Int. J. Res. Ind. Eng. Vol. 9, No. 3 (2020) 235–246
DOI:10.22105/riej.2020.259298.1155

International Journal of Research in Industrial Engineering
www.riejournal.com

Breast Cancer Detection Using Machine Learning Algorithms

R. Shastri*, N. Pradeep, K. K. Rao Mangalore, B. Rajpal, N. Prasad

Department of MCA, School of Computer Science and IT, Jain (deemed-to-be) University, Bengaluru, India.
(*Corresponding Author's Email Address: ravishastri9031@gmail.com)

A B S T R A C T

Breast cancer has been the riskiest malignancy among ladies around the world. Nearly 2 million new cases were diagnosed in 2018. The main problem in the detection of breast cancer is to find how tumors turn into malignant or benign and we can do this with the help of machine learning techniques as they provide an appropriate result. According to research, an experienced physician can diagnose cancer with 79% accuracy while using machine learning techniques provides an accuracy of 91%. In this work, machine learning techniques have been applied which include K-Nearest Neighbors algorithm (KNN), Support Vector Machine (SVM), and Decision Tree Classifier (DT). To predict whether the cause is benign or malignant we have used the breast cancer dataset. The SVM classifier gives more accurate and precise results as compared to others, and this classifier is trained with the larger datasets.

Keywords: Machine learning, K-nearest neighbors, Support vector machine, Decision tree classifier, Jupyter.

Article history: *Received: 11 April 2020 Reviewed: 09 May 2020 Revised: 14 July 2020 Accepted: 23 August 2020*

1. Introduction

The enhancement of science and technologies has made life more comfortable than in older days. The emerging technologies like neutrosophic shortest path [1-5], transportation problem [6-8], uncertainty problem [9,-14], fuzzy shortest path [15-19], Powershell [20], wireless sensor network [21-28], computer language [29,30], neural network [31], routing [32], image processing [33] have made the products more intelligent and self-healing based. Smart city applications like smart water [34, 35], smart grid, smart parking, smart resource management, etc. are based on IoT and IoE [36-39] technologies. According to research, an experienced physician can diagnose cancer with 79% accuracy while using machine learning techniques provides an accuracy of 91%. According to the World Health Organization (WHO) bosom malignancy is situated as the topmost infection worldwide and it is extending bit by bit in most of the countries. According to cancer.net statistics, an approximate 43,000 deaths will happen because of breast cancer this year. To diagnose breast cancer, doctors use many tests which include imaging tests, biopsy, analyzing the biopsy sample, genomic test to predict recurrence risk, and blood test. The arrival of new medical technologies and a large amount of data have triggered the path for the development of

new techniques in the detection of breast cancer. Also, retrieving information from an enormous amount of data is a very difficult task. Machine Learning (ML) classifiers can be used to handle this task. ML techniques have been used in the past few years for the growth of models to support effective decision making.

There are many ways to detect breast cancer. This task focuses on classifying the performance of the ML techniques which has been used in the project. And the performance can be evaluated based on the accuracy of classification, recall, and precision.

2. Related Works

Numerous examinations with many ideas and techniques are utilized in the field of breast cancer detection. Numerous scientists have introduced various techniques and calculations to recognize breast cancer. Here, we will talk about some of them.

Microwave radiometry was used by Barrett et al. [40] in 1977 to detect breast cancer. A microwave radiometer was used to measure the thermal radiation of the body. They combined infrared thermographic and microwave data which provided them with a 96% positive detection rate based on 30 cases.

The authors [41] discuss the errors which were missed by mammography while detecting breast cancer. Two major errors were disclosed:

- Poor radiographic technique.
- Shortage in radiographic criteria of cancer.

In another study [42], the applications of ANNs were applied to the survival analysis issue. The ANNs results were compared on different datasets that use morphometric features. The result indicates that ANNs were successful in predicting recurrence probability and separating patients with bad and good prognoses. In 2010 [43], Computer-Aided Diagnosis (CAD) system i.e. ultrasound imaging was used to detect breast cancer which provided more improved diagnosis accuracy results.

3. Literature Review

3.1. Different Researcher's Contributions

Some of the major contributions to breast cancer detection are discussed in the ***Table.1***.

Table 1. A literature review of Breast cancer detection.

Authors	Years	Different Techniques used in breast cancer detection
Gershon-Cohen et al.[44]	1961	The authors proposed a routine examination in the Radiology Department of the Albert Einstein medical Centre to detect cancer.
Stevens et al. [45]	1966	The authors proposed to use a mammographic survey by implementing a modified Egan technique to detect breast cancer.
Barrett et al. [40]	1977	The authors proposed to use microwave radiometry for determining breast cancer.
Martin et al. [41]	1979	The authors proposed the issues found in the mammography strategy which incorporates poor radiographic method, nonattendance of radiographic models of malignant growth, clear oversight by the radiologist, and absence of acknowledgement of unobtrusive radiographic signs.
Li et al. [46]	2001	The authors proposed (2-D) an anatomically realistic MRI-derived FDTD model of the cancerous breast for cancer detection.
Zou et al. [47]	2003	The authors proposed an electrical impedance technique to detect breast cancer.
Chi et al. [42]	2007	The authors proposed Artificial Neural Networks (ANNs) to predict the result of breast cancer.
Cheng et al. [43]	2010	The authors proposed to use the Computer-Aided Diagnosis (CAD) system i.e. ultrasound imaging which provides more improved diagnostic accuracy.

Additionally, the higher literature review reveals that there are gaps in the study of breast cancer detection. As such, the subsequent gaps are studied:

- Systems like mammography miss breast cancer detection due to poor radiographic technique.
- Large datasets affect the accuracy of algorithms used in the models.
- Sometimes cancer is not visible at the time of mammography.

Thus, it motivates us to provide a new model for society:

- Implementing machine learning algorithms to larger datasets helps to improve the accuracy of results.
- Using ml techniques gives more precise results than experienced physicians.
- It reduces overfitting.

4. Material and Methods

4.1. Dataset Used for Research

In this work, we have used the Breast cancer dataset which was retrieve using the UCI repository. The dataset includes information of 699 patients of Wisconsin hospitals and it also identifies the number of malignant and benign cases i.e. 249 and 450. Dataset consists of ten attributes, some of them are mentioned below:

- Clump Thickness
- Marginal Adhesion
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses
- Class

4.2. Classification Techniques

Classification is part of the supervised learning approach in which a program first learns from the data i.e. input data and then it uses this learning for the classification of new observations. In other words, a training dataset is employed to obtain appropriate boundary conditions which are used to identify each target class, once such boundary conditions are determined, and to predict the target class is the next task.

ML as a field of study is worried about calculations that gain from models. There are various sorts of characterization assignments that you may experience in AI and particular ways to deal with the model that might be utilized for each.

4.2.1. K-nearest neighbor algorithm (KNN)

KNN is an algorithm that keeps a record of all cases and categories of new cases based upon comparison measures. In 1970 this algorithm was used as a non-parametric technique. This algorithm is one of the basic procedures utilized in machine learning. It is a technique favoured by numerous individuals in the industry because it is simple to use and takes low calculation time.

Pros.

- Easy to use.
- Quick calculation time.

Cons.

- Accuracy relies on the nature of the data and must locate an ideal k value.
- Poor at ordering information focus on a limit where they can be arranged somehow.

4.2.2. Support vector machine (SVM)

In machine learning, Support Vector Machine Algorithms (SVMs) are flexible and powerful supervised algorithms. They are used for regression and classification. In classification problems, it is generally used. SVMs was first introduced in 1960 but later it was improved in 1990. It differs from other ML algorithms because it uses a unique way of implementing. In scikit-learn SVMs support sparse and dense vectors as input.

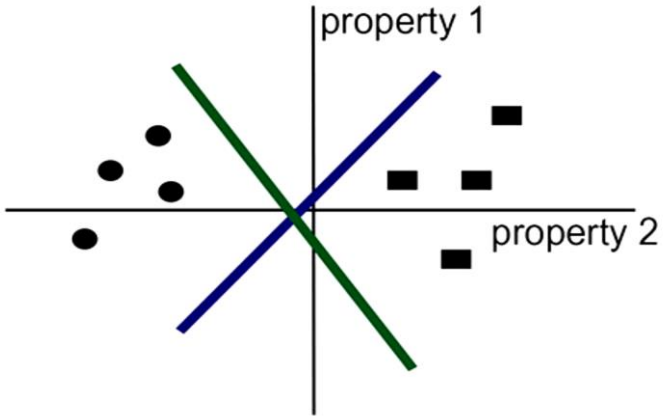

Figure 1. SVMs generated hyper-planes.

4.2.3. Decision tree classifier

The decision tree comes under the supervised machine learning algorithm. To create a classification model decision tree classifier uses a decision tree. It consists of nodes, edges/branches, and leaf nodes.

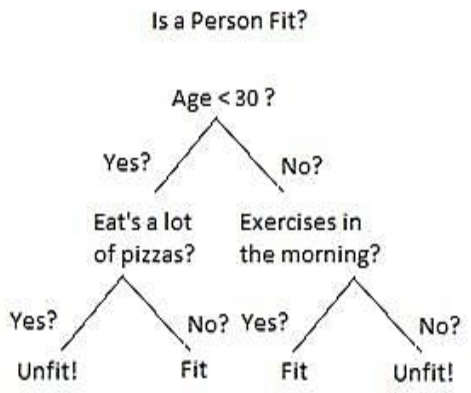

Figure 2. Decision tree classification for a persons's fitness**.**

5. Experiment

5.1. The Proposed Method

In this work, various machine learning algorithms were used to find out a benign and malignant tumour by importing the dataset. By using the matplotlib library, different types of maps were plotted to find the actual number of affected cases. After plotting the maps, the dataset was split into two parts i.e. training and testing part, as it will help us train our models to check their accuracy. Two variables were used after splitting the dataset i.e. dependent and independent variable (X and Y). We have applied feature scaling because the machine learning algorithm does not understand the unit. So by applying it, we are making it unitless by putting them in a particular range. Though we are applying it on the (X) variable because it is an independent variable and the dependent variable (Y) is already in the range 0,1.

Then each model is trained by importing them from sklearn.linear_model by making a classifier of the models. We have also acquired the classification report using the sklearn.metrics which reports the models.

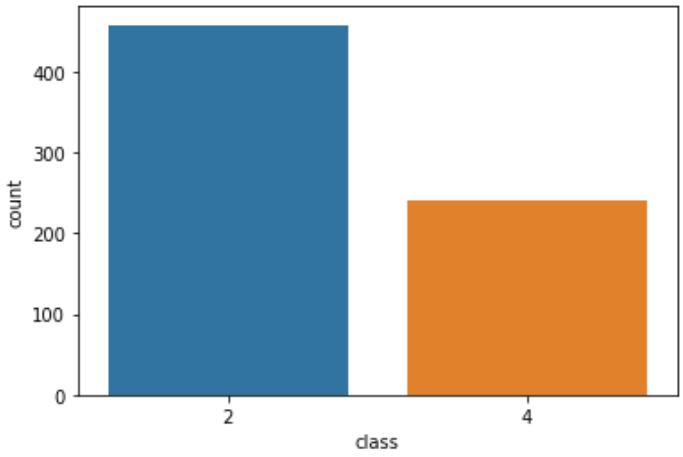

Figure 2. The number of benign and malignant cases; 2- benign 4-malignant.

6. Result and Discussion

This part describes the results of the classifiers which has been used in this paper. It includes the classification report which consists of accuracy, recall, and precision.

6.1. Accuracy

Accuracy tells how well the classifier can predict the correct cases into their actual category. To find accuracy, the number of right predictions are divided by the total number of instances in the dataset. ***Table 2*** shows the accuracy values of the models.

$$Accuracy = \frac{\text{Correct Predictions}}{\text{All Predictons}}$$

Table 2. Accuracy values of models.

	KNN	SVM	DTC
Accuracy	95%	96%	93%

6.2. Precision

Precision tells us that it handles the positive predictions and on the other hand it does not describe much about the negative predictions. To find precision, true positives are divided by true positives + false positives. ***Table 3*** shows the precision values of the classifiers. It is also known as positivity.

Precison = $\frac{tp}{tp+fp}$, where tp = True Positives and fp = False Positives

Table 3. Precision values of classifiers.

	KNN	SVM	DTC
Benign	96%	97%	95%
Malignant	94%	94%	92%
Average	95%	96%	93%

6.3. Recall

Sensitivity is used to find recall values true positive values are divided by true positives + false negatives. ***Table 4*** shows the recall value of the classifiers.

Recall = $\frac{tp}{tp+fn}$**,** where tp = True Positives and fn = False Negatives**.**

Table 4. Recall values of classifiers.

	KNN	SVM	DTC
Benign	96%	96%	96%
Malignant	94%	95%	90%
Average	95%	96%	94%

7. Conclusion

In this work of breast cancer prediction, we have applied machine learning models which have put up a great performance. The main intent of the paper was to find malignant and benign cases and to provide more accurate results as compared to the experienced physicians. And the results above have shown that the models have provided a better performance while experimenting. We have also described the performance comparison of the used models. It shows that SVM has the highest performance in the field of accuracy, precision, and recall. In the last few years, machine learning methods have been utilized in the field of clinical science to settle an increasing number of complex clinical issues. This model can be deployed in an application for the well-being of society.

References

[1] Broumi, S., Dey, A., Talea, M., Bakali, A., Smarandache, F., Nagarajan, D., ... & Kumar, R. (2019). Shortest path problem using Bellman algorithm under neutrosophic environment. *Complex & intelligent systems*, *5*(4), 409-416.

[2] Kumar, R., Dey, A., Broumi, S., & Smarandache, F. (2020). A study of neutrosophic shortest path problem. In *Neutrosophic graph theory and algorithms* (pp. 148-179). IGI Global.

[3] Kumar, R., Edalatpanah, S. A., Jha, S., Broumi, S., Singh, R., & Dey, A. (2019). *A multi objective programming approach to solve integer valued neutrosophic shortest path problems*. Infinite Study.

[4] Kumar, R., Edalatpanah, S. A., Jha, S., & Singh, R. (2019). *A novel approach to solve gaussian valued neutrosophic shortest path problems*. Infinite study.

[5] Kumar, R., Edaltpanah, S. A., Jha, S., Broumi, S., & Dey, A. (2018). *Neutrosophic shortest path problem*. Infinite Study.

[6] Pratihar, J., Kumar, R., Dey, A., & Broumi, S. (2020). Transportation problem in neutrosophic environment. In *Neutrosophic graph theory and algorithms*. IGI Global.

[7] Kumar, R., Edalatpanah, S. A., Jha, S., & Singh, R. (2019). A Pythagorean fuzzy approach to the transportation problem. *Complex & intelligent systems*, *5*(2), 255-263.

[8] Pratihar, J., Kumar, R., Edalatpanah, S. A., & Dey, A. (2020). Modified Vogel's approximation method for transportation problem under uncertain environment. *Complex & intelligent systems*, 1-12.

[9] Gayen, S., Jha, S., Singh, M., & Kumar, R. (2019). On a generalized notion of anti-fuzzy subgroup and some characterizations. *International journal of engineering and advanced technology*, *8*, 385-390.

[10] Gayen, S., Smarandache, F., Jha, S., & Kumar, R. (2020). Interval-valued neutrosophic subgroup based on interval-valued triple t-norm. In *Neutrosophic sets in decision analysis and operations research* (pp. 215-243). IGI Global.

[11] Gayen, S., Smarandache, F., Jha, S., Singh, M. K., Broumi, S., & Kumar, R. (2020). Introduction to plithogenic subgroup. In *Neutrosophic graph theory and algorithms* (pp. 213-259). IGI Global.

[12] Gayen, S., Smarandache, F., Jha, S., Singh, M. K., Broumi, S., & Kumar, R. (2020). Soft Subring Theory Under Interval-valued Neutrosophic Environment. *Neutrosophic sets and systems*, *36*(1), 16.

[13] Gayen, S.; Smarandache, F.; Jha, S.; Kumar, R (2020). Introduction to interval-valued neutrosophic subring. *Neutrosophic sets and systems*, *36*, pp 220-245.

[14] Gayen, S., Smarandache, F., Jha, S., Singh, M. K., Broumi, S., & Kumar, R. (2020). Introduction to plithogenic hypersoft subgroup. *Neutrosophic sets and systems*, *33*(1), 14.

[15] Kumar, R., Edalatpanah, S. A., & Mohapatra, H. (2020). Note on "Optimal path selection approach for fuzzy reliable shortest path problem". *Journal of intelligent & fuzzy systems*, (Preprint), 1-4.

[16] Kumar, R., Jha, S., & Singh, R. (2020). A different approach for solving the shortest path problem under mixed fuzzy environment. *International journal of fuzzy system applications (IJFSA)*, *9*(2), 132-161.

[17] Kumar, R., Jha, S., & Singh, R. (2017). Shortest path problem in network with type-2 triangular fuzzy arc length. *Journal of applied research on industrial engineering*, *4*(1), 1-7.

[18] Kumar, R., Edalatpanah, S. A., Jha, S., Gayen, S., & Singh, R. (2019). Shortest path problems using fuzzy weighted arc length. *International journal of innovative technology and exploring engineering*, *8*(6), 724-731.

[19] Singh, A., Kumar, A., & Appadoo, S. S. (2019). A novel method for solving the fully neutrosophic linear programming problems: Suggested modifications. *Journal of intelligent & fuzzy systems*, *37*(1), 885-895.

[20] Mohapatra, H., Panda, S., Rath, A., Edalatpanah, S., & Kumar, R. (2020). A tutorial on powershell pipeline and its loopholes. *International journal of emerging trends in engineering research*, *8*(4), 975-982.

[21] Mohapatra, H., Rath, S., Panda, S., & Kumar, R. (2020). Handling of man-in-the-middle attack in wsn through intrusion detection system. *International journal of emerging trends in engineering research*, *8*, 1503-1510.

[22] Mohapatra, H., Debnath, S., & Rath, A. K. (2019). Energy management in wireless sensor network through EB-LEACH. *International journal of research and analytical reviews (IJRAR)*, 56-61.

[23] Mohapatra, H., Rath, A. K., Landge, P. B., Bhise, D., Panda, S., & Gayen, S. A. (2020). Comparative analysis of clustering protocols of wireless sensor network. *International journal of mechanical and production engineering research and development*, *10*, 8371-8386.

[24] Mohapatra, H., & Rath, A. K. (2020). A survey on fault tolerance based clustering evolution in wsn. *IET Networks*, 9(4), 145-155.

[25] Mohapatra, H., Debnath, S., Rath, A. K., Landge, P. B., Gayen, S., & Kumar, R. (2020). An efficient energy saving scheme through sorting technique for wireless sensor Network. *International journal*, *8*(8), 4278-4286.

[26] Mohapatra, H., & Rath, A. K. (2020). Fault tolerance in WSN through uniform load distribution function. *International journal of sensors, wireless communications and control* **,** *10*. https://doi.org/10.2174/2210327910999200525164954

[27] Mohapatra, H., & Rath, A. K. (2019). Fault tolerance through energy balanced cluster formation (EBCF) in WSN. In *Smart innovations in communication and computational sciences* (pp. 313-321). Springer, Singapore.

[28] Mohapatra, H., & Rath, A. K. (2019). Fault tolerance in WSN through PE-LEACH protocol. *IET wireless sensor systems*, *9* (6), 358-365(7).

[29] Mohapatra, H (2018). *C Programming: Practice*.Amazon.

[30] Mohapatra, H., & Rath, A. K. (2020). *Fundamentals of software engineering.* BPB.

[31] Mohapatra, H. (2009). *HCR by using neural network* (Master's thesis; M.Tech_s Desertion, Govt. College of Engineering and Technology, Bhubaneswar).

[32] Panda, M., Pradhan, P., Mohapatra, H., & Barpanda, N. K. (2019). Fault tolerant routing in heterogeneous environment. *International journal of scientific & technology research*, *8*, 1009-1013.

[33] Nirgude, V. N., Nirgude, V. N., Mahapatra, H., & Shivarkar, S. A. (2017). Face recognition system using principal component analysis & linear discriminant analysis method simultaneously with 3d morphable model and neural network BPNN method. *Global journal of advanced engineering technologies and sciences*, *4*, 1-6.

[34] Mohapatra, H., & Rath, A. K. (2020, October). Nub Less Sensor Based Smart Water Tap for Preventing Water Loss at Public Stand Posts. In *2020 IEEE Microwave Theory and Techniques in Wireless Communications (MTTW)* (Vol. 1, pp. 145-150). IEEE.

[35] Mohapatra, H., & Rath, A. K. (2020). IoT-based smart water. In *IOT Technologies in Smart-Cities: From Sensors to Big Data, Security and Trust* (pp. 63-82). DOI: 10.1049 /PBCE128E_ch3

[36] Mohapatra, H. (2020). Offline drone instrumentalized ambulance for emergency situations. *International journal of robotics and automation*, *9*, 251-255.

[37] Mohapatra, H., & Rath, A. K. (2019). Detection and avoidance of water loss through municipality taps in India by using smart taps and ICT. *IET wireless sensor systems*, *9*(6), 447-457.

[38] Panda, H., Mohapatra, H., & Rath, A. K. (2020). WSN-Based Water Channelization: An Approach of Smart Water. In *smart cities—opportunities and challenges* (pp. 157-166). Springer, Singapore.

[39] Rout, S. S., Mohapatra, H., Nayak, R. K., Tripathy, R., Bhise, D., Patil, S. P., & Rath, A. K. (2020). Smart Water Solution for Monitoring of Water Usage Based on Weather Condition. *International journal*, *8*(9).

[40] Barrett, A. H., Myers, P. C., & Sadowsky, N. L. (1977). Detection of breast cancer by microwave radiometry. *Radio science*, *12*(6S), 167-171.

[41] Martin, J. E., Moskowitz, M., & Milbrath, J. R. (1979). Breast cancer missed by mammography. *American journal of roentgenology*, *132*(5), 737-739.

[42] Chi, C. L., Street, W. N., & Wolberg, W. H. (2007). Application of artificial neural network-based survival analysis on two breast cancer datasets. In *AMIA annual symposium proceedings* (Vol. 2007, p. 130). American Medical Informatics Association.

[43] Cheng, H. D., Shan, J., Ju, W., Guo, Y., & Zhang, L. (2010). Automated breast cancer detection and classification using ultrasound images: A survey. *Pattern recognition*, *43*(1), 299-317.

[44] Gershon-Cohen, J., & Berger, S. M. (1961). Detection of breast cancer by periodic X-ray examinations: a five-year survey. *JAMA*, *176*(13), 1114-1116.

[45] Stevens, G. M., & Weigen, J. F. (1966). Mammography survey for breast cancer detection. A 2-year study of 1,223 clinically negative asymptomatic women over 40. *Cancer*, *19*(1), 51-59.

[46] Li, X., & Hagness, S. C. (2001). A confocal microwave imaging algorithm for breast cancer detection. *IEEE Microwave and wireless components letters*, *11*, 130–132.

[47] Zou, Y., & Guo, Z. (2003). A review of electrical impedance techniques for breast cancer detection. *Medical engineering & physics*, *25*, 79–90.

[48] Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated breast cancer diagnosis based on machine learning algorithms. *Journal of healthcare engineering*. https://doi.org/10.1155/2019/4253641

[49] Hussain, L., Aziz, W., Saeed, S., Rathore, S., & Rafique, M. (2018). Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies. *17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE)* (pp 327–331).

[50] Chaurasia, V., & Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. *International journal of innovative research in computer and communication engineering (An ISO 3297: 2007 Certified Organization) Vol*, *2*.

[51] Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *5th international conference on electronic devices, systems and applications (ICEDSA)* (pp. 1–4).

[52] Alarabeyyat, A., & Alhanahnah, M. (2016, August). Breast cancer detection using k-nearest neighbor machine learning algorithm. *9th international conference on developments in esystems engineering (DeSE)* (pp. 35-39). IEEE.

[53] Aruna, S., & Rajagopalan, S. P. (2011). A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. *International journal of computer applications*, *31*(8).

[54] Kelly, K. M., Dean, J., Comulada, W. S., & Lee, S. J. (2010). Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. *European radiology*, *20*(3), 734-742.

[55] Adam, A., & Omar, K. (2006). Computerized breast cancer diagnosis with Genetic Algorithm and Neural Network. *Proc. of the 3rd international conference on artificial intelligence and engineering technology (ICAIET)* (pp. 22-24).

[56] Mu, T.; Nandi, A. K (2005). Detection of breast cancer using v-SVM and RBF networks with self-organized selection of centres. *3rd IEE international seminar on medical applications of signal processing* (47-52).

[57] Yao, X., & Liu, Y. (1999, July). Neural networks for breast cancer diagnosis. *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)* (Vol. 3, pp. 1760-1767). IEEE.

[58] Colak, S. B., Van der Mark, M. B., t Hooft, G. W., Hoogenraad, J. H., Van der Linden, E. S., & Kuijpers, F. A. (1999). Clinical optical tomography and NIR spectroscopy for breast cancer detection. *IEEE Journal of selected topics in quantum electronics*, *5*(4), 1143-1158.

[59] Reeder, S., Berkanovic, E., & Marcus, A. C. (1980). Breast cancer detection behavior among urban women. *Public health reports*, *95*(3), 276.

[60] Gershon-Cohen, J., & Hermel, M. B. (1969). Modalities in breast cancer detection: Xeroradiography, mammography, thermography, and mammometry. *Cancer*, *24*(6), 1226-1230.