Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 191 (2021) 487–492

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Edge IA-IoT for Smart Agriculture (SA2IOT)
August 9-12, 2021, Leuven, Belgium

Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis

Mohammed Amine Naji [a,*], Sanaa El Filali[b]
Kawtar Aarika[c], EL Habib Benlahmar[d], Rachida Ait Abdelouhahid[e], Olivier Debauche[f]
[a,b,c,d,e] *Faculty of Science Ben M'sik, Hassan 2 University, Casablanca, Morocco*
[f] Faculty of Engineering, University of Mons, Mons, Belgium

Abstract

Each year number of deaths is increasing extremely because of breast cancer. It is the most frequent type of all cancers and the major cause of death in women worldwide. Any development for prediction and diagnosis of cancer disease is capital important for a healthy life. Consequently, high accuracy in cancer prediction is important to update the treatment aspect and the survivability standard of patients. Machine learning techniques can bring a large contribute on the process of prediction and early diagnosis of breast cancer, became a research hotspot and has been proved as a strong technique. In this study, we applied five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbours (KNN) on the Breast Cancer Wisconsin Diagnostic dataset, after obtaining the results, a performance evaluation and comparison is carried out between these different classifiers. The main objective of this research paper is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective whit respect to confusion matrix, accuracy and precision. It is observed that Support vector Machine outperformed all other classifiers and achieved the highest accuracy (97.2%).All the work is done in the Anaconda environment based on python programming language and Scikit-learn library.

Keywords: Breast Cancer; Prediction; Diagnostic ; SVM; Random Forest; Logistic regression ; C4.5; k-NN; Accuracy; Precision;

* Corresponding author. Tel.: Tel.: +212-634-32-83-78;
E-mail address: aminenaji55@gmail.com

10.1016/j.procs.2021.07.062

1. Introduction

Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide, according to statistics released by the International Agency for Research on Cancer (IARC) in December 2020. In the past two decades, the overall number of people diagnosed with cancer nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020 [1]. Today, one in 5 people worldwide will develop cancer during their lifetime. Projections suggest that the number of people being diagnosed with cancer will increase still further in the coming years, and will be nearly 50% higher in 2040 than in 2020. The number of cancer deaths has also increased, from 6.2 million in 2000 to 10 million in 2020. More than one in six deaths is due to cancer. This reinforces the need to invest in both the fight against cancer and cancer prevention. The successful introduction of information and communication technologies (ICT) in medical practice is an important stake in the renovation of the health system and more precisely in cancer care. Actually, Big data has revolutionized the size of data and also creating value from it Big data has made a big change in BI by analyzing large amount of unstructured, heterogeneous, non-standard and incomplete healthcare data. It does not only forecast but also helps in decision making and is increasingly noticed as breakthrough in ongoing advancement with the goal is to improve the quality of patient care and reduces the healthcare cost. Data mining algorithms applied in healthcare industry play a significant role due to their high performance in predicting, diagnosis of the diseases, reducing costs of medicine, making real time decision to save people's lives. The Most common Data mining modeling goals are classification and prediction which uses several algorithms for the prediction of breast cancer. This paper mainly gives a comparison between the performance of five classifiers: Support Vector Machine (SVM)), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbors (KNN Network) which according to research community are among the most influential data mining algorithms and among the top 10 data mining algorithms [2]. Our objective is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision and sensitivity. The rest of this paper is organized as follows .section 2 introduces methods and results of previous research on breast cancer diagnosis. Section 3 describes the proposed methodology for our work. Section 4 presents and explains in detail the experiments results. Section 5 concludes the paper.

2. Related Works

A large number of machine learning algorithms are available for prediction and diagnosis of breast cancer. Some of the machine learning algorithm are Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbors (KNN Network) etc. A lot of researcher have realized research in breast cancer by using several dataset such as using SEER dataset, Mammogram images as dataset, Wisconsin Dataset and also dataset from various hospitals. By exploiting these dataset authors extract and select various features and complete their research. These are some significant research. The author Sudarshan Nayak [3], demonstrates the use of various supervised machine learning algorithms in classification of breast cancer from using 3D images and find out that SVM is the best based on his overall performance. On the other side, we find that B.M. Gayathri [4], work on comparative study of Relevance vector machine which provides Low computational cost while comparing with other machine learning techniques which are used for breast cancer detection, and explain how RVM is better than other machine learning algorithms for diagnosing breast cancer even the variables are reduced and achieved 97% accuracy. Hiba Asri [5], demonstrated that Support vector Machine (SVM) proves its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate with an accuracy of 97.13%. in recent works, we find that Youness khoudfi and Mohamed Bahaj [6], similarly proposed a comparison between Machine learning algorithms and they found the SVM is the best classifier with an accuracy of 97.9% compared with K-NN, RF and NB, they are based on Multilayer perception with 5 layers and 10 times cross validation using MLP. The author Latchoumiet TP [7] Found a classification value of 98.4% proposing an optimization weighting of the particle swarm (WPSO) based on the SSVM for the classification. Ahmed Hamza Osman [8] proposed a solution for the diagnosis of Wisconsin breast cancer (WBCD) with a prediction of 99.10% found by the SVM algorithm by combining a clustering algorithm with an efficient probabilistic vector support machine. Our research is focused on assessing such machine learning algorithms and

approaches in order to conclude the best methodology for breast cancer prediction and diagnosis.

3. Methodology

The main objective of our experiment is to identify the effective and predictive algorithm for the detection of breast cancer, therefore we applied machine learning classifiers Support Vector Machine (SVM), Random Forests, Logistic Regression, Decision tree (C4.5), K-Nearest Neighbors (KNN) on Breast Cancer Wisconsin Diagnostic dataset and evaluate the results obtained to define which model provides a higher accuracy.

The proposed architecture is detailed in figure 1.

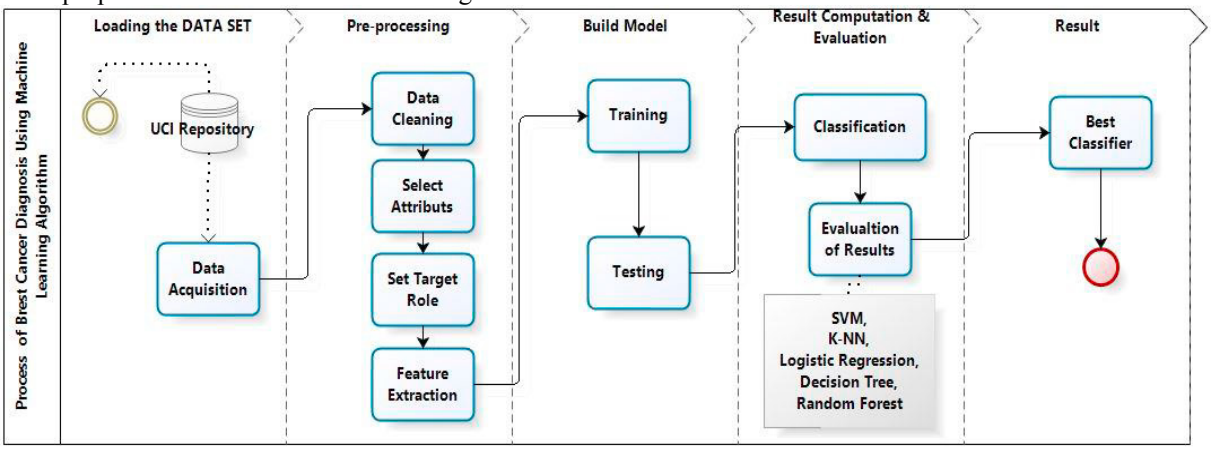

Fig. 1. Process Flow Diagram.

Our methodology begins with data acquisition followed by pre-processing, which contains four steps viz: data cleaning, select attributes, set target Role and features extraction. The prepared data is used to build machine learning algorithms that can predict the breast cancer for a new set of measurements. To evaluate the algorithms performances, we show the model new data for which we have labels. This is usually done by splitting the labeled data we have collected into two parts whit Train_test_split method. 75% of the data is used to build our machine learning model, and is called the training data or training set. 25% of the data will be used to access how well the model works and is called test data, test set. After testing the models we compare the obtained results to select the algorithm that provides the high accuracy and identify the most predictive algorithm for the detection of breast cancer.

3.1. Machine Learning Algorithms

In our project, the predictive analysis of the machine learning algorithms is achieved. The machine learning algorithms applied in our project are:

- Support Vector Machine (SVM) is a classifier which divides the datasets into classes to find a maximum marginal hyper plane (MMH) via the nearest data points [9].
- Random forests or random decision forests are an ensemble method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set
- k-Nearest Neighbors (K-NN) is a supervised classification algorithm. It takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point, which is its nearest neighbors, and has those neighbors vote [10].

- Logistic regression is a very powerful modeling tool, is a generalization of linear regression [11]. Logistic Regression is used to assess the likelihood of a disease or health condition as a function of a risk factor (and covariates). Both simple and multiple logistic regression , assess the association between independent variable(s) (Xi) -- sometimes called exposure or predictor variables — and a dichotomous dependent variable (Y) -- sometimes called the outcome or response variable. It is used primarily for predicting binary or multiclass dependent variables.
- Decision Tree C4.5 is a predictive modeling tool that can be applied across many areas. It can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions [12].

3.2. Dataset acquisition

In our study, we use Breast Cancer Wisconsin Diagnostic dataset from University of Wisconsin Hospitals Madison Breast Cancer Database [13]. The features of dataset are computed from a digitized image of a breast cancer sample obtained from fine-needle aspirate (FNA). The characteristics of the cell nuclei present in the image are determined from these features. Breast Cancer Wisconsin Diagnostic has 569 instances (Benign: 357 Malignant: 212), 2 classes (62.74% benign and 37.26% malignant), and 11 integer-valued attributes (-Id -Diagnosis -Radius -Texture -Area -Perimeter -Smoothness -Compactness -Concavity -Concave points -Symmetry -Fractal dimension).

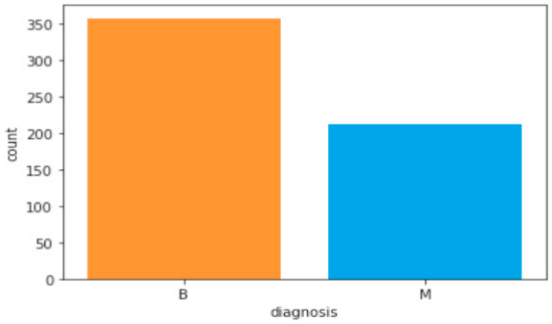

Fig. 2. WISCONSIN BREAST CANCER DIAGNOSTIC DATASETS.

3.3. Experiment Environment

All experiments on the machine learning algorithms described during this paper were conducted using Scikit-learn library and Python programming language. Scikit-learn also known as sklearn is a free software machine learning library for the Python programming language. [14] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

4. Results And Discussion

After applying Machine Learning Algorithms on Breast Cancer Wisconsin Diagnostic dataset.We used Confusion Matrix, Accuracy, Precision, Sensitivity, F1 Score, AUC as performance metrics to evaluate and compare the models and identify the best algortihm for the brest cancer Prediction. Confusion Matrix is the way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is a table with two dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, and “False Negatives (FN)”. Accuracy is most common performance metric for classification algorithms. It defined as the number of correct predictions made as a ratio of all predictions made. Precision, used in document retrievals, may be defined as the number of correct documents returned by our ML model. Sensitivity may be defined as the number of positives returned by your ML model. F1 score gives us the harmonic mean of precision and Sensitivity. Mathematically, F1 score is the weighted

average of the precision and Sensitivity.
Table 1 and figure 2 show the accuracy percentage for Wincson Breast Cancer Diagnostic datasets. From the results of training set and testing set we can see that all the classifiers have varying accuracies but SVM always has higher accuracy testing set (97.2%) than the other classifiers.

Table 1. Accuracy percentage for breast cancer diagnostic dataset.

Algorithms	Accuracy Training Set (%)	Accuracy Testing Set (%)
SVM	98.4%	97.2%
Radom Forest	99.8%	96.5%
Logistic Regression	95.5%	95.8%
Decision Tree	98.8%	95.1%
K-NN	94.6%	93.7%

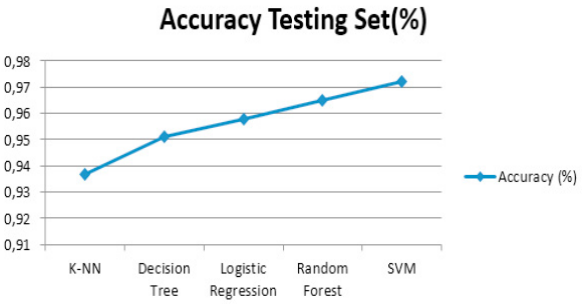

Fig. 3. Comparative graph of different classifiers

Since confusion matrices are a useful way to assess the classifier, each row in Table 2 represents the rates in an actual class while each column displays the predictions. Table 3 present the calculated performance measures of classification models based on confusion matrix results, precision sensitivity f1 score for benign and malignant.

Table 2. Confusion Matrix.

	Malignant	Benign	
SVM	201	11	Malignant
	1	356	Benign
Random Forest	196	16	Malignant
	7	350	Benign
Logistic Regression	201	11	Malignant
	5	352	Benign
C4.5	195	17	Malignant
	22	335	Benign
KNN	201	11	Malignant
	7	350	Benign

Table 3. Classifiers performances

Algorithms	Precision	Sensitivity	F-Measure	Class
SVM	0.98	0.94	0.96	Benign
	0.97	0.99	0.98	Malignant
Random Forests	0.96	0.94	0.95	Benign
	0.97	0.98	0.97	Malignant
Logistic Regression	0.98	0.91	0.94	Benign
	0.95	0.99	0.97	Malignant
Decision Tree	0.94	0.92	0.93	Benign
	0.96	0.97	0.96	Malignant
K-NN	0.92	0.91	0.91	Benign
	0.95	0.96	0.95	Malignant

Table 2 Confusion matrix shows that Support Vector Machine predicts correctly 556 cases out of 569 cases constituted of 201 malignant cases that are actually malignant and 356 benign cases that are actually benign, and 11 cases incorrectly predicted including 11 cases of malignant class predicted as benign and 1 case of benign class predicted as malignant. That is why the accuracy of Support Vector Machine is better than other classification techniques. From the results of table we can see that the percentages of precision 0.98%, sensitivity 0.94%, F-Measure 0.96% of SVM is higher than that of other classifiers. SVM always outperforms than the other classifiers in performance for two class malignant and benign in Breast Cancer Wisconsin Diagnostic dataset cancer.
The ROC curves of each machine learning algorithms are presented on Fig.4. ROC curve is an important metric for the performance of classifiers. The area under ROC curve (AUC) is computed. The area is bigger, the performance of the classifier is better. The Support Vector Machine has the highest AUC score 0.96% while the AUC score of Decison tree 0.94% is the lowest as shown in table 4.

Table 4. The Area under roc curve (AUC).

Algorithms	AUC (%)
SVM	0.966
Random Forests	0.960
Logistic Regression	0.947
Decision Tree	0.945
K-NN	0.952

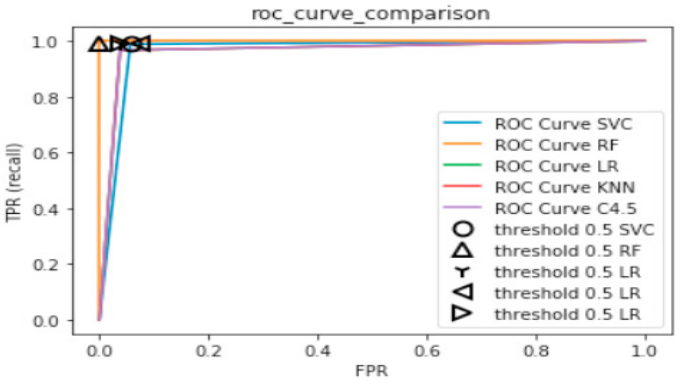

Fig. 4. ROC curve

5. **Conclusion**

On the Wisconsin Breast Cancer Diagnostic dataset (WBCD) we applied five main algorithms which are: SVM, Random Forests, Logistic Regression, Decision Tree, K-NN, calculate, compare and evaluate different results obtained based on confusion matrix, accuracy, sensitivity, precision, AUC to identify the best machine learning algorithm that are precise, reliable and find the higher accuracy. All algorithms have been programmed in Python using scikit-learn library in Anaconda environment. After an accurate comparison between our models, we found that Support Vector Machine achieved a higher efficiency of 97.2%, Precision of 97.5%, AUC of 96.6% and outperforms all other algorithms. In conclusion, Support Vector Machine has demonstrated its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of accuracy and precision. It should be noted that all the results obtained are related just to the WBCD database, it can be considered as a limitation of our work, it is therefore necessary to reflect for future works to apply these same algorithms and methods on other databases to confirm the results obtained via this database, as well as, in our future works, we plan to apply our and other machine learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy.

References

[1] ‘WHO | Breast cancer’, *WHO*. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).

[2] Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithmsdemystified/1144. Accessed December 29, 2015.

[3] S. Nayak and D. Gope, "Comparison of supervised learning algorithms for RF-based breast cancer detection," 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona, 2017, pp.

[4] B.M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Chennai, 2016, pp. 1-5.

[5] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, ‘Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis’, *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[6] Y. khoudfi and M. Bahaj, Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification, 978-1-5386- 4225-2/18/$31.00 ©2018 IEEE.

[7] L. Latchoumi, T. P., & Parthiban, “Abnormality detection using weighed particle swarm optimization and smooth support vector machine,” Biomed. Res., vol. 28, no. 11, pp. 4749–4751, 2017.

[8] A. H. Osman, “An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 158–165, 2017.

[9] Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565.

[10] Larose DT. Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.

[11] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer-Verlag;2001.

[12] Quinlan JR. C4.5: Programs for Machine Learning.; 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1.

[13] “UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set.”

[14] Fabian Pedregosa and all (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.