Analysis of Classification Algorithms for Breast Cancer Prediction

S. P. Rajamohana, K. Umamaheswari, K. Karunya and R. Deepika

Abstract According to global statistics, breast cancer is the second of all the fatal diseases that cause death. It will cause an adverse effect when left unnoticed for a long time. However, its early diagnosis provides significant treatment, thus improving the prognosis and the chance of survival. Therefore, accurate classification of the benign tumor is necessary in order to improve the living of the people. Thus, precision in the diagnosis of breast cancer has been a significant topic of research. Even though several new methodologies and techniques are proposed machine learning algorithms and artificial intelligence concepts lead to accurate diagnosis, consequently improving the survival rate of women. The major intent of this research work is to summarize various researches done on predicting breast cancer and classifying them using data mining techniques.

Keywords Breast cancer prediction · Classification · Decision tree · Feature selection · K-NN · Random forest · SVM

1 Introduction

A malignant tumor developed from breast cells is termed as breast cancer, and it is the most commonly observed type of tumor in women. Almost 70% of deaths due to cancer arise in poverty-stricken and developing countries. Breast cancer affects about

S. P. Rajamohana (✉) · K. Umamaheswari · K. Karunya · R. Deepika
Department of Information Technology, PSG College of Technology,
Coimbatore 641004, India
e-mail: monamohanasp@gmail.com

K. Umamaheswari
e-mail: umakpg@gmail.com

K. Karunya
e-mail: karunyakandimuthu@gmail.com

R. Deepika
e-mail: deepika.rajendran97@gmail.com

N. Sharma et al. (eds.), *Data Management, Analytics and Innovation*, Advances in Intelligent Systems and Computing 1042, https://doi.org/10.1007/978-981-32-9949-8_36

10% of women at some point during their lives. As of late, the death rate continues to expand, with 88% of survival following 5 years from determination and 80% following 120 months from conclusion. There exists a popular method, called knowledge discovery in databases (KDD) which is used by medical researchers to predict the disease outcome by discovering patterns and relationships among the various number of variables stored in the database as historical data. In order to maintain the natural working mechanism of the body, a balance must be maintained between the growth and death rate of cells. But sometimes there is an abnormal and rapid growth of cells which can lead to cancer. Among numerous types of cancer, breast cancer is the major cause for the death of the women worldwide [1]. It is proved that there is a reduction of 38–48% in the mortality rate. There might be several reasons for breast cancer, including age, bosom density, obesity, and changing food habits like alcohol, causing undesirable effects. It is evident from the recent statistics tha, currently the scenario has become worse. The major symptom is a lump observed underarm or in the mammary gland that continues to exist even after the menstrual cycle of a woman. Out of the lump formed, most of them are painless, some may give prickly sensation. The other common symptoms are redness, swollen lymph nodes, or thickening or puckering off the skin. Milk secreting and duct cells are the main cells that cause tumors by draining the milk into the nipple from lobules. A proportion of growth developed from fibrous or fatty tissue can lead to breast cancer. This kind of gene mutation related to cancer is very popular among woman. The treatment for breast cancer includes chemotherapy, hormonal therapy, and radiation therapy. The efficiency result of treatment is comparatively low and requires more attention for preventive measures and control in the current research world. Many breast cancer charities are organizing campaigns such as National Breast Cancer Awareness Month (NBCAM) on the eighth month of every year to bring about awareness of breast cancer in society. This annual international health campaign assists in raising funds for research in breast cancer.

Breast Cancer is categorized as follows:

(i) **Benign (Noncancerous)** The benign cases are those which are noncancerous and non-life threatening. But they could turn into cancerous cells easily. However, these cells could be easily separated.
(ii) **Malignant (Cancerous)** Malignant case leads to abnormal cell growth invading nearby tissues. It is life-threatening in most cases.

Many researchers proposed various data mining algorithms that are deployed for the diagnosis of breast cancer. Among which, feature extraction and classification algorithms employed in it lead to the design of an efficient system. These techniques provide a significant process for extracting the key features which can lead to a proper diagnosis. It is experimentally proven that machine learning and deep learning algorithms are efficient when compared to conventional approaches [2].

2 Breast Cancer Overview

The twentieth century is also called as the cancer century because hundreds of cancer types were discovered in this century. After decades of hard work in analyzing various types of cancers, doctors are now able to identify the causes of these diseases, preventive measures to adopt, and type of treatment to be given. Among all types of cancers existing, breast cancer is rampant among women, very rarely in men. Heredity is a major factor among multiple factors that can cause breast cancer. Nearly 15–20% of women identified with breast cancer has had a recorded occurrence of the same through their generations. In extremely rare cases, a gene called p53 is responsible for breast cancer. The severity of breast cancer was 16 times more than average in families having this type of gene. The number of families with this gene is about 100 all over the world. Researchers have noticed the double risk of breast cancer in individuals producing wet wax of ear glands than in those producing dry wax. Some of the most common symptoms and types are listed in Tables 1 and 2.

Table 1 Breast cancer factors with symptoms

Breast cancer risk factors	Symptoms
Age	• Redness • Swollen lymph nodes • Thickening or puckering of the skin • Scaling of nipple
Family history	
Genetics	
Breast cancer history	• Dimpled skin • Change in the texture of the skin • Breast or nipple pain
Exposed to radiation in the chest or face below the age of 30	
Race/ethnicity	
Being overweight	
Menstrual history	
Drinking alcohol	

Table 2 Breast cancer types

Types of breast cancer	Description
Metastatic breast cancer	Cancer cells break and spread from the original tumor to other parts by means of the lymphatic system or through the blood
Phyllodes tumors of the breast	These are rare case tumors which require surgery to reduce the risk as they can grow rapidly fast. Phyllodes tumors are malignant and borderline benign
Lobular carcinoma in situ	In this type, the growth of cells in milk producing glands is abnormal and increases the person's risk of developing invasive breast cancer
Inflammatory breast cancer	Instead of lump, this type of cancer is found with swelling and reddening in the area of the breast and spreads rapidly to other areas with symptoms worsening
Invasive lobular carcinoma	Also known as infiltrating lobular carcinoma. It starts in duct cells of milk carrying area and spreads beyond it

3 Related Works

There are several studies on breast cancer prediction based on machine learning algorithms.

A. A Novel Approach for Breast Cancer Detection Using Data Mining Techniques [2014]

Chaurasia and Pal [3] compared the performance of the supervised learning classifiers like Naive Bayes, decision tree, and SVM-RBF kernel and simple CART. The dataset that is used for classification is the Wisconsin breast cancer dataset that comprises 11 attributes. The attributes include sample code number, clump thickness, uniformity of cell size, cell shape, and marginal adhesion. The experiment's results proved that the SVM-RBF kernel has the maximum accuracy of about 96.84.

B. Artificial Neural Networks Applied to Survival Prediction in Breast Cancer [2000]

Lundin et al. [4] constructed the classification model using neural network and statistical analysis. The dataset that is used consists of the attributes like age, primary tumor size, axillary nodal status, mitotic count, and tumor necrosis. The dataset contains the details of 951 breast cancer patients, predicted 5-, 10-, and 15-year breast cancer by using ANN and logistic regression models. The accuracy obtained from the constructed model is about 94.35% which is a considerable accuracy.

C. Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis [2009]

Choi et al. [5] constructed the classification model using artificial Bayesian network. The dataset that is used consists of the attributes such as age at diagnosis, a clinical extension of the tumor, and the number of primary tumors. The dataset consists of the details of 2, 94,275 breast cancer patients. This dataset uses 15 attributes of which seven are primary attributes, seven derived, and one target attribute. The accuracy obtained from the constructed model using the artificial Bayesian network is 88.8%.

D. Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods [2004]

Delen et al. [6] refined a prediction model with the help of ANN, decision tree, and logistic regression to envision breast cancer and to determine the survival rate by analyzing the SEER cancer incidence database. The SEER breast cancer dataset consists of 4, 33,272 records and 72 variables. The accuracy obtained from the model is 89.2%.

E. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence [2013]

Ahmed et al. [7] analyzed the use of decision trees, SVM, and artificial neural network on breast cancer prediction. The dataset used contains population characteristics and includes 22 input variables. The sample dataset is the data collected from 1189 women. The experimental results found that the SVM produced the least error rate with greater accuracy of 95.7%.

F. Breast Cancer Prediction Using DT-SVM Hybrid Model [2015]

Sivakami [8] developed a model for the breast cancer prediction which uses the Wisconsin breast cancer dataset. The dataset encompasses 699 records, containing 458 belonging to the benign class and the remaining to malignant class. The data is collected from the needle extracts of patients' breasts. Before the prediction process, preprocessing must be done. Preprocessing fills up missing values with attribute's mean value in the dataset. The nine traits of the dataset include uniformity in size of the cell, the shape of cell, and thickness of clump. The implemented algorithm in this work is based on SVM and decision tree. The accuracy obtained from the classification model is 91%, and the error rate is 2.58. The number of precisely classified instances is 459 and imprecisely classified instances are 240.

G. Combining Bagging and Boosting [2007]

Kotsiant et al. [9] worked on various ensemble approaches such as bagging, boosting, and combination of both with a variety of base learners. They are C4.5, Naïve Bayes, OneR, and decision stump. These algorithms used datasets from the UCI repository dataset. The accuracy obtained is about 93.47%.

H. Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors [2017]

Islam et al. [10] used the most common dataset, Wisconsin breast cancer (WBC) dataset obtained from the repository of UCI machine learning. This dataset consists of 699 instances, where the cases are labeled as either benign or malignant. Classification is performed using SVM and K-NN. The accuracy obtained from the model using SVM and K-NN is 98.57% and 97.14%, respectively.

I. Heterogeneous Classifiers Fusion for Dynamic Breast Cancer Diagnosis Using Weighted Vote-Based Ensemble [2014]

This work proposed an ensemble approach by combining various classifiers including decision tree with Gini index and information gain, Naïve Bayes, SVMn and memory based learner. The classification is decided on the basis of weighted voting. Different datasets are collected from the public repository. Classification accuracy is enhanced using different preprocessing techniques and feature selections. The experimental results conclude that proposed approach contributed to major improvement than other existing classifiers. The accuracy obtained from the ensemble model is about 97.2%, the precision is 100%, and the recall value is 98.60%.

4 Proposed Methodology

4.1 Data Source

The publicly available breast cancer database is used. The database [11] constitutes about 570 records. The dataset consists of numeric attributes. Most of the research papers referred to the 32 attributes present in the dataset for breast cancer prediction.

4.2 Data Preprocessing

The dataset applied in the proposed work is a clinical dataset which may contain many inconsistent, incomplete, or missing data. Such data reduces the accuracy of the model. Hence, preprocessing is an important step that needs to be carried out. To remove the inconsistencies in the dataset, several preprocessing techniques were applied. Also, normalization techniques were applied to handle the missing data. Thus, preprocessing techniques tend to increase the accuracy of the model constructed. The workflow of the proposed model is shown in Fig. 1.

4.3 Random Forest

For the classification of malignant and benign types of cancer, random forest algorithm can be employed. Based on various types of randomization, random forest has

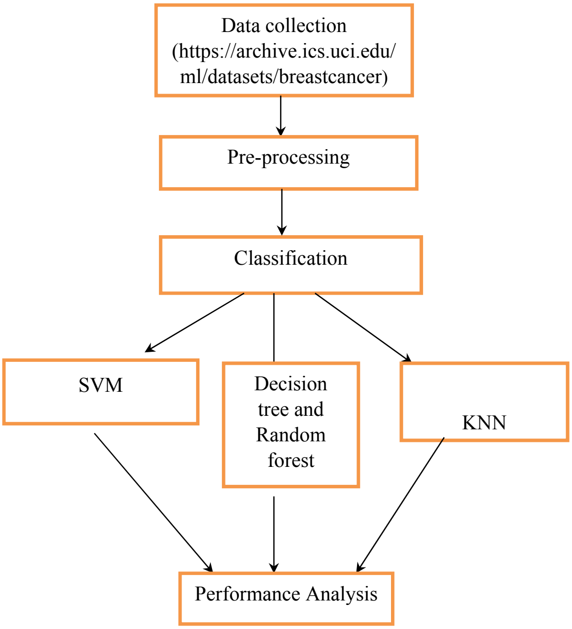

Data collection (https://archive.ics.uci.edu/ml/datasets/breastcancer)
Pre-processing
Classification
SVM
Decision tree and Random forest
KNN
Performance Analysis

Fig. 1 Workflow of the proposed model

been built as it is an ensemble of decision trees. As the random forests are very flexible rather it is widely used [12]. This supervised learning algorithm creates forests using many trees. The accuracy depends on the number of trees. One of the major advantages with random forest is that it could be functioned as both regression and classification. It can also handle the missing values and it will not overfit in case of a greater number of trees [13]. The random forest takes the test features and predicts the outcome of the randomly created trees based on rules and then stores the result. The votes for each predicted target is calculated, since each tree results in different prediction. Finally, the target receiving high vote will be considered as the final prediction. The random forest processes a huge amount of data at very high speed. The random forest has each tree in binary structure form, which is created based on top-down approach. Generally, the convergence in the random forest algorithm is very fast. The most important parameters are the depth and number of trees. Increasing depth led to an increase in the performance [14]. Thus, random forest is considered as the best classification algorithm on the basis of the processing time and accuracy. This algorithm is implemented on the dataset and the accuracy obtained is 97.34%. The overall view of the random forest algorithm is laid out in Fig. 2.

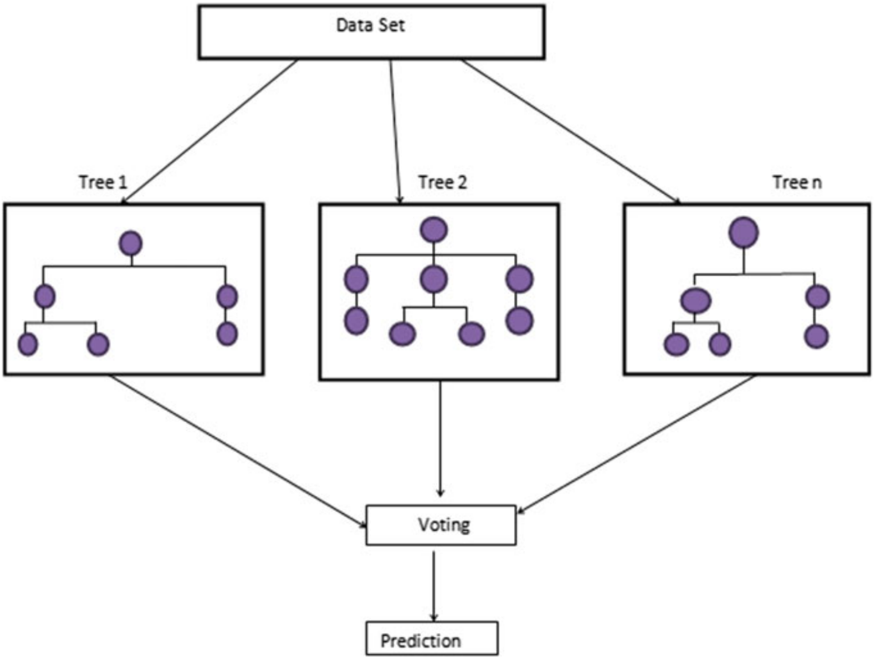


Fig. 2 Random forest

4.4 Support Vector Machine (SVM)

Another powerful supervised learning model is the support vector machine that examines the data for classification and regression analysis using the support of associated learning algorithms. SVM has been specialized for a certain range of problems and gained success in the field of pattern recognition specifically in bioinformatics and cancer diagnosis. In support vector machine, all the data points are drafted in an N-dimensional space [15]. SVM performs linear classification. In addition to this, it also performs nonlinear classification by mapping the inputs implicitly at the range of high-dimensional feature space. The main advantage of SVM is considered to be a unique technique called kernel trick, where lower dimensional space is converted to higher dimensional space and classified. In other words, a hyperplane or set of hyperplanes can be constructed with support vector machine constructs in a high- or infinite-dimensional space for the purpose of classification, regression, or outlier detection [16]. In general, the hyperplane with the largest functional margin achieves good separation as it lowers the generalization error of the classifier. Support vector machine gives an accuracy of 97% when employed on the dataset. The hyperplane used for classification is illustrated in Fig. 3.

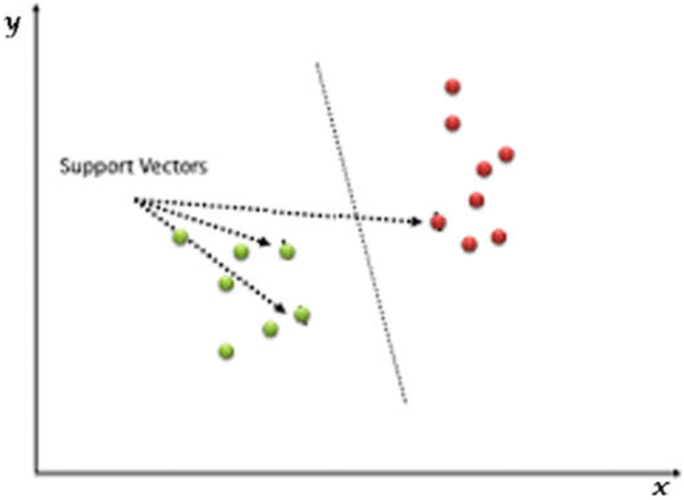


Fig. 3 Support vector machine

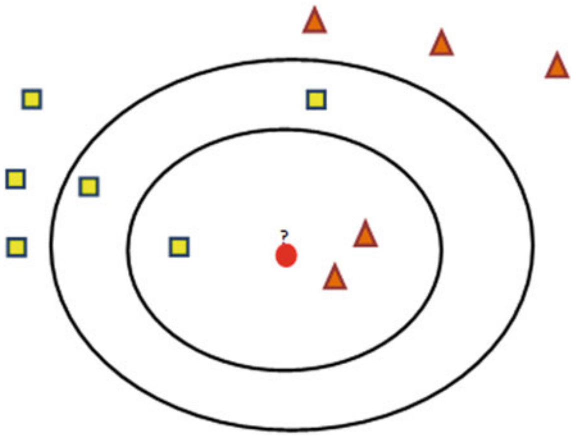
4.5 K-Nearest Neighbor

A nonparametric method, K-nearest neighbors algorithm (K-NN) is another algorithm to be utilized for various purposes like regression and classification. The output of the classification from k-NN is a class label which describes to which class or group it belongs to [17]. In K-NN, the membership is assigned based on the majority vote by the neighbors which is decided by the K value. In other words, each object has been assigned to one class that is most common among the neighbors. For example, if $K = 2$, then the query point is assigned to the class where two nearest neighbors belong to. Being the simplest machine learning algorithm, the explicit training step is not essential. In the training step, neighbors are chosen from the set of objects for which their corresponding classes are known. The algorithm is very sensitive to the local data [18]. Euclidean distance for continuous variables and Hamming distance for discrete variables are most widely used. However, the accuracy can be improved by using specialized algorithms like Large Margin Nearest Neighbor or Neighborhood Components Analysis. The K value is chosen based on the data. Choosing an appropriate K value is significant because that decides if the data is classified correctly. Heuristic techniques like hyperparameter optimization are used because the larger K value reduces the effect of noise but makes less distinct boundaries between classes [19]. The performance gets degraded as the noise in the data increases. The level of accuracy achieved from K-NN is about 95% with an appropriate K value. K-NN is shown in Fig. 4.

4.6 Decision Trees

These are classification algorithms where the attributes in the dataset are recursively partitioned. Decision trees contain many branches and leaf nodes. All the branches tell the conjunction of the attributes that leads to the target class or class labels [20].

Fig. 4 K-nearest neighbor



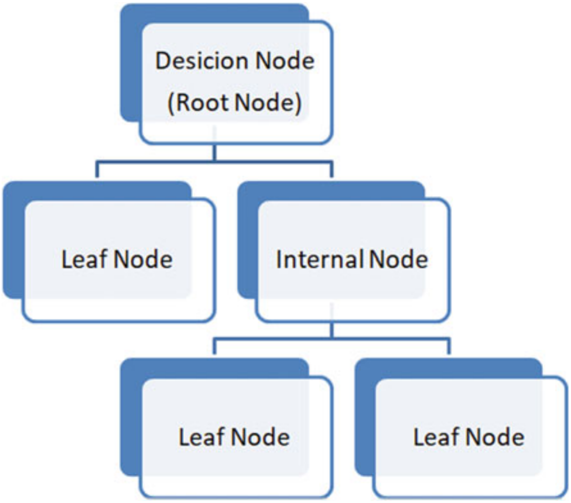


Fig. 5 Decision tree

The leaf nodes contain class labels or the target class that tells to which class tuple it belongs to [21]. There are various decision tree algorithms which can be used for classification of the data. Some algorithms include C4.5, C5, CHAID, ID3, J48, and CART [11, 22].

The Decision Tree can be built as

- The attribute splits decide the attribute to be selected.
- The decisions whether to continue for splitting or to represent as a terminal node is made.
- The assignment of the terminal node to a class.

Information gain, gain ratio, Gini index, etc. are the impurity measures to decide attribute splits done on the tree. After pruning, the tree is checked against noise and overfitting. As a result, the tree becomes an optimized tree. The main advantage of having a tree structure is that it is very easy to understand and interpret. The algorithm is also very robust to the outliers also. The structure of the decision tree is shown in Fig. 5.

5 Experimental Results

Initially, the entire dataset was preprocessed and normalized to handle all the missing values. The mandatory features are extracted using statistical methods. Similar to the experimental results shown in the table above, for the breast cancer prediction, random forests can give a better result that helps the people to get a systematic way of treatment and save their life and hence decreasing the mortality rate. Support vector machines are also equally good with an accuracy of about 91%.

Table 3 Performance analysis of various machine learning algorithms

Algorithm	Accuracy (%)	Precision (%)	Recall (%)
Support vector machine	91	82	81
Decision trees	90.6	83	79
K-nearest neighbor	87	78	76.8
Random forests	93.34	89	86.3

The decision tree and K-NN also give us a good result in predicting breast cancer with an accuracy of 90.6% and 87%, respectively. Hence, a good and accurate model can be built using the algorithms mentioned above. In conclusion, random forest has proven that it is very efficient for breast cancer prediction with an accuracy of 93.34% for diagnosis is exhibited in Table 3. It achieves the finest performance with respect to precision and recall.

6 Conclusion

Most of the Indian women die due to breast cancer [23]. Proper diagnosis of breast cancer is very important in the medical domain. Many models are built using machine learning and data mining methodologies to analyze huge voluminous data. But the challenge with the model is its accuracy and precision for the medical industry. The key for proper treatment and cure is significantly dependent on detecting breast cancer at its early stages [24]. This paper describes how several machine learning algorithms like random forests, decision trees, K-NN, and SVM are employed to model the exact diagnosis of breast cancer for an organized treatment which can save lots of life. The experimental results show the effectiveness of various machine learning algorithms and it proves to be efficient in breast cancer classification thus providing a major breakthrough in clinical diagnosis and treatment of the same. As mentioned earlier, the SVM, K-NN, decision tree, and random forest have their own advantages. Hence, the careful usage of the above will definitely lead to better results. Thus, the appropriate use of a proper algorithm eliminates the risk of death and alleviates the survival rate in women.

References

1. Ferlay, J., Héry, C., Autier, P., Sankaranarayanan, R.: Global burden of breast cancer. In: Li, C. (ed.) Breast Cancer Epidemiology, pp. 1–19. Springer, New York (2010)
2. Sharma, A., Kulshrestha, S., Daniel, S.: Machine learning approaches for breast cancer diagnosis and prognosis. In: Proceedings of the International Conference on Soft Computing and Its Engineering Applications, Changa, India, 1–2 December (2017)

3. Chaurasia, V., Pal, S.: Data mining techniques: to predict and resolve breast cancer survivability. Int. J. Comput. Sci. Mobile Comput. **3**(1), 10–22 (2014)
4. Lundin, M., Lundin, J., Burke, H.B., Toikkanen, S., Pylkkänen, L., Joensuu, H.: Artificial neural networks applied to survival prediction in breast cancer. Oncology **57**(4), 281–286 (1999)
5. Choi, J.P., Han, T.H., Park, R.W.: A hybrid Bayesian network model for predicting breast cancer prognosis. J. Korean Soc. Med. Inform. **15**(1), 49–57 (2009)
6. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. **34**(2), 113–127 (2005)
7. Ahmad, L.G., Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M., Razavi, A.R.: Using three machine learning techniques for predicting breast cancer recurrence. J Health Med. Inform. **4** (124), 3 (2013)
8. Sivakami, K., Saraswathi, N.: Mining big data: breast cancer prediction using DT-SVM hybrid model. Int. J. Sci. Eng. Appl. Sci. (IJSEAS) **1**(5), 418–429 (2015)
9. Kotsianti, S.B., Kanellopoulos, D.: Combining bagging, boosting and dagging for classification problems. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 493–500. Springer, Berlin (2007)
10. Islam, M.M., Iqbal, H., Haque, M.R., Hasan, M.K.: Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 226–229. IEEE
11. Barghout, L.: Spatial-taxon information granules as used in iterative fuzzy-decision-making for image segmentation. In: Pedrycz, W., Chen, S.M. (eds.) Granular Computing and Decision-Making, pp. 285–318. Springer, Cham (2015)
12. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 37–44. IEEE (2006)
13. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
14. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007, pp. 1–8. IEEE
15. Marcano-Cedeño, A., Quintanilla-Domínguez, J., Andina, D.: WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst. Appl. **38**(8), 9573–9579 (2011)
16. TP, L., Parthiban, L.: Abnormality detection using weighed particle swarm optimization and smooth support vector machine. Biomed. Res. (0970–938X), 28(11) (2017)
17. Mirkes, E.: KNN and Potential Energy (Applet). University of Leicester, Leicester (2011)
18. Breast Cancer Organization: http://www.breastcancer.org/symptoms/
19. Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Amsterdam (2011)
20. Devi, R.D.H., Devi, M.I.: Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer. Int. J. Adv. Eng. Technol. **VII**(II), 98 (2016)
21. Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., Colliot, O.: Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. Med. Image Anal. **15**(5), 729–737 (2011)
22. Bhargava, N., Sharma, G., Bhargava, R., Mathuria, M.: Decision tree analysis on j48 algorithm for data mining. Proc. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**(6),1114–1119 (2013)
23. Wolberg, W.H., Mangasarian, O.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine
24. Blockeel, H., Struyf, J.: Efficient algorithms for decision tree cross-validation. J. Mach. Learn. Res. **3**, 621–650 (2002)