Journal of Theoretical and Applied Information Technology
15th February 2021. Vol.99. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BREAST CANCER DIAGNOSIS USING MACHINE LEARNING AND ENSEMBLE METHODS ON LARGE SEER DATABASE

[1]HAJAR SAOUD, [2]ABDERRAHIM GHADI, [3]MOHAMED GHAILANI

[1,2]LIST Laboratory, Abdelmalek Essaâdi University, Tangier, Morocco

[3]LabTIC Laboratory, Abdelmalek Essaâdi University, Tangier, Morocco

E-mail: [1]saoudhajar1994@gmail.com, [2]ghadi05@gmail.com, [3]ghalamed@gmail.com

ABSTRACT

Machine learning is a subdomain of artificial intelligence that has proved its performance in the medical fields, especially in the classification of the diseases. In previous researches we tried to classify breast cancer into its two categories using several machine learning algorithms, some algorithms have proved their performance but others have produced a weak accuracy. In this study, we will try to improve the accuracy of weak machine learning algorithms using the normalization/ standardization and the ensemble methods like: voting, stacking, bagging and boosting in the classification of breast cancer disease using the large SEER database and the python library. The goal of this paper is not only the improvement of the classifiers accuracy, but also the proposition of new architecture of breast cancer diagnosis based on SEER database features for predicting breast cancer in the earlier stage and with the right way. All the examined techniques have proved their performance in the improvement of the accuracy of classification of breast cancer, Specially Voting technique. It obtained the higher accuracy except the case of voting all classifiers, but it was enhanced by the normalization/ standardization of features.

Keywords: *SEER, Machine learning, Ensemble methods, Breast cancer, Diagnosis.*

1. INTRODUCTION

Breast cancer is a dangerous disease that threatens the health of women in all of the world, it touches 1 woman from 8, and it presents the second cause of mortality by cancers after lung cancer[1]. The early detection of breast cancer can decrease the rate of mortalities and increase the duration of survivability of patients. So, the development of aide-diagnosis solution has become a necessity to reduce the number of mortalities.

Machine learning techniques have approved there performance in medical field; they can be used in diagnosis of disease and prognosis also in drug development and epidemiology [2]. So, they will be effective tools in diagnosis of breast cancer.

SEER (Surveillance, Epidemiology, and End Results) program database [3] provided by National Cancer Institute (NCI), cover 34.6 % of the population of the United State of cancer incidence, it contains patient demographics, primary tumor site, tumor morphology, stage at diagnosis, first course of treatment, and cancer survivability. SEER program offer SEER*Stat software to access to the data. This database has already used in several research to develop models in cancer diagnosis, prediction recurrence and patient survivability. So, in this research we will use it to develop an ensemble classification models using the ensemble methods: voting, stacking, bagging and boosting and the 15 selected attributes for data extracted from 2008 to 2015 and 11 selected attributes for 2016 to improve the accuracy of breast cancer diagnosis and classification.

In past researches [4], [5], [6] and [7] we tried to evaluate the performance of machines learning algorithms in the diagnosis of breast cancer using non-massive databases. Some algorithms have proved their performance others have produced low accuracy. Therefore, the idea of this paper is to test the performance of machines learning techniques on large dataset like SEER database and also to improve their accuracy using ensemble methods, also we tried to show the effect of normalization and standardization of data in the improvement of the performance of some machine learning algorithms. All those experimentations will be done by the python libraries: Pandas (for data pre-processing) and Scikit-learn for the

classification and performance evaluation using the Jupyter Notebook.

The rest of this paper is structured as follows. In part two we cited some researches that have used the SEER dataset; in part three we explained our methodology and the materials used, then in part four we executed some machine learning in classification of breast cancer and also we tried to improve their performance by the techniques of normalization/standardization, then we tried also to improve their accuracy by ensemble methods in part five, a comparison of our results with existing work was done in part six and finally conclusion.

2. RELATED WORKS

Up to now, several searches have been carried out with SEER dataset and machines learning techniques, not only for the diagnosis of cancer but also in the prediction of the cancer recurrence and the duration of survivability of patients, and all of them showed the performance of machine learning techniques in the domain of cancers predictions. J48 and priority based decision tree algorithms are applied for breast cancer classification, the results show that priority based decision tree algorithm gives higher accuracy 98.51% [8]. The classification of breast cancer into two categories “Carcinoma in situ” and “Malignant potential” was made by C4.5, The accuracy obtained in training phase 94% and in testing phase 93% [9]. The three machine learning techniques Decision tree, Support Vector Machine and Random Forest are examined for the early diagnosis and prevention of the breast cancer. The original dataset was divided into 10 groups to apply the three machines learning algorithms in all of these groups. The higher accuracy was obtained by Random Forest in all of groups [10].

Three machines learning techniques, Decision Tree (DT), Support Vector Machine (SVM) and Artificial Neural Network (ANN) were performed to predict breast cancer recurrence for cancer patients. The higher accuracy was given by Decision Tree with 94.15 % followed by Support Vector Machine 91.95% then Artificial Neural Network with 90.86% [11]. The same three machines learning techniques Decision Tree (c4.3), Support Vector Machine (SVM) and Artificial Neural Network (ANN) were trained for predicting breast cancer recurrence but with higher accuracies, 93,6% for Decision Tree 94,7% for Artificial Neural Network and 95,7% for Support Vector Machine [12].

Comparative study of machines learning techniques approaches that are employed in the modeling of cancer progression with different input features, the review presents the performance of machines learning techniques in both prediction of cancer recurrence and survival [13]. An ensemble of machines learning techniques, logistic regression (LR), support vector machines (SVM), random forest (RF) and deep learning (DL), are examined to predict survival of pancreatic neuroendocrine tumors (PNETs). All algorithms gave accuracy more than 80% that is better than the AJCC stage system for PNETs cases in the SEER database[14].To predict 10-year breast cancer patient survival some machine learning algorithms are trained like Logistic Regression (LR), Naive Bayes, and C4.5 Decision Tree. The obtained accuracies are 76.29% for Logistic Regression, 59.71% for Naive Bayes, and 77.43% for C4.5 Decision Tree. Therefore, C4.5 Decision Tree proved to be the most accurate predictor of patient survival in ten years in this research [15]. Several supervised machines learning algorithms are applied to predict lung cancer patient survival, among them linear regression, Decision Trees, Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and a custom ensemble [16]. To predict 2-year colorectal cancer survivability several machines learning algorithms are used like logistic regression, random forest, AdaBoost, and neural network. The importance of ethnicity on model performance was investigated, the models proved their performance in single-ethnicity populations better than mixed-ethnicity populations [17].

3. MATERIAL AND METHODOLOGY

In this section, we will present the SEER database and methodology followed for breast cancer diagnosis using the proposed techniques.

3.1 Seer Database

The massive SEER (Surveillance, Epidemiology, and End Results) database provided by the National Cancer Institute, it collects data of cancer incidence, diagnosis, treatment, survival and mortality of all types of cancers from population-based cancer registries and it cover 34.6% of the population of the United State. The last submission of SEER database is 2019 submission, it contains the data from 1975 to 2017 and it covers more than 10,985,942 cases. We will work with the submission of 2018 and we will extract only the data of breast cancer from 2008 to 2016 to execute

the proposed techniques where the number of missing data is reduced.

3.2 Proposed Architecture

Figure 1 represents the proposed architecture:

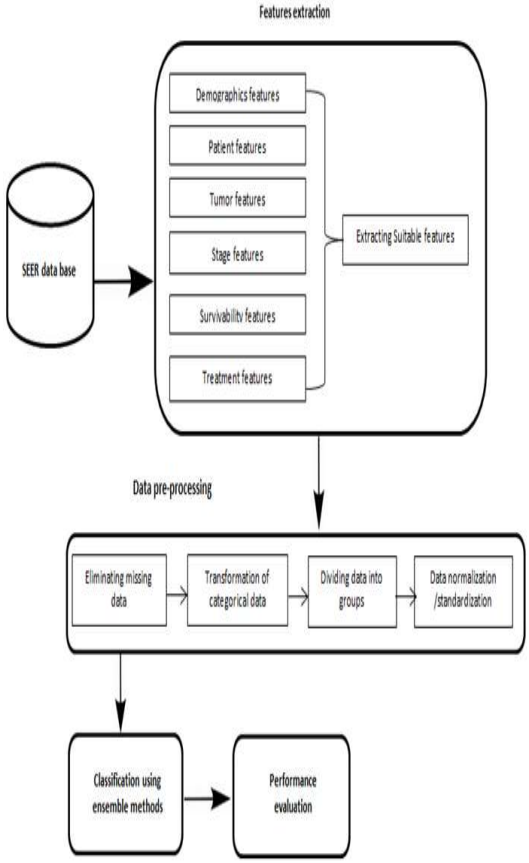

Figure 1: Proposed architecture

3.2.1 Data extraction

The data extracted from SEER database is breast cancer data for female sex and the selected years are form 2008 to 2016 in which the number of missing data is reduced. Figure 2 present the executed query in SEER*Stat software.

```
{Site and Morphology.Site recode ICD-O-3/WHO 2008} = ' Breast'
AND {Race, Sex, Year Dx, Registry, County,Sex} = ' Female'
AND {Race, Sex, Year Dx, Registry, County,Year of diagnosis} = '2008','2009','2010','2011','2012','2013','2014','2015','2016'
```

Figure 2: The executed query.

The output of this query is table of data displayed in a results matrix that has the extension (.slm), which was transformed to CSV file to execute our process. The total number of breast cancer extracted data is 699412 cases, 81386 of 2016 and 618026 of 2008-2015.

15 key attributes was selected for the years 2008 to 2015 and 11 key attributes for the year 2016, the database contain both numeric (continuous) and categorical (discrete) attributes.

Table 1: Extracted attributes

Features	Values	Description
Patient ID	Number	Number that identify a person uniquely.
Age recode with <1 year olds	19 age groups(<1 year, 1-4 years, 5-9 years, ..., 85+ years)	Age recode contain the age grouping based on age at diagnosis.
Age at diagnosis	000-130:Actual age in years 999:Unknown age	Age of the patient at diagnosis for breast cancer.
Race/ethnicity	White, Black, American Indian, Aleutian,Unknown,...	Race of the patient.
Marital status at diagnosis	Single, Married, Separated, Divorced, Widowed, Unmarried or domestic, partner, Unknown	Patient's marital status at the time of diagnosis.
Laterality	Right: origin of primary. Left: origin of primary. Only one side involved, right or left origin unspecified. Bilateral involvement, lateral origin unknown; stated to be single primary. Paired site, but no information concerning laterality.	Laterality describes the side of a paired organ or side of the body on which the reportable tumor originated.
CS extension(2004-2015)	Number	Information on extension of the tumor.
CS lymph nodes (2004-2015)	Number	Information on involvement of lymph nodes.
CS tumor size (2004-2015)	Number	Information on tumor size.
CS Tumor Size/Ext Eval	Number	

(2004-2015)		
CS Reg nodes Eval (2004-2015)	Number	The number of regional nodes evaluated.
Tumor Size Summary (2016+)	Number	In (2016+) all variables of tumor evaluation are summarized in this variable.
Primary Site	Code	This variable identifies the site in which the primary tumor originated.
First malignant primary indicator	Yes No	Variable identify if there is first malignant primary indicator.
Total number of In Situ/malignant tumors for patient	00-98: Valid values 99: (unknown)	Count the total number of cancers that patients have.
Behavior recode for analysis	In situ Malignant	Type of tumor.

3.2.2 Data preprocessing

This step is divided into four tasks: eliminating missing data, transforming categorical data to integer, dividing the data by group of years and finally data normalization and standardization to improve the accuracy of classification.

a. Eliminating missing data

The first step in data preprocessing is eliminating missing data, we used the pandas python library, the missing value of categorical data is identified by Unknown and for continuous data by 999 or 99. The Total number of data after eliminating missing values is 593004 (97548 in situ and 495456 malignant), (522483) data of 2008-2015 and (70521) data of 2016.

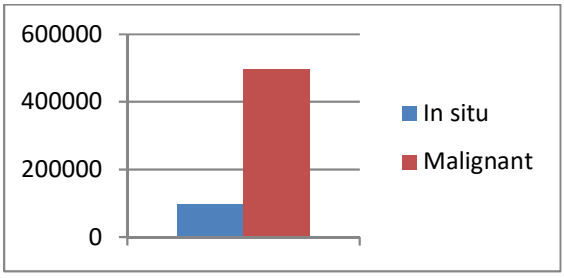

Figure 3: Behavior recode for analysis class distribution.

b. Transformation of categorical data

This step consists to transform categorical data to integer format using also pandas python library, which our predictive models can better understand. Like for example Behavior recode for analysis has two possible values (In situ or Malignant) are transformed to 0 and 1, the same thing for the others categorical data.

c. Dividing data into groups

Due to the large number of data extracted we divided the data, by years of diagnosis, into 9 groups to evaluate the performance of executed algorithms in the classification of large data of breast cancer. The total number of data in each group is the following: (60992 for 2008, 62894 for 2009, 62002 for 2010, 63735 for 2011, 65377 for 2012, 67723 for 2013, 68741 for 2014, 71019 for 2015 and 70521 for 2016).

d. Data normalization /standardization

Normalization and standardization are two techniques of data preprocessing which make features in the same scale, so that no one has more influence than the others on classification. The difference between them that normalization scale features between minimum and maximum values and standardization rescale data to have a mean of 0 and a standard deviation of 1.

3.2.3 Classification

The final step is classification in which we will apply the selected algorithms using the scikit learn this library provides many classification algorithms and facilitate the use of them; this step is divided into two tasks: first, we will evaluate the performance of the selected machine learning algorithms in the classification of large breast cancer dataset then we will show the impact of normalization/standardization in the improvement of classification accuracy.
Second, we will test the capacity of ensemble methods in the improvement of the accuracy of machine learning algorithms that get a low accuracy.

4. CLASSIFICATION AND MACHINES LEARNING ALGORITHMS

Classification is a supervised learning process that categories data into classes using machine learning classifiers. In this paper we will try to classify the breast cancer into its two categories using machine learning algorithms, the

data is divided into training and testing data. We executed the classifiers into training data then we examined their performance in testing data. Some classifiers have approved their performance, but for others we will try to improve their performance by two techniques. First, through the normalization/standardization techniques, then by ensemble methods.

4.1 K-Nearest Neighbors

K-Nearest neighbors (KNN) is one of the top 10 machine learning algorithm [18], from the category of Lazy Learning that can be used in classification also in regression. k-nearest neighbors tries to classify the unknown sample of testing data based on the known classification of its neighbors from training data by calculating the distance between them [19] , The KNN search in the training data the k closest simples to unknown test simple, then the classification of test simple can be defined based on those closest simples.

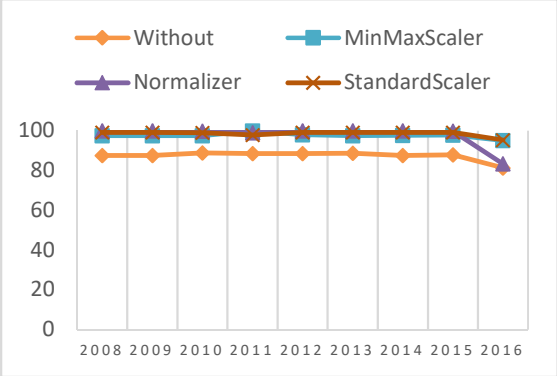

Figure 4: Accuracy of KNN.

The figure above shows the results of the examination of k-nearest neighbors (KNN) in breast cancer database. The KNN algorithm shows a low accuracy without doing any normalization/standardization of data. All the three techniques Normalizer, MinMaxScaler and StandardScaler have improved the accuracy of classification of the KNN, but the higher improvement was done by Normalizer for the years 2008-2015 except 2011. For example, for the year 2008 the improvement was more than 12%, and for 2016 the higher improvement was done by StandardScaler with an improvement of 13.97% followed by MinMaxScaler, the Normalizer does not give a big improvement.

4.2 Naive Bayes (NB)

Naive Bayes is a simple Bayesian classifier that is based on the Bayes Theorem with a strong independence between the features. The naïve Bayes model is easy in the construction and can be used in huge set of data [18]. Naive Bayesian classifier assumes that the existence of a feature in a class is independent of the existence of others features.

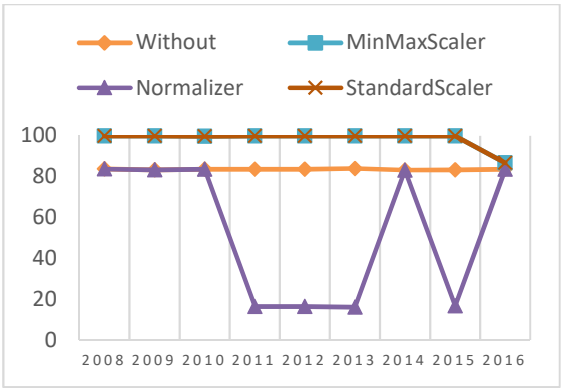

Figure 5: Accuracy of NB.

After examining Naive Bayes performance in testing data, also it's gives low accuracy which vary between 83.16% for 2004 and 83.68% for 2008. The higher improvement was given by both MinMaxScaler and StandardScaler. The Normalizer does not give any improvement and in some case it decreased more the accuracy of classification especially in the years 2011, 2012 and 2013.

4.3 Decision Tree

Decision Tree is learning method used in both classification and regression. It is similar to flowchart [20] where the internal nodes represent the test on the attributes, the branches represents the result of the test, and the leaf contain the prediction results. They are two ways for building decision tree, from top to bottom and from bottom to top.

The most popular decision Trees algorithms are: ID3, C4.5, C5, J48 and CART.

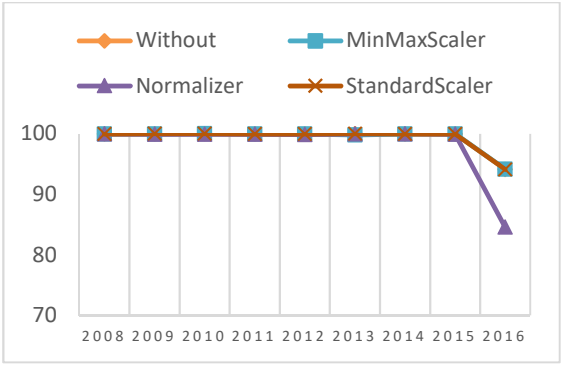

Figure 6: Accuracy of DT.

4.4 Random Forest

Random Forest is an algorithm that combines many decision trees algorithms and merge them into one forest. The principle of random forest that each decision tree built randomly will be trained on a subset of data, and then the classification will be taken by voting the result of predictions following the Bagging principle.

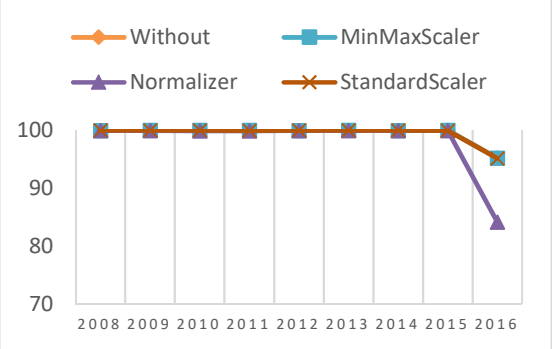

Figure 7: Accuracy of RF.

Decision Tree and Random Forest, those two algorithms show their performance without needing to any pre-processing of data as presented in figures 6 and 7. And also, the three techniques of Normalization/Standardization have improved more the accuracy of classification.

4.5 Logistic Regression

Logistic regression [21] is one of the generalized linear models much used in machine learning. Logistic regression predicts the probability of a result that can take two values from a set of predictor variables. Logistic regression is mainly used for prediction and also to calculate the probability of success.

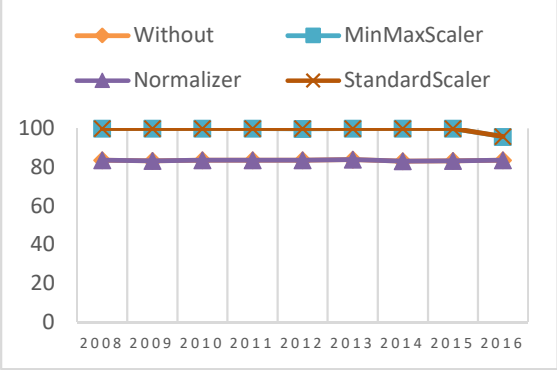

Figure 8: Accuracy of LR.

Logistic regression and Multi-layer Perceptron algorithms give a lower accuracy compared with Decision Tree and Random Forest as shown in figures 8 and 9, but MinMaxScaler and StandardScaler have improved their performance, for some year the improvement was more than 16%. The Normalizer does not give a big improvement.

4.6 Multi-Layer Perceptron

Multi-layer Perceptron algorithm is an artificial neural network model, composed of many layers. The input layer receive the information and the output layer gives the decision, and between them an ensemble of hidden layer. Those layers are composed of number of variable named neurons that are similar to the neurons of the human brain.

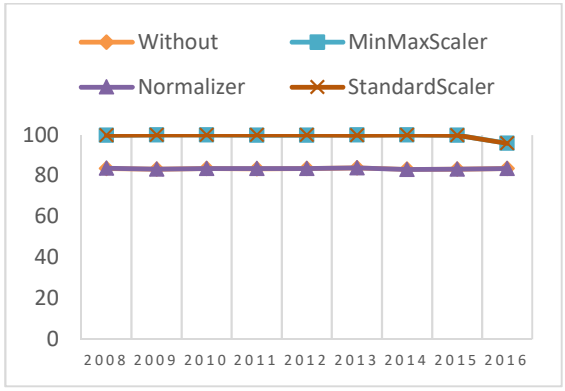

Figure 9: Accuracy of MLP.

In this section we examined the sex machines learning algorithms k-nearest neighbors, Naive Bayes, Decision Tree, Random Forest, Logistic regression and Multi-layer Perceptron. Some of them have a good performance without needing to normalizing data or doing any ensemble methods like Decision Tree and Random Forest, others give lower accuracy, in which we apply the normalization techniques. For the most MinMaxScaler and StandardScaler give more interesting result, the Normalizer did an improvement but not big like other, but for k-nearest neighbors it worked well. So we conclude, that the Normalization/Standardization have a good impact in the performance of machine leaning classifiers. In the next section, we will try to improve the accuracy of poor algorithms using ensemble methods.

5. ENSEMBLE METHODS

Ensemble methods are an ensemble of techniques that aim to produce better prediction performance using multiple models, by combining

between them. So, we will use those techniques to improve the classification of breast cancer.

5.1 Voting

Voting algorithm is a technique that combines between an ensemble of classifiers to improve the accuracy of classification. The principle of voting technique that each machine learning technique gives classification or output then the vote of those outputs will be taken as classification.

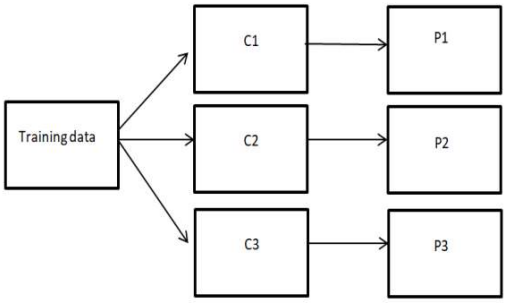

Figure 10: Voting technique.

If we take the example of 3 classifiers C1, C2 and C3 the prediction of each classifier successively will be P1, P2 and P3. The final prediction will be:

PF = mode {P1, P2, P3}.

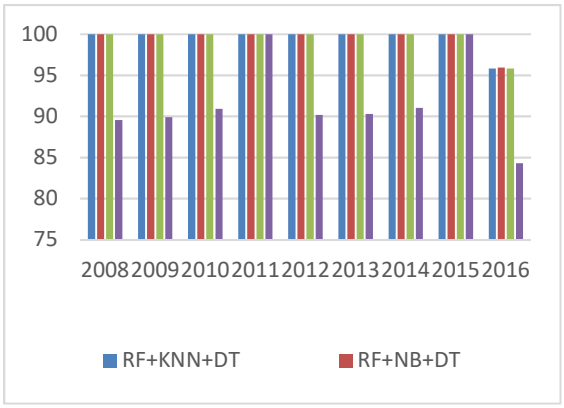

Figure 11: Accuracy of Voting technique.

The Voting technique improves the accuracy of weak algorithm by combining them with the efficient algorithm. In this work four combinations were done: voting the KNN with RF and DT, NB with DT and RF, NB and KNN with RF and DT and finally the combination of all classifiers (KNN, RF, NB, LR and MLP with RF and DT). The combination of KNN with strong algorithms improve its accuracy by more than 12%, and combination of NB with DT and RF improve its accuracy by more than 16%, also the combination of NB and KNN with RF and DT give the same results. But, the assembling of all classifiers does not give a big difference, so we tried to improve the performance of this assembling by Normalization/Standardization of data.

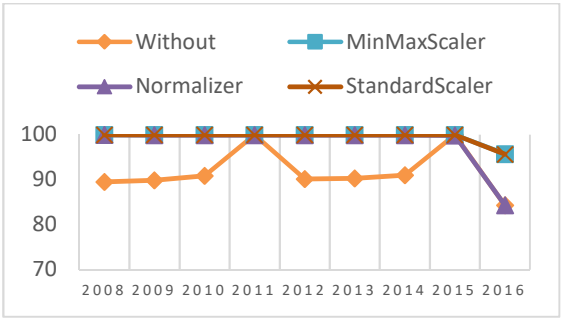

Figure 12: Improving Accuracy by Normalization/Standardization.

As shown in figure 12 the Normalization/Standardization techniques have a great impact in improving the accuracy of assembling KNN, RF, NB, LR and MLP with RF and DT.

5.2 Bagging

Bagging algorithm, shorthand of the combination of bootstrapping and aggregating, also known as Bootstrap Aggregating. This method improves the accuracy of classification by decreasing the variance and reducing the overfitting. The principle of bagging technique that it divides the data into subsets (Bootstrap) from training data, then it apply the classifier into each subset. Once the prediction of each subset is generated the algorithm uses the technique of averaging or voting to get the final prediction.

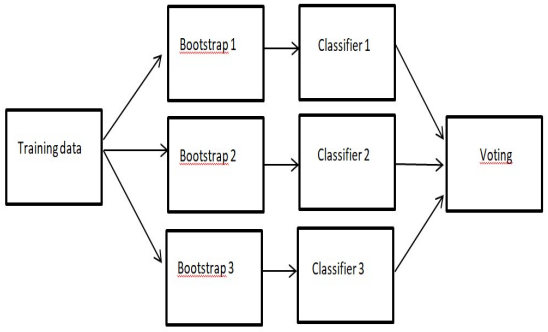

Figure 13: Bagging technique.

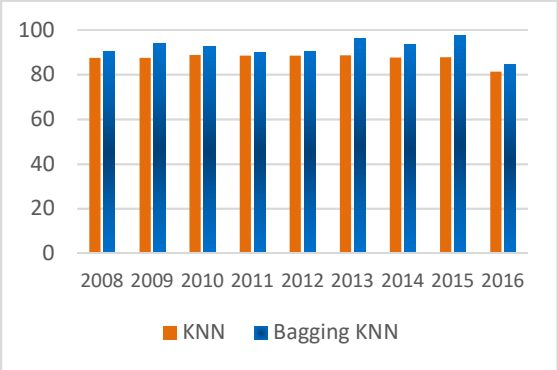

Figure 14: Accuracy of Bagging KNN.

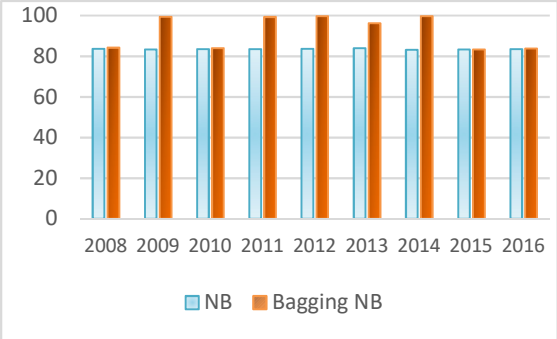

Figure 15: Accuracy of Bagging NB.

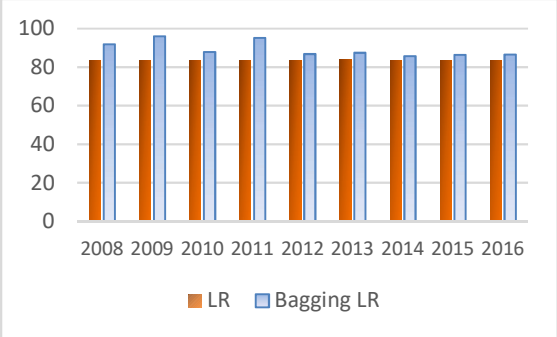

Figure 16: Accuracy of Bagging LR.

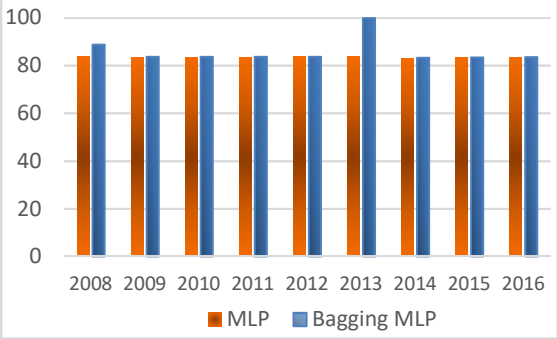

Figure 17: Accuracy of Bagging MLP.

A comparison of the results given by single weak algorithms and Bagging algorithms was presented in figures 14, 15, 16 and 17. It shows that the Bagging technique has brought an improvement for all the algorithms. For the KNN the maximum improvement was 10% in the year 2015, more than 16% for NB, about 13% for LR and 16% for MLP in the year 2013.

5.3 Boosting

Boosting algorithm is used specially in transforming weak algorithms into strong algorithms by reducing the bias and the variance. The principle of boosting algorithm that it trains weak learners sequentially, that mean that in each step the new subset is generated from the wrong classified elements, each step try to correct its predecessor. The famous types of Boosting are: AdaBoost, Gradient Boosting and XGBoost

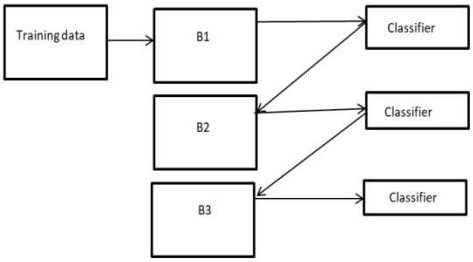

Figure 18: Boosting technique.

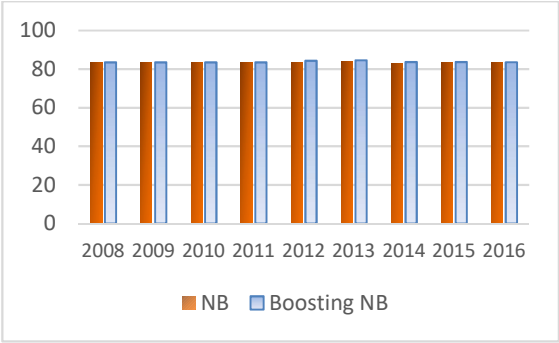

Figure 19: Accuracy of Boosting NB.

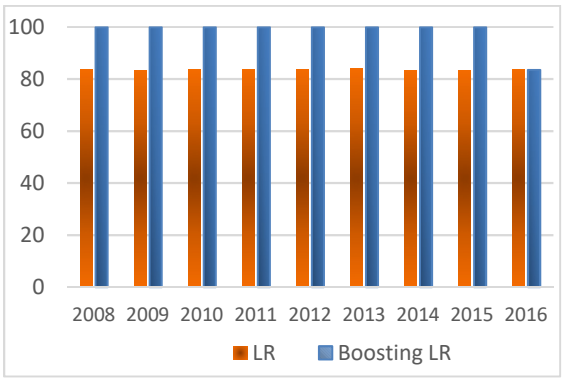

Figure 20: Accuracy of Boosting LR

Boosting technique was applied for NB and LR, for LR it shown a higher improvement in contrary of Boosting NB as presented in figures 19 and 20. A comparison between Bagging NB and Boosting NB was done in figure 21 and between Bagging LR and Boosting LR in figure 22, for LR Boosting techniques worked well than Bagging in contrary of NB in which Bagging show more good results in almost of all the years.

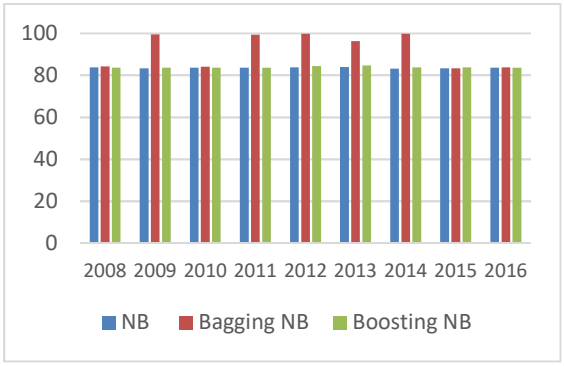

Figure 21: Comparison between Bagging and Boosting of NB

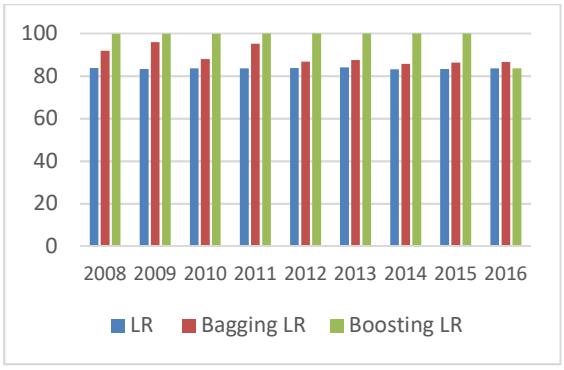

Figure 22: Comparison between Bagging and Boosting of LR.

5.4 Stacking

Stacking algorithm has a different paradigm from bagging and boosting. The principle of stacking that it combines multiple classifiers with meta-classifier to improve the accuracy of prediction. It contains two levels: in level 0 the classifiers are trained on the training data and in level 1 the meta-classifier is trained on the output of the level 0.

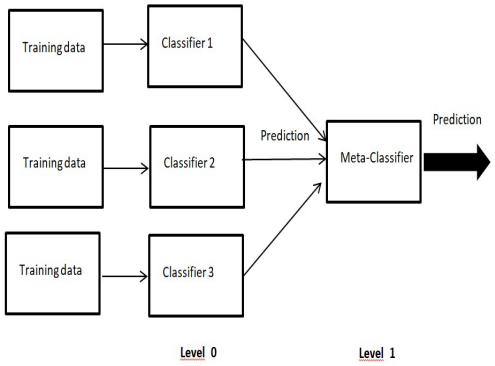

Figure 23: Stacking technique .

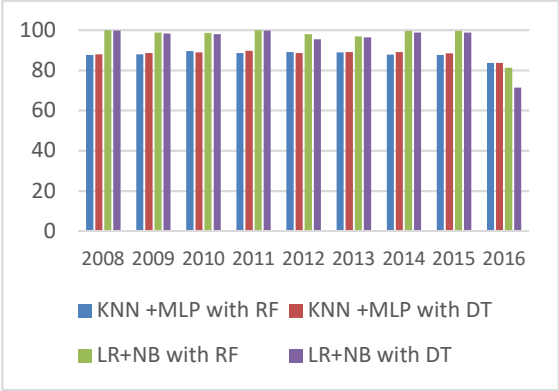

Figure 24: Accuracy of Stacking.

As already said, stacking combines multiple classifiers with meta-classifier to improve the accuracy of classification, the meta classifier taken in this step are DT and RF grace to their performance in the classification of breast cancer. The good result was given when stacking LR+NB with RF and LR+NB with DT except the year 2016, in which KNN+MLP with RF and KNN+MLP with DT worked more good. All the ensemble methods have improved the accuracy of the classification of the weak algorithms. For some algorithms voting technique was better for improving their accuracy like KNN and NB, Bagging for NB and MLP for some years, Boosting for LR and Stacking for LR and NB.

6. COMPARISON WITH EXISTING WORK

Our proposed methods are compared with others researches, some of them used the same

SEER database with different features and different algorithms [8], [9], [10] and [22]. Others worked with ensemble methods but with others breast cancer databases [23] and [24].

The Table 2 shows that our proposed method Voting Naive Bayes, Decision Tree and Random Forest with the selected features give better results compared with the others researches.

Table 2: Comparative study with the existing work of breast cancer classification

Work	Proposed Method	Accuracy
Ours	Voting Naive Bayes, Decision Tree and Random Forest	99.99%
Assiri et al. [23]	Majority-based voting mechanism	99.42%
Mathew1 et al.[24]	Stacking Naive Bayes with Logistic Regression and SMO	97.8%
Farooqui et al. [10]	Random Forest	73%
Wang et al.[22]	Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE)	97.10% (for WBC dataset) and 76.42% for (SEER database).
Hamsagayathri el al. [8]	Priority Based Decision Tree	98.51%
Rajesh et al. [9]	C4.5	93%

7. CONCLUSION

To conclude, in this paper we tried to examine and improve the performance of machine learning techniques in the classification of massive breast cancer database like SEER database using the python library scikit learn that facilitate for us the use and the implementation of the executed algorithms. First, we tested the performance of several machine learning techniques in the classification of large SEER breast cancer database, the KNN, NB, LR and MLP techniques show low accuracy in contrary of DT and RF that proved their performance, then we tried to improve the performance of those weak algorithms by Normalization/standardization techniques and ensemble methods like: Voting, Stacking, Bagging and boosting. The improvement by Normalization/standardization techniques goes until 16% for MinMaxScaler and StandardScaler and 12% for Normalizer. And when using ensemble methods the improvement by Bagging goes until 15,97% , by stacking goes until 16,23%, by Boosting goes until 16,77% and by voting goes until 16,83%. So, the higher improvement and the higher accuracy were given by voting technique. The result shows that the Normalization/standardization and ensemble methods have a big impact in the improvement of classification accuracy of the weak algorithms.

There are some limitations in this work. First, the proposed methods are not examined on others breast cancer dataset. Second, Bagging and Boosting in same cases didn't give a good improvement.

In future work, those prosed models can be tested on others breast cancer datasets, or on others cancers datasets. Also, features selection techniques can be used to select relevant features and others combination of machine learning algorithms can be done. And finally this research can be a good start for classifying breast cancer from medical image.

ACKNOWLEDGEMENTS:

H. Saoud acknowledges financial support for this research from the "Centre National pour la Recherche Scientifique et Technique" CNRST, Morocco.
Also we acknowledge National Cancer Institute (NCI) for providing as the access to SEER cancer database.

REFRENCES:

[1] 'U.S. Breast Cancer Statistics', Breastcancer.org, Jan. 27, 2020. https://www.breastcancer.org/symptoms/understand_bc/statistics (accessed Apr. 15, 2020).

[2] 'Ascent of machine learning in medicine | Nature Materials'. https://www.nature.com/articles/s41563-019-0360-1 (accessed Sep. 14, 2019).

[3] 'Surveillance, Epidemiology, and End Results Program'. https://seer.cancer.gov/ (accessed Sep. 14, 2019).

[4] H. Saoud, A. Ghadi, M. Ghailani, and B. A. Abdelhakim, 'Application of Data Mining Classification Algorithms for Breast Cancer Diagnosis', *in Proceedings of the 3rd International Conference on Smart City Applications - SCA '18*, Tetouan, Morocco, 2018, pp. 1–7, doi: 10.1145/3286606.3286861.

[5] H. Saoud, A. Ghadi, M. Ghailani, and B. A. Abdelhakim, 'Using Feature Selection Techniques to Improve the Accuracy of Breast Cancer Classification', *in Innovations in Smart Cities Applications Edition 2, M. Ben Ahmed, A. A. Boudhir, and A. Younes, Eds. Cham: Springer International Publishing*, 2019, pp. 307–315.

[6] 'Proposed approach for breast cancer diagnosis using machine learning | *Proceedings of the 4th International Conference on Smart City Applications*'. https://sci-hub.tw/https://dl.acm.org/doi/abs/10.1145/3368756.3369089 (accessed Apr. 15, 2020).

[7] H. Saoud, A. Ghadi, and M. Ghailani, 'Hybrid Method for Breast Cancer Diagnosis Using Voting Technique and Three Classifiers', *in Innovations in Smart Cities Applications Edition 3*, Cham, 2020, pp. 470–482, doi: 10.1007/978-3-030-37629-1_34.

[8] P. Hamsagayathri and P. Sampath, 'Priority based decision tree classifier for breast cancer detection', i*n 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, Jan. 2017, pp. 1–6, doi: 10.1109/ICACCS.2017.8014598.

[9] K. Rajesh, D. S. Anand, and P. Student, 'Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm', vol. 1, no. 2, p. 6.

[10] N. A. Farooqui, 'A STUDY ON EARLY PREVENTION AND DETECTION OF BREAST CANCER USING THREE-MACHINE LEARNING TECHNIQUES', *Int. J. Adv. Res. Comput. Sci.*, p. 7, 2018.

[11] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, 'Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review', *ACM Comput. Surv.*, vol. 49, no. 3, pp. 1–40, Oct. 2016, doi: 10.1145/2988544.

[12] A. Lg and E. At, 'Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence', *J. Health Med. Inform.*, vol. 04, no. 02, 2013, doi: 10.4172/2157-7420.1000124.

[13] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, 'Machine learning applications in cancer prognosis and prediction', *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[14] Y. Song, S. Gao, W. Tan, Z. Qiu, H. Zhou, and Y. Zhao, 'Multiple Machine Learnings Revealed Similar Predictive Accuracy for Prognosis of PNETs from the Surveillance, Epidemiology, and End Result Database', *J. Cancer,* vol. 9, no. 21, pp. 3971–3978, 2018, doi: 10.7150/jca.26649.

[15] D. Solti and H. Zhai, 'Predicting Breast Cancer Patient Survival Using Machine Learning', *in Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13*, Wshington DC, USA, 2007, pp. 704–705, doi: 10.1145/2506583.2512376.

[16] C. M. Lynch et al., 'Prediction of lung cancer patient survival via supervised machine learning classification techniques', *Int. J. Med. Inf.*, vol. 108, pp. 1–8, Dec. 2017, doi: 10.1016/j.ijmedinf.2017.09.013.

[17] S. Li and T. Razzaghi, 'Personalized Colorectal Cancer Survivability Prediction with Machine Learning Methods', *ArXiv190103896 Cs Stat*, Jan. 2019, Accessed: Sep. 11, 2019. [Online]. Available: http://arxiv.org/abs/1901.03896.

[18] X. Wu et al., 'Top 10 algorithms in data mining', *Knowl. Inf. Syst.,* vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

[19] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, 'k-Nearest Neighbor Classification', *in Data Mining in Agriculture*, vol. 34, New York, NY: Springer New York, 2009, pp. 83–106.

[20] J. Han and M. Kamber, 'Data Mining : Concepts and Techniques', p. 772.

[21] H. Yusuff, N. Mohamad, U. K. Ngah, and A. S. Yahaya, 'BREAST CANCER ANALYSIS USING LOGISTIC REGRESSION', p. 9, 2012.

[22] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, 'A support vector machine-based ensemble algorithm for breast cancer diagnosis', *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, Jun. 2018, doi: 10.1016/j.ejor.2017.12.001.

[23] A. S. Assiri, S. Nazir, and S. A. Velastin, 'Breast Tumor Classification Using an Ensemble Machine Learning Method', *J. Imaging*, vol. 6, no. 6, p. 39, May 2020, doi: 10.3390/jimaging6060039.

[24] T. E. Mathew, K. S. A. Kumar, and K. S. Kumar, 'Breast Cancer Diagnosis using Stacking and Voting Ensemble models with Bayesian Methods as Base Classifiers', p. 14, 2020.