A Comparative Study of Machine Learning Techniques to Predict Types of Breast Cancer Recurrence

Meryem Chakkouch[1], Merouane Ertel[2], Aziz Mengad[3], Said Amali[4]
Informatics and Applications Laboratory (IA)-Faculty of Sciences, Moulay Ismail University Meknes, Morocco[1, 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]
Informatics and Applications Laboratory (IA), FSJES Moulay Ismail University, Meknes, Morocco[4]

***Abstract*—The prediction of breast cancer recurrence is a crucial problem in cancer research that requires accurate and efficient prediction models. This study aims to compare the performance of different machine learning techniques in predicting types of breast cancer recurrence. In this study, the performance of logistic regression, decision tree, K-Nearest Neighbors, and artificial neural network algorithms was compared on a breast cancer recurrence dataset. The results show that the artificial neural network algorithm outperformed the other algorithms with 91% accuracy, followed by the decision tree (DT) algorithm and K-Nearest Neighbors (kNN) also performed well with accuracies of 90.10% and 88.20%, respectively, while the logistic regression algorithm had the lowest accuracy of 84.60%. The results of this study provide insight into the effectiveness of different machine learning techniques in predicting types of breast cancer recurrence and could guide the development of more accurate prediction models.**

Keywords—Breast cancer; machine learning; recurrence prediction; classification multi-classes; logistic regression; decision tree; K-Nearest Neighbors; artificial neural network

I. Introduction

Breast cancer is one of the most common types of cancer in women worldwide, with approximately 2.3 million new cases diagnosed in 2020 alone [1]. Although treatment options for breast cancer have advanced significantly in recent years, thanks to improved surgical techniques, chemotherapy and radiation therapy, the risk of recurrence remains a significant concern for patients and their clinicians.

Recurrence of breast cancer can occur in a variety of forms, including local, regional and distant recurrence. Local recurrence is when the cancer recurs in the same area where it was initially treated, while regional recurrence is when the cancer recurs in the lymph nodes of the axillary or supraclavicular region [2]. Distant metastases are when the cancer spreads to distant organs such as the lungs, liver, or bone. Each type of recurrence has its own clinical characteristics and treatment considerations, which makes customization of treatment particularly important for each patient.

In recent years, there has been increasing interest in developing predictive models that can accurately identify the likelihood of different types of breast cancer recurrence, allowing for more personalized treatment plans and better patient outcomes [3]–[9]. Predictive models are based on a variety of factors, such as the characteristics of the initial cancer, the patient's age, the stage of the cancer and the type of treatment received. Tools such as the Online Recurrence Score (Oncotype DX) [10], Metastatic DNA Prognosis Score (MammaPrint) [11], and Molecular Profiling Signature Score (EndoPredict) are available to help clinicians assess the risk of recurrence in patients with breast cancer [12]. These predictive models can help clinicians decide whether a patient requires adjuvant chemotherapy, radiation therapy, or hormone therapy, or whether she should simply be monitored closely. This can help avoid unnecessary treatment and reduce potentially harmful side effects of treatment.

In this paper, a multi-classification model is proposed, with the aim of producing predictions about the types of recurrence in breast cancer patients. This model is based on several important parameters, such as initial TNM status of the tumor, estrogen and progesterone receptors (ER, PR), HER2 expression levels, and previous treatments received by the patients. Additionally, new variables such as physical activity, diet type, and post-treatment psychology are incorporated to improve risk assessment and patient management. Indeed, previous research has demonstrated the importance of considering psychological and behavioral aspects of patients to better predict the risk of recurrence.

In the second section of this article, the predictive variables included in the model will be presented. A description of various data coding techniques such as data collection, preprocessing, cleaning, and transformation of the pathological, biological, and clinical dataset will be provided in the third section. The fourth section will explain the proposed multi-classification technique, including Logistic Regression (LR), K-Nearest Neighbors (kNN), Decision Tree (DT), and Neural Network (NN). Finally, the performance of the proposed model in predicting recurrence types for breast cancer patients will be evaluated.

II. Materials and Methods

A. Data Collection

Clinical and pathological data were collected from electronic medical records of patients treated for breast cancer between 2015 and 2022 at a single center in the Meknes, Morocco. The dataset included 1189 patients who underwent surgery, radiation therapy, and/or chemotherapy - the follow-up of at least 60 months.

The dataset included 19 features, including tumor size, age, hormone receptor status, histological grade, lymph node status, human epidermal growth factor receptor 2 (HER2) status, progesterone receptor (PR) status, estrogen receptor (ER) status, chemotherapy, Targeted therapies, radiation therapy, hormonotherapy, healthy eating, physical activity, type of psychosocial stress and type of recurrence. The important factors of our study are outlined in Table I.

B. Data Preprocessing

In this study, the data set was preprocessed to handle missing values, categorical variables, and feature selection. The Python programming language and the Pandas library were used to preprocess the data, as well as to analyze the medical records of cancer patients to identify factors associated with cancer recurrence. Several preprocessing techniques were used to prepare the data for machine learning, as shown in Fig. 1. First, missing or invalid values were verified and replaced where appropriate. Second, the categorical variables were recoded such as 'POSTMENOPAUSAL' and ‘CHEMIOTHERAPY (ANTHRACYCLINES / TAXANE)', ‘TYPE OF PSYCHOSOCIAL STRESS’ into numeric variables for easy analysis. The 'LYMPH_NODES' variable was also transformed into an ordinal variable, assigning values from 0 to 3 to represent the different levels of lymph node involvement. Additionally, variables such as AGE_DIAGNOSIS, TUMOR_SIZE, and KI67 were scaled into a common range to avoid bias in the analysis.Then, one-hot encoding was employed to convert categorical variables such as 'HER2,' 'ER,' 'PR,' 'SURGERY_TYPE,' 'TARGETED THERAPIES,' 'RADIOTHERAPY,' 'HEALTHY EATING,' and 'PHYSICAL ACTIVITY' into binary variables to enhance model accuracy. Fourth, a new target ordinal variable representing types of breast cancer recurrence (No, local, regional, or distant) was created, as shown in Table I.

Finally, feature screening was performed to identify the most important variables in predicting types of breast cancer recurrence. This reduced the dimensionality of the data improved the efficiency of the analysis.

TABLE I. Code and Value of the Variables used in this Study

Variable_Name		Code or Value
AGE_DIAGNOSIS		≤ 25; 25 to 35; 36-45; 46 ≥ years old
POSTMENOPAUSAL		0 < 49 ; 1 ≥ 50)
TUMOR_SIZE		≤2 ; 3-4 ; ≥5
LYMPH_NODES		0 = "No" ; 1 = "1–3" ; 2 = "4–9" ; 3 = "0>9"
TUMOUR_GRADE		1 ; 2 or 3
HER2		0 ; 1
ER		0 ; 1
PR		0 ; 1
KI67		≤ 20; 21 to 40; 41-60; 61 ≥
SURGERY_TYPE		0= “lumpectomy”; 1= “mastectomy”
CHIMIOTHERAPY	ANTHRACYCLINES	0=”No” ;1=”EC 50”; 2=”FEC 100”;3= “AC60”
	TAXANE	0=”No” ;1=” PACLITAXEL ”; 2=” DOCETAXEL ”
TARGETED THERAPIES		0= “No”; 1= “ TRASTUZUMAB ”
RADIOTHERAPY		0= “No”; 1= “ RADIOTHERAPY ”
HORMONOTHERAPY		0= “No”; 1= “ Tamoxifen ”; 2=”Aromatase inhibitors”; 3=”Tamoxifene + Aromatase Inhibitors” ;4=”Aromatase Inhibitors + Tamoxifene” .
HEALTHY EATING		0= “No”; 1= “ Yes ”
PHYSICAL ACTIVITY		0= “No”; 1= “ Yes ”
TYPE OF PSYCHOSOCIAL STRESS		0=”No”; 1= “Cognitive stress”; 2=”Familial stress”; 3=”Work-related stress”; 4=”Environmental stress”; 5=”Event-related stress”.
TYPE OF RECURRENCE		No=”0000” ; Local=”0100”; Regional=”0010” or Distant=”0001”)

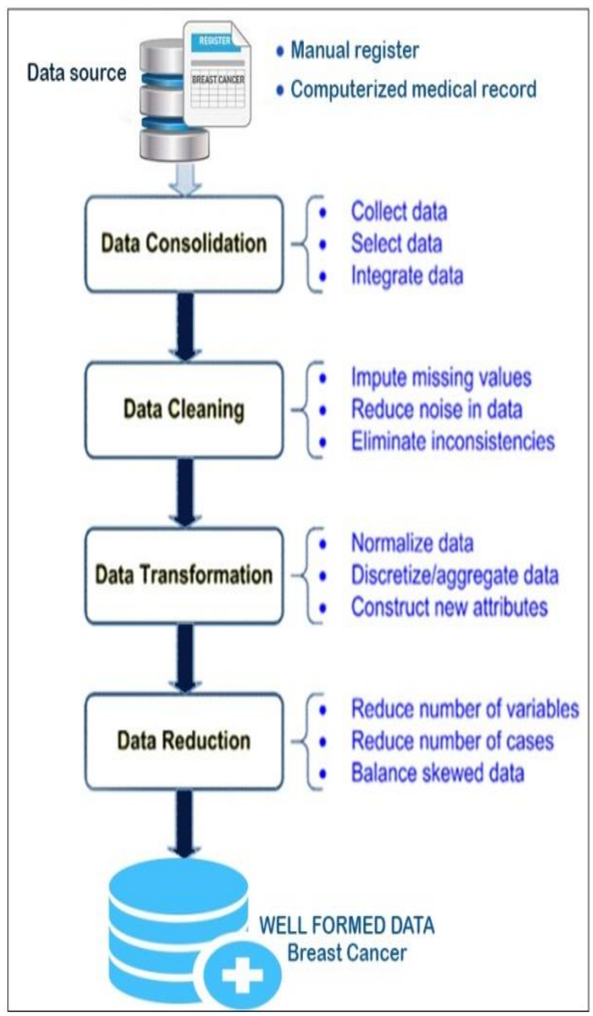


Fig. 1. The typical data mining procedure used in this study.

C. Machine Learning Models

The forms of recurrence in breast cancer patients were categorized into classes in this study using classification models based on demographic, biopsychological, and clinical traits. Fig. 2 depicts the complete experimentation procedure.

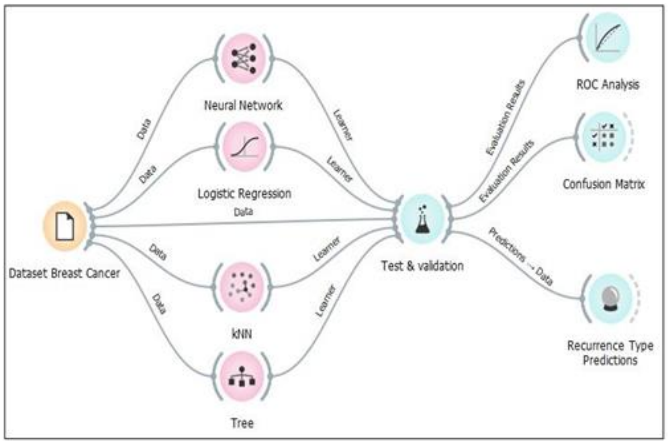


Fig. 2. Model machine learning use.

In this paper, the performance of four machine learning techniques was compared: logistic regression, decision tree, K-Nearest Neighbors, and artificial neural network. The Python programming language and the Scikit-learn library were used to implement the models. The models were trained and tested using a 10-fold cross-validation method [13].

1) Artificial neural networks: Artificial neural networks are a machine learning technique based on the architecture of interconnected neurons inspired by the functioning of the human brain [14]. They are very useful in multi-class classification, where several categories are to be predicted. In the case of recurrence in breast cancer patients, neural networks can be used to predict whether a patient is likely to experience a recurrence of the disease.

Once the data is prepared, the neural network model is constructed. It can consist of several layers of neurons, each with a specific number of neurons. The input layer is the one that receives the input data, while the output layer is the one that gives the classification results [15]. The intermediate layers are called hidden layers and are responsible for learning the relationships between the input features and the output classes.

The learning process is done using a gradient back-propagation algorithm, which calculates the gradients of the loss function with respect to the model weights and adjusts the weights to minimize the loss function [16]. The loss function can be a cross-entropy, which measures the distance between the model predictions and the actual classes.

2) Logistic regression: Logistic regression is a statistical technique commonly used in modeling binary data, where the dependent variable can take only two possible values[17]. However, it is possible to apply logistic regression to multi-class classification problems, such as the classification of recurrence types in breast cancer patients.

To use logistic regression in this context, there are several approaches. One is to apply multinomial logistic regression, which models the probability of each class of recurrence. In this case, the model will produce several equations for each class that describe the relationships between the input variables and the probabilities of each class.

Another approach was used, binary logistic regression for each recidivism class, treating each class as a separate binary variable. In this case, several logistic regression models must be fitted for each recurrence class, and the final class for each patient is determined based on the probabilities predicted by each model.

3) K-Nearest Neighbors (K-NN): K-Nearest Neighbors (K-NN) is a classification algorithm used to predict the category of a new sample based on the training data samples closest to that sample[18]. In the case of multi-class classification of recurrence types in breast cancer patients, the K-NN algorithm can be used to predict whether a given patient is likely to experience local recurrence, distant recurrence, simultaneous local and distant recurrence, or no recurrence. The above steps can be applied after preprocessing the data as follows:

Distance definition: the distance can be calculated using Euclidean distance or other distance measures appropriate for the types of variables used.

Determining the K number: the K value can be chosen by cross-validation using different values for K and choosing the one that gives the best performance.

Category prediction: once the closest data samples have been identified, the majority of their categories can be used to predict the most likely recurrence category for the patient.

4) Decision tree: A decision tree can be an effective tool for multiclass classification of recurrence types in breast cancer patients[19]. The decision tree is a machine learning model that progressively divides a dataset into smaller subsets based on specific criteria, until each subset can no longer be divided. Each division of the dataset is based on a specific variable, which can be a continuous or categorical variable. At the end of the process, the decision tree provides a set of decision rules that can be used to predict the target classes.

D. Performance Indicators

In this study, the performance of each model was evaluated using the confusion matrix to calculate the following metrics: precision, sensitivity, specificity, and AUC-ROC.

1) Confusion matrix: A multiclass confusion matrix is used to evaluate the performance of a classification model that predicts classes belonging to more than two categories. In this case, assuming that the four classes are: "No recurrence," "Local recurrence," "Loco-regional recurrence," and "Remote recurrence," the confusion matrix will look like this, as shown in Table II.

TABLE II. CONFUSION MATRIX FOR MULTI-CLASS CLASSIFICATION OF RECURRENCE TYPE

Actual class / Predicted class	No recurrence	Local recurrence	Locoregional recurrence	Distant recurrence
No recurrence	**True Positives**	False Negatives	False Negatives	False Negatives
Local recurrence	False Positives	**True Positives**	False Negatives	False Negatives
Locoregional recurrence	False Positives	False Positives	**True Positives**	False Negatives
Distant recurrence	False Positives	False Positives	False Positives	**True Positives**

In this confusion matrix, TP stands for True Positives, TN stands for True Negatives, FP stands for False Positives and FN stands for False Negatives [20].

- True Positives (TP): The model correctly predicted the recurrence class for breast cancer patients who actually had a recurrence of that type.
- True Negatives (TN): The model correctly predicted the "No Recurrence" class for breast cancer patients who did not experience a recurrence.
- False Positives (FP): The model incorrectly predicted that a patient had a recurrence, when in fact there was no recurrence.
- False Negatives (FN): The model incorrectly predicted that there was no recurrence for a patient who actually had a recurrence.

The Sum of each row and column gives us useful statistics to evaluate the performance of the model, including:

Accuracy (1) is the proportion of the total number of correct predictions made by the model divided by the total number of predictions made by the model. It is typically represented as a percentage and is one of the most commonly used performance metrics for classification models [21]. The formula for accuracy is:

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

Sensitivity (or recall) (2) for each class, which measures the proportion of true positives among all true positive cases of this class.

$$\text{Recall}_{\text{class}} = \frac{\text{TP}_{\text{class}}}{\text{TP}_{\text{class}} + \text{FN}_{\text{class}}} \quad (2)$$

Specificity (3) for each class, which measures the proportion of true negatives among all the true negative cases in that class is given as,

$$Specificity_{class} = \frac{TN_{class}}{FP_{class} + TN_{class}} \quad (3)$$

Precision(4) for each class, which measures the proportion of true positives among all predicted positive cases in that class is given as,

$$Precision_{class} = \frac{TP_{class}}{TP_{class} + FP_{class}} \quad (4)$$

F1-Score (5) for each class, which is a measure of the combined accuracy and sensitivity is measured as,

$$\boldsymbol{F1 - Score} = \frac{2 * TP_{class}}{2 * TP_{class} + FN_{class} + FP_{class}} \quad (5)$$

2) The Roc and AUC curve: Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) curves are commonly used evaluation tools in binary classification to assess the performance of classification models [22].

In the case of four-class classification of recurrence types in breast cancer patients, these curves can be used to evaluate the ability of a classification model to distinguish between different types of recurrence.

The ROC curve plots the true-positive rate (sensitivity) against the false-positive rate (1 - specificity) for different classification cut-off values. The AUC represents the area under the ROC curve and measures the overall ability of the model to discriminate between classes. The closer the AUC is to 1, the better the performance of the classification model.

In the case of four-class classification, several ROC curves can be plotted for each class. The overall performance of the model can be evaluated by calculating a weighted average of the AUC for each class.

Using these curves, physicians and researchers can evaluate the performance of classification models and select the best model for predicting recurrence types in breast cancer patients.

III. RESULTS

A. Analysis of Result

The study evaluated the performance of four classification algorithms - Neural Network, Decision Tree, kNN, and Logistic Regression - in predicting the type of recurrence in breast cancer. The evaluation metrics used to assess the performance of these algorithms were AUC, CA (Classification Accuracy), F1 score, Precision, Recall, and Specificity. The variables used to predict the type of recurrence were age at diagnosis of breast cancer, primary tumor size (TS), postmenopausal status, number of involved axillary lymph nodes, histological grade of the tumor, type of surgery, cellular marker of proliferation (Ki67),PR status, ER status, HER2 status, healthy eating, physical activity, type of psychosocial stress . Including these variables improved the performance of the classification algorithms. The study used 10-fold cross-validation to ensure that the results were representative and not overfit to the data, see Table III.

Table III provides the results of four different classification algorithms (A, B, C and D) for four classes (0000, 0100, 0010 and 0001). The columns represent the predictions of the algorithms and the rows represent the actual classes. The values in the cells indicate the number of times each class was predicted for each actual class.

For Algorithm A (neural network), it correctly predicted class 0000 for 904 times, class 0100 for 105 times, class 0010 for 24 times and class 0001 for 49 times. There are 107 incorrect predictions in total for algorithm A.

For Algorithm B (decision tree), it correctly predicted class 0000 for 904 times, class 0100 for 99 times, class 0010 for 27 times and class 0001 for 41 times. There are 118 incorrect predictions in total for algorithm B.

For Algorithm C (kNN), it correctly predicted class 0000 for 900 times, class 0100 for 110 times, class 0010 for 17 times and class 0001 for 22 times. There are 140 incorrect predictions in total for algorithm C.

For Algorithm D (logistic regression), it correctly predicted class 0000 for 903 times, class 0100 for 63 times, class 0010 for 8 times and class 0001 for 32 times. There are 183 incorrect predictions in total for algorithm D.

Looking at the overall results, it is evident that Algorithm A gave the best result with the least number of incorrect predictions, while Algorithm D gave the worst result with the highest number of incorrect predictions. The other algorithms (B and C) gave intermediate results. However, further analysis of the data, including measures of precision, recall, and F1-score for each class, would be required to provide a more complete and accurate assessment of the performance of the algorithms.

TABLE III. MULTI-CLASS CONFUSION MATRIX OF THE APPLICABLE CLASSIFICATION MODELS

A : Neural Network - Classifier

		Predicted				
		0000	**0100**	**0010**	**0001**	∑
Current	**0000**	**904**	7	10	8	**929**
	0100	7	**105**	1	11	**124**
	0010	23	1	**24**	7	**55**
	0001	6	21	5	**49**	**81**
∑		**940**	**134**	**40**	**75**	**1189**

B : Decision Tree – Classifier

		Predicted				
		0000	**0100**	**0010**	**0001**	∑
Current	**0000**	**904**	7	7	11	**929**
	0100	3	**99**	0	22	**124**
	0010	22	0	**27**	6	**55**
	0001	19	18	3	**41**	**81**
∑		**948**	**124**	**37**	**80**	**1189**

C : kNN – Classifier

		Predicted				
		0000	**0100**	**0010**	**0001**	∑
Current	**0000**	**900**	19	2	8	**929**
	0100	2	**110**	0	12	**124**
	0010	36	0	**17**	2	**55**
	0001	34	25	0	**22**	**81**
∑		**972**	**154**	**19**	**44**	**1189**

D : Logistic Regression –Classifier

		Predicted				
		0000	**0100**	**0010**	**0001**	∑
Current	**0000**	**903**	12	2	12	**929**
	0100	44	**63**	2	15	**124**
	0010	40	7	**8**	0	**55**
	0001	30	15	4	**32**	**81**
∑		**1017**	**97**	**16**	**59**	**1189**

B. Performance Evaluation

The four algorithms' performances were compared using classification metrics; see Table IV for details.

TABLE IV. PERFORMANCE METRICS OF MACHINE LEARNING MODELS

Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Specificity**
Neural Network	0.976	0.910	0.907	0.905	0.910	0.887
Decision Tree	0.899	0.901	0.898	0.897	0.901	0.863
kNN	0.934	0.882	0.868	0.873	0.882	0.778
Logistic Regression	0.941	0.846	0.826	0.822	0.846	0.652

Based on the table, we can see that the neural network model has the highest AUC (0.976) and the highest classification accuracy (0.910). It also has high precision (0.905), high recall (0.910), and high specificity (0.887), indicating that it performs well in both detecting positive instances and avoiding false positives and false negatives.

The decision tree model has a lower AUC (0.899) and classification accuracy (0.901) than the neural network, but still performs relatively well. It has a similar F1 score (0.898) and precision (0.897) to the neural network, but slightly lower recall (0.901) and specificity (0.863).

The kNN model has a lower AUC (0.934) and classification accuracy (0.882) than both the neural network and decision tree models. It has a lower F1 score (0.868) and precision (0.873) than the other two models, but a similar recall (0.882) and lower specificity (0.778).

The logistic regression model has the lowest AUC (0.941) and classification accuracy (0.846) of all the models. It also has the lowest F1 score (0.826), precision (0.822), recall (0.846), and specificity (0.652). This suggests that the logistic regression model may not be as effective at distinguishing between the positive and negative classes and may have a higher rate of false positives and false negatives than the other models.

C. ROC and AUC Curve

Knowing the associated Receiver Operating Characteristic (ROC) curve, True Positive Rate (TPR), and False Positive Rate (FPR) is important when evaluating the performance of classification models. In this study, it can be concluded that all machine learning classifiers predict with >89% accuracy on the type of recurrence in Boobs cancer patients, this shows that these classification algorithms work well to classify different types of recurrence. Using the following iteration type codes (0000 - 0100 - 0010 - 0001), the neural network obtained the maximum AUC (area under the curve) of the ROC (see Figure 3).

In the curve cube graph (Fig. 3), there are four ROC curves, each representing the performance of a different classifier for a different class. The classifiers include a neural network, a decision tree, logistic regression, and K-Nearest Neighbors (KNN).

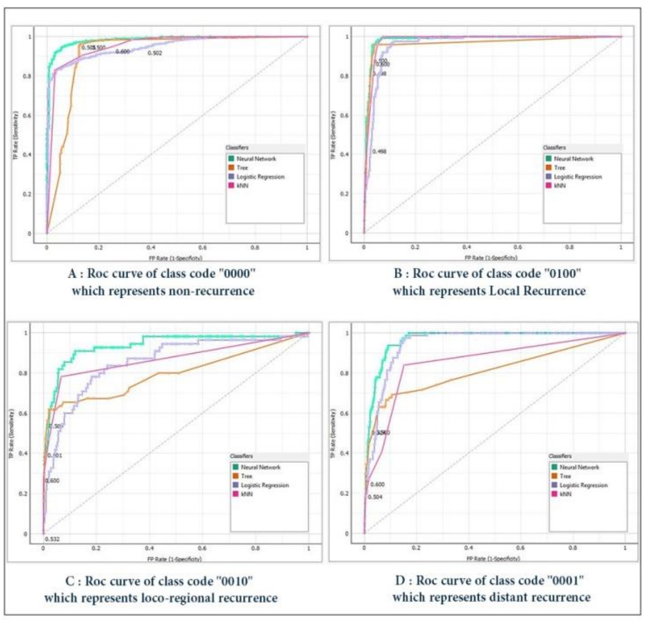
For class code "0000," which represents non-recurrence, the ROC curve shows that the neural network and logistic regression have the highest sensitivity (TP rate) of 0.8, while the decision tree and KNN have a sensitivity of 0.6. The false positive rate (FP rate) is relatively low across all classifiers, indicating that they have good specificity.

For class code "0100," which represents local recurrence, the ROC curve shows that the neural network has the highest sensitivity of 0.6, followed by the decision tree and logistic regression with a sensitivity of 0.4, and KNN with a sensitivity of 0.2. The false positive rate is relatively high for all classifiers, indicating that they have poor specificity.

For class code "0010," which represents loco-regional recurrence; the ROC curve shows that the neural network and logistic regression have the highest sensitivity of 0.8, followed by the decision tree with a sensitivity of 0.6, and KNN with a sensitivity of 0.4. The false positive rate is relatively low for all classifiers, indicating that they have good specificity.

For class code "0001," which represents distant recurrence, the ROC curve shows that the neural network has the highest sensitivity of 1, followed by logistic regression with a sensitivity of 0.6, the decision tree with a sensitivity of 0.4, and KNN with a sensitivity of 0.2. The false positive rate is relatively high for all classifiers, indicating that they have poor specificity.

Overall, the ROC curves in the curve cube graph demonstrate that the neural network and logistic regression classifiers perform better than the decision tree and KNN classifiers, particularly for classes that are more difficult to classify, such as local recurrence and distant recurrence. The curve cube graph is a useful visualization tool for comparing the performance of multiple classifiers for different classes in a multi-class classification problem.



A : Roc curve of class code "0000" which represents non-recurrence

B : Roc curve of class code "0100" which represents Local Recurrence

C : Roc curve of class code "0010" which represents loco-regional recurrence

D : Roc curve of class code "0001" which represents distant recurrence

Fig. 3. AUC - ROC curve of classifiers used in the prediction of types of recurrence in patients with breast cancer.

IV. Discussion

This paper presents four different types of classifiers (neural network, decision tree, kNN, and logistic regression) for predicting recurrence type in breast cancer patients. The classifiers were evaluated for their ability to correctly predict recurrence type (0000, 0100, 0010, or 0001) for a data set of 1189 patients. The results show that the neural network classifier performed the best, correctly predicting recurrence type with an overall accuracy of 91%. The decision tree classifier was the second best performer with an accuracy of 90,1%, followed by kNN with 88,2% and logistic regression with 84,6%.

Indeed, the article also points out that incorporating biopsychological variables into recidivism prediction studies is important for improving understanding of recidivism risk in individual patients. This may lead to more personalized treatment decisions and better monitoring after initial treatment. In addition, regular psychological and social assessments can help identify patients who need additional support to improve their quality of life and overall well-being.

These findings underscore the importance of considering not only medical factors, but also psychological and social factors in the management of breast cancer. By incorporating these factors into clinical decision making, physicians can improve the quality of care for breast cancer patients.

Overall, the results of this study indicate that these classifiers can be used to accurately predict the type of recurrence in breast cancer patients. However, further studies are needed to confirm the effectiveness of these models and determine their utility in clinical practice. Nevertheless, this study is an important contribution to the field of medical research, as it shows the potential of machine learning techniques to improve outcomes for cancer patients.

A. Limitations

Validation of our results using other datasets would be necessary to generalize our findings to other populations. Additionally, our study focused only on predicting the type of breast cancer recurrence and did not consider other factors such as disease-free survival or overall survival. Further studies may investigate the potential use of machine learning techniques to predict these outcomes.

V. Conclusion

The study used a dataset of 1189 patients with breast cancer and applied various machine learning techniques, including logistic regression, decision tree, K-Nearest Neighbors and artificial neural network, to predict the type of recurrence. The results showed that the artificial neural network outperformed the other models in terms of accuracy, precision, recall, and F1-score.

The high performance of the artificial neural network can be attributed to its ability to capture complex non-linear relationships between the features and the target variable. The logistic regression model, which is a linear model, achieved a lower performance than the other models, indicating that non-linear relationships exist between the features and the target variable.

The findings of this study have significant implications for breast cancer patients and clinicians. Accurate prediction of the type of recurrence can help to guide treatment decisions and improve patient outcomes. Machine learning techniques can provide a valuable tool in predicting breast cancer recurrence and may lead to more personalized treatment plans.

However, it is essential to note that machine learning techniques are not without limitations. One potential limitation is the need for large datasets to train the models effectively. Additionally, machine learning models may not always be transparent, and it may be challenging to understand how the models arrive at their predictions. Further investigation is needed to address these limitations and improve the clinical application of machine learning techniques in predicting breast cancer recurrence.

REFERENCES

[1] "The Global Cancer Observatory - March, 2021. Page 1."

[2] B. Holleczek, C. Stegmaier, J. C. Radosa, E.-F. Solomayer, and H. Brenner, "Risk of loco-regional recurrence and distant metastases of patients with invasive breast cancer up to ten years after diagnosis – results from a registry-based study from Germany," BMC Cancer, vol. 19, no. 1, p. 520, Dec. 2019, doi: 10.1186/s12885-019-5710-5.

[3] E. Merouane, A. Said, and E. F. Nour-eddine, "Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers," International Journal of Advanced Computer Science and Applications, vol. 13, no. 2, p. 6, 2022.

[4] A. G. V. Bitencourt et al., "MRI-based machine learning radiomics can predict HER2 expression level and pathologic response after neoadjuvant therapy in HER2 overexpressing breast cancer," EBioMedicine, vol. 61, p. 103042, Nov. 2020, doi: 10.1016/j.ebiom.2020.103042.

[5] N. Adnan, C. Lei, and J. Ruan, "Robust edge-based biomarker discovery improves prediction of breast cancer metastasis," BMC Bioinformatics, vol. 21, no. S14, p. 359, Sep. 2020, doi: 10.1186/s12859-020-03692-2.

[6] H. Alkhadar, M. Macluskey, S. White, I. Ellis, and A. Gardner, " Comparison of machine learning algorithms for the prediction of five - year survival in oral squamous cell carcinoma," J Oral Pathol Med, vol. 50, no. 4, pp. 378‐384, Apr. 2021, doi: 10.1111/jop.13135.

[7] R. Bhardwaj and N. Hooda, "Prediction of Pathological Complete Response after Neoadjuvant Chemotherapy for breast cancer using ensemble machine learning," Informatics in Medicine Unlocked, vol. 16, p. 100219, 2019, doi: 10.1016/j.imu.2019.100219.

[8] 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.

[9] 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.

[10] 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.

[11] G. Altan, "Deep Learning-based Mammogram Classification for Breast Cancer," ijisae, vol. 8, no. 4, pp. 171–176, Dec. 2020, doi: 10.18201/ijisae.2020466308.

[12] P. Chandrakar, A. Shrivas, and N. Sahu, "Design of a Novel Ensemble Model of Classification Technique for Gene-Expression Data of Lung Cancer with Modified Genetic Algorithm," EAI Endorsed Transactions on Pervasive Health and Technology, vol. 7, no. 25, p. 167845, Jan. 2021, doi: 10.4108/eai.8-1-2021.167845.

[13] 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.

[14] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," Journal of Pharmaceutical and Biomedical Analysis, vol. 22, no. 5, pp. 717–727, Jun. 2000, doi: 10.1016/S0731-7085(99)00272-1.

[15] P. J. Drew and J. R. T. Monson, "Artificial neural networks," vol. 127, no. 1, p. 9, 2000.

[16] M. Ertel and S. Amali, "'Predicting the efficacy of chemotherapy applied to breast cancer by machine learning'. (2021). 1st International Meeting on The health system: Managing the health crisis between public action and future prospects. April 07-08, Meknes, Morocco.," 2021.

[17] E. Bisong, "Logistic Regression," in Building Machine Learning and Deep Learning Models on Google Cloud Platform, Berkeley, CA: Apress, 2019, pp. 243–250. doi: 10.1007/978-1-4842-4470-8_20.

[18] S. E. Buttrey, "Nearest-neighbor classification with categorical variables," Computational Statistics & Data Analysis, vol. 28, no. 2, pp. 157–169, Aug. 1998, doi: 10.1016/S0167-9473(98)00032-2.

[19] S. Murugan, B. M. Kumar, and S. Amudha, "Classification and Prediction of Breast Cancer using Linear Regression, Decision Tree and Random Forest," in 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore: IEEE, Sep. 2017, pp. 763–766. doi: 10.1109/CTCEEC.2017.8455058.

[20] D.-Z. Du, P. Pardalos, and J. Wang, Eds., Discrete Mathematical Problems with Medical Applications, vol. 55. in DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 55. Providence, Rhode Island: American Mathematical Society, 2000. doi: 10.1090/dimacs/055.

[21] J. McCarthy, "Artificial Intelligence, Logic and Formalizing Common Sense," in Philosophical Logic and Artificial Intelligence, R. H. Thomason, Ed., Dordrecht: Springer Netherlands, 1989, pp. 161–190. doi: 10.1007/978-94-009-2448-2_6.

[22] 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.