Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Breast cancer detection using random forest classifier

Ch. Hari Chandana *, G. Bala Krishna

Department of Computer Science, CVR College of Engineering, Hyderabad, India

ARTICLE INFO

Article history:
Received 28 December 2020
Received in revised form 28 December 2020
Accepted 17 January 2021
Available online xxxx

Keywords:
Breast cancer
Random forest classifier
Node frequency
Ranking
Out of bag analysis

ABSTRACT

Breast Cancer is a menacing disease and is commonly seen in women. Based on the severity, breast cancer is classified into duplet variants. One is Benign type of breast cancer, which can be detected at early stages and can be cured with the help of medication. Other is Malignant type of breast cancer, which shows severe affect and might lead to death. To detect breast cancer at early stages, wide variety of algorithm techniques are used such as Navie Bayes, Convolution Neural Network, KNN, adaptive voting ensemble machine learning algorithm and so on. Most latest algorithm that is under practice is adaptive voting ensemble machine learning algorithm. In this algorithm, Wisconsin Breast Cancer dataset and CNN algorithm is used to classify images and for object detection. But the major drawback of ensemble machine learning algorithm is lack of accuracy. It is proved that Neutral Network works more effective on humans mostly in analyzing data and to perform pre-diagnosis without medical knowledge. In this paper, we propose Random Forest Classifier algorithm to achieve more accuracy.

Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

1. Introduction

The prevailing disease in this world is cancer and this is more threatening among the women. Numerous expire due to malignant growth. Distinguishing the malignant growth physically takes lot of time and it is hard for the doctor to detect. So distinguishing the malignant growth through different programmed indicative methods is very vital. There is different strategy and calculation accessible for recognizing disease, for example, Support Vector Machine, Naive Bayes, KNN and Convolution Neural System is most recent calculation. CNN and deep learning calculation are mostly utilized for pictures and article recognition. In this paper we use UC Irvine repository to train and test the data where dual classes are attained. One is least harmful tumor and the other is threatening which is carcinogenic. Numerous scientists are yet carrying out research for identifying disease in the beginning period. The early stage malignant growth is costly to complete its treatment and numerous analyst are yet attempting to building up a appropriate analysis framework for identifying the tumor. So the treatment can be done at early stages Fig 1.

Later this paper is assembled as follows: Section 2 includes related work to proposed method. Section 3 provides the overview of proposed method Random Forest Classifier Algorithm. Section 4 presents architecture and control flow of recommended method. Section 5 presents results obtained by using the present algorithm. Section 6 includes future scope and conclusion of proposed method. Finally, Section 7 provides References.

2. Related work

A lot of research work has been done to diagnosis breast cancer at early stages with maximum accuracy. An automatic diagnosis system was proposed by Murat Karabatak and M. Cevdet Ince [1] for detecting breast cancer and achieved positive rate of 95.6%. Ali Keles et al. [2] implemented a algorithm to diagnose cancer depending on neuro-fuzzy regulation and achieved 96% classifier accuracy rate. Hui-Ling Chen et al. [3] obtained classifier accuracy of 96.87% by using rough set supporting vector machine (RS_SVM) classifier. Kemal Polat and Salih Güne [4] conducted diagnosis on breast cancer utilizing least square SVM and obtained 98.53% classification accuracy. Sahan et al. [5] created a hybrid procedure considering fuzzy immune system as base and k nearest was used to get accuracy of 98.14%. A. Marcano [6] created an algorithm named AMMLP considering the biological property of neurons as base and obtained total classification accuracy of 99.26%. All the above mentioned observations were tried out using Wisconsin Breast Cancer (WBC) Dataset which is quite different from WBCDD and WBCPD.

* Corresponding author.
E-mail address: hari97april@gmail.com (Ch. Hari Chandana).

https://doi.org/10.1016/j.matpr.2021.01.461
2214-7853/© 2021 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

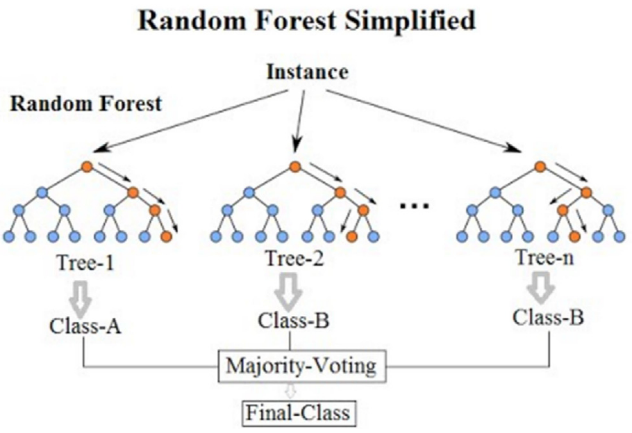


Fig. 1. Classification of Random Forest.

Below mentioned few researches conducted based on WBCDD and WBCPD. T. Mu and A. K. Nandi (2007) [7] conducted a research based on WBCDD, by including both radial networks and maps to attained 98% accuracy. C. Li, W. Liu (2011) [8] created a three-stage algorithm. In first stage, an arbitrary transformation technique is used to enhance classification related information from the algorithm variables for a tiny scale data set. In second stage, optimal subset of all the features are obtained by applying Principle Component Analysis (PCA) on the newly transformed, which eventually are used as inputs to support vector machine and obtained 96.35% accuracy.

3. Existing method

In the early stages of research breast cancer detection is based on low energy X- ray mammography was in practice. Later Magnetic Resonance images, Ultra Sound images are also preferred. In pre-processing stage various methods such as Linearization, Image thinning, Image gray scale extending, discrete wavelets, real valued or complex valued continuous filters, fuzzy filters were applied by many researchers. Similarly, transformation techniques were used to extract physical features or textural features of an image. Transformation techniques Euclidean Distance Transform, Fourier Transform, Discrete Wavelet Transform were widely used. There are several classifiers to in process obtain the decision of finding the presence or absence of the cancer cells. The classification results varies based on the research work to find the whether it is benign or malignant. Some researchers provide the stages of cancer. But most of the research works was based on single classifier. In general the majority voting method was also applied to find the optimized result. Parameter tuning is not needed all the time after tuning individual classifier. A weighted voting method is another approach to find the optimal solution.

3.1. Disadvantages

- Lack of accuracy when compared to neural networks which works more effective on humans in data analysis and diagnosis.
- By using number of filters we are detecting the cancer which increases the cost of the system and also expert person need to be there to detect the cancer.

4. Proposed method

In recent days, highly developed machine learning techniques had been used in a wide range to detect breast cancer at early stages. While performing diagnosis, the data gathered from the patient, data analysis and decisions taken by experts plays vital role to acquire better results. Diagnosis performed using different algorithm techniques can avoid human errors in analysis data and cancer detection at early stages. Proposed method Random Forest Classifier algorithm had around 99.7% accuracy in classifying the data and to obtain better results. Also, the suggested method can be used to analyze other cancerous problems which has high rate of training data. In our proposed method we used Wisconsin Breast Cancer Data set (WBC-DD). This is widely used by researchers who use different algorithm techniques to cure breast cancer, as it is used to compare our system with other references related to the same issue.

RF tree differs from CART, as RF uses dual level randomization procedure. In first stage, RF develops tree as similar to CART using part of original information. In second stage, instead of splitting tree using all the inputs, RF selects random variables at each node and uses only selected variables to find out the best split in the whole tree. Few researchers conclude that RF tree is more efficient in low data sample applications. Whereas, if the tree is deeply grown or if the original data sample is huge, RF tree yields low bias and eventually reduces variance.

4.1. Software development lifecycle

The proposed objective uses repetitive evolution lifecycle, where members of the user project are implemented through a iteration of steps which is shown in Fig. 2. Basic functionality is performed in the first phase or first iteration and followed by performing diagnostics on the errors identified. This is followed by adding new functionality to the error free functionality that is generated in previous iteration Fig 3 Fig 4 Fig 5 Fig 6 Fig 7 Fig 8.

The five stages of software development life cycle of proposed method are:

- Requirement Gathering
- Requirement Analysis
- Architecture implementation
- Software Implementation and
- Product Maintenance

The above mentioned five steps can be executed one after the other in a sequential manner where the output of each step is fed to the next step, adds new functionality or additional attempts and generates outcomes that lessen the preceding effort. It also checks the traceability to go back to previous steps. At every stage,

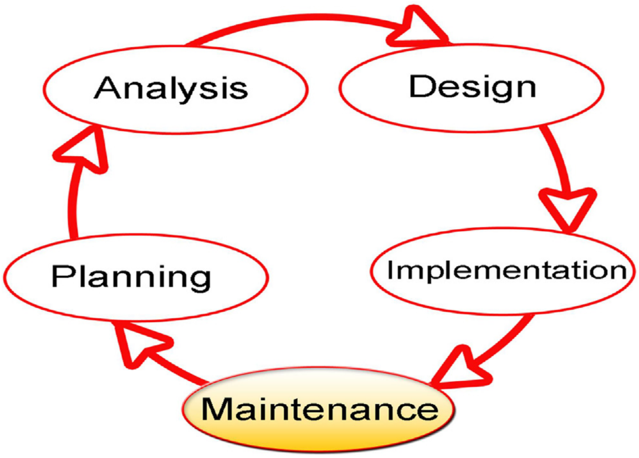


Fig. 2. Software Development Life cycle of proposed method.

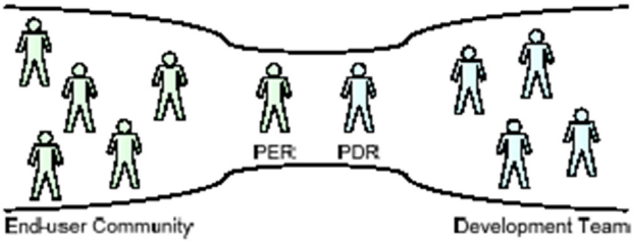


Fig. 3. PER-PDR Relationship.

implementation and validation is performed and generates deliverables.

A wide range of software development efforts got wasted when the process of automation comes into picture. Automation process made the team to mainly concentrate on features with low priority instead of high priority tasks. This made team to take additional efforts to get better results. This is the principal source of huge proportion of failure cases which made to abandon development efforts and make the development team to use continual prototype.

4.2. Roles and responsibilities of PDR AND PER

The frequentative lifecycle has dual important factors that operates jointly to discuss product concerns between users and the developers.

4.3. Primary End-user representative (PER)

This is a representative to end user. PER acts as the primary point of contact to the end-user. The PER is also responsible to collect review from the end user and update to development team promptly.

4.4. PER-PDR relationship

The PER and PDR plays important role in delivering deliverables on time. The PER has basic domain knowledge that is required to have basic understanding of the issues related to the application development process that is going to be implemented and also be the representative of end-user community. The PDR also plays similar role like PER regarding the application development process and in maintaining healthy relationship with the additional group associates of the coding team together.

The main aim of this iteration model is to develop healthy work relationship between end user and the software development team. This approach is similar to concept of Agile methodologies. As it is difficult to maintain work relationship with the whole development team, it is suggested to have good relationship between team leads and end-user community.

In general, it’s a hard task for the whole development team to be in touch with end user community. It might get complicated when the members in the end user community or the development team increases. So, as per iterative model standards, it is required to maintain communication between PER, PDR and end user community consistently. This allows the PER and PDR to resolve issues when there are different requirements raised for same application.

4.5. Input design

Input design is also called as architecture implementation. The prominent aim of current stage is as given below:

- Use inputs in productive manner.
- To attain the highest possible accuracy rate.
- To make sure the inputs are acceptable by end user.

4.6. Input stages

Before information gets stored in database, below mentioned are main input stages

- Record Data
- Upload Data
- Pre-process Data,
- Data transformation
- Data control
- Data analyses
- Data validation and
- Data correction

5. Architecture of Random Forest classification

Random tree is grown as explained below.

1. Initially, the training set and testing set are differentiated from the original data set.
2. In the next step, new data set named “in bag” will be formed from the training set using bootstrap method as shown below. In general, the in bag data set and training data set will contain same number of samples. But the only difference between these two is, in bag data set contains the duplicates or replacements of training data set. This method is referred as “boot strapping”.
3. Another data set named “Out Of Bag (OOB) will be derived from the training data set. Based on Boot strapping technique, OOB data set contains one-third of the training data set samples as shown in the above figure. Out of Bag data set is also known as” Left-Over data“.

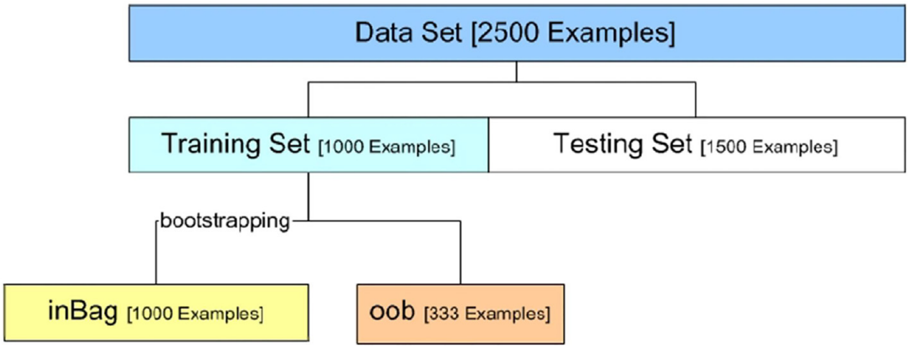


Fig. 4. RF Tree Development.

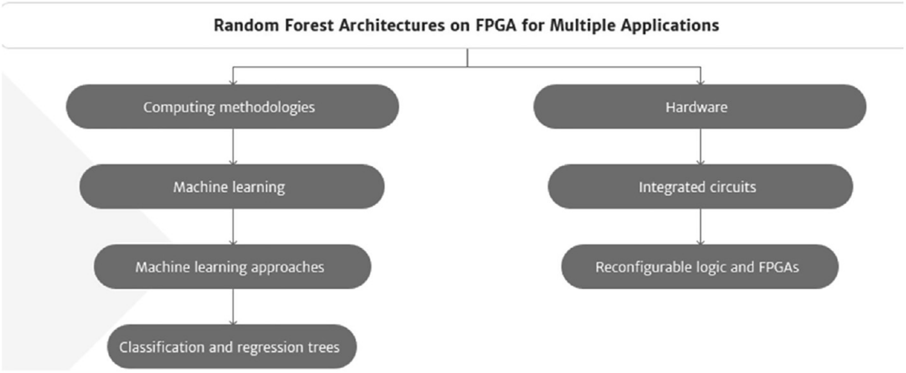

Random Forest Architectures on FPGA for Multiple Applications
Computing methodologies
Machine learning
Machine learning approaches
Classification and regression trees
Hardware
Integrated circuits
Reconfigurable logic and FPGAs

Fig. 5. Random Forest Classifier Architecture for multiple applications.

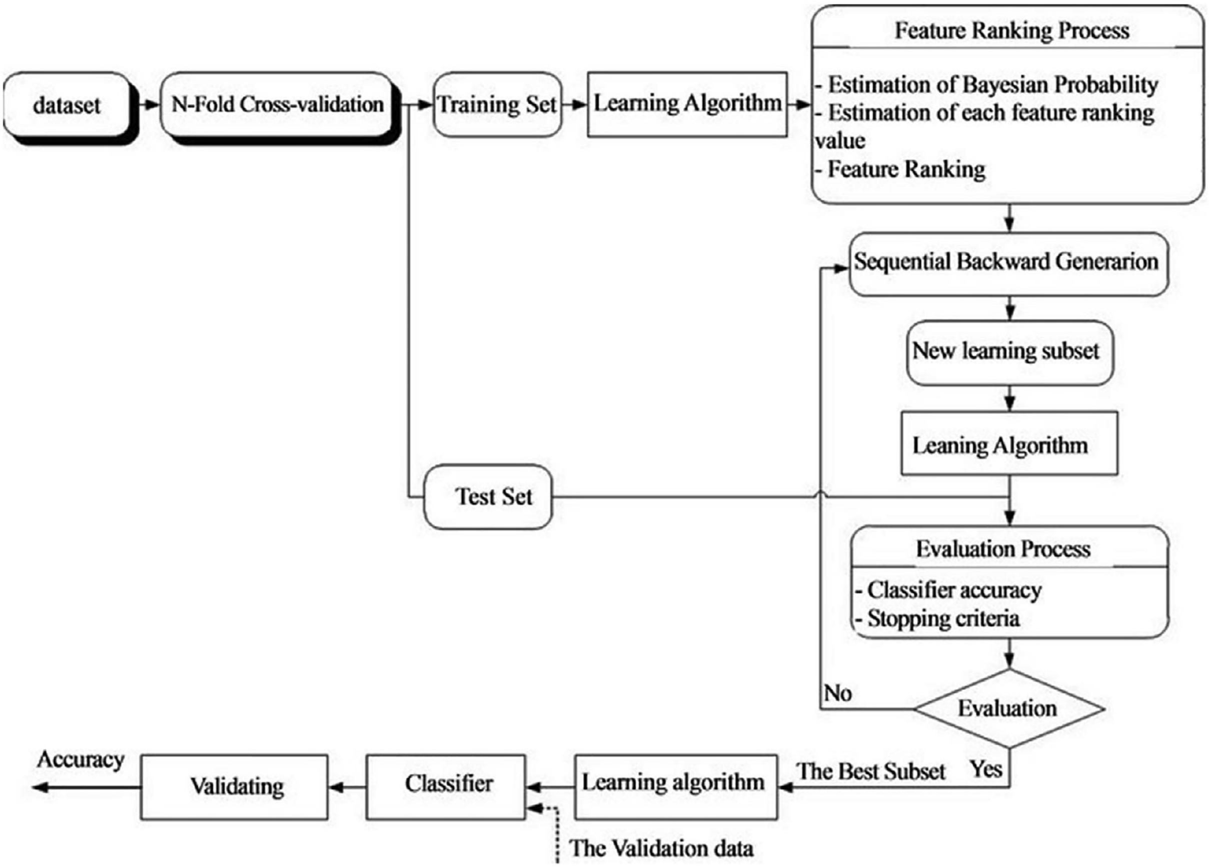


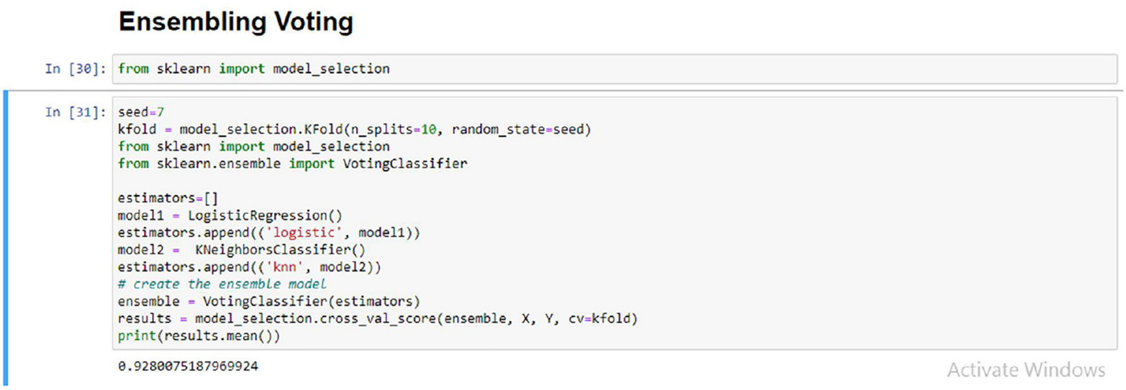
Fig. 6. Architecture of Random Forest Classifier.

Random variables are selected at each node and uses only selected variables to find out the best split in the whole tree.

The above procedure will be repeated until all the possible trees are derived from the parent tree. Later, when tree development is completed, Out Of Bag (OOB) samples are used to verify each individual tree and is applicable throughout forest. Out Of Bag error will be estimated based on average misclassification of all the trees. Performance of the machine and the variable weights can be derived based on the error estimates. Below figure explains random forest classifier architecture for multiple applications.

The proposed method can be explained in two phases to improve accuracy. In the first phase, as mentioned above Random Forest classifier algorithm will be trained first and fed to training set to test. This testing procedure is followed to select random variables from the whole tree and allocate rankings to each node. The feature of allocating ranks to each node will be performed based on Bayesian probability. Bayesian probability, allocates ranking to each node and arrange all nodes in ascending order. Node with least ranking will be eliminated first and the elimination process continues by matching accuracy rate pre and post removal of node. Overall, in first stage, bunch of randomly chosen nodes will be generated. Output of the first phase will be fed again to train the classifier in order to improve the classifier accuracy.

N-fold cross validation is one of approach where the initial data will be arbitrarily divided to N equal groups with same proportions and apply RF algorithm N number of times. Every time, One of the group in N groups will be considered as Test group and design will be applied on left over groups (i.e. N-1). Average of errors in each



Ensembling Voting

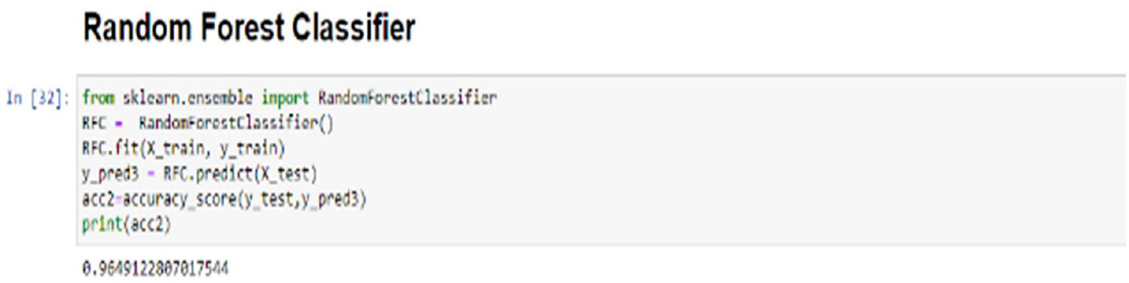
```
In [30]: from sklearn import model_selection

In [31]: seed=7
         kfold = model_selection.KFold(n_splits=10, random_state=seed)
         from sklearn import model_selection
         from sklearn.ensemble import VotingClassifier

         estimators=[]
         model1 = LogisticRegression()
         estimators.append(('logistic', model1))
         model2 =  KNeighborsClassifier()
         estimators.append(('knn', model2))
         # create the ensemble model
         ensemble = VotingClassifier(estimators)
         results = model_selection.cross_val_score(ensemble, X, Y, cv=kfold)
         print(results.mean())

         0.9280075187969924
```

Fig. 7. Ensemble Voting Algorithm give 92% accuracy.



Random Forest Classifier

```
In [32]: from sklearn.ensemble import RandomForestClassifier
         RFC =  RandomForestClassifier()
         RFC.fit(X_train, y_train)
         y_pred3 = RFC.predict(X_test)
         acc2=accuracy_score(y_test,y_pred3)
         print(acc2)

         0.9649122807017544
```

Fig. 8. Random Forest Classifier gives 96% accuracy.

set will be calculated and one set with least error will be selected and learn model parameters.

6. Results of Random Forest classifier to diagnose breast cancer

6.1. 7.Future scope

There are few other possible improvements that can be made to Random Forest Classifier algorithm so as to enhance more accuracy. The following improvements are beyond the scope of this paper and are mentioned under future scope section.

- In the proposed method, Out Of Box (OOB) error analysis is performed to provide rankings and calculate the tree strength. Alternative to OOB, margin of the current forest classifier is considered to calculate the weight of the tree. This can be useful to improve classifier accuracy.
- Bootstrap datasets are used as part of current strategy to find diversity based dynamic pruning. Alternatively, one can also find diversity using OOB datasets.

CRediT authorship contribution statement

Ch. Hari Chandana: Conceptualization, Methodology, Software, Visualization, Writing - original draft. **G. Bala Krishna:** Data curation, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Karabatak, M.C. Ince, An expert system for detection of breast cancer based on association rules and neural network, Expert Syst. Appl. 36 (2) (2009) 3465–3469, https://doi.org/10.1016/j.eswa.2008.02.064.

[2] A. Keleş, A. Keleş, U. Yavuz, Expert system based on neuro-fuzzy rules for diagnosis breast cancer, Expert Syst. Appl. 38 (5) (2011) 5719–5726, https://doi.org/10.1016/j.eswa.2010.10.061.

[3] H.-L. Chen, B.o. Yang, J. Liu, D.-Y. Liu, A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, Expert Syst. Appl. 38 (7) (2011) 9014–9022, https://doi.org/10.1016/j.eswa.2011.01.120.

[4] K. Polat, S. Güneş, Breast cancer diagnosis using least square support vector machine, Digital Signal Process. 17 (4) (2007) 694–701, https://doi.org/10.1016/j.dsp.2006.10.008.

[5] S. Sahana, K. Polat, H. Kodaz, S. Günes, A new hybridmethod based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis, Comput. Biol. Med. 377 (2007) 415–423, https://doi.org/10.1016/j.compbiomed.2006.05.003 [Citation Time(s):1].

[6] A. Marcano-Cedeño, J. Quintanilla-Domínguez, D. Andina, WBCD breast cancer database classification applying artificial metaplasticity neural network, Expert Syst. Appl. 38 (8) (2011) 9573–9579, https://doi.org/10.1016/j.eswa.2011.01.167.

[7] T. Mu, A.K. Nandi, Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier, J. Franklin Inst. 344 (3-4) (2007) 285–311, https://doi.org/10.1016/j.jfranklin.2006.09.005.

[8] D.-C. Li, C.-W. Liu, S.C. Hu, A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets, Artif. Intell. Med. 52 (1) (2011) 45–52, https://doi.org/10.1016/j.artmed.2011.02.001.