# Advancing breast cancer prediction: machine learning, data balancing, and ant colony optimization

**Abd Allah Aouragh[1], Mohamed Bahaj[1], Fouad Toufik[2]**
[1]MIET Laboratory, Faculty of Sciences and Techniques, Hassan 1st University, Settat, Morocco
[2]Computer Sciences Laboratory, Higher School of Technology, Mohammed V University, Rabat, Morocco

**Article Info**

***Article history:***

Received Feb 10, 2024
Revised Jun 8, 2024
Accepted Jun 26, 2024

***Keywords:***

Adaptive synthetic sampling
Ant colony optimization
Breast cancer
Machine learning
Synthetic minority over-sampling technique

**ABSTRACT**

Breast cancer constitutes a significant threat to women's health worldwide. The World Health Organization (WHO) reports around 2.3 million new cases each year, making this disease the primary reason for cancer-related fatalities among women. In light of this alarming situation, developing innovative tools for early detection and optimal treatment is imperative, as it directly addresses the pressing need to enhance our capabilities in the quest to overcome breast cancer. This study fits in with this approach, introducing a comparative assessment of multiple machine learning algorithms and integrating data preprocessing, data balancing and feature selection techniques. The studied Coimbra dataset, composed of 116 records and including 10 medical characteristics, exhibited promising performance in all classification metrics, reaching an accuracy of 89.74%, and an area under the receiver operating characteristic curve (AUC-ROC) of 89.68%. These findings highlight the significant potential of our approaches to improve breast cancer treatment and detection systems, providing health practitioners with more efficient resources.

***Corresponding Author:***

Abd Allah Aouragh
MIET Laboratory, Faculty of Sciences and Techniques, Hassan 1st University
Settat, Morocco
Email: abdallahaouragh@gmail.com

## 1. INTRODUCTION

Breast cancer is a prevalent global ailment, ranking as the most frequent cancer among women, constituting more than a third of all new cancer diagnoses within the female population [1]. The symptoms of this disease are diverse, including the presence of nodules, changes in skin texture, chest pain, unusual nipple discharge, and alterations in breast size or shape [1], [2]. Every year, the World Health Organization (WHO) registers around 2.3 million new breast cancer cases, making it the primary contributor to cancer-related fatalities among women. A proper diagnosis, based on various techniques such as mammography, ultrasound and biopsy, is essential to determine the nature and severity of the disease [1], [2]. Numerous treatments exist for breast cancer management, encompassing surgical interventions, radiation therapy, chemotherapy, and targeted therapeutic approaches. These treatment options, which vary according to the stage of the disease and its specific characteristics, offer multiple approaches to address the individual requirements of each patient [2], [3]. However, beyond the complexity of treatment, the most important factor in improving breast cancer management remains its early recognition [2], [3]. An early diagnosis of the disease facilitates the initiation of more effective interventions, thereby maximizing the chances of recovery [2], [3].

In the fight against breast cancer, machine learning, which is a subset of artificial intelligence, empowers computers to acquire knowledge and make decisions without the need for explicit programming. The early detection strategy based on machine learning takes advantage of the growing availability of varied medical data [4], [5]. The integration of machine learning algorithms, along with advanced data balancing and feature selection techniques, revolutionizes the assessment of clinical data, notably enhancing the sensitivity of breast cancer testing [6]–[8]. These advancements, specifically tailored for breast cancer [9], [10] mark a notable advancement in healthcare. They facilitate personalized procedures tailored to the unique characteristics of individual patients, thereby improving treatment effectiveness and enhancing outcomes for women worldwide [10]–[12].

In this context, Ghani *et al*. [13] conducted a study to identify essential attributes in the Coimbra breast cancer dataset. They used the recursive feature elimination method and concluded that body mass index (BMI), age, homeostatic model assessment (HOMA) index, Resistin, and blood glucose are the best biomarkers for breast cancer. They then evaluated various machine learning algorithms, concluding that artificial neural networks offered the best accuracy, reaching 80.00%. Mishra *et al*. [14] employed the genetic algorithm (GA) to detect biological indicators relevant to breast cancer prediction from routine data and blood tests. The features identified by GA were integrated into various machine learning classifiers. The integration of GA with the gradient boosting algorithm demonstrated enhanced performance, with an accuracy of 79.08%, an area under the receiver operating characteristic curve (AUC-ROC) of 76.05%, and an F1-score of 79.61%. Barwal *et al*. [15] explored a hybrid approach combining k-nearest neighbors (KNN) with singular value decomposition (SVD) and grey wolf optimization (GWO) techniques to enhance breast cancer identification in its early phases. The hybrid model demonstrates improved performance in the automatic categorization of breast cancer, with an accuracy of up to 87.8%. Khatun *et al*. [16] exploited four algorithms to analyze inflammatory-stage breast cancer and reveal the main determining features. Simple logistic regression (SLR), Naïve Bayes (NB), multilayer perceptron (MLP), and random forest (RF) methods were deployed, achieving accuracy rates of 75%, 70%, 85%, and 68%, respectively.

Nanglia *et al*. [17] proposed a heterogeneous machine learning ensemble technique for early breast cancer identification. This approach is based on the stacking technique, combining three distinct classifiers: decision tree (DT), support vector machines (SVM), and KNN. The efficiency of this metaclassifier was evaluated by comparing it with its base classifiers and other individual models. The resulting ensemble model attained a maximum accuracy of 78%. Rasool *et al*. [18] introduced exploratory data techniques, encompassing distribution, feature correlation, and hyperparameter optimization while developing four distinct predictive models designed to increase breast cancer diagnostic accuracy. These approaches were applied to the breast cancer Coimbra dataset (BCCD) and the Wisconsin diagnostic breast cancer (WDBC) dataset. The diagnostic capabilities of the models were significantly improved, achieving a higher accuracy of 76.42%, and an improved F1-score of 76.83% for the polynomial SVM with the BCCD. Rani *et al*. [19] investigated the early prediction of prostate and breast cancer using eight classification classifiers. To improve predictive performance, normalization and feature selection techniques were employed. After applying the feature selection method analysis of variance (ANOVA), most classifiers performed well. Notably, concerning breast cancer with the Coimbra dataset, KNN reached an accuracy of 80%. Alfian *et al*. [20] examined the combination of SVM and an extremely randomized tree algorithm (extra-trees) for the timely identification of breast cancer. The extra-trees algorithm was utilized for feature selection, filtering out irrelevant attributes, while the SVM was employed to classify breast cancer status. This integrated approach yielded the highest accuracy, reaching 80.23%.

From previous research, it is evident that machine learning provides an innovative approach to early breast cancer identification and recognition. While researchers have explored various techniques, no universal method suitable for all cases has yet emerged. Previous studies have primarily focused on developing and evaluating machine learning models using diverse datasets, examining the effects of various traditional feature selection techniques such as recursive feature elimination and ANOVA. However, these studies often overlook the importance of data balancing techniques and their influence on predictive accuracy. Although machine learning algorithms have been investigated for breast cancer diagnosis, few studies explicitly address the effectiveness of advanced data balancing techniques like synthetic minority over-sampling technique (SMOTE) and adaptive synthetic sampling (ADASYN) in improving model sensitivity and specificity. Furthermore, while feature selection methods have been extensively studied, they mainly involve traditional approaches rather than metaheuristic techniques. In our study, we bridged this gap by evaluating the effectiveness of ant colony optimization (ACO), a metaheuristic method, alongside data balancing techniques like SMOTE and ADASYN. This exploration of ACO's integration with data balancing techniques and their combined impact on model performance, along with the assessment of five machine learning algorithms, represents a novel contribution to breast cancer prediction. Importantly, our methodology has shown superior effectiveness in comparison to previous research endeavors.

The subsequent sections of this article are arranged as follows: section 2 elucidates the materials and methods in depth. Section 3 outlines and discusses the results, with an analysis of the impact of the employed techniques. Finally, section 4 summarizes the principal discoveries and suggests potential avenues for future investigation.

## 2. MATERIALS AND METHODS

### 2.1. Proposed method

In the context of our research centered around the forecasting of breast cancer employing machine learning algorithms, our methodology followed a structured workflow. We initiated the process by acquiring the Coimbra breast cancer dataset, a rich source of information specific to breast cancer. Data preprocessing involved crucial steps such as data splitting, outlier handling, and normalization. The data imbalance challenge was addressed through a comprehensive evaluation of balancing techniques such as SMOTE and ADASYN. For the identification of the most pertinent features for classification, we have chosen to utilize the ACO algorithm. Subsequently, a variety of machine learning techniques, including SVM, KNN, RF, logistic regression, and DT, were thoroughly evaluated. Figure 1 illustrates the methodological approach adopted in this research dedicated to the prediction of breast cancer.

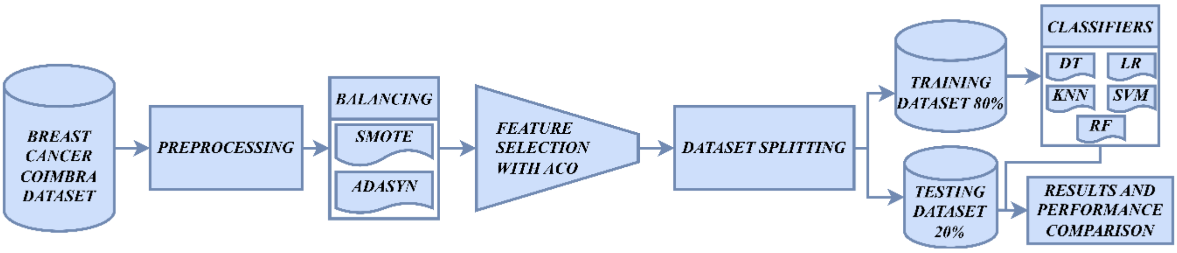

Figure 1. Proposed methodology

### 2.2. Dataset

Datasets are integral to machine learning projects, serving as the cornerstone for model development and evaluation. In our breast cancer prediction project, we rely on the BCCD [21], a valuable resource for healthcare professionals. This dataset facilitates research by offering a comprehensive collection of anthropometric and biological specifications, including glucose levels, age, BMI, insulin concentrations and other relevant indicators, obtained from a cohort of individuals diagnosed with breast cancer. Originating from the UCI machine learning repository, the BCCD comprises 116 records, providing a robust foundation for breast cancer analysis and prediction [21]. Table 1 illustrates the dataset's composition. The Coimbra breast cancer dataset comprises two classes: breast cancer patients, which make up 54.31% (64 cases) and non-breast cancer patients, representing 45.69% (52 cases) of the dataset. However, this class imbalance can introduce bias into the machine learning models. Therefore, it's crucial to address this issue to ensure that the models produce accurate and unbiased predictions for both classes, thus improving the reliability of the analysis.

Table 1. Coimbra dataset composition

| Attribute | Description |
|---|---|
| Age | Age of patients |
| BMI | A metric of body fat derived from weight and height |
| Glucose | Blood glucose levels, a crucial metabolic marker |
| Insulin | Insulin levels, a hormone involved in glucose regulation |
| HOMA | A technique used to evaluate insulin resistance and beta-cell function |
| Leptin | Leptin levels, a hormone that regulates appetite and energy balance |
| Adiponectin | Adiponectin levels, a protein linked to metabolic regulation |
| Resistin | Resistin levels, a protein associated with insulin resistance |
| Monocyte chemoattractant protein-1 (MCP-1) | A cytokine implicated in inflammation |
| Classification | Target variable, 0: healthy, 1: affected |

### 2.3. Balancing dataset

Dataset balancing is crucial in machine learning projects to address imbalances between different classes of data, thereby enhancing the performance of predictive models. Balancing the classes ensures that the model does not favor the majority class, leading to improved generalization of new data and more accurate decision-making. In our study, we opted for the SMOTE and ADASYN oversampling techniques due to their well-established effectiveness in addressing class imbalances. These techniques are widely adopted in various domains, including healthcare, where they have proven to enhance the effectiveness of predictive models by ensuring a balanced representation of minority classes [7], [8], [22]. Additionally, they have been extensively validated in the literature for their ability to generate synthetic samples that closely resemble the original data distribution, thereby facilitating robust model training and evaluation.

#### 2.3.1. Synthetic minority over-sampling technique

SMOTE is a machine learning approach designed to tackle class imbalance in datasets. It works by generating synthetic examples for minority classes based on existing ones. Specifically, SMOTE selects a sample from the minority class and identifies its KNN. New examples are then created by combining the initial sample with one or more of its neighbors. SMOTE's advantages include expanding the dataset without overfitting, as it introduces synthetic samples in underrepresented feature spaces. This enhances the model's ability to generalize and accurately predict minority classes [22]–[24].

#### 2.3.2. Adaptive synthetic sampling

ADASYN is a machine learning technique designed to tackle class imbalance in datasets by oversampling minority classes. Unlike SMOTE, ADASYN adjusts the density of synthetic examples generated based on the difficulty of classifying each minority example. It adaptively generates synthetic examples in regions where the density of minority instances is lower, thus enhancing the relevance of the synthetic examples produced. ADASYN effectively deals with class imbalances by better representing the variability of minority examples and improving model proficiency towards unbalanced classes [8], [22]–[24].

### 2.4. Ant colony optimization for feature selection

Feature selection is a pivotal step in machine learning, involving the identification of the most relevant features from a dataset to construct predictive models [25]. It aims to reduce data dimensionality by eliminating redundant attributes while retaining essential information for prediction tasks. ACO is a method inspired by the optimization capabilities observed in ant colonies, enabling them to find the most efficient paths to food sources. In feature selection, ACO is utilized to discover an optimal subset of features that maximizes predictive model performance. It employs a stochastic search process where potential solutions (feature sets) are explored and evaluated based on a specified performance metric, guided by artificial pheromones to identify promising solutions [25], [26].

ACO offers several significant advantages over traditional techniques like grid search. Firstly, it efficiently explores the search space of feature subsets, even in complex scenarios, leading to the discovery of promising solutions. Secondly, ACO exhibits robustness to noisy data and high-dimensional problems, making it suitable for real-world applications [25]. Moreover, it identifies non-redundant and informative feature subsets, thereby enhancing predictive model performance by mitigating overfitting and facilitating generalization. Lastly, ACO's computational efficiency surpasses that of grid search due to its intelligent and efficient exploration of the search space through stochastic search mechanisms inspired by ant behavior [26].

### 2.5. Machine learning algorithms

Machine learning algorithms are essential tools in various domains, including predictive analysis and medical classification, facilitating the extraction of valuable insights from complex data where variable relationships are intricate [7], [8], [10], [27]. In our study, we meticulously selected multiple machine learning algorithms for breast cancer prediction based on their performance in medical diagnosis, adaptability to complex classifications and reputation within the machine learning research community [8], [10], [28].

#### 2.5.1. Decision tree

DT algorithm is a supervised learning technique utilized for classification tasks. It constructs a DT from training data, with each node representing a feature, each branch indicating a decision based on that feature and each leaf representing a class or predicted value. DT offers simplicity, interpretability and robustness to noise and outliers. It can manage both categorical and numerical data, capture non-linear interactions between features and is relatively insensitive to missing data [10], [27], [28].

#### 2.5.2. Logistic regression

LR is a supervised learning model utilized for binary classification tasks, predicting binary response variables from explanatory variables through a logistic function transformation. LR is valued for its simplicity, fast execution, interpretable coefficients and capacity to provide class probabilities for each observation [8], [10]. Moreover, it demonstrates robustness to outliers compared to certain other algorithms [27], [28].

#### 2.5.3. K-nearest neighbors

KNN model is a supervised learning approach utilized for classification purposes. It assigns the predominant class among the KNN in the feature space to an unlabeled entry. KNN is conceptually simple and easy to implement, as it does not require explicit model training [8], [10]. However, it can be sensitive to noisy data or high dimensionality in the feature space, which may affect its performance [27], [28].

### 2.5.4. Support vector machines

SVM is a supervised learning technique utilized for classification tasks. It searches for the optimal hyperplane to separate different data categories in a high-dimensional space, maximizing the margin for clear class separation [8], [10]. SVM can handle non-linearly separable data by using kernels to transform it into a higher-dimensional feature space. It is also proficient in handling high-dimensional datasets and robust to overfitting due to built-in regularization [27], [28].

### 2.5.5. Random forest

RF is a supervised learning classification algorithm that operates by aggregating multiple DT. Each tree is trained on a random subset of data and features, with final predictions obtained through averaging or majority voting [8], [10]. RF excels at handling complex datasets with non-linear relationships and is resistant to overfitting due to its ensemble approach. Additionally, it demonstrates robustness to outliers and missing data [27], [28].

## 3. RESULTS AND DISCUSSION

For the software and hardware components of this study, the algorithms were coded in Python and run on a computer with a powerful AMD Ryzen 7 5700G processor. We chose to use the Jupyter development environment and we employed libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-Learn to create interactive notebooks facilitating visualization of our model results.

To assess the effectiveness of our models, we employed the confusion matrix, which comprises the following essential components: true positives (TP), representing the number of correctly predicted positive instances where the model identifies cases as positive. true negatives (TN), denoting the number of correctly predicted negative instances where the model accurately identifies cases as negative. False positives (FP), indicating the number of negative instances incorrectly classified as positive, while false negatives (FN) signify the number of positive instances erroneously classified as negative [29], [30]. Using these components, along with several standard machine learning metrics, we were able to comprehensively assess the proficiency of our models in predicting breast cancer. The following metrics were used in our study:
Accuracy measures the ratio of accurately predicted instances to the total number of instances [29], [30].

$$accuracy = \frac{TP+TN}{TP+TN+ FP+FN} \tag{1}$$

Precision determines the ratio of positive observations correctly identified among all observations identified as positive [29], [30].

$$precision = \frac{TP}{TP + FP} \tag{2}$$

Recall determines the ratio of positive observations identified among all possible positive observations [29], [30].

$$recall = \frac{TP}{TP + FN} \tag{3}$$

The F1-score is a metric that balances both precision and recall [29], [30].

$$F1 - score = 2 * \frac{precision*recall}{precision+recall} \tag{4}$$

Finally, AUC-ROC measures a model's ability to distinguish between classes by computing the area under the ROC curve [29], [30].

The performance of each algorithm across different phases has been carefully reviewed and is presented in detail in Tables 2 to 6. These tables provide a comprehensive view of each algorithm's performance through the various stages of analysis, enabling a thorough assessment of their effectiveness in predicting breast cancer. While previous research has extensively explored machine learning models and feature selection methods for breast cancer prediction, they often overlook the effectiveness of advanced data balancing techniques like SMOTE and ADASYN. These techniques significantly improve model sensitivity and specificity by addressing dataset imbalance, where minority class samples are underrepresented. Moreover, integrating these data balancing techniques with feature selection methods like ACO enhances predictive modeling effectiveness by ensuring representative feature selection and optimizing model performance.

Table 2. Original dataset

| | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | AUC-ROC (%) |
|---|---|---|---|---|---|
| DT | 51.43 | 50.00 | 64.71 | 56.41 | 51.80 |
| LR | **68.57** | **65.00** | **76.47** | **70.27** | **68.79** |
| KNN | 40.00 | 40.00 | 47.06 | 43.24 | 40.20 |
| SVM | 62.86 | 64.29 | 52.94 | 58.06 | 62.58 |
| RF | 65.71 | 64.71 | 64.71 | 64.71 | 65.69 |

Table 3. Balanced SMOTE dataset

| | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | AUC-ROC (%) |
|---|---|---|---|---|---|
| DT | 64.10 | 60.00 | 66.67 | 63.16 | 64.29 |
| LR | 76.92 | 71.43 | **83.33** | 76.92 | 77.38 |
| KNN | 58.97 | 53.85 | 77.78 | 63.64 | 60.32 |
| SVM | **79.49** | **75.00** | **83.33** | **78.95** | **79.76** |
| RF | 71.79 | 65.22 | **83.33** | 73.17 | 72.62 |

Table 4. Balanced ADASYN dataset

| | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | AUC-ROC (%) |
|---|---|---|---|---|---|
| DT | 64.86 | 61.11 | 64.71 | 62.86 | 64.85 |
| LR | 72.97 | 66.67 | **82.35** | **73.68** | 73.68 |
| KNN | 75.68 | 78.57 | 64.71 | 70.97 | 74.85 |
| SVM | **78.38** | **84.62** | 64.71 | 73.33 | **77.35** |
| RF | 71.79 | 64.71 | 68.75 | 66.67 | 71.33 |

Table 5. Balanced SMOTE dataset + ACO

| | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | AUC-ROC (%) |
|---|---|---|---|---|---|
| DT | 76.92 | 69.57 | 88.89 | 78.05 | 77.78 |
| LR | 82.05 | 78.95 | 83.33 | 81.08 | 82.14 |
| KNN | 82.05 | 73.91 | **94.44** | 82.93 | 82.94 |
| SVM | **89.74** | **88.89** | 88.89 | **88.89** | **89.68** |
| RF | 84.62 | 80.00 | 88.89 | 84.21 | 84.92 |

Table 6. Balanced ADASYN dataset + ACO

| | Accuracy (%) | Precision (%) | Recall (%) | F1-score (%) | AUC-ROC (%) |
|---|---|---|---|---|---|
| DT | 78.38 | 73.68 | 82.35 | 77.78 | 78.68 |
| LR | 79.49 | 81.25 | 72.22 | 76.47 | 78.97 |
| KNN | 75.68 | 66.67 | **94.12** | **78.05** | 77.06 |
| SVM | **81.08** | **85.71** | 70.59 | 77.42 | **80.29** |
| RF | 76.92 | 68.42 | 81.25 | 74.29 | 77.58 |

For the original dataset, summarized in Table 2, the logistic regression algorithm showed the best performance in all the metrics evaluated, yielding an accuracy of 68.57%, a precision of 65.00%, a recall of 76.47%, an F1-score of 70.27%, and an AUC-ROC of 68.79%. After applying the SMOTE technique to balance the dataset, Table 3 reveals a significant improvement in all evaluation metrics, with an average increase of 12%. This time, SVM outperformed the other algorithms by recording the best values across all metrics, achieving an accuracy of 79.49%, a precision of 75.00%, a recall of 83.33%, an F1-score of 78.95%, and an AUC-ROC of 79.76%. It is also notable that RF and logistic regression achieved the same highest score for precision, with 83.33%. After applying the ADASYN method to balance the dataset, the results presented in Table 4 also demonstrate a significant improvement in all evaluation metrics, with an average increase of 14%. This time, SVM outperformed the other algorithms by recording the best values in several metrics, achieving an accuracy of 78.38%, a precision of 84.62%, and an AUC-ROC of 77.35%. Regarding recall and F1-score, logistic regression recorded the leading performance, with values of 82.35% and 73.68%, respectively. After the application of SMOTE data balancing combined with the ACO method for feature selection, the results presented in Table 5 reveal the predominance of SVM, which demonstrated outstanding performance. Notably, the ACO method for feature selection identified the following features as the most influential for SVM: ('Age', 'BMI', 'HOMA', 'Resistin'). The SVM algorithm registered a remarkable accuracy of 89.74%, a precision of 88.89%, an F1-score of 88.89%, and an AUC-ROC of 89.68%. In terms of recall, the highest score was recorded by KNN, reaching a score of 94.44%. For KNN, the ACO method identified the following features as the most influential: ('Age', 'Glucose', 'Leptin', 'Resistin'). After applying ADASYN data balancing combined with the ACO method for feature selection, the results presented in Table 6 once again highlight the supremacy of SVM. The SVM algorithm performed remarkably well, with an accuracy of 81.08%, a precision of 85.71%, and an AUC-ROC of 80.29%. Concerning the recall and the F1-score, the best results were obtained by KNN, with values of 94.12% and 78.05%, respectively. Notably, for SVM, the ACO method identified the following best features: ('BMI', 'Insulin', 'Resistin'). Similarly, for KNN, the ACO method identified the following best features: ('Age', 'BMI', 'Glucose', 'Resistin').

Our study reveals significant improvements in predictive performance after applying SMOTE, ADASYN and ACO techniques. Notably, the SVM algorithm consistently outperformed other algorithms across different phases, demonstrating the highest accuracy (89.74%), precision (88.89%), AUC-ROC (89.68%), and F1-score (88.89%) when utilizing features ('Age', 'BMI', 'HOMA', 'Resistin') in combination with the SMOTE technique. The top selected features ('Age', 'BMI', 'HOMA', 'Resistin') represent a comprehensive set of factors associated with breast cancer risk and pathogenesis. Age is a well-established risk factor, with incidence increasing with age. BMI reflects overall adiposity and is linked to postmenopausal breast cancer risk. HOMA measures insulin resistance, a factor implicated in breast cancer development. Resistin, an adipokine associated with obesity and insulin resistance, plays a role in breast cancer pathogenesis. These features provide insights into the complex interplay of metabolic and hormonal factors in breast cancer diagnosis. By comparing our findings with previous research, we highlight the superior performance of our methodology and its potential to contribute to more accurate breast cancer prediction models. While our study comprehensively evaluates the effectiveness of SMOTE, ADASYN and ACO techniques, additional research is needed to validate their performance across diverse datasets and clinical settings. Moreover, the generalizability of our findings may be restricted by the specific characteristics of the Coimbra dataset, suggesting the need for further investigation using larger and more diverse datasets. Additionally, additional studies are required to explore the impact of these techniques on specific subtypes of breast cancer and their potential to improve personalized treatment strategies. In summary, our study showcases the efficacy of SMOTE, ADASYN, and ACO techniques in improving the performance of machine learning algorithms for breast cancer prognosis. These results carry significant implications for enhancing the accuracy and reliability of breast cancer diagnostics, ultimately contributing to improved patient outcomes and personalized treatment strategies. Figure 2 provides a graphical representation of all the discussed metrics.

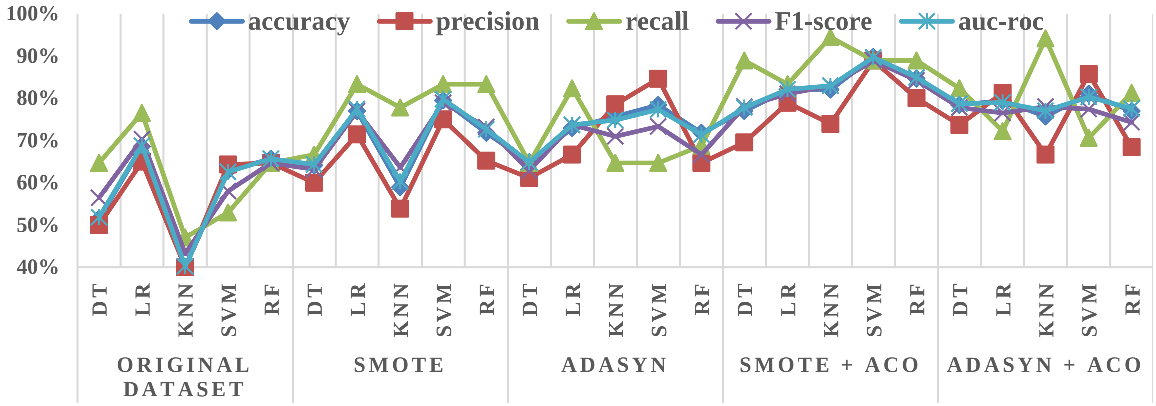

Figure 2. Metrics of different algorithms

## 4. CONCLUSION

In conclusion, our study presents significant implications for both the research field and the broader community. In the research field, our findings highlight the efficiency of machine learning models, particularly in the context of early breast cancer detection using the Coimbra dataset. By demonstrating the efficacy of various models and the critical role of techniques such as SMOTE, ADASYN and ACO in addressing data imbalance and feature selection challenges, our research contributes to advancing the understanding of predictive modeling in healthcare. Additionally, our study allows for the determination of the features most implicated in breast cancer, providing valuable insights for future research in this area. Notably, the SVM algorithm demonstrated outstanding performance with an accuracy of 89.74%, an F1-score of 88.89%, and an AUC-ROC of 89.68%. Regarding the community, our findings have tangible implications for improving healthcare outcomes related to breast cancer. The integration of advanced machine learning techniques offers the potential to enhance screening processes, enable early detection and optimize treatment strategies. This has direct implications for patients, as early detection can lead to more timely interventions, improved prognosis and better overall quality of life. Moreover, our study emphasizes the importance of interdisciplinary collaboration between healthcare professionals and machine learning experts. By fostering such collaborations, we can leverage expertise from both fields to drive innovations in breast cancer detection and management, ultimately benefiting patients and communities worldwide.

## ACKNOWLEDGEMENTS

The authors wish to express their gratitude to the professors of the MIET Laboratory at the Faculty of Sciences and Techniques, Hassan First University, Settat, for their invaluable support and guidance, which significantly contributed to the success of this research.

## REFERENCES

[1] B. Smolarz, A. Z. Nowak, and H. Romanowicz, “Breast cancer—epidemiology, classification, pathogenesis and treatment (review of literature),” *Cancers*, vol. 14, no. 10, p. 2569, May 2022, doi: 10.3390/cancers14102569.

[2] “Breast Cancer,” *Atlas of Clinical Positron Emission Tomography, Second edition*. Accessed: Jun. 08, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[3] O. Ginsburg *et al*., “Breast cancer early detection: a phased approach to implementation,” *Cancer*, vol. 126, no. S10, pp. 2379–2393, Apr. 2020, doi: 10.1002/cncr.32887.

[4] C. J. Haug and J. M. Drazen, “Artificial Intelligence and machine learning in clinical medicine, 2023,” *New England Journal of Medicine*, vol. 388, no. 13, pp. 1201–1208, Mar. 2023, doi: 10.1056/nejmra2302038.

[5] L. Rubinger, A. Gazendam, S. Ekhtiari, and M. Bhandari, “Machine learning and artificial intelligence in research and healthcare,” *Injury*, vol. 54, pp. S69–S73, May 2023, doi: 10.1016/j.injury.2022.01.046.

[6] B. S. Abunasser, M. R. J. AL-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, “Literature review of breast cancer detection using machine learning algorithms,” in *AIP Conference Proceedings, AIP Publishing*, 2023, doi: 10.1063/5.0133688.

[7] A. Ghavidel and P. Pazos, “Machine learning (ML) techniques to predict breast cancer in imbalanced datasets: a systematic review,” *Journal of Cancer Survivorship*, Sep. 2023, doi: 10.1007/s11764-023-01465-3.

[8] A. A. Aouragh and M. Bahaj, “Advancing breast cancer diagnosis with machine learning: exploring data balancing, feature selection and bayesian optimization,” in *2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, IEEE, Nov. 2023, doi: 10.1109/cloudtech58737.2023.10366058.

[9] B. Sahu and A. Panigrahi, “Efficient role of machine learning classifiers in the prediction and detection of breast cancer,” *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3545096.

[10] A. A. Jasim, A. A. Jalal, N. M. Abdulateef, and N. A. Talib, “Effectiveness evaluation of machine learning algorithms for breast cancer prediction,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 3, pp. 1516–1525, Jun. 2022, doi: 10.11591/eei.v11i3.3621.

[11] S. H. Abdulla, A. M. Sagheer, and H. Veisi, “Breast cancer segmentation using K-means clustering and optimized region-growing technique,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 1, pp. 158–167, Feb. 2022, doi: 10.11591/eei.v11i1.3458.

[12] C. Aroef, Y. Rivan, and Z. Rustam, “Comparing random forest and support vector machines for breast cancer classification,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, p. 815, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14785.

[13] M. U. Ghani, T. M. Alam, and F. H. Jaskani, “Comparison of classification models for early prediction of breast cancer,” in *2019 International Conference on Innovative Computing (ICIC)*, IEEE, Nov. 2019, doi: 10.1109/icic48496.2019.8966691.

[14] A. K. Mishra, P. Roy, and S. Bandyopadhyay, “Genetic Algorithm based selection of appropriate biomarkers for improved breast cancer prediction,” in *Intelligent Systems and Applications, Springer International Publishing*, 2019, pp. 724–732, doi: 10.1007/978-3-030-29513-4_54.

[15] R. K. Barwal, N. Raheja, M. Bhiyana, and D. Rani, “Machine learning-based hybrid recommendation (SVOF-KNN) model for breast cancer coimbra dataset diagnosis,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 1s, pp. 23–42, Jan. 2023, doi: 10.17762/ijritcc.v11i1s.5991.

[16] T. Khatun *et al*., “Performance analysis of breast cancer: a machine learning approach,” in 2021 *Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, Sep. 2021, doi: 10.1109/icirca51532.2021.9544879.

[17] S. Nanglia, M. Ahmad, F. A. Khan, and N. Z. Jhanjhi, “An enhanced predictive heterogeneous ensemble model for breast cancer prediction,” *Biomedical Signal Processing and Control*, vol. 72, p. 103279, Feb. 2022, doi: 10.1016/j.bspc.2021.103279.

[18] A. Rasool, C. Bunterngchit, L. Tiejian, M. R. Islam, Q. Qu, and Q. Jiang, “Improved machine learning-based predictive models for breast cancer diagnosis,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 6, p. 3211, Mar. 2022, doi: 10.3390/ijerph19063211.

[19] S. Rani, T. Ahmad, and S. Masood, “Comparative analysis of breast and prostate cancer prediction using machine learning techniques,” in *Lecture Notes in Networks and Systems, Springer Nature Singapore*, 2022, pp. 643–650, doi: 10.1007/978-981-19-2821-5_54.

[20] G. Alfian *et al*., “Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method,” *Computers*, vol. 11, no. 9, p. 136, Sep. 2022, doi: 10.3390/computers11090136.

[21] F. Patrcio *et al*., “Breast cancer Coimbra,” *UCI Machine Learning Repository*, 2018, doi: 10.24432/C52P59.

[22] S. Susan and A. Kumar, “The balancing trick: optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art,” *Engineering Reports*, vol. 3, no. 4, Oct. 2020, doi: 10.1002/eng2.12298.

[23] S. M. J. Moghaddam and A. Noroozi, “A novel imbalanced data classification approach using both under and over sampling,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2789–2795, Oct. 2021, doi: 10.11591/eei.v10i5.2785.

[24] I. Dey and V. Pratap, “A comparative study of SMOTE, Borderline-SMOTE and ADASYN oversampling techniques using different classifiers,” in 2023 3rd *International Conference on Smart Data Intelligence (ICSMDI)*, IEEE, Mar. 2023, doi:

10.1109/icsmdi57622.2023.00060.

[25] N. Nayar, S. Gautam, P. Singh, and G. Mehta, "Ant colony optimization: a review of literature and application in feature selection," in *Lecture Notes in Networks and Systems*, *Springer Nature Singapore*, 2021, pp. 285–297, doi: 10.1007/978-981-33-4305-4_22.

[26] H. Almazini, K. R. Ku-Mahamud, and H. F. Almazini, "Enhanced feature clustering method based on ant colony optimization for feature selection," *International Journal of Advances in Intelligent Informatics*, vol. 9, no. 1, p. 79, Mar. 2023, doi: 10.26555/ijain.v9i1.987.

[27] I. Ibrahim and A. Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 10–19, Mar. 2021, doi: 10.38094/jastt20179.

[28] A. A. Aouragh, M. Bahaj, and N. Gherabi, "Comparative study of dimensionality reduction techniques and machine learning algorithms for Alzheimer's disease classification and prediction," in 2022 IEEE 3rd *International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, IEEE, Dec. 2022, doi: 10.1109/icecocs55148.2022.9983211.

[29] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, Jul. 2020, doi: 10.1016/j.aci.2018.08.003.

[30] G. Varoquaux and O. Colliot, "Evaluating machine learning models and their diagnostic value," in *Neuromethods, Springer US*, 2023, pp. 601–630, doi: 10.1007/978-1-0716-3195-9_20.

## BIOGRAPHIES OF AUTHORS

**Abd Allah Aouragh** is currently pursuing his Ph.D. at the MIET Laboratory, Faculty of Sciences and Techniques, Hassan First University, Settat, Morocco. His research is focused on advancing medical diagnosis support systems, with a particular emphasis on leveraging machine learning, deep learning and computer vision techniques to develop innovative solutions. He can be contacted at email: abdallahaouragh@gmail.com.

**Mohamed Bahaj** received his Ph.D. in mathematics and computer science from the University Hassan 1st, Morocco and currently serves as a Full Professor in the Department of Mathematics and Computer Sciences at the University Hassan 1st, Faculty of Sciences and Technology, Settat, Morocco. With a robust academic background, he has contributed over 60 peer-reviewed papers, spanning areas such as intelligent systems, ontologies engineering, partial and differential equations, numerical analysis and scientific computing. He has provided valuable peer review services for various journals and mentored several Ph.D. students in computer sciences and mathematics. He actively engages in workshops, seminars and academic forums to enhance teaching methodologies and research practices. He can be contacted at email: mohamedbahaj@gmail.com.

**Fouad Toufik** received his Ph.D. in computer science from the University Hassan 1st, Settat, Morocco. He currently holds the position of Professor of Computer Sciences at the Higher School of Technology SALE, Mohammed V University, Morocco. With expertise in artificial intelligence, big data and database architectures, his research interests lie at the intersection of these fields. His academic pursuits aim to advance knowledge and innovation in these areas, contributing to the development of cutting-edge technologies. He can be contacted at email: toufik.fouad@gmail.com.