Hybrid Method for Breast Cancer Diagnosis Using Voting Technique and Three Classifiers

Hajar Saoud[1](✉), Abderrahim Ghadi[1], and Mohamed Ghailani[2]

[1] LIST Laboratory, University of Abdelmalek Essaadi (UAE), Tangier, Morocco
saoudhajar1994@gmail.com, ghadi05@gmail.com
[2] LabTIC Laboratory, University of Abdelmalek Essaadi (UAE),
Tangier, Morocco
ghalamed@gmail.com

Abstract. Breast cancer is one of the most dangerous types of cancer in women sector; it infects one woman from eight during her life and one woman from thirty die and the rate keeps increasing. The early prediction of breast cancer can make a difference and reduce the rate of mortalities, but the process of diagnosis is difficult due to the varying types of breast cancer and due to its different symptoms. So, the proposition of decision-making solution to reduce the danger of this phenomenon has become a primordial need. Machine learning techniques have proved their performance in this domain. In previous work we tested the performance of several machine learning algorithms in the classification of breast cancer such as Bayesian Networks (BN), Support Vector Machine (SVM) and k Nearest Neighbor (KNN). In this work, we will combine those classifiers using the voting technique to produce better solution using Wisconsin breast cancer dataset and WEKA tool.

Keywords: Breast cancer · Diagnosis · Voting technique · Classification · WEKA

1 Introduction

Breast cancer is a hard disease and its diagnosis is sometime difficult, the patient should pass through several tests starting with clinical examination ending with extracting and analyzing biological simples of breast cancer, the proposition of decision-making solution here has become a primordial need to reduce the process of diagnosis and also to reduce the rate of mortalities. In this paper, we tried to propose a solution for breast cancer diagnosis using machine learning due to their performance in the medical field.

In previous work we tried to classify breast cancer using several classifiers such as Bayes Network (BN), Support Vector Machine (SVM), k-nearest neighbors algorithm (Knn), Artificial Neural Network (ANN), Decision Tree (C4.5) and Logistic Regression in [1] the higher accuracies are given by Bayes Network (BN) and Support Vector Machine (SVM) 97.28%. Then we tried to improve those accuracy in [2] by using the technique of feature selection Best First, the accuracy of Bayes Network (BN) has increased to 97.42% but the accuracy of Support Vector Machine (SVM) has decreased

M. Ben Ahmed et al. (Eds.): SCA 2019, LNITI, pp. 470–482, 2020.
https://doi.org/10.1007/978-3-030-37629-1_34

to 95.279%. So, we should search for others solutions that can improve more the accuracy of classification of breast cancer.

The objective of this work is to improve the accuracy of breast cancer classification using voting technique that aim to combine between classifiers. First, we did a combination between Bayes Network (BN) and Support Vector Machine (SVM) but there is no improvement. Consequently, we added K Nearest Neighbors algorithm (Knn) and the accuracy of classification has improved.

The rest of this paper is structured as follows. Part two is a presentation of breast cancer. Part three gives a vision about similar research. Part four is a theoretic presentation of machine learning algorithms. Part five give the definition of voting technique. In part six we will explain our proposed approach. Part seven shows the experiments performed by WEKA software on Wisconsin breast cancer dataset and results of these experiments and finally conclusion and perspectives in part eight.

2 Breast Cancer

Breast cancer can be defined as an abnormal production of cells in the breast that form in the form of cancerous masses, these masses are called tumors. Cancer cells can stay in the breast these types of cancer are called non-invasive, they lead to healing and do not produce metastatic cases. The other type of breast cancer is called invasive. These are dangerous type of cancers that can spread to the other organs of the body and can lead to metastatic cases.

2.1 Types of Breast Cancer

The types of breast cancers are invasive and non-invasive, Ductal Carcinoma In Situ is a non-invasive type the others are invasive [3] (Table 1):

Table 1. Types of breast cancer.

Type	Description
Ductal Carcinoma In Situ	Ductal Carcinoma In Situ is the most common type of breast cancer in the non-invasive cancer category in women. As the name suggests, it is formed inside the breast lactation channels
Ductal carcinoma	This type of cancer is also formed in the lactation ducts, but cancer cells pass through the canal wall
Lobular carcinoma	In this type of cancer the cancerous cells appear in the lobules grouped in the lobes. Afterwards, they cross the wall of the lobules and disperse in the surrounding tissues
Inflammatory carcinoma	Is a rare type of cancer that is known by a breast that can turn red, swollen and hot
Paget's disease	Is also a rare type of cancer that is manifested by a small nipple wound that does not heal
Other carcinomas	(medullary, colloidal or mucinous, tubular, papillary). Are the rarest types of breast cancer

2.2 Diagnosis of Breast Cancer

The process of the diagnosis of breast cancer is difficult due to the varying types and symptoms of breast cancer and also the patient should pass through several steps [4] starting with physical examination, it is a palpation of the breast that can determine the signs of appearance of cancer. The next step is medical imaging, it allows the detection of tumor masses and also it provides details on the clinical examination, there are several types of medical imaging among them: Mammography, Ultrasonography and MRI, The choice of one of these techniques is made according to the case of the patient. A diagnosis can only be decided after having studied biological samples at the microscopic level of the lesions that appeared in the medical imaging, the choice of the sampling method is according to the characteristics of the lesion, the exciting techniques are Aspiration or Cytological Puncture, Biopsy and Macrobiopsia (Fig. 1).

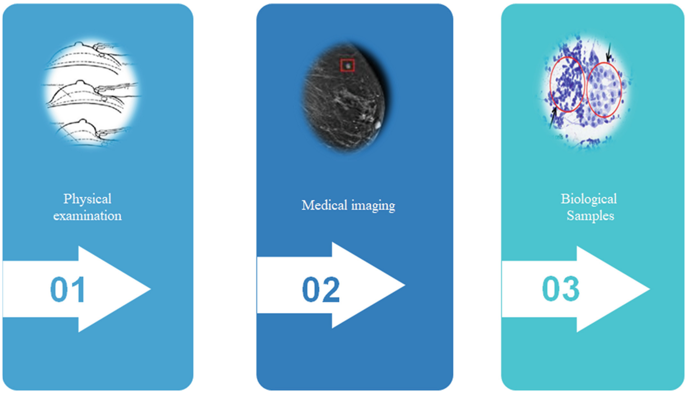

Fig. 1. Process of diagnosis of breast cancer.

The image obtained from the microscopic level will be studded at the same time with others images and features. So, the proposition of decision making solution will be an interesting thing to reduce the number of steps of the diagnosis also to avoid any error in the diagnosis. The machine learning techniques will be powerful tools due to their performance in the domain of medicine.

3 Related Works

Several approaches are proposed in the domain of cancer diagnosis and also for others diseases using machine learning algorithms, voting technique and others techniques like bagging, stacking and boosting. In this paragraph we will cite same of them:

Khuriwal and Mishra in [5] they proposed an adaptive ensemble voting method using Artificial Neural Network (ANN) and Logistic Regression (LR), the database used is Wisconsin Breast Cancer database. They achieved 98.50% in accuracy.

Kumar et al. in [6] they compared the performance of machine learning techniques in the classification of breast cancer using Wisconsin Breast Cancer database then they combined those techniques using voting technique. The three techniques tested in this research are Naïve Bayes, SVM and j48.

Latha and Jeeva in [7] they examined the ensemble algorithms bagging, boosting, stacking and majority voting for prediction of the heart disease using Cleveland heart dataset from the UCI machine learning repository.

Leon et al. in [8] they analyzed the influence of several voting methods on the performance of K Nearest Neighbor and Naïve Bayes algorithms used for datasets with different levels of difficulty.

Rishika and Sowjanya in [9] they aim to compare the performance of Decision Tree, Neural Network and Naive Bayes, then they tried to combine between them using stacking approach.

Sri Bala and Rajya Lakshmi in [10] they implemented four models Adaboosting, bagging and stacking or blending on preliminary classifiers to improve the accuracy of the classification of breast cancer. So, the totals of built models are 12.

4 Machine Learning Algorithms

The machine learning techniques that we will see in this paper are Bayesian Network (BN), Support Vector Machines (SVM) and k-nearest neighbors algorithm (Knn). We will examine each algorithm separately than we will combine between them to improve the accuracy of classification of the breast cancer.

4.1 Bayesian Network (BN)

Bayesian Network [11], also called (Bayesian belief network), is directed acyclic graph (DAG) composed of nodes and edges, the nodes represent variables and edges represent the probabilistic dependencies between those variables. Bayesian Network combines principles of statistics, graph theory, probability theory and computer science.

4.2 Support Vector Machines (SVM)

Support Vector Machines is supervised learning model, which is always known by the notion of hyperplane, this hyperplane is a line that divide a plan into two spaces each space represent a class. Taking training data the Support Vector Machines well search an optimal hyperplane that will separate the data into two dimensional spaces.

4.3 K Nearest Neighbors Algorithm (KNN)

The k-nearest neighbors classifier is a supervised machine learning algorithm that can be used in both classification and regression. The k-nearest neighbors classifier capture the idea of similarity (called also distance). So, the principle of k-nearest neighbors that it calculates the distance between a given test tuple and others tuples to search the K closest tuples, these tuples are named (k nearest neighbors).

5 Voting Classifier Technique

Voting classifiers is a technique used in classification; it aims to combine between 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 (Fig. 2).

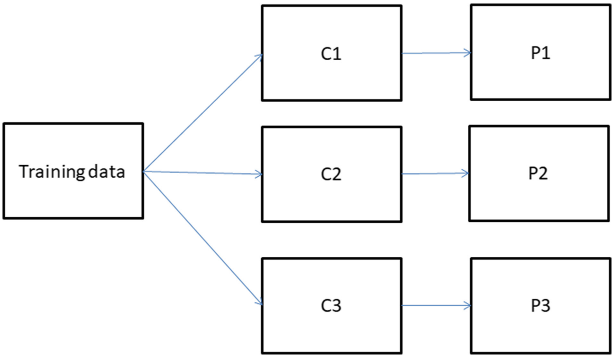

Fig. 2. Voting classifiers technique.

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:

$$P_F = \text{mode}\ \{P1, P2, P3\}.$$

6 Proposed Method

In our proposed method we will improve the accuracy of the classification of the three machine learning algorithms Bayes Network (BN) and Support Vector Machine (SVM) and K Nearest Neighbors algorithm (KNN) by using the voting technique, that aim to combine between them to improve the accuracy of classification. Figure 3 represents the process of the proposed method first we choose Wisconsin breast cancer dataset, then we did the pre-processing of data to eliminate missing data and finally we passed to the step of classification.

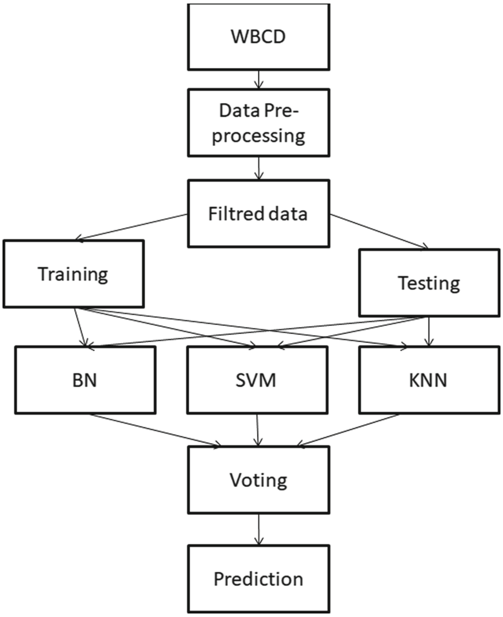

Fig. 3. Process of proposed method.

7 Experimentation and Results

7.1 Description of the Dataset

The database that we used in this research is the Wisconsin breast cancer dataset available in UCI machine learning repository [12]. It contains 699 records (458 benign tumors and 241 malignant tumors). It is composed of 11 variables 10 predictor variables and one result variable that shows whether the tumor is benign or malignant. The predictive attributes vary between 0 and 10. The value 0 corresponds to the normal state and the value 10 corresponds to the most abnormal state.

The table above presents the description of the 11 attributes of the Wisconsin breast cancer dataset (Table 2):

Table 2. Attributes of WBCD.

Attributes	Description
Id	A code for the identification of each line
Clump thickness	The benign cells are grouped in monolayers whereas the cancer cells are grouped in multilayers
Uniformity of cell size	The size of the cancer cells
Uniformity of cell shape	The shape of the cancer cells
Marginal adhesion	Cancer cells can lose their tights; this is a sign of malignant cancer
Single epithelial cell size	Single Epithelial Cell Size
Bare nuclei	The nuclei are not surrounded by the rest of the cell in benign tumors
Bland chromatin	Cancer cells have coarse chromatin
Normal nucleoli	In cancer cells, the nucleoli are transforming into protuberant, but the nucleoli are small
Mitoses	Cell growth
Class	If the cancer is a benign tumor or malignant tumor

7.2 WEKA Tool

The tool that we used to apply the machine learning algorithms on the breast cancer database is WEKA [13], because WEKA is a collection of open source machine learning algorithms, which allows realizing the tasks of data mining to solve real world problems. It contains tools for data preprocessing, classification, regression, grouping, and association rules. Also it offers an environment to develop new models.

7.3 K-Fold Cross-validation

To evaluate the performance of machine learning algorithms based on breast cancer data we used the K-fold cross validation test method. This method aims to divide the database in two sets, the training data to run the model and the testing data to evaluate the performance of the model. This is the most used method in the evaluation of machine learning techniques (Fig. 4).

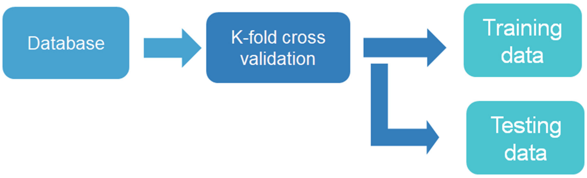

Fig. 4. Process of K-fold Cross-validation.

7.4 Confusion Matrix

Confusion matrix gives the possibility to evaluate the performance of each classifier by calculating its Accuracy, Sensitivity and Specificity. It contains information about real classifications or (current) and predicted (Table 3):

Table 3. Confusion matrix.

	Predicted benign	Predicted malignant
Actual benign	TP (true positives)	FN (false negatives)
Actual malignant	FP (false positives)	TN (true negatives)

TP: the cases predicted as benign tumors, they are in fact benign tumors.

TN: the cases predicted as malignant tumors, they are in fact malignant tumors.

FP: the cases predicted as benign tumors but in the reality they are malignant tumors.

FN: the cases predicted as malignant tumors but in the reality they are benign tumors.

From the confusion matrix we can calculate:

- accuracy $= \dfrac{TP + TN}{TP + FP + TN + FN}$
- Sensitivity $= \dfrac{TP}{TP + FN}$
- Specificity $= \dfrac{TN}{TN + FP}$

7.5 Bayesian Network (BN)

The accuracy obtained by Bayesian Network (BN) is 97.28%, 680 from 699 are well classified instances and 19 are incorrectly classified instance that represent the 2.71%. Table 4 represents the confusion matrix of Bayesian Network (Figs. 5, 6 and Table 5):

Table 4. Confusion matrix of BN.

	Predicted benign	Predicted malignant
Actual benign	442	16
Actual malignant	3	238

Table 5. Results of BN.

	TP rate	FP rate	Precision	Recall	F-measure	ROC area
benign	0,965	0,012	0,993	0,965	0,979	0,991
malignant	0,988	0,035	0,937	0,988	0,962	0,991

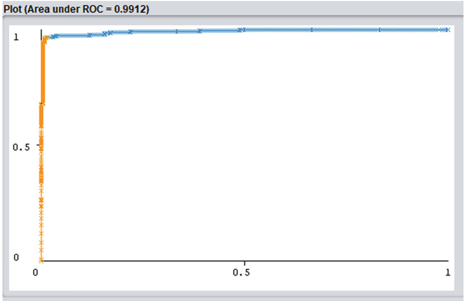

Fig. 5. ROC curve of benign for BN.

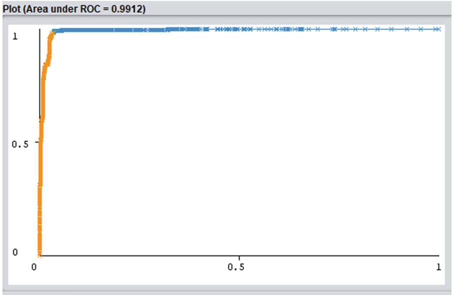

Fig. 6. ROC curve of malignant for BN.

7.6 Support Vector Machines (SVM)

The accuracy obtained by Support Vector Machines (SVM) is 97.28% using the Puk as kernel function, 680 from 699 are well classified instances and 19 are incorrectly classified instance that represent the 2.71%, the same results as Bayesian Network (BN). Table 6 represents the confusion matrix of Support Vector Machines (SVM) (Figs. 7, 8 and Table 7):

Table 6. Confusion matrix of SVM.

	Predicted benign	Predicted malignant
Actual benign	442	16
Actual malignant	3	238

Table 7. Results of SVM.

	TP rate	FP rate	Precision	Recall	F-measure	ROC area
Benign	0,967	0,017	0,991	0,967	0,979	0,975
Malignant	0,983	0,033	0,940	0,983	0,961	0,975

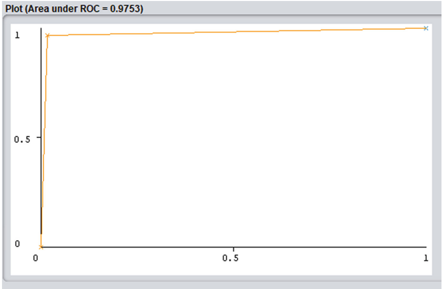

Fig. 7. ROC curve of benign for SVM.

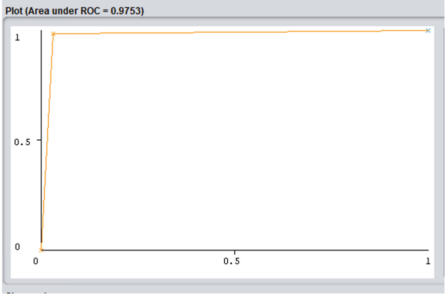

Fig. 8. ROC curve of malignant for SVM.

7.7 BN-SM

The accuracy obtained by BN-SVM is 96.99% there is no improvement, 678 from 699 are well classified instances and 21 are incorrectly classified instance that represent the 3%. Table 8 represents the confusion matrix of BN-SVM (Figs. 9, 10 and Table 9):

Table 8. Confusion matrix of BN-SVM.

	Predicted benign	Predicted malignant
Actual benign	445	13
Actual malignant	20	221

Table 9. Results of BN-SVM.

	TP rate	FP rate	Precision	Recall	F-measure	ROC area
Benign	0,974	0,037	0,980	0,974	0,977	0,991
Malignant	0,963	0,026	0,951	0,963	0,957	0,991

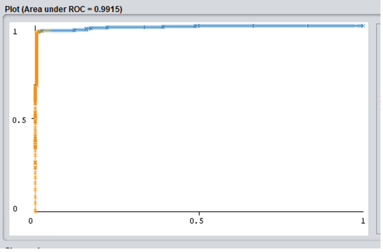

Fig. 9. ROC curve of benign for BN-SVN.

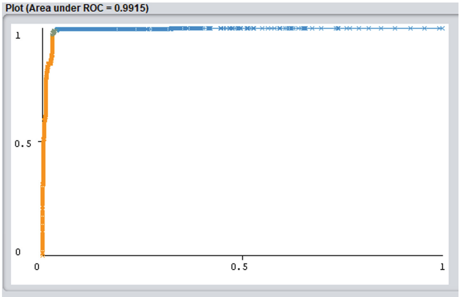

Fig. 10. ROC curve of malignant for BN-SVM.

7.8 K Nearest Neighbors Algorithm (KNN)

The accuracy obtained by k-nearest neighbors algorithm (Knn) is 95.27%, 666 from 699 are well classified instances and 30 are incorrectly classified instance that represent the 4.72%. Table 10 represents the confusion matrix of k-nearest neighbors (KNN) (Figs. 11, 12 and Table 11):

Table 10. Confusion matrix of KNN.

	Predicted benign	Predicted malignant
Actual benign	445	13
Actual malignant	20	221

Table 11. Results of KNN.

	TP rate	FP rate	Precision	Recall	F-measure	ROC area
Benign	0,972	0,083	0,957	0,972	0,964	0,974
Malignant	0,917	0,028	0,944	0,917	0,931	0,974

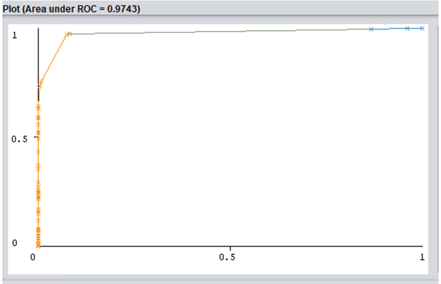

Fig. 11. ROC curve of benign for KNN.

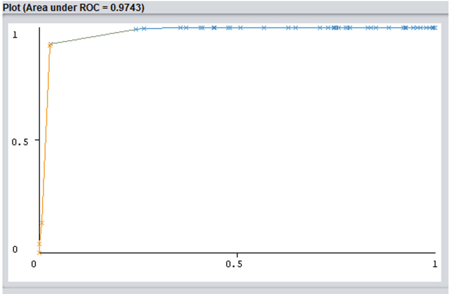

Fig. 12. ROC curve of malignant for KNN.

7.9 BN-SVM-KNN

The accuracy obtained by the proposed combination of the three algorithms by voting techniques is 97.56%, 682 from 699 are well classified instances and 17 are incorrectly classified instance that represent the 2.43%. Table 12 represents the confusion matrix of BN-SVM-KNN (Figs 13, 14 and Table 13):

Table 12. Confusion matrix of BN-SVM-KNN.

	Predicted benign	Predicted malignant
Actual benign	445	13
Actual malignant	4	237

Table 13. Results of BN-SVM-KNN.

	TP rate	FP rate	Precision	Recall	F-measure	ROC area
Benign	0,972	0,017	0,991	0,972	0,981	0,990
Malignant	0,983	0,028	0,948	0,983	0,965	0,990

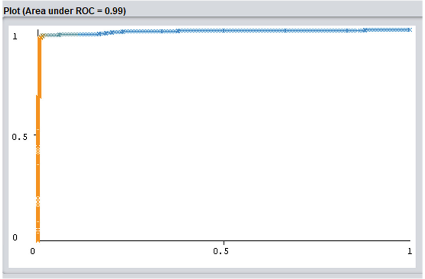

Fig. 13. ROC curve of benign for BN-SVN-KNN.

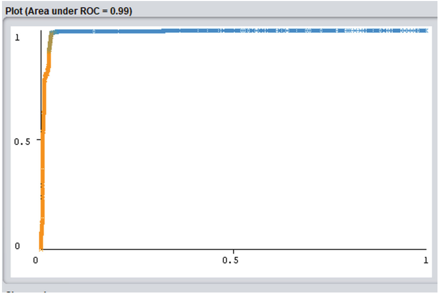

Fig. 14. ROC curve of malignant for BN-SVM-KNN.

Table 14 resumes the obtained results by each algorithm:

Table 14. Results of all models.

	Accuracy	Well classified instance	Wrong classified instance	Time taken
BN	97.28%	680	19	0.02 s
SVM	97.28%	680	19	0.22 s
BN-SVM	96.99%	678	21	0.09 s
KNN	95.27%	666	30	0 s
BN-SVM-KNN	97.56%,	682	17	0.03 s

8 Conclusion

To conclude, in this paper we tried to classify breast cancer into its two types benign or malignant using machine learning algorithm and the voting technique. First we examined each algorithm, Bayes Network (BN), Support Vector Machine (SVM) and k-nearest neighbors algorithm (KNN) separately then we tried to combine between them to improve the accuracy of the classification of breast cancer using the voting technique the accuracy produced 97.56%. The database of breast cancer in which the algorithms are tested is Wisconsin breast cancer dataset available in UCI machine learning repository using the WEKA tool.

Acknowledgement. H. Saoud acknowledges financial support for this research from the “Centre National pour la Recherche Scientifique et Technique” CNRST, Morocco.

References

1. Saoud, H., et al.: Application of data mining classification algorithms for breast cancer diagnosis. In: Proceedings of the 3rd International Conference on Smart City Applications - SCA 2018, pp. 1–7. ACM Press, Tetouan (2018). https://doi.org/10.1145/3286606.3286861

2. Saoud, H., et al.: Using feature selection techniques to improve the accuracy of breast cancer classification. In: Ben Ahmed, M., et al. (ed.) Innovations in Smart Cities Applications, edn. 2. pp. 307–315. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-11196-0_28
3. Le cancer du sein. https://www.passeportsante.net/fr/Maux/Problemes/Fiche.aspx?doc=cancer_sein_pm. Accessed 19 Oct 2019
4. Le diagnostic. https://rubanrose.org/cancer-du-sein/depistage-diagnostics/diagnostic. Accessed 19 Oct 2019
5. Khuriwal, N., Mishra, N.: Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm. In: 2018 IEEMA Engineer Infinite Conference (eTechNxT), pp. 1–5. IEEE, New Delhi (2018). https://doi.org/10.1109/ETECHNXT.2018.8385355
6. Kumar, U.K., et al.: Prediction of breast cancer using voting classifier technique. In: 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), pp. 108–114. IEEE, Chennai (2017). https://doi.org/10.1109/ICSTM.2017.8089135
7. Latha, C.B.C., Jeeva, S.C.: Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked **16**, 100203 (2019). https://doi.org/10.1016/j.imu.2019.100203
8. Leon, F., et al.: Evaluating the effect of voting methods on ensemble-based classification. In: 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA), pp. 1–6. IEEE, Gdynia (2017). https://doi.org/10.1109/INISTA.2017.8001122
9. Rishika, V., Sowjanya, A.M.: Prediction of breast cancer using stacking ensemble approach 11
10. Int. J. Adv. Res. Comput. Sci. Softw. Eng
11. Mahmood, A.: Structure learning of causal bayesian networks: a survey 6
12. UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). Accessed 19 Oct 2019
13. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. https://www.cs.waikato.ac.nz/ml/weka/. Accessed 19 Oct 2019
14. Saoud, H., Ghadi, A., Ghailani, M.: Analysis of evolutionary trends of incidence and mortality by cancers. In: Ben Ahmed, M., Boudhir, A. (eds.) Innovations in Smart Cities and Applications. SCAMS 2017. Lecture Notes in Networks and Systems, vol. 37. Springer, Cham (2018)