A Breast Cancer Detection Problem using various Machine Learning Techniques in the Context of Health Prediction System

RAFALIA Najat[1], ETTAZI Haitam[1], ABOUCHABAKA Jaafar[1]
[1]Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

Abstract. Today, breast cancer is one of the most common diseases that can cause certain complications, sometimes worst-case scenario is death. Thus, there is an urgent need for a diagnosis tool that can help doctors detect the disease at an early stage and recommend the necessary lifestyle changes to stop the progression of the disease; the likelihood of developing cancer at a young age has also been greatly increased by environmental changes in our everyday lives. Machine learning is an urgent need today to enhance human effort and offer higher automation with fewer errors. In this article, a breast cancer detection and prediction system is developed based on machine learning models (SVM, NB, AdaBoost). The achieved accuracies of the developed models are as follows: SVM achieved an overall score of 98.82%, NB achieved an overall score of 97.71%, and finally, AdaBoost achieved an overall score of 97.71%.

Index Terms— Machine Learning, NB, Environmental changes SVM, AdaBoost, breast cancer, detection, prediction.

1 Introduction

Distinct habitat types have distinct effects on a range of cancer risk factors (exposure to UV radiation, pollution, and habitat fragmentation). Machine learning has been used to map different habitat types in different ecosystems, as well as human-caused environmental deterioration such nuclear fallout, hydrocarbon spill pollution, and habitat fragmentation. The increasing application of machine learning in habitat mapping will advance our understanding of how cancer affects various organisms and landscapes.

Depending on the kind of habitat (exposure to ultraviolet radiation, pollution, and habitat fragmentation), a number of cancer risk variables are affected differently. In addition to mapping different habitat types across different ecosystems, machine learning has also been used to map environmental degradation brought on by humans, such as habitat fragmentation and contamination from petroleum spills. Research on how cancer affects species and ecosystems will be improved by the growing application of machine learning in habitat mapping.

The use of machine learning in disease prediction has become increasingly popular in the medical and health fields due to its ability to process large amounts of data quickly and efficiently. By using precise techniques to extract patterns from medical data such as genomic data, patient data, and medical images, machine learning can assist healthcare professionals in identifying risk factors, diagnosing diseases, and predicting their development. These models can be customized to predict the likelihood of a patient developing a particular disease, identify risk factors, diagnose diseases, predict the course of a disease, and even help select the best treatment for a particular patient.

Breast cancer is a serious and prevalent disease that affects millions of women worldwide. Early detection and diagnosis of breast cancer is critical for improving a patient's chances of recovery and clinical outcomes. Machine learning algorithms can be applied to extract significant features from medical data, such as mammography images, breast tissue biopsies, and patient history. By analysing this data, an accurate prediction model can be developed to estimate the probability of a patient developing breast cancer. These models can also help identify high-risk patients who may require closer monitoring or earlier screening.

However, it is important to note that the use of machine learning in disease prediction raises ethical and confidentiality concerns. Medical data is often sensitive and must be protected against misuse, which is why it is crucial to implement security measures to ensure data is used ethically and responsibly. This is particularly important in the case of medical data, which contains sensitive and personal information about patients. Healthcare professionals must take the necessary steps to protect the confidentiality of patient data and prevent it from being misused or falling into the wrong hands.

Therefore, the use of machine learning in disease prediction has great potential to transform the field of medicine, particularly in the early detection and diagnosis of breast cancer. However, it is essential to strike a balance between the benefits of this technology and the ethical and confidentiality concerns that come with it. By implementing proper security measures and working alongside healthcare professionals, machine learning can be used effectively to improve patient outcomes and save lives.

2 Related works

Research related to the detection of breast cancer has multiplied during the last decade. Much work has been directed towards detecting the presence of tissues breast cancer and classification of tumours. The approaches used come from several areas: probability and statistics, connectionism, as well as other tools from artificial intelligence and cognitive science. Therefore, taxonomy of recent classification approaches in the context of breast cancer has been established.

In this initial part of the literature review, we focus on probabilistic and statistical approaches and on statistical approaches used as classifiers for breast cancer detection. We focus on probabilistic and statistical approaches used as classifiers for breast cancer detection. Some Statistical and probabilistic methods appear in the literature, often proposing improved versions of classical approaches such as Bayesian networks and the karma method. Improved versions of classical approaches such as Bayesian networks and

the k rule and the nearest neighbour rule. In [2] the authors propose an approach based on a generalization of the nearest neighbour rule. Generalization of the nearest neighbour rule for classification in the context of breast cancer breast cancer screening. It is a method that represents a non-parametric classifier but whose performance depends on the vectors of the distributions of the meansU1and the covariance matrices. If, in addition, these distributions are Gaussian in nature, the performance of this approach becomes interesting. The approach has been implemented and tested on two databases 20 breast cancer databases, WDBC (Wisconsin Diagnosis Breast Cancer) and WBC (Wisconsin Breast Cancer).

The results obtained through the experiments were highlighted in comparison with those obtained using the conventional obtained by using the conventional k-nearest neighbour rule. Through the classification of the different partitions, the authors in [2] show that the method used is robust and performs better than the classical k-nearest neighbour rule than the classical k-nearest neighbour rule. The best recognition rate obtained is of 98.1% for the WDBC database and 97% with WBC. We note that the method proposed in [2] represents a classifier that is both simple to implement and of general application because it is non-parametric, but its performance depends on the vectors of mean distributions and covariance matrices. Moreover, the classification results obtained are interesting insofar as it is a binary classification binary classification (benign class/malignant class). However, in real problems related to breast cancer, we are often dealing with a larger number of tumour classes.

The authors in [1] offer three adjustments and explain why they are more effective in certain circumstances. We propose two distinct actions to achieve this. By accounting for their surroundings, the authors calculate context-dependent anities, or point anities, in the first. The second method, the conductivity method, aims to enhance the block structure of the anity matrix. We combine these two, beginning from extremely weak anities, and build a unique block-diagonal structure. When grouping spectral images in the final step, K-means is typically employed. Our third improvement is the usage of our K-lines method. On synthetic and real-world data sets, respectively, our techniques beat rival clustering algorithms.

We derive multiplicative updates from the authors' solution for the nonnegative quadratic programming issue in support vector machines in [3].

The authors of [4] demonstrated how the SVM model may be transformed into a fractional programming (FP) issue that is concave-convex or pseudo-concave.

The Markov Blanket Bayesian Classifier, introduced by the authors of [5], is a method for building probabilistic classifiers. Three different Bayesian classifiers are empirically compared to the MBBC method in this study.

Two look ahead-based approaches for decision tree induction at any time were provided by the authors in [6], allowing for trade-offs between tree quality and learning time.

A multi-expert system for the automated categorization of clustered micro-calcifications was developed by the authors in [7], and it produces results that are realistic.

In order to categorize micro-calcifications in mammougrams, the authors in [8] used two techniques, the Ant colony and the Swarm optimization.

The Ant colony's total performance was enhanced in the research by [9], and this metaheuristic outperformed its earlier iteration.

The work in [10] recommends a rule-discovery method called Ant-Miner (Ant Colony-Based Data Miner), which is along the same lines. The goal of Ant-Miner is to extract classification rules from data.

The Improved Quickreduct approach, which has been established for data preparation, is integrated with ant mining in the system that the authors of [11] have constructed. Using a shared data set, the performance of the recommended system was assessed; it exceeds the original Ant Miner method.

Finally, the author of [12] suggested and wrote a book in 1996 that was regarded as a handbook to intelligent systems and served as the motivation for several articles and research that dealt with all of the approaches employed in various types of intelligent systems.

3 Proposed Approach

Most of the chronic diseases are predicted by machine learning systems. It accepts the structured type of data as input to the machine learning model. This system is used by end-users i.e., patients/any user. In this system, the user will enter all the variant

numerical cells data in which the patient will provide throughout the given medical diagnostic record. These numerical cell's data then will be given to the machine learning model to predict the disease. Algorithms are then applied to which gives the best accuracy. The system will predict the state of the patient's tumor based on the symptoms if it's malign or benign. This system leverages Machine Learning Technology to predict diseases based on symptoms. It employs various algorithms, including Naïve Bayes for disease prediction, ADAboost for classification, logistic regression for feature extraction, and SVM for dataset partitioning. The ultimate result of this system is the breast cancer prediction generated by the most out preformed model within the proposed algorithms.

3.1 Methodology

In order to assess the performance of the experiment, we begin by defining TP (true positive), TN (true negative), FP (false positive), and FN (false negative). TP represents the number of correctly predicted results that match the criteria, TN indicates the number of results that do not match the criteria, FP represents the number of results that are incorrectly predicted to meet the criteria, and FN denotes the number of results that are incorrectly predicted to not meet the criteria. From these values, we can derive four key measurements: recall, precision, accuracy, and F1 measures, which are calculated as follows: The formulas for calculating precision, recall and F1 score is presented below in equations (1, 2, 3 and 4):

$$Precision=TP/(TP+FP) \quad (1)$$

$$Recall =TP/(TP+FN) \quad (2)$$

$$F1score=2*(Precision*Recall)/(Precision+Recall)=TP/(TP+1/2(FP+FN)) \quad (3)$$

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad (4)$$

Such that:

- TP=number of true positives
- FN=number of false negatives
- TN=number of true negatives
- FP=number of false positives

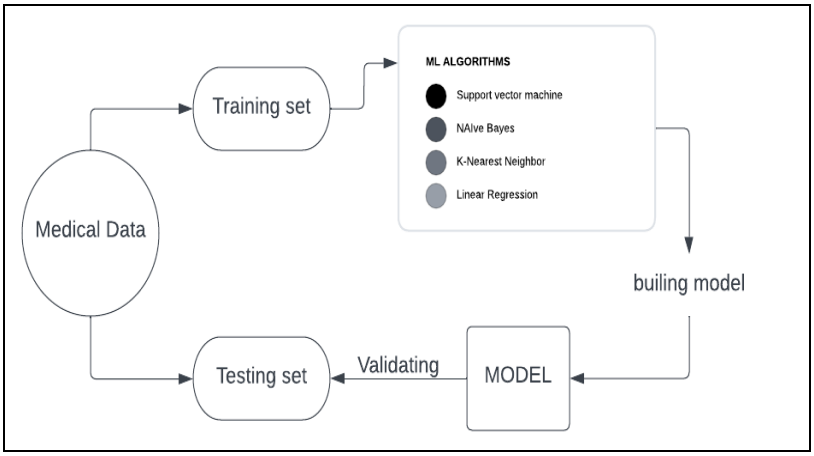

Fig. 1. System Architecture

3.2 Algorithm Techniques

-NAIVE BAYES: The Naive Bayes Classifier is a well-known machine learning algorithm commonly employed in supervised learning tasks, specifically for classification purposes. It proves to be especially valuable in text classification problems.

-SVM (Support Vector Machine): Support vector machines, or support vector machine (SVM), are supervised machine learning models focused on solving problems of discrimination and regression math.

-AdaBoost (Adaptive Boosting): is a machine learning algorithm that combines multiple weak classifiers to create a strong ensemble classifier. It is primarily used for binary classification tasks but can be extended to multi-class problems as well.

4 Numerical Evaluation

At this stage, we made a comparison between the results of the selected algorithms SVM, ADABoost, and NB.

In the table below we have compared the main algorithm's accuracy, F1-score, recall and precision in order to maintain a sustainable choice on the best preformed algorithm within the used ones.

Comparison table of used algorithm's accuracy down below:

Table 1. Classification report of the ML algorithms used

Model	AS (%)		Classification report			
			Precision	Recall	F1-score	Support
NB	97.71	B	0.98	0.98	0.98	118
		M	0.96	0.96	0.96	57
SVM	98.28	B	0.98	0.99	0.99	118
		M	0.98	0.96	0.97	57
AdaBoost	97.14	B	0.98	0.99	0.99	118
		M	0.98	0.96	0.97	57

The SVM showed the highest accuracy of 98.95%.

The summary provides an overview of the prediction outcomes in a classification problem, presenting the counts of correct and incorrect predictions categorized by each class. This summary is crucial in constructing the confusion matrix as it unveils the sources of confusion in the classification model's predictions. By analyzing the confusion matrix, one gains insight not only into the overall errors made by the classifier but, more significantly, the specific types of errors that occur.

As shown in the figure below the model scored an accuracy of 71 true negative against 2 of false positive.

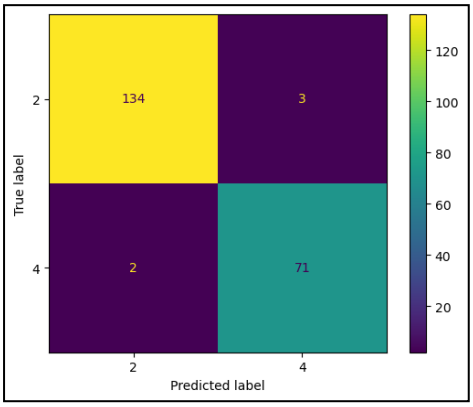

Fig. 2. Heat Map of confusion matrix of NB

As shown in the figure below the model scored an accuracy of 70 true negative against 3 of false positive.

Fig. 3. Heat Map of confusion matrix of AdaBoost

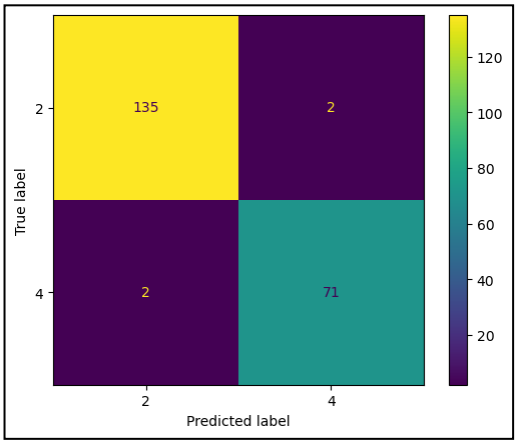

Fig. 4. Heat Map of confusion matrix of SVM

As shown in the figure above the model of SVM scored an accuracy of an overall accuracy of 98% true negative against 3 of false positive.
The Support Vector Machine scores the highest accuracy of 98.95 % in the testing data.
In the figure below a precise capture of the SVM model accuracy.

```
Model: SVM
Accuracy score: 0.9828571428571429
```

Fig. 5. SVM Accuracy

While evaluating machine learning models, accuracy is a useful metric to consider, but it's not the only factor that matters. There are several other aspects to consider when selecting an SVM model, such as precision, recall, F1 score, and the specific problem you are trying to solve.
Once the data undergoes pre-processing and cleaning, there is a possibility of obtaining a misleadingly high accuracy score. In such instances, it becomes crucial to evaluate additional metrics such as precision, recall, and F1 score to ensure that the model does not exhibit bias towards the dominant class.
it's important to assess if the SVM model is overfitting or underfitting. Overfitting occurs when the model performs well on the training set but poorly on the testing set. Underfitting, on the other hand, occurs when the model is too simple and does not capture the complexity

of the data. In our case we have on the matter by training the data on both the negative and the positive aspects of the influence of each feature on the target, which is noted on the table above in the result section.
Based on the observations from the boxplot graph, the SVM boxplot has a higher median, a narrower IQR, and fewer outliers compared to the ADABoost boxplot, it suggests that SVM is more consistent in breast cancer detection than AdaBoost. This means that SVM consistently performs better or more consistently achieves high evaluation metrics (e.g., accuracy, precision, recall) across different subsets of the dataset or during multiple runs.
This interpretation is based on the results that have already obtained through evaluation metrics for both algorithms and constructed the boxplot graph using those metrics.

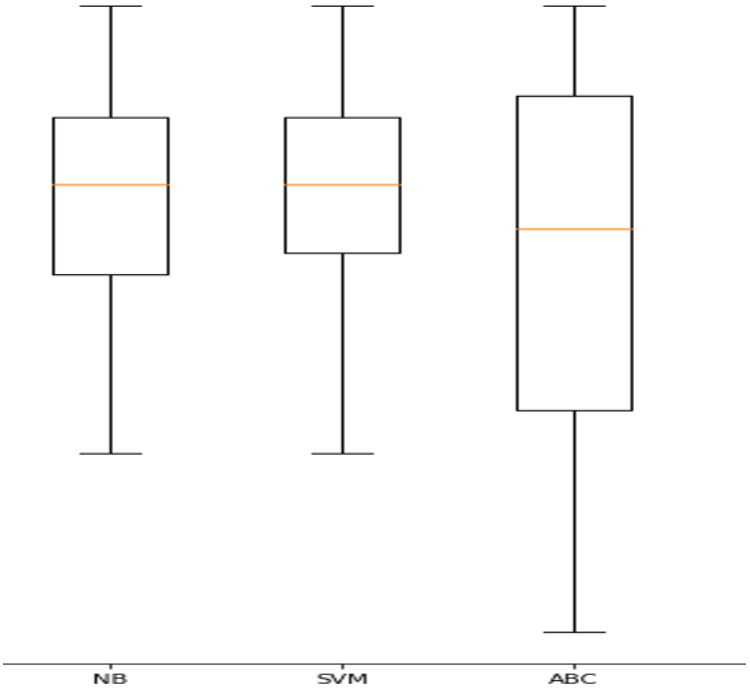

Fig. 6. Performing comparison between the SVM & NB and ADABoost based on the boxplot graph

We considered as well the specific problem we are trying to solve which is a classification problem. A high accuracy score may not be the most important metric for certain applications. As far as we are concerned, breast cancer diagnosis, aims to minimize false negatives even if it means increasing false positives.
In summary, while a 98.95% accuracy score may seem impressive, considering other metrics is very important in determining the appropriate model.

5 Discussion

Breast cancer detection is a complex problem that can be approached using different types of data. However, working with numerical data on the cells has some advantages over working with mammographic data.

The numerical data on the cells provides more detailed information about the tissue at a microscopic level. This can be useful in identifying specific characteristics of cancer cells, such as their shape, size, and texture, that may not be visible on mammographic images. Furthermore, the numerical data on the cells can be easily quantified and analyzed using statistical and machine learning techniques. These techniques can be used to identify patterns and correlations between different features of the cells and to develop predictive models for breast cancer diagnosis.

In Addition, mammographic data can be subject to noise and variability due to factors such as breast density, positioning, and imaging techniques. In contrast, numerical data on the cells can provide more consistent and reliable information, making it easier to develop accurate and reproducible models for breast cancer detection.

The collection of the numerical data can be obtained using minimally invasive techniques such as fine-needle aspiration or core biopsy. This means that the data can be collected with minimal discomfort or risk to the patient, making it a more practical and accessible approach to breast cancer diagnosis.

As a we can conclude the choice of working with the numerical data on cells rather than mammographic data which may provide important information for breast cancer detection, but numerical data on the cells has several advantages in terms of providing more detailed information, being easier to analyze, providing more consistent and reliable information, and being obtained using minimally invasive techniques.

6 Conclusion and perspectives

Breast cancer detection using machine learning is a promising and effective method to improve the early diagnosis and management of this disease. Machine Learning algorithms can be trained to recognize characteristics of malignant tumors and predict the risk of developing breast cancer.

The benefits of this method are numerous, including improved diagnostic accuracy, fewer false positives and false negatives, lower costs associated with screening tests, and more efficient healthcare management.

However, it should be noted that detecting breast cancer using machine learning should not be considered a one-size-fits-all solution. This method should be used as a complement to existing screening methods, such as mammography and clinical examination, and should be implemented in a comprehensive healthcare context.

In summary, the use of machine learning in the medical field, especially in the detection of breast cancer, remains an innovative technique, but it must be used in addition to existing methods and must be integrated into a global approach to health care.

There are various perspectives on the use of machine learning for a medical prediction system, particularly in the context of breast cancer detection. Some of these perspectives include:

-Improved accuracy: Machine learning algorithms can analyze large datasets of medical information of the patient, to detect patterns and predict the likelihood of breast cancer development. This can lead to improved accuracy in diagnosis, as well as earlier detection and treatment of the disease.

-Personalized medicine: By using machine learning algorithms to analyze patient data, medical professionals can develop personalized treatment plans that are tailored to each individual patient's needs. Which leads to more effective treatments and better patient outcomes?

-Ethical and confidentiality concerns: The use of machine learning in medical prediction systems raises ethical and confidentiality concerns regarding the collection, storage, and use of patient data. It is essential to implement proper security measures and ensure that patient data is protected and used ethically and responsibly.

-Potential biases: Machine learning algorithms can be trained on biased data, which can lead to inaccurate or discriminatory predictions. It is important to ensure that machine learning models are developed and trained on diverse and representative datasets to reduce the risk of bias.

Overall, the use of machine learning for a medical prediction system, such as breast cancer detection, has great potential to improve accuracy, personalize treatment plans, and increase efficiency in the diagnosis and treatment of the disease. However, it is crucial to address ethical, confidentiality, and bias concerns to ensure that the technology is used responsibly and for the benefit of patients.

Future study will aim to establish a set of fundamental framing principles necessary for comprehending the complex evidence supporting (or not supporting) the growing body of evidence linking specific environmental toxins to an increased risk of breast cancer. Following are the framing concepts: (a) low-dose and non-monotonic responses; (b) interactions among environmental toxins; (c) gene-environment interactions and epigenetic modifications; (d) cell-cell interactions and the Tissue Organization Field Theory; and (e) exposure period. We conclude with a visual representation of the complexity of risk factors for breast cancer, with an emphasis on environmental factors.

Also as a second perspective the aim will be to establish a set of fundamental framing principles necessary for comprehending the complex evidence supporting (or not supporting) the growing body of evidence linking specific environmental toxins to an increased risk of breast cancer. Following are the framing concepts: (a) low-dose and non-monotonic responses; (b) interactions among environmental toxins; (c) gene-environment interactions and epigenetic modifications; (d) cell-cell interactions and the Tissue Organization Field Theory; and (e) exposure period. We conclude with a visual representation of the complexity of risk factors for breast cancer, with an emphasis on environmental variables.

References

1. Fisher Igor, et Poland Jan, 2005. « Amplifying the block matrix structure for spectral clustering ». Technical Report, IOSIA, pp. 03-05.
2. Subhash c., Bagui, Sikha Bagui, Kuhu Pal, et Nikhil R, Pal, 2003. « Breast cancer detection using ranI< nearest neighbor classification rules ». Elsevier Pattern recognition, vol 36, pp. 25-34.
3. Sha, Fei et al. "Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector Machines." NIPS (2002).
4. Huang Kaizhu, Yang Haiqin, King Irwin, Lyu Michael R, Chan Laiwan, 2004. « Biased minimax probability machine for medical diagnosis ». The 8th International Symposium on Artificial Intelligence and Mathematics, pp. 4-6.
5. Madden Michael G., 2002. « Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm». CoRR, cs. LG/02l1003.
6. Esmeir Saher, Markovitch Shaul, 2004. « Lookahead-based Algorithms for anytime induction of decision trees ». 21 th international conference on machine learning, vo169, pp. 33.
7. De Santo Massimo, Molinara Mario, Tortorella Francesco, Vento Mario, 2003. « Automatic classification of c1ustered microcalsifications by a multiple expert system ». Elsevier Pattern Recognition, 36, pp. 1467-1477.
8. Karnan M., Thangavel K., Ezhilarasu P., 2008. « Ant colony optimization and a new partic1e swarm optirnizaion algorithm for classification of microcalcifications in mammograms ». The 6th International Conference on Advanced Computing and Communication.
9. Liu Bo, Abbass Hussein A, McKay Bob, 2004. « Classification mIe discovery with ant colony optimization ». IEEE Computational intelligence bulletin, Vol.3 No. 1.
10. Parepinelli R. S., Lopes H. S., Freitas A, 2002. « An Ant Colony Algorithm for Classification Rule Discovery ». Data Mining: Heuristic Approach: Idea Group Publishing, H. A a. R. S. a. C. Newton Edition.
11. Jaganathan P., Thangavel K., Pethalakshmi A, Karnan M., 2007. « Classification mIe discovery wit ant colony optimization and improved quick reduct algorithm ». lAENG International journal of computer science, 33-1, IJCS_33_L9.
12. Negnevitsky Michael. Artificial Intelligence, 2005. « A Guide to Intelligent Systems ». Addison Wesley, Second Edition.