dissemination. Additionally, this model can be used to assess the extent of invisible metastases and predict future metastatic growth to help select patients who will most benefit from adjuvant therapy.
Another study by [20] aimed to compare modeling performance and classification frameworks to predict a priori response to neoadjuvant chemotherapy (NAC) in breast cancer (BC) patients. Machine learning (ML) algorithms such as the K-Nearest Neighbors algorithm, Random Forest (RF) algorithm, Naive Bayes algorithm, Support Vector Machine, and Multi-layer Perception model are used in clinical and pathological research. They analyze and compare data from 431 patients with BC. They use a traditional multivariate logistic regression (MLR) model. The results showed that certain ML classifiers can perform better than MLR predictive modeling, as evidenced by the increase in area under the curve (AUC). The study says an AI-driven predictive model could better select

a patient's NAC options by determining the likelihood of her NAC success before she starts treatment provided evidence. Despite the limitations of the retrospective study design, this study contributed to the utilization of new technology to help balance the efficacy and toxicity control of her NAC.

4. Materials and methods

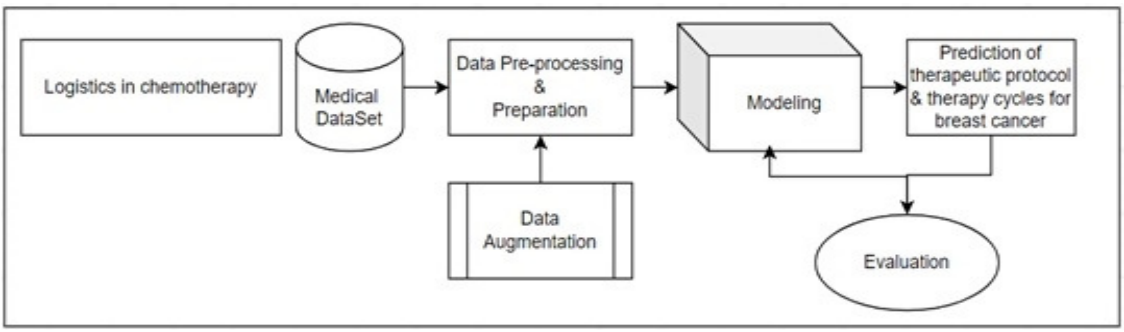

Figure 1. Materials and methods used

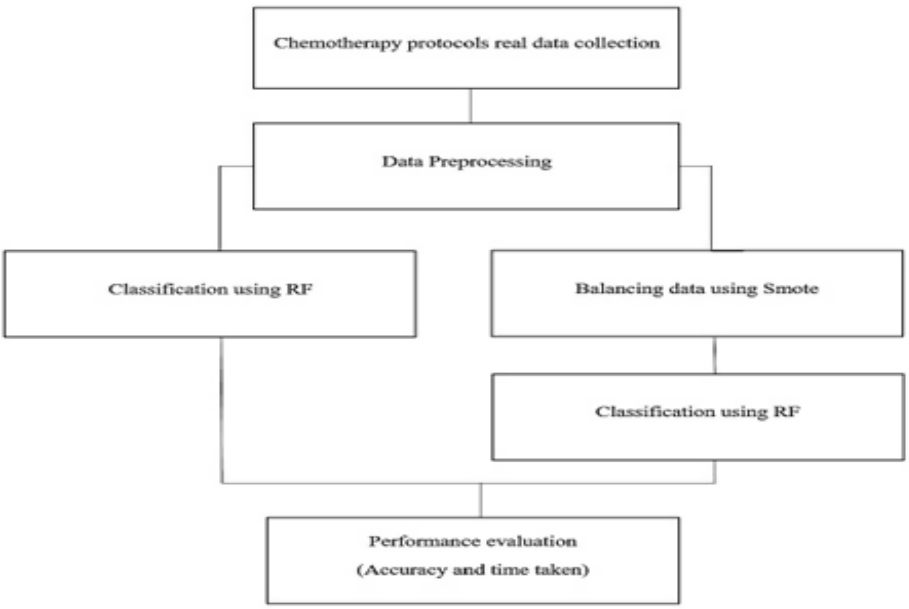

Figure 2. Proposed pipeline for prediction of therapeutic protocol and number of cycles for each patient

4.1. Logistics in chemotherapy

Anticancer agents, also called anti-neoplastic agents, are any drugs that are effective in treating malignant or cancerous conditions. There are several major classes of anticancer drugs. These include alkylating agents, anti-metabolites, natural products, and hormones. The logistics of these chemotherapeutic molecules require a prior prediction of the adequate molecule for each patient as well as the number of care cycles. Thus, the quantity of

products that hospitals obtain to create their stocks is limited and the loss of expired products is also reduced to a minimum. In order to reach this in collaboration with the University Hospital Center, Marrakesh, Morocco, our objective is to perform several models related to ML and DL, tested on a real-world database, which will also undergo a data augmentation or a data slitting because of its imbalanced nature.

4.2. Medical dataset construction

Variables obtained or derived from the database include: age, performance state, CT/pt, Cn/pn, Cm (Metastatic sites), RH, Her2, Classification, pCR, Therapeutic strategy, Protocol, Number of cycles, Therapy adjuvant anti HER2, Adjuvant post neoadjuvant, Maintenance. The data may influence or facilitate breast cancer-related information, including patient demographics, clinical and histopathology characteristics, patient treatment options, and the decision-making process of whether to initiate chemotherapy. were collected on certain determinants of In the setting of neoadjuvant, adjuvant, or palliative therapy. Treatment details were recorded for each patient (i.e. schedule, number of cycles, amount of chemical administered). Performance status (PS) of each patient was assessed and recorded on a scale of 0 to 3 on the Eastern Cooperative Oncology Group (ECOG) scale 0 represents a fully functional patient and 3 represents a bedridden patient. Tumor stage was assessed and classified according to the American Joint Committee on Cancer (AJCC) classification system [23].

4.3. Data pre-processing

The database used in this research was collected from nearly 601 patients hospitalized at the Mohamed V hospital center in Marrakesh city (Morocco) [23]. Data related to 451 patients was used to train DL algorithms, while data from the remaining 150 patients was used to test DL algorithms. The first phase involved collecting data for pre-processing and applying classification. Data pre-processing is known as a data mining technique and transforms raw data into an understandable format. Real-world data is often inconsistent or incomplete, contains many errors, and is not secure.

Data pre-processing is a proven method to solve such problems: it prepares raw data for further processing. As for our concern, we pre-processed the UCI dataset using the standardization method. This step is very important as the quality and quantity of data collected directly determines the accuracy of the predictive model.

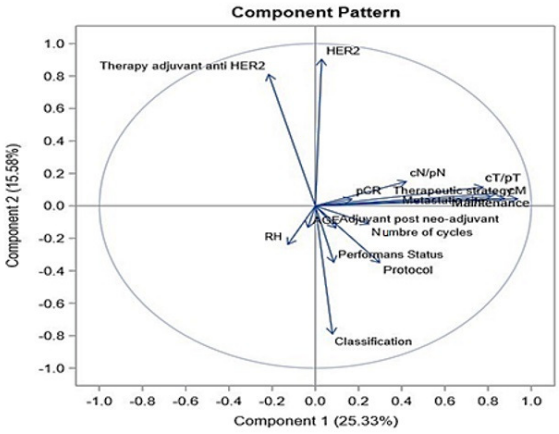

Figure 3. Correlation circle on our constructed dataset

Correlation diagrams in figure 3 show the relationships between different variables. It can be concluded that maintenance, treatment strategy, metastatic site, cM and cT/pT are the main variables that are positively correlated, while HER2 and adjuvant anti-HER2 are the main variables that are positively correlated. On the other hand, classification was the main negatively correlated variable [21]. The results of this study will be very useful for the development of machine learning models to predict breast cancer treatment in the Moroccan context [22, 23]. This

is significant in cervical analysis because it can help identify essential factors. For example, two characteristics may be positively correlated, which means that both variables increase or decrease together. Alternatively, one characteristic may be inversely correlated with another, meaning that one variable increases while the other decreases [24].

4.4. Data Augmentation on Imbalanced Datasets

The method Synthetic Minority Over-sampling Technique (SMOTE) provides synthetic samples by interpolating between the all minority classes, addresses class imbalance: SMOTE can help address the class imbalance problem by generating synthetic samples for the minority class, making the dataset more balanced, and improves model performance: By making the dataset more balanced, SMOTE can help in order to increase the performance of machine learning models, especially when the minority class is important.

Algorithm 1 SMOTE(T, N, K)

procedure SMOTE(T, N, K)
  Input: Number of minority class samples T; Amount of SMOTE $N\%$; Number of nearest neighbors k
  Output: $(N/100) \times T$ synthetic minority class samples
  if $N < 100$ **then**
    Randomize the T minority class samples
    $T = (N/100) \times T$
    $N = 100$
  end if
  $N =$ int$(N/100)$ ▷ The amount of SMOTE is assumed to be in integral multiples of 100.
  $k =$ Number of nearest neighbors
  $numattrs =$ Number of attributes
  $Sample[\,][\,]$: array for original minority class samples
  $newindex$: keeps a count of the number of synthetic samples generated, initialized to 0
  $Synthetic[\,][\,]$: array for synthetic samples
  for $i = 1$ to T **do**
    Compute k nearest neighbors for i, and save the indices in the $nnarray$
    Populate($N, i, nnarray$)
  end for
  procedure POPULATE($N, i, nnarray$) ▷ Function to generate the synthetic samples.
    while $N \neq 0$ **do**
      Choose a random number between 1 and k, call it nn.
      for $attr = 1$ to $numattrs$ **do**
        Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
        Compute: $gap =$ random number between 0 and 1
        $Synthetic[newindex][attr] = Sample[i][attr] + gap \times dif$
      end for
      $newindex++$
      $N = N - 1$
    end while
  end procedure
end procedure

SMOTE stands for Synthetic Minority Over Sampling Techniques and is used to generate synthetic samples of minority classes in imbalanced data sets. Due to the imbalance in the data set, the number of samples in the minority class will be significantly lower than the minority class, resulting in poor classification performance.
The SMOTE algorithm uses three input parameters T, N, and K. T is the number of samples of the minority class

in the original dataset, N is the proportion of synthetic samples generated, and K is the number of nearest neighbor samples used. What is generated must be a synthetic sample. If N is less than 100%, the minority class sample is randomized. This is because only a random proportion of the minority class is excluded. Calculate the number of synthetic samples generated as (N/100) * T and set N to 100. The algorithm then computes the K nearest neighbors for each sample in the minority class and stores their indices in an array called nnarray. This algorithm calls the populate function to generate synthetic samples. The Populate function takes three input parameters, N, i, and nnarray. where N is the number of synthetic samples generated, i is the index of the current minority class sample, and nnarray contains the K nearest neighbor indices of the current minority class sample. The population function generates a synthetic sample by randomly selecting one of the K neighbors of the current minority class sample and selecting the neighborhood by calculating the difference between the attribute values of the current sample.
The difference is then multiplied by a random number between 0 and 1 and added to the attribute values of the current sample to create a new synthetic sample. The algorithm iterates the padding function N times to generate N synthetic samples. Finally, the algorithm returns the original samples of the minority class and synthetic samples that are used to train the machine learning model.

4.5. Data preparation

The database was processed in three different ways (table 1). The first one is called normal, in which the database is processed entirely. the disadvantage of this method is that the database is imbalanced. The second way is called classified, in which the data is classified in three levels. Only two protocols occupy the two thirds of all protocols [23]. Thus, we chose to put them each in a separate class, and all the other protocols in a third class. Now, the database becomes balanced. As for the last one, it is said to be augmented: the data will be amplified thanks to the SMOTE algorithm.The SMOTE algorithm is a data mining technique that enables a data miner to over sample all the minority class, that can achieve potentially better performance without loss of information.

Table 1. The proposed database technics

Data Procedures	**Methods**
Procedure 1	Data treatment on 3 classes (0; 1; 2)
Procedure 2	Data treatment on 16 classes (0; 1; . . . ; 15)
Procedure 3	Increased data with SMOTE algorithm on 16 classes

Procedure 1 concerns 3 class which are:
• Class 1: AC60 + Paclitaxel weekly.
• Class 2: EC100 + Docetaxel.
• Class 3: AC60+Paclitaxel, TC, AC60+Docetaxel, OOS, AC60, Paclitaxel, Paclitaxel weekly-Trastruzumab-Pertuzumab, Capecitabine, EC100, Paclitaxel+Trastuzumab+Pertuzumab, Carboplatine-Gemcitabine, Docetaxel-trastuzumab-Pertuzumab, Paclitaxel weekly-Bevacizumab, and TCH.
Procedure 2 and 3 concern 16 class which are:
• Class 1: AC60 + Paclitaxel weekly.
• Class 2: EC100 + Docetaxel.
• Class 3: AC60+Paclitaxel.
• Class 4: TC.
• Class 5: AC60+Docetaxel.
• Class 6: AC60.
• Class 7: Paclitaxel.
• Class 8: Paclitaxel weekly-Trastruzumab-Pertuzumab.

• Class 9: Capecitabine.
• Class 10: EC100.
• Class 11: Paclitaxel+Trastuzumab+Pertuzumab.
• Class 12: Carboplatine-Gemcitabine.
• Class 13: Docetaxel-trastuzumab-Pertuzumab
• Class 14: Paclitaxel weekly-Bevacizumab.
• Class 15: TCH.

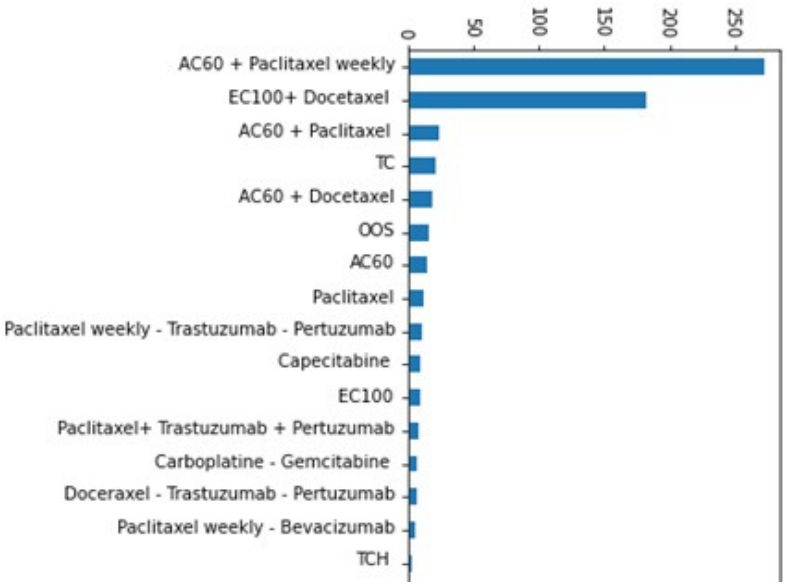

Figure 4. Number of cases treated for each therapeutic protocol

Note that our objective is to predict the adequate therapeutic protocol presented is Fig. 4 with the number of treatment cycles.

4.6. Evaluation and results

To evaluate the performance of the proposed pipeline. Precision, recall, precision, and F-measure were used. Accuracy and F-measure were used to accurately compare the classification performance. The F-measure is the harmonic mean of precision and recall and lies between 0 and 1. A value of 1 indicates perfect precision and recall (Equations 1,2,3,4,5,6 and 7).

$$\text{Accuracy} = \frac{TP + TN}{\text{All}} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Specificity (TNR)} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{Sensitivity (TPR)} = \frac{TN}{TP + FN} \tag{5}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity + Specificity}}{2} \tag{6}$$

$$\text{F1 Score} = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{7}$$

Our proposed pipeline was compared with the following state-of-the-art methods: ExtraTrees (ET), RandomForest (RF), XGBoost (XGB), LightGBM (LGBM), Bagging (BG), DecisionTree (DT), KNeighbors (KNN), LabelSpreading (LS), LabelPropagation (LP), Nu-Support Vector Classification (NuSVC), Support Vector Classification (SVC), Perceptron (P), Stochastic Gradient Descent (SGD), Ridge Classifier Cross-Validated (RidgeCV), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Ridge (R), Linear Support Vector Classification (LinearSVC), Calibrated Classifier Cross-Validated (CalibratedCV), AdaBoost (AB), Bernoulli Naive Bayes (BernoulliNB), ExtraTree (ET), Nearest Centroid (NC), Passive Aggressive (PA), Gaussian Naive Bayes (GaussianNB), Dummy, Quadratic Discriminant Analysis (QDA).

4.6.1. Normal dataset

According to the results, when we test ML and DL algorithms on raw data base, Decision Tree, XGB, Nearest Cendroid, Bagging and Random Forest give best results in accuracy terms, accuracy balanced, F1 score and time taken. These five algorithms that give best results achieved respectively score of 74%, 74%, 36%, 74%, 69%. In addition to that, in terms of F1 score, Decision Tree, XGB, Nearest Cendroid, Bagging and Random Forest achieved respectively scores of 72%, 73%, 35%, 73%, 66%. When it comes to the required time to finalise tasks, those five algorithms did give respectively scores of 0,01 second, 0.21 second, 0.01 second, 0.02 second, 0.17 second. We also tested the performance of RNN (Recurrent Neural Network), a DL algorithm chosen for its specification on tabular data, and it achieved an inferior score compared to ML algorithms with a value of 65% for accuracy (Figure 5).

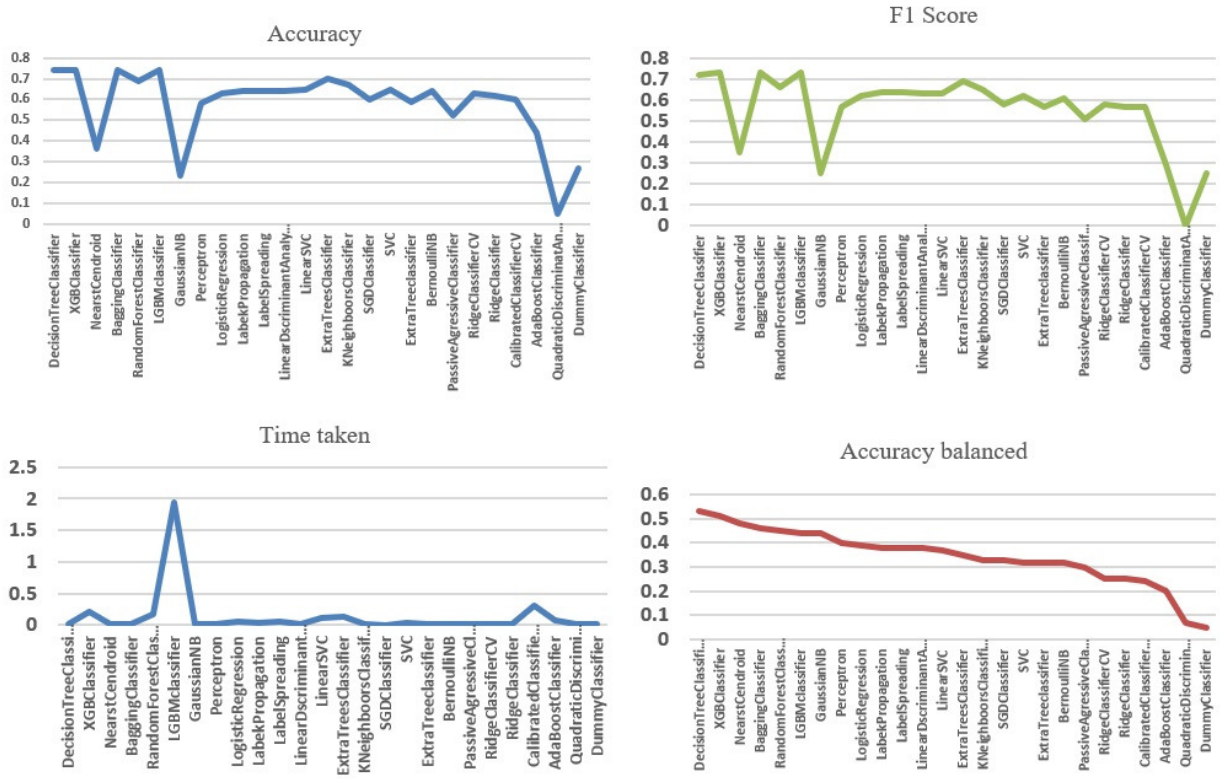

Figure 5. Results obtained on raw data

4.6.2. Splitted dataset (3 classes)

According to obtained results, it was observed that, when we test only ML algorithms on three classes classified data base, Extra Trees, Random Forest, XGB, LGBM, Bagging gives better results in terms of accuracy, accuracy balanced, F1 score and time taken. These five algorithms that give best results achieved respectively scores of 85%, 84%, 84%, 83%, 81%. In addition to that, in terms of F1 score, Extra Trees, Random Forest, XGB, LGBM, Bagging achieved respectively scores of 85%, 84%, 84%, 83%, 81%. When it comes to the time necessary to finalise tasks, those five algorithms did give respectively scores of 0,10 second, 0.11 second, 0.62 second, 0.62 second, 0.03 second (Figure 6).

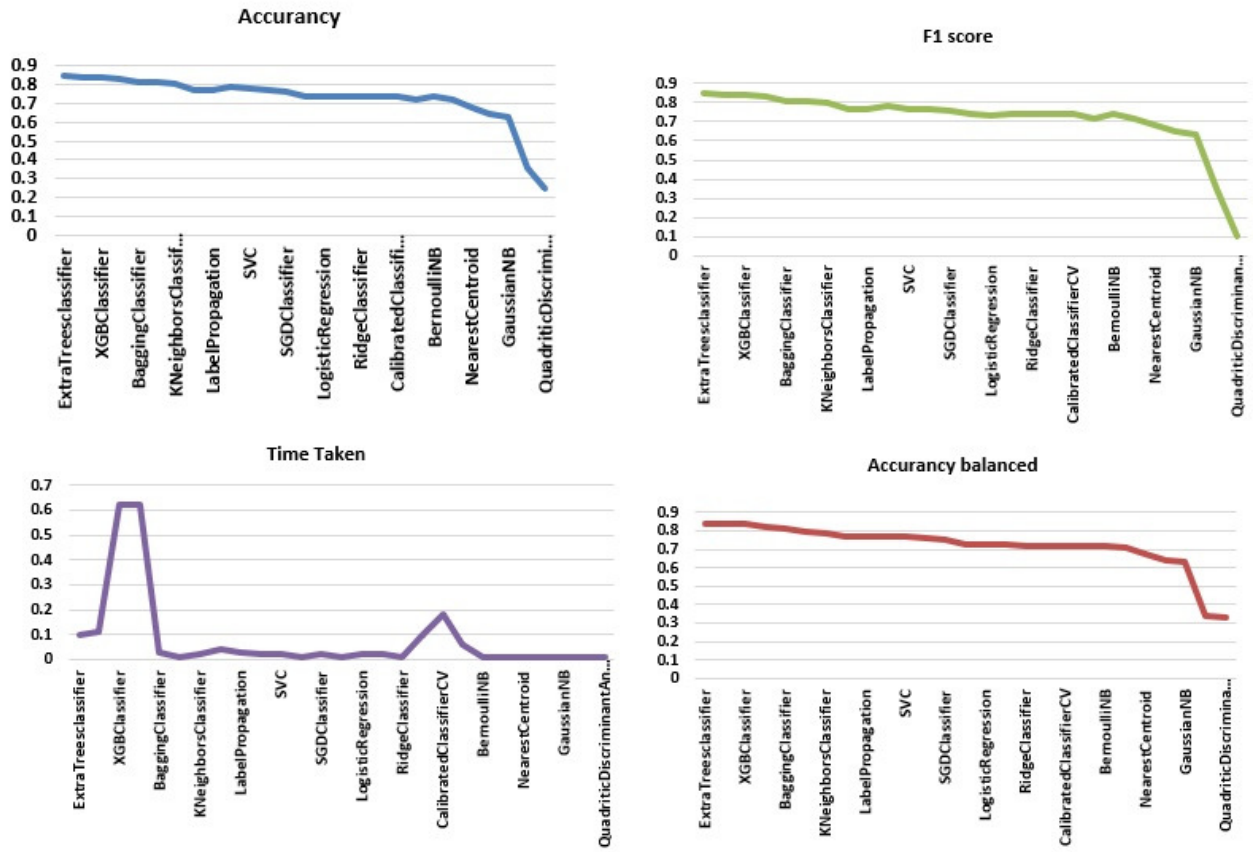

Figure 6. Results obtained on 3 classes classified data base

4.6.3. Amplified data set (SMOTE)

According to the results, it was observed that, when we test ML algorithms on amplified data base using SMOTE, Random Forest, XGB, Extra Trees, LGBM, Label Propagation give better results in terms of accuracy, accuracy balanced, F1 score and time taken. These five algorithms that give best results achieved respectively scores of 95%, 95%, 95%, 95%, 93%. In addition to that, in terms of F1 score Random Forest, XGB, Extra Trees, LGBM, Label Propagation achieved respectively scores of 95%, 95%, 95%, 95%, 93%. When it comes to the time necessary to finalise tasks, those five algorithms did give respectively scores of 0,21 second, 0.51 second, 0.21 second, 0.51 second, 0.52 second.

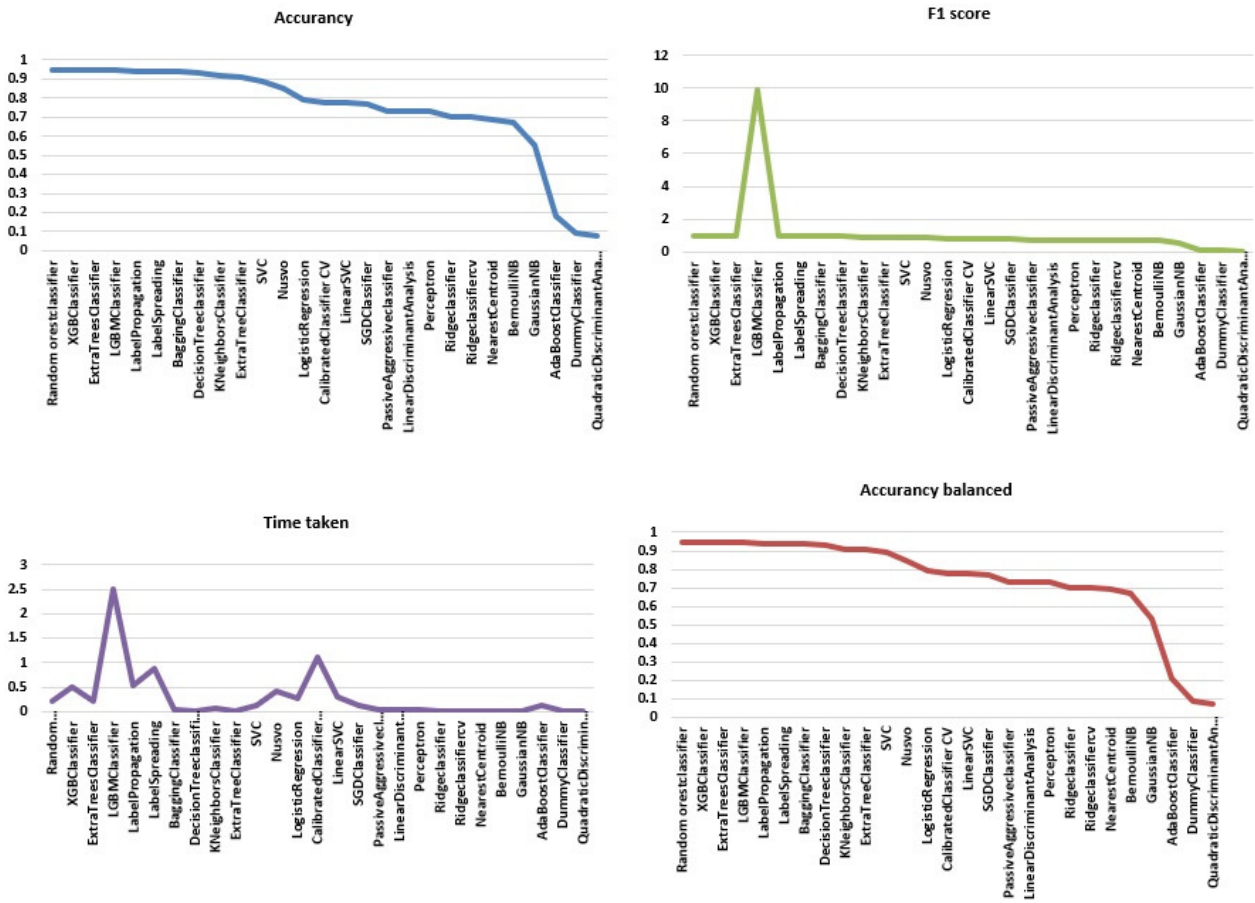

Figure 7. Results obtained on amplified data base using SMOTE

Our work involved a real database collected from the hospital services of the University Hospital Center (UHC) of Marrakesh, meaning that it is naturally restricted and imbalanced. First, we tested our algorithms on the raw database without any additional treatment. The results show that this learning program allowed relatively low precision for the 5 best algorithms (precision less than 75%). However, it was noted that the problem lies in the fact that, amongst the 601 data related to the patients with breast cancer, 2 thirds of the cases shared two similar therapeutic protocols (AC60 + paclitaxel weekly, EC100 + docetaxel), while the remaining cases underwent the remaining therapeutic protocols (15 protocols different than the 2 mentioned before). To solve the problem of imbalanced data, two possibilities have been tested to deduce the most suitable one. The first possibility was to create classes (3 classes in total) to distribute two major therapeutic protocols, namely AC60 + paclitaxel weekly and EC100 + docetaxel in two separate classes and create a third class for minor protocols (15 in total).
The second possibility was data augmentation using SMOTE which allows a data miner to create more samples from the minority class, leading potentially to better classifier performance, without losing any data. The ML algorithms Random-Forest and XGB are consistently ranked among the top 5 for the three treatments of the database, whether it was the raw data base, the splitted data base, or the augmented data base.
The database object of our study is a restricted and tabular database, so the DL algorithm does not manage to be efficient in learning in this kind of data. Eventually, we totally focused on more specified ML algorithms in this genre with different database processing. First, by classifying it into three classes of 200 items since some protocols were in the majority of them, and also by amplifying it, that is to say by creating the identical items with

a few differences for the minority classes. The aim of it was to have classes with the same number of items and, eventually, the database becomes balanced. The ML algorithms Randoms forest classifier shows better results for the different treatment of the data base.
Concerning data classification, Birba [24] (2020) reported that data splitting into different classes leads to an optimization of the accuracy of prediction compared to normal data training especially for a restricted data base. On the other hand, for a fairly developed database, data splitting does not affect performance and accuracy.

5. Conclusion and future work

In this paper, we proposed a simple and effective pipeline to classify and predict therapeutic protocol of breast cancer and number of treatment cycle in the case of a very small training data (601 samples). To increase the robustness of the classifier, we opted for three data treatment. We did a data amplification for the minority items to re-balance the database then we did a data splitting based on the therapeutic protocols. We noticed that data treatment (data augmentation and data splitting) gives higher accuracy's.
In the medical field, and when it comes to predicting a therapeutic protocol, accuracy becomes the most important factor to consider. In our case, the increase in data allowed a marked improvement in precision. Thus, and considering that medical data related to cancer is extremely limited, we demonstrated that the use of data augmentation or even data splitting resolves the problems linked to imbalanced and limited database.
Our work indicates that using Smote can increase the accuracy of the Random Forest method in predicting chemotherapy regimens. However, the use of SMOTE for data augmentation presents some cons :
• Can generate noisy data: The synthetic data generated by SMOTE may not accurately reflect the true distribution of the minority class, leading to noisy data and potentially decreasing model performance.
• May lead to overfitting: In some cases, SMOTE may lead to overfitting, where the model becomes too specific to the synthetic data generated by SMOTE and does not generalize well to new data.
• Requires careful parameter tuning: SMOTE requires careful parameter tuning to ensure that the synthetic data generated is representative of the minority class and does not introduce bias into the dataset.
Recommendations for future research are hybrid samples such as Smote-Tomek Link and Smote-ENN. To address the problem of unbalanced data in breast cancer chemotherapy regimens and improve the performance of classification methods, experimenting with GAN and VAE for data augmentation is superior to Smote alone and provides the necessary Save time.

REFERENCES

1. Tendances mondiales en oncologie 2022 . s. d. Consulté le 25 octobre 2022. https://www.iqvia.com/fr-fr/locations/france/newsroom/2022/07/tendances-mondiales-en-oncologie-2022.
2. Santé : Pourquoi la pénurie de médicaments s'intensifie en cette fin d'année , 2022)
3. Breast Cancer World Health Organisation . 2021. 2021. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
4. Arnold, Melina, Eileen Morgan, Harriet Rumgay, Allini Mafra, Deependra Singh, Mathieu Laversanne, Jerome Vignat, et al. 2022. Current and Future Burden of Breast Cancer: Global Statistics for 2020 and 2040 . The Breast 66 (décembre): 15-23. https://doi.org/10.1016/j.breast.2022.08.010.
5. La revue d'gibier : le coût mondial du cancer atteint 900 milliards de dollars — Les Echos . 2013. 2013. https://www.lesechos.fr/2013/10/la-revue-dgibier-le-cout-mondial-du-cancer-atteint-900-milliards-de-dollars-345814.
6. Global treatment costs of breast cancer by stage: A systematic review - PMC . s. d. Consulté le 20 octobre 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258130/.
7. Alshebly, O., Abdullah , S. N. (2023). Proposed Two-Steps Procedure of Classification High Dimensional Data with Regularized Logistic Regression. Statistics, Optimization Information Computing. https://doi.org/10.19139/soic-2310-5070-1846
8. Ayers, M., W.F. Symmans, J. Stec, A.I. Damokosh, E. Clark, K. Hess, M. Lecocke, et al. 2004. Gene Expression Profiles Predict Complete Pathologic Response to Neoadjuvant Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy in Breast Cancer . Journal of Clinical Oncology 22 (12): 2284-93. https://doi.org/10.1200/JCO.2004.05.166.
9. Basu, Kanadpriya, Ritwik Sinha, Aihui Ong, et Treena Basu. 2020. Artificial Intelligence: How is It Changing Medical Sciences and Its Future? Indian Journal of Dermatology 65 (5): 365-70. https://doi.org/10.4103/ijd.IJD$_4$21$_2$0.
10. Filali, Y., Abdelouahed, S., Aarab, A. (2019). An Improved Segmentation Approach for Skin Lesion Classification. Statistics, Optimization Information Computing, 7(2), 456-467. https://doi.org/10.19139/soic.v7i2.533
11. Mani, Subramani, Yukun Chen, Xia Li, Lori Arlinghaus, A Bapsi Chakravarthy, Vandana Abramson, Sandeep R Bhave, Mia A Levy, Hua Xu, et Thomas E Yankeelov. 2013. Machine Learning for Predicting the Response of Breast Cancer to Neoadjuvant

Chemotherapy . Journal of the American Medical Informatics Association 20 (4): 688-95. https://doi.org/10.1136/amiajnl-2012-001332.
12. Moussaoui, H., Nabil El Akkad, Mohamed Benslimane. (2023). A Hybrid Skin Lesions Segmentation Approach Based on Image Processing Methods. Statistics, Optimization Information Computing, 11(1), 95-105. https://doi.org/10.19139/soic-2310-5070-1549
13. Sugimoto, Masahiro, Shiori Hikichi, Masahiro Takada, et Masakazu Toi. 2021. Machine Learning Techniques for Breast Cancer Diagnosis and Treatment : A Narrative Review . Annals of Breast Surgery 0 (janvier): 0-0. https://doi.org/10.21037/abs-21-63.
14. Rawal, Ramik. 2020. breast cancer prediction using machine learning 7 (5): 13.
15. Asri, Hiba, Hajar Mousannif, Hassan Al Moatassime, et Thomas Noel. 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis . Procedia Computer Science 83: 1064-69. https://doi.org/10.1016/j.procs.2016.04.224.
16. Konstantinos Vougas and al.,2017. Deep Learning and Association Rule Mining for Predicting Drug Response in Cancer. A Personalised Medicine Approach doi: https://doi.org/10.1101/070490;
17. He, Tiancheng, Mamta Puppala, Chika F. Ezeana, Yan-siang Huang, Ping-hsuan Chou, Xiaohui Yu, Shenyi Chen, et al. 2019. A Deep Learning–Based Decision Support Tool for Precision Risk Assessment of Breast Cancer . JCO Clinical Cancer Informatics, no 3 (décembre): 1-12. https//doi.org/10.1200/CCI.18.00121.
18. Nicolò, Chiara, Cynthia Périer, Melanie Prague, Carine Bellera, Gaëtan MacGrogan, Olivier Saut, et Sébastien Benzekry. 2020. Machine Learning and Mechanistic Modeling for Prediction of Metastatic Relapse in Early-Stage Breast Cancer . JCO Clinical Cancer Informatics. https://doi.org/10.1200/CCI.19.00133
19. Sidey-Gibbons, Chris, André Pfob, Malke Asaad, Stefanos Boukovalas, Yu-Li Lin, Jesse Creed Selber, Charles E. Butler, et Anaeze Chidiebele Offodile. 2021. Development of Machine Learning Algorithms for the Prediction of Financial Toxicity in Localized Breast Cancer Following Surgical Treatment . JCO Clinical Cancer Informatics, no 5 (décembre): 338-47. https://doi.org/10.1200/CCI.20.00088.
20. Nicholas Meti and al., 2020. Machine Learning Frameworks to Predict Neoadjuvant Chemotherapy Response in Breast Cancer Using Clinical and Pathological Features DOI https://doi.org/10.1200/CCI.20.00078
21. Houda Ait Brahim, Mariam Benllarch, Nada Benhima, Salah El-Elhadaj, Abdelmoutalib Metrane, Rhizlane Belbaraka. 2023. New Real Dataset Creation To Develop An Intelligent System For Predicting Chemotherapy Protocols . International Journal of Advanced Computer Science and Applications Vol. 14, No. 8, 2023. https://DOI: 10.14569/IJACSA.2023.0140886.
22. Krishnamoorthi, Raja, Shubham Joshi, Hatim Z. Almarzouki, Piyush Kumar Shukla, Ali Rizwan, C. Kalpana, et Basant Tiwari. 2022. A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques . Journal of Healthcare Engineering 2022: 1684017. https://doi.org/10.1155/2022/1684017.
23. Oncology Spending Worldwide 2011-2023 . 2022. Statista. 2022. https://www.statista.com/statistics/696208/oncology-costs-worldwide/.
24. Birba, Delwende Eliane. 2020. A Comparative Study of Data Splitting Algorithms for Machine Learning Model Selection , 29.