Machine Learning for a Medical Prediction System "Breast Cancer Detection" as a use case

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

Abstract. Breast cancer is a widespread and serious illness, highlighting the importance of an early detection tool that can provide prognostic information and suggest necessary lifestyle changes to prevent its advancement, also the environmental changes in our daily life have significantly enhance the chances of getting cancer at an early stage of our life. Machine learning has become an indispensable tool in addressing this pressing need, enhancing human capabilities and offering greater automation with reduced errors. In this article, a breast cancer detection and prediction system has been created, utilizing diverse machine learning models including KNN, LR, and XGBoost.

Index Terms— Machine Learning, KNN, Environmental changes XGBoost, LR, breast cancer, detection, prediction.

1 Introduction

A variety of cancer risk factors are altered differently by habitat type (exposure to ultraviolet light, pollution, habitat fragmentation). Machine learning has been used to map habitat types in a variety of ecosystems, but also human-induced environmental degradation such as hydrocarbon spill contamination, radioactive fallout, habitat fragmentation. The rising use of machine learning in habitat mapping will improve research into the effects of cancer on species and ecosystems.

Machine learning in disease prediction has gained popularity in healthcare due to its ability to process data efficiently. It shows promise in improving early detection and diagnosis of breast cancer, a prevalent and serious disease affecting millions of women worldwide.

Machine learning algorithms analyse various medical data, such as mammography images and patient history, to extract meaningful features. These algorithms can develop accurate prediction models to estimate the likelihood of a patient developing breast cancer. They also help identify high-risk individuals who may require closer monitoring or earlier screening.

Ethical and confidentiality concerns arise when using machine learning in disease prediction. Medical data contains sensitive information that must be protected against misuse. Proper security measures are crucial to ensure responsible use of data and maintain

patient privacy. Healthcare professionals must prioritize patient confidentiality and take steps to prevent unauthorized access or data breaches.

Despite these concerns, machine learning holds great potential in improving patient outcomes and saving lives through early breast cancer detection and diagnosis. Collaboration between healthcare professionals and data scientists is essential to address ethical considerations and implement robust security measures. By striking a balance between the benefits and ethical concerns, machine learning can effectively contribute to better patient outcomes.

2 Related works

Over the past decade, research on breast cancer detection has significantly increased. Various approaches have been explored, including probabilistic and statistical methods, as well as tools from artificial intelligence and cognitive science. This literature review focuses on the classification approaches used in breast cancer detection, specifically probabilistic and statistical methods.

In recent studies, several statistical and probabilistic approaches have been proposed as classifiers for breast cancer detection. These methods often aim to improve upon traditional approaches such as Bayesian networks, the k-nearest neighbour rule, and the karma method. For instance, [2] introduced a generalized version of the nearest neighbour rule for breast cancer classification. This non-parametric classifier's performance depends on the distributions of mean vectors and covariance matrices. When these distributions follow a Gaussian nature, the approach shows promising results.

The proposed method was implemented and tested on two breast cancer databases: WDBC (Wisconsin Diagnosis Breast Cancer) and WBC (Wisconsin Breast Cancer). The experimental results were compared with those obtained using the conventional k-nearest neighbour rule. The authors demonstrated that their method is robust and outperforms the classical approach. The recognition rate achieved was 98.1% for the WDBC database and 97% for WBC. The method presented by [2] is simple to implement, non-parametric, and applicable in various scenarios. However, its performance relies on the vectors of mean distributions and covariance matrices.

It is worth noting that the classification results obtained in this study were based on a binary classification problem, distinguishing between benign and malignant classes. In real-world breast cancer problems, there are often multiple tumour classes to consider.

In summary, recent research in breast cancer detection has explored probabilistic and statistical approaches. Studies have proposed improved versions of classical methods, such as Bayesian networks and the k-nearest neighbour rule. The method introduced by [2] demonstrated robustness and outperformed traditional approaches in terms of classification accuracy. However, future research should address the challenge of dealing with a larger number of tumour classes, which is often encountered in real-world breast cancer scenarios.

The authors in [1] suggest three modifications and demonstrate how they work better in these situations. To accomplish this, we suggest two separate actions. The authors compute

point anities in the first, which we refer to as context-dependent anity, by taking into account their neighborhoods. The block structure of the anity matrix is intended to be amplified in the second approach, the conductivity method. Despite starting with very weak anities, combining these two allows us to construct a distinct block-diagonal structure. K-means is frequently used for the last stage, clustering spectral pictures. As a third enhancement, we propose to use our K-lines algorithm. Our methods outperform competing clustering algorithms on synthetic and real-world data sets, respectively.

The authors in [3] proposed a solution for the nonnegative quadratic programming problem in support vector machines, we derive multiplicative updates.

The authors in [4] showed that the SVM model can be converted into a pseudo concave problem or a concave-convex fractional programming (FP) problem.

The authors in [5] proposed an algorithm for creating probabilistic classifiers has just been proposed: the Markov Blanket Bayesian Classifier. In this study, the MBBC algorithm is empirically compared to three other Bayesian classifiers.

The authors in [6] offered two look ahead-based methods for decision tree induction at any time, allowing for trade-offs between tree quality and learning time.

The authors in [7] introduced a multi expert system for automatic classification of clustered micro-calcifications which happen to give reasonable results overall.

The authors in [8] introduced a metaheuristic approach by combining to algorithms, the Ant colony and the Swarm optimization for classifying micro-calcifications in mammougrams.

In the study conducted by [9] improved the overall functioning of the Ant colony and performed better than the original form of this metaheuristic.

In the same line, the study in [10] suggests a rule-discovery technique named Ant-Miner (Ant Colony-Based Data Miner). Ant-Miner's objective is to derive categorization rules from data.

The authors in [11] implemented a system that integrates ant mining with the Improved Quickreduct technique that has been introduced for data preparation. The suggested system's performance was evaluated using a common data set, and it outperforms the original Ant Miner technique.

Finaly, the author in [12] proposed and authored a book in 1996 which was considered to be a guide to intelligent system where a load of papers and studies would get the inspiration from this book dealing with all sort of techniques used in all forms of intelligent systems.

3 Proposed Approach

Machine learning systems play a crucial role in predicting chronic diseases. These systems rely on structured data as input and are utilized by end-users, such as patients or any other user. To make predictions, users enter relevant numerical data from their medical diagnostic records. This data is then fed into a machine learning model, which employs algorithms to achieve the highest possible accuracy in disease prediction.

In the case of tumor detection, the system utilizes machine learning technology to analyze symptoms and determine whether a patient has a tumor. Specific algorithms are employed for different tasks within the system. The K-nearest neighbors (KNN) algorithm is utilized for classification purposes. Logistic regression is employed to identify features that have the most significant impact, and XgBoost Classifier, with Dimensionality Reduction technique

The final output of the system is the disease prediction made by the machine learning model. By leveraging these algorithms and techniques, the system aims to provide accurate predictions and assist in the early detection and diagnosis of chronic diseases.

3.1 Methodology

To calculate performance evaluation in the experiment, first, we denote TP, TN, FP and FN true positive (the number of results correctly predicted as required), true negative (the number of results not required), false positive (the number of results incorrectly predicted as required), false negative (the number of results incorrectly predicted as not required) respectively. As we can obtain four measurements: recall, precision, accuracy, and F1 measures as follows: At this stage, we made a comparison between the results of the algorithms that we selected KNN, XGBoost and LR. The evaluation metrics used were: accuracy, precision, recall and F1 score. The formulas for calculating precision, recall and F1 score is presented below in equations (1, 2, 3 and 4):

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

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

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

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

Such that:

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

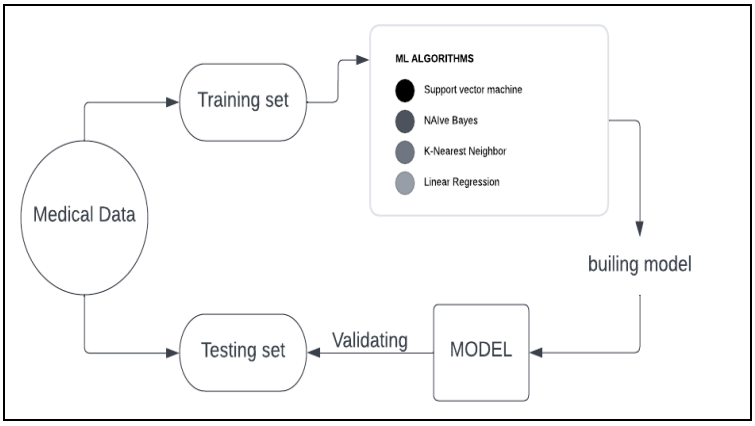

Fig. 1. System Arichitecture

3.2 Algorithm Techniques

-K Nearest Neighbour (KNN): K-nearest neighbour (KNN) algorithm is a simple and flexible machine learning algorithm commonly used for disease prediction in the healthcare system. It classifies diseases based on symptoms by identifying the nearest neighbours in the feature space. The algorithm assigns the new instance to the class of its closest neighbour(s) using a distance measure like Euclidean distance. KNN is intuitive but sensitive to outliers and requires proper selection of the K value. It offers a straightforward approach for disease prediction, utilizing feature similarity and nearest neighbour classification.
Overall, KNN is a suitable algorithm for disease prediction in the healthcare system, offering simplicity and flexibility. It can be used effectively in classifying diseases based on symptoms, utilizing the concept of feature similarity and nearest neighbour classification.

-LR (Logistic Regression): Logistic regression is a model statistic to study the relationships between a set of variables qualitative variables Xi and a qualitative variable Y. This is a model generalized linear using a logistic function as a link function.

-XGBoost Classifier: XGBoost, short for eXtreme Gradient Boosting, is a powerful library of gradient boosting algorithms designed for modern data science challenges and tools. It offers several notable advantages, including high scalability and parallelizability, fast execution speed, and superior performance compared to other algorithms. XGBoost also employs a regularized model formulation to address overfitting, resulting in improved overall performance.

The diagram illustrates the workflow of XGBoost, a popular machine learning algorithm. The shaded area represents the data used for both training and testing. Within the dashed lines, there are boxes representing the training and testing procedures, with "T" denoting trees and "GBM" representing gradient boosting machines.

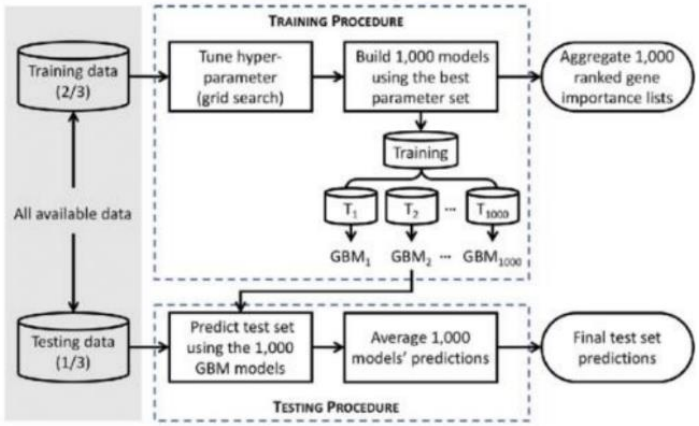

Fig. 2. XGBoost Classifier

The outputs from XGBoost are depicted by the two oval boxes on the right, positioned outside of the dashed box. The diagram provides a visual representation of the overall process and flow of XGBoost, highlighting the various steps involved in training and testing the model.

4 Numerical Evaluation

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
XGB	96.66	B	0.97	0.98	0.97	137
		M	0.96	0.95	0.95	73
KNN	98.00	B	0.98	0.99	0.99	118
		M	0.98	0.96	0.97	57
LR	97.00	B	0.97	0.99	0.98	118
		M	0.98	0.93	0.95	57

In table1 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.

4.1 Confusion matrix

The confusion matrix is a summary of prediction results in a classification problem. It provides the count values of correct and incorrect predictions, broken down by each class. This matrix is crucial as it illustrates the specific ways in which a classification model

becomes confused when making predictions. By analyzing the confusion matrix, one gains insight not only into the overall errors made by the classifier but also into the types of errors that occur. It helps in understanding the performance and effectiveness of the classification model.

Classification Rate or Accuracy is typically calculated as the ratio of the number of correct predictions to the total number of predictions made by a classification model. Mathematically, it can be expressed as:

Accuracy= (TP+TN) / (TP+TN+FP+FN)

	Predicted **No**	Predicted **Yes**
Actual **No**	**TP** True Positive	**FN** False Negative
Actual **Yes**	**FP** False Positive	**TN** True Negative

Fig. 3. Confusion Matrix

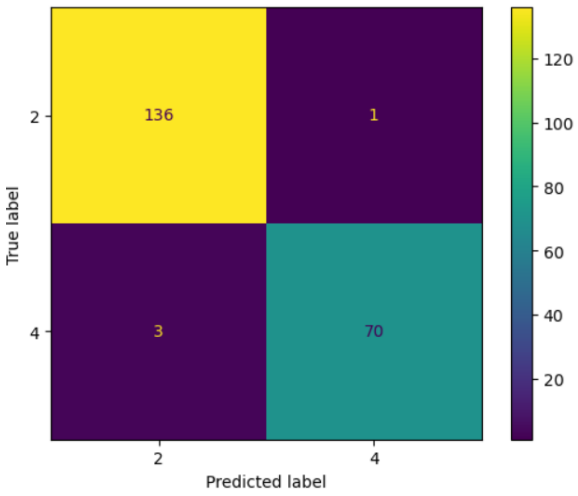

Fig. 4. Heat Map of confusion matrix of KNN

The KNN model is giving 3 type II errors and it is best.

By choosing KNN as the best model despite having some Type II errors, it suggests that the model focuses on achieving higher sensitivity and minimizing the risk of missing potential cases of breast cancer.

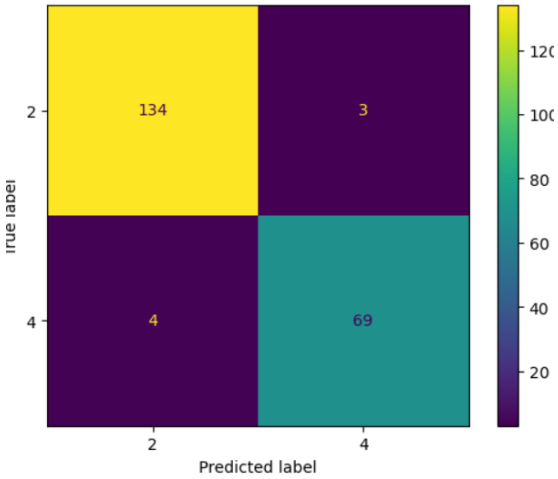

Fig. 5. Heat Map of confusion matrix of XGB

The XGB model is giving 4 cases of type II error.

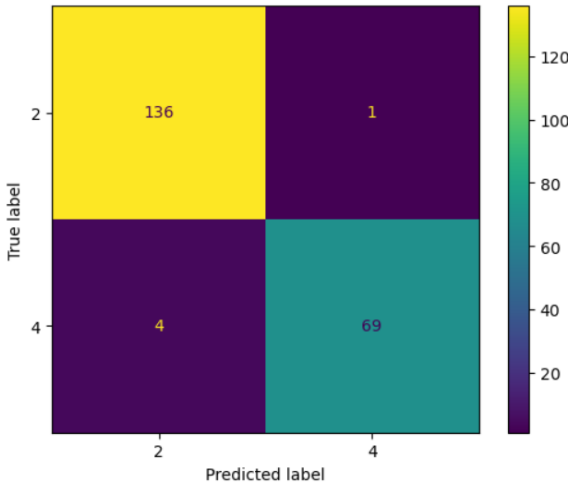

Fig. 6. Heat Map of confusion matrix of LR

The LR model is giving 4 cases of type II error.

5 Discussion

Breast cancer detection is a complex task that can utilize various types of data. However, utilizing numerical data on the cells offers several advantages over working with mammographic data. Numerical data provides in-depth information about the tissue at a microscopic level, enabling the identification of specific characteristics of cancer cells like shape, size, and texture, which may not be visible on mammograms. Additionally, numerical data can be easily quantified and analyzed using statistical and machine learning

techniques, facilitating the identification of patterns and correlations between different cell features and the development of predictive models for breast cancer diagnosis.

On the other hand, mammographic data can be prone to noise and variability caused by factors like breast density, positioning during imaging, and variations in imaging techniques. In contrast, numerical data on the cells offers more consistent and reliable information, making it easier to develop accurate and reproducible models for breast cancer detection.

Furthermore, collecting numerical data can be achieved through minimally invasive techniques such as fine-needle aspiration or core biopsy. This means that patients can undergo data collection with minimal discomfort or risk, making it a practical and accessible approach for breast cancer diagnosis.

To sum up, choosing to work with numerical data on cells rather than mammographic data provides valuable information for breast cancer detection. Numerical data offers advantages in terms of providing detailed information, ease of analysis, consistency, reliability, and the use of minimally invasive techniques for data collection.

6 Conclusion and perspectives

Breast cancer detection can be significantly enhanced through the application of machine learning techniques. Machine learning algorithms can be trained to identify specific characteristics of malignant tumors and predict the risk of developing breast cancer.

The advantages of using machine learning for breast cancer detection are manifold. It offers improved diagnostic accuracy, reducing the occurrence of false positives and false negatives. This leads to more reliable identification of breast cancer cases. Additionally, employing machine learning can potentially lower the costs associated with screening tests, making healthcare more cost-effective. Moreover, it can contribute to more efficient healthcare management by assisting medical professionals in making informed decisions and prioritizing cases based on risk assessment.

However, it is important to emphasize that machine learning should not be considered a standalone solution for breast cancer detection. Instead, it should be utilized in conjunction with existing screening methods like mammography and clinical examination. The integration of machine learning into a comprehensive healthcare framework ensures that it complements other diagnostic tools and contributes to a more holistic approach to breast cancer detection and management.

There are several perspectives to consider when examining the use of machine learning in a medical prediction system, particularly in the context of breast cancer detection:

Improved Accuracy: Machine learning algorithms can analyze large sets of patient medical data to identify patterns and predict the likelihood of breast cancer development. This can lead to improved accuracy in diagnosis, enabling earlier detection and treatment of the disease.

Personalized Medicine: By leveraging machine learning algorithms to analyze patient data, healthcare professionals can develop personalized treatment plans tailored to each individual patient's specific needs. This approach can result in 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 surrounding the collection, storage, and utilization of patient data. It is imperative to implement robust security measures and ensure that patient data is protected and handled ethically and responsibly.

Potential Biases: Machine learning algorithms can be trained on biased datasets, which may lead to inaccurate or discriminatory predictions. It is crucial to ensure that machine learning models are developed and trained on diverse and representative datasets to mitigate the risk of bias and promote fairness.

As a future work, We will aspire to provide a set of essential framing concepts required for understanding the complicated evidence supporting (or not supporting) an increasing understanding of the data indicating certain environmental toxicants in an elevated risk of developing breast cancer. These framing concepts are as follows: (a) low-dose and non-monotonic responses; (b) interactions amongst environmental toxins; (c) gene-environment interactions and epigenetic alterations; (d) cell-cell interactions and the Tissue Organization Field Theory; and (e) exposure time. We finish with a graphical model of the complexity of factors influencing breast cancer risk, with a focus on environmental influences.

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.